A Defense of the Public Health-Quarantine Model of Punishment in Light of Obligations of the State to the Wrongdoer

by

Eric Nicholas Bohner

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Arts

in

Philosophy

Waterloo, Ontario, Canada, 2017

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Punishment is traditionally justified retributively or consequentially; that is, with respect to the desert of the wrongdoer or the positive consequences of the punishment. State-sanctioned punishment (the kind of punishment that is administered by the state to those who break the law), I find, cannot be justified in these traditional ways. In the first two chapters of this thesis I take a detailed look at these traditional theories of punishment and point out several strong moral objections which provide reason to believe they do not sufficiently justify punishment. In the next chapter, I argue that punishment can only be morally justified as a last resort of keeping society safe, as described in the Public Health-Quarantine Model of Punishment (PHQ model). That is, punishment can only be justified insofar as it is *necessary* for ensuring the safety of individuals in a society. Even then, however, I argue in the fourth chapter that the state creates an obligation to the wrongdoer when it punishes her for breaking the law. This obligation, I argue, is best fulfilled by providing every reasonable opportunity to assist with the wrongdoer's rehabilitation into society.

By taking seriously my recommendation of ensuring the health of the wrongdoer, I provide an interpretation of the PHQ model which avoids the "mere means" objection to punishment. That is, it ensures that wrongdoers are not merely used as a means of promoting the general wellbeing of society, but are treated as autonomous individuals who deserve the respect that is required of all people. Most importantly, it ensures that dangerous wrongdoers who are incarcerated for the safety of society are helped as quickly and efficiently as possible, similar to the way in which we are obligated to treat those with dangerously infectious diseases who are quarantined to ensure they cannot unintentionally harm others. By ensuring wrongdoers are treated this way, the PHQ model of punishment cannot lose sight of the needs of the individual in

favour of the greater good for society. In that way, my argument that state-sanctioned punishment entails an obligation to the wrongdoers, actually strengthens the PHQ model from some common (and often damning) objections to the justification of punishment.

Acknowledgments

I acknowledge that this thesis took too long to write, and filled more pages than I originally intended. I would also like to thank my cat, Louie, for micromanaging my work schedule by making sure I took frequent breaks to pet, feed, and pay attention to her. I am sure I could not have completed this thesis in the time it took me without her help. In fact, I'm sure it could have been done much sooner.

I would also like to thank the people who *actually* helped make this thesis what it is: most importantly, my supervisor, Mathieu Doucet, whose constructive comments and patience with my work ethic ensured that I actually finished the thesis in the quality that you can see today. Also, my readers, Shannon Dea and Christopher Lowry, for their helpful comments. And, of course, thanks to Ananya Chattoraj, for putting up with my whining and complaining during this long and arduous process.

Table of Contents

Preface

Punishing individuals when they have done something wrong is relatively commonplace; perhaps even expected or to be desired. We punish children when they hit their siblings, we punish a person who drives recklessly in a school zone, and we punish dangerous criminals for causing monetary or physical harm to society or citizens. The reason why we punish is also usually very straightforward: the subject of our punishment has done something *wrong*, broadly construed, as something undesirable or unacceptable given social or moral norms. Our goals for what the punishment ought to achieve may differ greatly case by case – we may wish the wrongdoer to learn a lesson, or to exact our revenge on someone who has slighted us, or to deter others from wronging us in the future, or some other reason entirely – but at the very least, we can say that punishment is a common response to those who have done something wrong.

People begin to disagree, however, on when, exactly, we are justified in administering a punishment, or even who "we" is meant to represent. Indeed, the topic is widely debated, not only in philosophy, but in politics, businesses, and households alike. Questions arise, such as: when is it acceptable to punish? How severely are we allowed to punish? Should the outcomes of the punishment be taken into consideration? Which outcomes are acceptable and not acceptable? Is it okay to punish just because someone deserves it? What does it even mean to *deserve* a punishment? and Who is allowed to punish? These questions have far-reaching implications for the way our society functions as a whole. Not only are they philosophical questions that relate to questions about free will and determinism, agency, and moral theory, but they have dramatic, far reaching political and institutional consequences and implications. It is because many people accept a certain set of answers to those questions that we have courts and prisons, and feel entitled to incarcerate many thousands of people. If, for example, we determine that it is only

ever morally acceptable to punish an individual who has committed some wrong for the explicit purpose of making society safe, and *not* because he or she truly *deserves* some harm for her immoral actions, then it appears that our current justice system might be totally unjustified in the way that it punishes convicted criminals (and, indeed, I believe that it is).

The answers to these questions determine the kinds of punishment that are meted out, since moral justifications will provide strict guidelines for when and how a punishment ought to be carried out. It is important, then, to determine what makes a punishment justified, and then what practical applications this has on how we go about punishing wrongdoers. Of particular interest in this discussion – at least to me – is how this affects state-sanctioned punishment, since the state is supposed to be an indifferent, unbiased entity for ensuring justice within society. If that is the case, and given the enormous power that it has, then it seems incredibly important that the state is absolutely justified when it punishes one of its citizens lest it fail at the high standards to which we hold it. For this reason, the contents of this thesis tend to focus more specifically on state-sanctioned punishment rather than the punishment that might be between a parent and child or a dog trainer and her disobedient dog.

I discuss three kinds of justification for punishment in this thesis: forward-looking, backward-looking, and mixed justifications (which incorporate elements from both forward- and backward-looking justifications). They might also be referred to as consequentialist/utilitarian or retributivist/desert-based; and they refer to the moral imperatives that are being used to justify different theories of punishment. In chapter one, I briefly outline the most common theory of punishment for each kind of justification, providing an overview of its unique appeals and disadvantages. These are the forward-looking deterrence theory, the backward-looking retributivist/desert theory, and the mixed justification, moral education theory of punishment.

The deterrence theory of punishment is purely consequentialist (H.L.A. Hart 1959; Rawls 1955); that is, it claims that a punishment is only ever justified if it produces desirable consequences (where desirable consequences are usually considered to be a safer society with fewer crimes committed, or simply fewer criminals overall). In this way, the ends *always* justify the means, so that if a punishment produces better results than any of the alternatives (i.e., not punishing a wrongdoer), then the punishment is justified, and perhaps, morally required. This is a forward-looking justification of punishment because it is only concerned with future consequences and denies the value of considerations such as a victim's retribution or a wrongdoer's *deserving* to be harmed for her wrongs, at least inasmuch as it could be considered separate from causing the best possible outcome overall. Instead, the deterrence theory is more interested in punishing wrongdoers with the explicit intent to prevent future crimes by creating a deterrent for would-be criminals or would-be-again criminals. In a perfect world, the deterrence theory of punishment provides the most effective punishment such that the threat of punishment prevents (or deters) anyone from ever committing a crime and everyone would be better off for it. If that could be achieved, then the consequentialist would claim that the punishment (or even merely the *threat* of punishment) had been perfectly justified.

On the other end of the spectrum, retributivists (Kant 1788; Moore 1997; Zaibert 2006) claim that a punishment is justified only on backward-looking considerations. That is, a punishment can only be justified when it is given to an individual who *deserves* to be punished for committing some wrong. This justification asserts that when an individual acts in a way that is morally wrong, and when the relevant conditions are met such that she bears moral responsibility for her actions, it is then morally required that she suffer some hardship or harm for acting in that way. In this way, the retributivist claims that a punishment is justified only

when it is able to harm an individual in the way that she deserves for her actions, and as long as it does this, considerations about the future outcome of this punishment need not be considered.

Lastly, there exist numerous mixed justification of punishment which combine considerations for the future outcomes of a punishment, as well as ensuring that the individual is treated in the way that she deserves. One popular justification for punishment of this kind is the moral education theory of punishment proposed by Jean Hampton (1983). It takes the most important consideration to be the moral knowledge of the wrongdoer; similar to punishing a child to ensure that she knows hitting her sister is wrong, we punish criminals in order to teach that person our moral imperatives. Thus, we punish an individual not only to be able to prevent future crimes from occurring, nor only to ensure that she gets what she deserves, but because she deserves to be punished *and* because we hope that she learns the value in not doing it again. In this way, the moral education theory of punishment has elements that are both forward-looking and backward-looking, making it a mixed theory of punishment.

In chapter one I explain each of these justifications of punishment in more detail, outlining numerous arguments and justifications for the use of state sanctioned punishment. I provide a brief overview of the benefits that each theory of punishment has to offer, along with the kinds of punishments that are justified under each theory. Then, in chapter two, I examine some of the flaws with each theory of punishment, paying close attention to when the theories recommend punishment in ways we normally think cannot be justified. Ultimately, I argue that we actually have strong reasons for abandoning all three of the commonly held justifications of punishment.

In chapter three I introduce the public health-quarantine (PHQ) model of punishment (Caruso 2016; forthcoming), which I argue succeeds in justifying punishment because it only

aims to harm individuals as little as possible (sometimes not at all) while focusing on removing the underlying causes of crime so that more people are less likely to be dangerous or harmful to society. In extreme cases, when an individual *is* harmful or a danger to society, the PHQ model of punishment allows for incarcerating the individual in the same way that we would quarantine individuals with Tuberculosis or other dangerous and infectious diseases; it is a regrettable but necessary measure in order to protect society, but importantly does not propose any harmful consequences for the individual *except* that they are kept safely away from society. In chapter three I discuss in more detail the ways in which the PHQ model succeeds in justifying punishment where the other three theories have failed, showing that it is a preferable theory of punishment to the others.

In the final chapter, I expand on the PHQ model of punishment, suggesting that an important aspect of the model is that we must not neglect the health of the individual in favour of overall public health. While the importance of social programs and education, as well as the safety of society is not to be denied, special focus must be allocated for ensuring individuals who have committed crimes receive the help they need to become functioning members of society. In this chapter, I discuss how the PHQ model is able to deny the moral responsibility of individuals (therefore denying basic desert claims and the legitimacy of blame or praise for one's actions), while still justifying the necessary harm done to them by incarcerating harmful individuals. I then discuss the moral claim by Saul Smilansky (2011; 2016) that individuals who do not bear moral responsibility for their actions deserve compensation for being unfairly incarcerated, and that this undermines the possibility that the PHQ model could advance public health. I argue that the PHQ model is able to accommodated this requirement, arguing that, as long as we are able to negate the bad brute luck which caused individuals to become wrongdoers, this is sufficient in

compensating them for the unfair incarceration. I then discuss how it should not be a concern that this model of punishment *over*compensates individuals such that imprisonment becomes an enjoyable and preferred experience to life on the outside.

By taking my recommendation of ensuring the health of the wrongdoer, I provide an interpretation of the PHQ model which avoids the "mere means" objection to punishment. That is, it ensures that wrongdoers are not merely used as a means of promoting the general wellbeing of society, but are treated as autonomous individuals who deserve the respect that is required of all people. Most importantly, it ensures that dangerous wrongdoers who are incarcerated for the safety of society are helped as quickly and efficiently as possible, similar to the way in which we are obligated to treat those with dangerously infectious diseases who are quarantined to ensure they cannot unintentionally harm others. By ensuring wrongdoers are treated this way, the PHQ model of punishment cannot lose sight of the needs of the individual in favour of the greater good for society. In that way, my additional recommendations strengthen the PHQ model from some common (and often damning) objections to the justification of punishment.

Chapter 1

1. Introduction

This first chapter will be dedicated to expositing three of the most common justifications of punishment in the philosophical literature. In each section, I will briefly outline one theory of punishment, discussing its strengths and implications for state-sanctioned punishment. Then, in chapter 2, I will argue that each theory has independent moral reasons for thinking that there are key aspects in which the theory fails to justify punishment satisfactorily. While there are certainly more than three theories of punishment (indeed, in chapters 3 and 4 I will introduce a fourth theory), I will be looking at three distinct *types* of justification – desert-based, deterrence-based, and an expressivist view – such that if one can be shown to not be justified, then any relevantly similar theory of justice will likely fail to justify punishment in largely the same or similar ways. Since I cannot respond to every single theory of punishment, I will focus on the most common theories of each type in order to show that the kinds of justification that have been provided, so far, have been insufficient.

Any thorough examination of a topic, I think, should take some time to define its terms. In the case of punishment, however, it is not entirely obvious how to do so because so many philosophers have defined it differently (see, for example: Zaibert 2006; Hampton 1983; Hart 1959; Rawls 1955). For that reason, throughout the thesis, whenever I am discussing a particular philosopher's view of punishment, I will provide additional clarification – where necessary – of how the term is being used. For the moment, however, it will suffice to say that a punishment is an intentional harm which is inflicted in response to a wrongdoer's action in the form of some suffering or hardship in which the wrongdoer has no choice (i.e., she cannot opt out of receiving the punishment). For my own discussion of punishment, I am specifically interested in the state-

sanctioned variety. That is, I am interested in when and why it is morally permissible for the state to intentionally impose a harm on its citizens, and not in the kinds of punishment that a parent or teacher might give to a disobedient child.

In section 2 of this chapter, I will discuss retributivism, the view that punishment of a wrongdoer is only justified because she *deserves* something bad to happen to her only in virtue of the fact that she has knowingly done something wrong (Pereboom 2014, 157). Here, I will describe arguments that aim to morally justify retributive punishment. I will also explain the retributive account of *how* the state should punish; that is, the retributive account of when a punishment is justified – or required – and what such a punishment would look like). In section 3, I discuss the deterrence theory of punishment, the view in which a punishment is only justified when it produces (or is reasonably believed to produce) an effective deterrent for crime (i.e., it causes would-be wrongdoers to choose *not* to commit a crime they otherwise would have). In this section, I look at Rawls' defense of the deterrence theory in which he adopts a rule utilitarian view of punishment that is justified in principle by deterrence, but which requires additional moral reasons in order for the practice of punishment to be justified. Lastly, in section 4, I look at the moral education theory of punishment, the view that the goal of punishment should be to morally educate wrongdoers (and would-be wrongdoers in society) about the social norms and moral imperatives that exist, which will make those individuals behave morally in the future (Hampton 1984, 212).

Ultimately, this chapter will provide an overview of three different theories of punishment along with how philosophers have typically argued they are justified. The justifications are typically classified in three different ways: deterrence-based theories which are mostly forward-looking, desert-based theories which are mostly backward-looking, and a mixed

justification. Forward-looking, consequentialist justifications, like the deterrence theory, are primarily focused on the consequences of a punishment. That is, a punishment can only be justified when it produces the best outcomes (such as the public interest or the common good). Desert-based, backward-looking justifications, such as retributivism, claim that a punishment can only be justified when the wrongdoer receives what she deserves in a basic desert sense; that is, without any regard to the consequences of the punishment. And, of course, the mixed justification of punishment combines elements of desert and deterrence in justifying punishment, but only in achieving some larger goal. In the case of the moral education theory, that is to morally educate wrongdoers and to censure immoral acts. In my discussion of the three particular theories of punishment, then, I will be discussing one theory of each kind of justification. In chapter 2, I will then provide criticisms of these theories to show that not only can retributivism, deterrence, and moral education not justify punishment, but that there is reason to be sceptical of any of the justifications of punishment that utilize these kinds of justifications.

2. Retributivism

The retributivist believes that "a person who unjustifiably and inexcusably causes or risks harm to others or to significant social interests deserves to suffer for that choice, and he deserves to suffer in proportion to the extent to which his regard or concern for others falls short of what is properly demanded of him" (Berman 2008, 269). In other words, if a person acts while knowing that the action is morally wrong *and* is not subject to factors which would impede responsibility for her actions (e.g., coercion, evil genius mind control, Cartesian demons), then that person *deserves* to suffer for her actions. This view is particularly pervasive and intuitively appealing, especially among laypeople (Carlsmith et. al. 2002; Carlsmith 2008), but also in the

philosophical literature on punishment (see, for example: Morris 1968; Mackie 1982; Moore 1997; Dimock 1997; Zaibert 2006). For this reason, retributivism seems to be a natural place to start a chapter that reviews different justifications of punishment; starting with the oldest, most widely held view and working toward the newest views.

Punishment claims to be a *justified* act of deliberately harming or causing an individual to suffer. All else being equal, we ordinarily take actions of deliberate harm to be morally unjustified. So, if a punishment can be justified, there must be something special happening during acts of punishment which make it morally acceptable but that does not exist in merely harming another individual. The obvious and relevantly different feature of a punishment from all other acts of harm, of course, is that a punishment is a specific harm that is done *to a wrongdoer*. But it is not immediately clear which morally relevant imperatives are able to show us how harming a wrongdoer in particular makes the harm *good* (i.e., morally praiseworthy), so any sufficiently justified theory of punishment must tackle this difficulty.

2.1 Possible Justifications of Retributive Punishment

Michael Moore (1997, 101) claims that there are two strategies for justifying retributive punishment:

(1) Showing how retributive punishment "follows from some yet more general principle of justice that we think to be true"; and

(2) Showing that retributive punishment fits with a theory of punishment that "best accounts for those of our more particular judgments that we also believe to be true".

In the remainder of section 2, I will briefly look at several attempts to justify punishment by utilizing these strategies: first, by arguing that we deserve punishment when we feel guilty about

doing something wrong; second, by making the claim that wrongdoers owe a debt to society which can only be repaid by enduring some suffering; and third, asserting that when a wrongdoer suffers in the correct proportion for her wrongdoing, it is intrinsically good that she has done so (i.e., it is better that she suffers than not, given that she has done something wrong). I will briefly describe how retributivists have argued in favour of these claims before I discuss the implications that this has on punishment; for example, what kinds of punishment are acceptable, and how severe the punishments ought to be in order to be justified.

2.1.1 Guilt Implies Desert

An interesting way to try to show that punishment best accounts for other judgments which we hold to be true (i.e., utilizing the second strategy of justifying punishment) is expressed by Moore when he imagines that, if he had committed a terrible crime, he hopes that his response "would be that I would feel guilty unto death" (1997, 145) and that "it is elitist and condescending toward others not to grant them the same responsibility and desert you grant to yourself" (1997, 148-149). That is, we should assume individuals who *feel* guilty have the ability to know when they are *actually* guilty because they are the authority on whether or not their own subjective experiences are legitimate. Moore claims that we should not, therefore, question if a wrongdoer might be mistaken about her feelings of guilt; we should instead assume that she is aware of the cause of her feelings, and only feels *real, legitimate* guilt when there is due cause. In this way, Moore attempts to show that punishing the guilty is justified because guilty people feel that a punishment is deserved ("unto death" in some cases). Thus, the fact that we feel guilty when we have done something wrong is supposed to show that retributive punishment fits with a theory of punishment that "best accounts for those of our more particular

judgments that we also believe to be true", from Moore's second strategy of justifying retribution.

On its own, this is not a sufficient justification of punishment because we can always doubt the authenticity or legitimacy of a person's feelings of guilt. For example, a child may feel guilty (and responsible) after her mother died of cancer, but we would never claim that her mother's death is *actually* the child's fault. Even if we were to imagine ourselves in the shoes of the daughter, having experienced the same circumstances as her, it is very likely that *we* would feel guilty and responsible for our mother's death as well. Thus, it is not necessarily the case that feelings of guilt are unimpeachable and a clear indication that we *deserve* punishment. Utilizing the second method alone, then, is not sufficient for justifying retributive punishment, although it might show that retributive punishment can only be justified when a wrongdoer *ought* to feel guilty about her actions. In order to have a complete justification, this appeal to our "particular judgments" would have to be supplemented with other justifications that do not solely rely on our judgments that we usually assume to be true.

2.1.2 Debt to Society

Susan Dimock attempts to justify punishment by showing that retributive punishment follows from a more general principle of justice: that it is our duty or responsibility to maintain the trust of society by obeying the law. By not obeying the law, we are breaking the trust of society which causes us to owe a debt that can only be repaid by enduring retributive punishment (i.e., suffering). As Dimock says, "the harm of legal offenses is to be found in their violation of the conditions of basic trust in society, and we restore trust by punishing the offender" (Dimock

1997, 40). In this way, Dimock claims that retributive punishment is justified only when it effectively restores trust in society.

Dimock argues that a social contract provides the necessary requirements for trust to exist between parties, and that, in a functional[1] way, it is the central purpose of law to enable trust between individuals to exist (ibid., 45). Moreover, the kind of objective trust that society provides can *only* exist while individuals in society are obeying the law; whenever one individual disobeys the law, this reduces our reason to believe that others in society are likely to be trustworthy, reducing the amount of trust overall. "Knowing that our fellows have allowed us to be victimized without complaint and protest and condemnation, or that they are unwilling to assist us in providing protection against further abuse, would make trust of them, and not just of our violator, less objectively reasonable" (ibid., 51). Therefore, in order to maintain the trust in society, and to repair any damage that might have been caused to it by a wrongdoer, Dimock asserts that censure must occur for the wrongdoer's actions.

When an individual commits a crime, she harms society itself by making it harder for individuals to trust others in society. Since the wrongdoer has agreed to maintain the trust of society by being a part of that society,[2] she then owes a debt to repair the trust that she has broken. The only way to restore that trust, Dimock claims, is to retributively punish the wrongdoer. This is so that "when trust between members of society has been violated, trust in the law as capable of maintaining the conditions of trust must be reaffirmed" and punishing the criminal "serves to reestablish that trust and demonstrates that individuals need not adopt

[1] Here I mean "functional" in the sense that it is a functionalist view of government in which government's functional purpose is to provide the means of trust between parties.

[2] Dimock takes a contractarian position regarding membership of society. This is important because it implies the duty of the citizen to maintain society's standards and norms.

recourse to anticipatory violence as a means of protecting their interests against those who are willing to harm" (ibid., 54).

This justification of retributivism utilizes the first strategy of Morris's to argue that justice demands that wrongdoers be retributively punished in order to maintain the trust that is essential to society. It does not, however, provide very strong reasons for believing that retributive punishment is the *only* way in which the objective trust of society can be restored; at best, it offers us the suggestion that people will only be satisfied that trust has been objectively restored when there has been retributive punishment for the wrongdoer. It seems, though, that this may simply be empirically false: that members of society could be assured that they do not have reason to doubt the trust in society through other methods. For instance, the Alliance for Safety and Justice reports that victims of violent crime want to see less spending on prisons, shorter prison sentences, and greater focus on the rehabilitation of criminals (2016). This does not prove that individuals would be reassured in the trust of society through means other than retributive punishment, but it might suggest that others exist. For this reason, I think that if retributive punishment can be morally justified and shown to be morally *required*, that some argument must be made in favour of the goodness of punishment in and of itself, without reference to consequences that could apparently be satisfied through methods other than retributive punishment.

2.1.3 Intrinsic Goodness of Deserved Punishment

Leo Zaibert attempts to provide us with an argument for the intrinsic value of punishment, claiming that when a wrongdoer suffers in the correct proportion for her wrongdoing, it is intrinsically good that she has done so. That is, given that the wrongdoer has done something

wrong, harming that person in the amount deserved actually produces a morally good outcome. In this way, Zaibert would like to say that retributive punishment is morally better than no punishment at all, no matter what the consequences would be in either case.

In Zaibert's theory of punishment, the driving normative force of the argument is in the claim that it is intrinsically good when one gets what one deserves. In order to see how Zaibert argues for this, I must quickly discuss Organic Wholes (or organic unity), which are described by G.E. Moore: "the value of a whole must not be assumed to be the same as the sum of the values of its parts" (Moore 1903, §18). Essentially, even if the intrinsic value of each individual action may be negative, the value of the *whole* (i.e., the sum of the individual actions) may yet be positive. Relating this to retributivism, a wrongdoer may act in a way that is wrong and therefore deserve to be punished. The punishment in and of itself is *prima facie* intrinsically bad because it causes a harm to another individual; however, when the punishment is given retributively (i.e., with the correct proportion) to an individual who deserves exactly that punishment, an intrinsically good thing has occurred. The organic whole, then, is intrinsically good, even while its individual parts are intrinsically bad.

Zaibert's argument for an intrinsically good organic whole in the case of punishment comes from his treatment of an example taken from Michael Moore's *Placing Blame: A Theory of Criminal Law* (1997). The example is meant to show that it can be intrinsically good to punish even when no further beneficial consequences can be obtained:[3]

> Consider Michael Moore's sensible condemnation of that sordid spectacle of "fraternity boys" throwing parties outside the prison's gates while executions take place. … Retributivists may recommend the punishment of the fraternity boys celebrating at the prison's gates independently of any consequences that this punishment might have. Consequentialists can perhaps do it as well, but, unlike retributivists, they do not have an

[3] i.e., the punishment will not necessarily deter future crimes or provide any other intrinsically valuable consequences, other than the intrinsic good which arises from giving the wrongdoer the punishment he deserves.

> obvious argument available. Consequentialists, moreover, face an uphill battle in the sense that to punish those who celebrate punishments which are sanctioned by consequentialist rationales might send confusing messages. In other words, if punishment X is justified because it brings about consequence Y, celebrating the infliction of punishment X will, in principle, contribute to bringing about more consequences Y … For those of us who believe that the fraternity boys are doing something wrong when they throw parties outside prison gates where executions take place, retributivism provides a much clearer rationale for punishing them than does consequentialism.

The rationale for the retributivist is simply that the fraternity boys have done something *morally wrong* by celebrating the execution of some criminals, and therefore deserve to be punished. The consequentialist, on the other hand, must provide a story about *why* it is worse for the greater benefit of society, and how punishing those individuals will help society overall. The tricky part about the example is that the fraternity boys appear to be promoting a state sanctioned punishment, so it is not intuitively obvious how their celebrating is actually worse for society than if they were to show compassion for the wrongdoers being executed. The retributivists, however, have a much simpler time explaining why the fraternity boys should be punished, as they are "concerned centrally with the intrinsic goodness of the organic whole whereby the deserving get what they deserve, [and] are from the start sensitive to taking seriously the intrinsic goodness and badness of certain actions, and of the way in which these discrete actions combine to form organic wholes which themselves can be of greater or smaller value than the value of their constituent parts" (Zaibert, p. 213). Essentially, the fraternity boys, in celebrating the execution of criminals, are doing something intrinsically bad and are therefore deserving of punishment. To punish them, then, will be morally good as it will bring about an organic whole that is intrinsically good (i.e., good in and of itself and not dependent on the consequences of the punishment).

Zaibert sums up his theory of punishment in this way: "To be a retributivist is not merely to claim that desert is a necessary condition for the infliction of just punishment, nor quite to claim that it is a sufficient condition either. To be a retributivist is to recognize that deserved punishment is an intrinsic good" (p. 214). Lastly, he adds the caveat that "the fact that something is intrinsically *good* does not make it the case that to bring it about is, willy-nilly, the *right* thing to do" (p. 214). That is to say that, even though it is a good thing to punish the deserving, there are a host of other factors which might confound the appropriateness of punishing the deserving in certain situations. Therefore, punishment is intrinsically good (when it is deserved), but not always the *right* course of action to take (i.e., providing the *most* desirable outcome, even if it might provide *one* desirable outcome). Zaibert, in contrast to Dimock, is not making the strong positive desert claim that wrongdoers ought to be punished and it is morally good if they are, but rather makes the claim that when retributively punishing the deserving, it is merely an intrinsic good. Or, put another way, Zaibert supports the negative desert claim that it is never justified for the state to punish the undeserving (e.g. the innocent), but wrongdoers lose the right to not be punished by the state and it *may* be permissible to punish, provided further justification. Zaibert's position does not necessitate the instantiation of the punishment merely on the grounds that it is an intrinsic good, but he does provide us with an argument for showing that sometimes causing a harm can be *good*. This, I think, provides the strongest reason in favour of retributivism out of the three I have discussed in this chapter.

2.3 Retributivism Conclusion

In discussing retributivism, I have discussed several different strategies for justifying retributive punishment, which attempt to provide moral reasons showing that it is better for a wrongdoer to

be harmed than not harmed, without making any reference to the consequences of the punishment. That is, the retributivist attempts to justify punishment only insofar as it is good in virtue of the fact that it is deserved or otherwise morally required. By arguing that punishment is justified in this way, we can determine that any *kind* of punishment can be justified, as long as it is administered in the amount deserved. Fines, imprisonment, community service, menial tasks, or even capital punishment are all *kinds* of punishment that the retributivist position could potentially recommend. For example, we might think that a more severe crime, like murder, deserved to be punished with life in prison, while a less severe crime, like speeding, might be punished with a simple fine. But as long as the punishment harmed the wrongdoer in the amount deserved, then any punishment could, in principle, be justified. Usually, retributivists argue that the punishment deserved is proportional to the wrong committed. For example, for Dimock, the deserved punishment would correlate to the amount of damage that the individual has done to the trust of society; the more society's trust had been damaged, the more severe the punishment the wrongdoer deserved.

The proportionality of the punishment is a key aspect of the retributivist position. In the following section, I will describe a theory of punishment that is not at all concerned with the severity of punishment proportional to the severity of the crime. The deterrence theory of punishment will provide an alternative goal for what punishment hopes to achieve, which will show how different theories of punishment can greatly change the kinds of punishment that are recommended. While the retributivist theory is principally concerned with ensuring that the wrongdoer is punished in the amount that she *deserves*, we will see that the deterrence theory is not concerned with this at all, instead favouring punishments which bring about the best consequences.

3. Deterrence Theory

Retributivism is often measured up against, and contrasted with, the consequentialist theory of punishment as the other side of the debate to the strictly deontological, desert-based retributivist position. Which makes it strange, then, when we learn that there are very few strict consequentialists about theories of punishment; most contemporary philosophers who oppose retributivism offer mixed theories that incorporate both consequentialist and deontological considerations in order to justify punishment. Yet any paper that defends retributivism will always point out that there exists a "classic debate"[4] between retributive and utilitarian theories of punishment that has been ongoing for time immemorial. At least lately, then, it seems like one side of the debate has been talking to an empty room. Nevertheless, a survey of justifications of punishment would not be complete without a brief discussion to at least show why very few people strictly adhere to the deterrence theory.

In this section of the chapter, I will outline the main features of a deterrence theory of punishment before discussing some of the limitations that are normally discussed which work against the theory's justification. The substantive part of this section will be used to discuss John Rawl's defence of a deterrence theory of punishment (1955), where I will look at his arguments in favour of justifying the *institution* of punishment with utilitarian values, but not the individual instances of a punishment. This will ultimately provide an alternative justification of punishment from the retributivist theory, and, to conclude, I will discuss the differences in how each theory recommends we punish wrongdoers.

[4] Joel Feinberg, "The Classic Debate", *The Philosophy of Law* 4th Edition, eds, Joel Feinberg and Hyman Gross (Belmont, California; Wadsworth Publishing Co., 1991). 624-629

3.1 Deterrence Overview

The deterrence theory of punishment usually takes as its primary goal maximally increasing the safety of a society by reducing recidivism and deterring crime in general. The theory claims that any actions which bring about this goal, including punishment, are justified. In theory, a punishment is able to do this by providing a strong incentive *not* to act in such a way which will result in the punishment occurring. This is easily shown in an example: Rob wants a hundred dollars, and he happens to be in a convenience store. The store, he knows, has a hundred dollars in the cash register, and the cashier has recently gone to the bathroom. Rob could easily open the cash register and take a hundred dollars. Since Rob knows that robbing a store is illegal, however, he might be punished for theft (assuming the necessary contingencies are in place that make his apprehension by the police a possibility). Rob does not want to be punished for theft, so he chooses not to act on his desire for a hundred dollars by robbing the store. Due to the threat of punishment Rob has been deterred from robbing the store.

The mechanisms at play in deterring a crime are simple enough to intuitively grasp:

(1) the negative consequence (i.e., the punishment) must outweigh the positive consequences of committing the crime,

(2) the punishment seems likely to occur to the individual,

(3) the individual should be able to rationally consider those factors and come to the conclusion that the action is not in their best interest.

A purely utilitarian view will not insist that the punishment is always enforced for the best possible outcome in all situations. There may indeed be instances where alternative methods to punishment produce better results, and in those situations punishment cannot be justified to

increase the safety of a society. For the purpose of this discussion, however, it is useful to discuss a deterrence theory of punishment where punishing wrongdoers *can* be the most effective method of providing a safe society. In chapter 2, I will be pointing out several reasons that we might think that punishment is *rarely* – or at least less frequently than one might assume – a justified way of producing a safe society.

Although not a view held by many philosophers, the deterrence theory of punishment is often espoused by politicians and others who want to "crack down" on violence and other crimes. It is a frequent and – as I will argue later – unfortunately held belief that if we make punishments more severe, it will deter the wrongdoers of the world from continuing to do wrong *and the world will be a better place*. In the following few pages I will quickly outline some of the limitations and arguments against deterrence as a theory of punishment. I will then briefly examine one influential argument from John Rawls (1955) which defends a rule utilitarian view of the deterrence theory from these limitations and provides a potential way for the theory to remain justified, at least in principle, if not in practice.

3.2 Rule Utilitarian Justification of Deterrence

The strict utilitarian claims that the ends *always* justify the means, and that it is *morally* good if the best possible outcomes are achieved (for utilitarians like Bentham (1789) and Mill (1843; 1859; 1861), this meant bringing about the greatest good, where "good" is simply happiness). In regards to punishment, this might mean that whatever actions are required to bring about the safest society (i.e., make people the happiest overall) are morally obligated. However, this means that it might sometimes be morally justifiable to punish the innocent and punish more harshly than a wrongdoer proportionally deserved for the crime committed, simply because it

would bring about the best results. For example, framing an innocent person for murder, and punishing her severely (e.g., capital punishment), may provide an effective deterrent to other would-be murderers. But, since we usually hold the intuition that punishing individuals can only be morally justified when they have *actually* committed a crime, and *only* with the severity that the crime calls for (i.e., proportional to the severity of the crime), then if the deterrence theory recommends punishing in ways that go against that intuition, we feel that it cannot be justified.

Rawls defends the deterrence theory against this objection by making a distinction: "one must distinguish between justifying a practice as a system of rules to be applied and enforced, and justifying a particular action which falls under these rules; utilitarian arguments are appropriate with regard to questions about practices, while retributive arguments fit the application of particular rules to particular cases" (Rawls, 1955, p. 5). Instead of arguing that a utilitarian theory like the deterrence theory must be changed or have caveats which ensure the guilt of the individual and a correctly proportional punishment, Rawls proposes that "stating utilitarianism in a way which accounts for the distinction between the justification of an institution and the justification of a particular action falling under it" (ibid. p. 10) is sufficient to assuage our worries regarding the possible tyranny of deterrence.

By making this distinction, Rawls argues that we can justify punishment *in general* because of its ability to deter, but punishing in particular cases requires additional considerations that a consequentialist theory cannot provide. For example, we might say that the practice of punishing wrongdoers is justified because it shows others that there are bad consequences for wrongdoing, which deters them from committing wrongs, but we cannot make an example out of one particular wrongdoer (or innocent person) by severely harming her, even if it could be shown to maximally deter everyone else from committing a crime. The reason for this is because of our

other moral concerns for the individual that are not being addressed by the utilitarian objective: we do not think it is morally acceptable to harm someone when they have not done anything to deserve a particular treatment.

This position is a rule utilitarian account of the deterrence theory of punishment: it justifies an intentional infliction of harm as punishment to the extent that it will reduce the overall level of crime, therefore increasing overall happiness, but this justification is constrained by other considerations we deem to be valuable. This means that, since we *do* have other considerations we hold to be morally true (namely: it is not morally acceptable to harm the innocent or to punish a wrongdoer too harshly), then rule utilitarianism prohibits punishing in those circumstances. Thus, Rawls provides us with an account of the deterrence theory of punishment that circumvents our initial concerns for the justification of the theory. As I will show in the next section, however, it is still not quite sufficient for giving us a completely justified account of the deterrence theory of punishment.

3.3 Deterrence Conclusion

Deterrence is a hugely popular justification of punishment among politicians and people who generally think that the threat of harsh punishments is effective in preventing crime. Since the deterrence theory is able to justify punishing innocent people, however, the philosophical community has been much more hesitant to proclaim themselves proponents of the theory. Still, Rawls has shown that the theory is able to provide a justification for the *institution* of punishment as a deterrent. In the next chapter, I will discuss several other reasons we may have for believing that the deterrence theory cannot be justified, or at least should be abandoned in favour of some other theory.

In the next section, I will look at the moral education theory. This theory, similar to Rawls' version of the deterrence theory, has elements of both the forward-looking and backward-looking justification of punishment. However, while Rawls is principally interested in justifying punishment on its ability to deter, the moral education theory has the advantage of expressing our disapproval of morally bad actions. This difference is important, since the deterrence theory is not interested in expressing *that* we think an action is wrong, only that *if* you act in a particular way, *then* you will be punished.

4. Moral Education Theory

Section 2 and Section 3 discuss two diametrically opposed justifications of punishment. One in which punishment is only ever justified when it is given to a wrongdoer *in the amount deserved* (i.e., in the basic desert sense, without any regard to the consequences of the punishment), and the other, in which it is justified only when it produces the best consequences. This is only, however, insofar as we are trying to justify the institution of (state-sanctioned) punishment and not the act itself. As far as the act of punishment itself, retributivists like Zaibert claim that consequentialist considerations are required to determine when the punishment is *right* as opposed to *good* (Zaibert 2006, 215), while consequentialists like Rawls admit that retributive concerns may also be needed to determine when punishing is *right* on a case by case basis (Rawls 1955, 5). It seems, then, that even those who claim to be strictly retributivist or strictly consequentialist with regards to punishment, often admit of some value to other considerations. For this reason, it should be no surprise that the moral education theory utilizes both retributive and consequentialist reasons in order to justify punishment.

This alternative to the retributive and deterrence justifications of punishment gauges whether a punishment is justified on its ability to provide moral knowledge. That is, the object of a punishment is to teach the wrongness a particular act, either to the wrongdoer who committed the act, or to society as a whole. In that way, the punishment contains both a backward-looking and a forward-looking aspect: the wrongdoer deserves to be punished because she acted immorally, and the end goal of the punishment is that she becomes morally educated as a result. More than merely this, however, is that the wrongness of the action itself is expressed to the wrongdoer and others. Censure, then, plays a significant role in the punishment as well. To close out this chapter, I will discuss the role that punishment plays in the moral education theory by discussing Jean Hampton's account of the theory, as well as the key benefits of using punishment in order to express the wrongness of an action.

4.1 Moral Education

As a basic premise for justifying state-sanctioned punishment, the moral education theory supposes that laws are made based on ethical or moral imperatives. That is, we make laws which express ethical requirements such as "don't murder", or for moral reasons, for example "drive on the right" to ensure all drivers' safety (Hampton, p. 210). For the purpose of discussion, let us assume that all laws successfully follow this guide in order to bypass concerns about unjust or immoral laws. In this way, breaking the law is immoral, not only because the act of breaking the law is wrong, but because the action itself (regardless of the law) is also immoral. Even if this is not actually the case in the real world, we can likely agree that this *ought* to be the case.

The threat of negative consequences for breaking the law is a non-moral incentive to obey the law. That is, there is a prudential deterrent (on top of the moral considerations) that is

provided from threatening to punish for committing a crime. And following through with the threat – actually punishing the wrongdoer – is a way to "make good" on that threat and can be justified if it helps to deter additional crimes (Hampton, p. 211). Provided a conception of law as "rules of obligation"[5], the laws are meant to express the moral boundary between permissible and impermissible actions, and provide non-moral reasons for following the law.

More than merely conditioning society to obey the law due to the negative consequences of committing a crime, however, the moral education theory asserts that "punishment is intended as a way of teaching the wrongdoer that the action she did (or wants to do) is forbidden because it is morally wrong and should not be done for that reason" (Hampton, p. 212). Since individuals are able to reflect upon the reasons for these laws, they are intended to instruct an individual in identifying the moral boundary that is being crossed in breaking the law. So not only do individuals not want to break a law because of the threat of punishment; upon reflection, they are also supposed to realize that breaking the law is *morally* wrong, and therefore should not be done.

In describing the moral (and legal) boundary between right and wrong, the moral education theory does not only provide instruction to the wrongdoer who is punished, but also to others in society. Consider laws against discrimination in hiring practices based on race. Punishing those who refuse to hire people of a particular race, then, is not only useful in showing the hiring manager that her unfair treatment of a particular group of people is wrong, but also others in society to identify the practice as immoral. Since others are able to reflect on the

[5] Hart, *The Concept of Law* (1961) discusses some laws as "rules of obligation" in instances when it is used to produce rules that must be obeyed under threat of some punishment (e.g. "drive the speed limit, or receive a speeding ticket").

punishment of the individual, it is supposed that they are able to determine that it actually *is* wrong to discriminate in such a way, and that a punishment for doing so is deserved.

The punishment of wrongdoers, for the moral education theory, is used to teach society the moral norms that ought to exist by clearly defining the boundary between moral and immoral actions (i.e., those actions that are punishable versus those that are not). In addition to this educational aspect, however, is the expression of condemnation of an action which is lacking without the punishment. In particular, Kahan argues that the public feels that imprisonment is a particularly effective method of expressing this, while alternatives are rejected "not because [the public] perceives that these punishments won't work or aren't severe enough, but because they fail to express condemnation as dramatically and unequivocally as imprisonment" (Kahan 1996, 592).

There are certainly ways of expressing that some particular actions are immoral without punishment. For example, if we can reasonably expect laws to describe what we find to be moral or immoral, then simply providing everyone with a list of the laws would ostensibly be an effective method of expressing our disapproval of certain actions. It would not, however, be able to provide a scale for *how much* we wish to condemn an action. We could not, for example, say that "murder is -10g [goodness], while petty theft is -2g." Expressing the moral value of actions in this way would certainly fail to accomplish what we hope to achieve. Instead, the infliction of a punishment might be able to accomplish this goal much better, since we can punish to a lesser or greater degree. Murder, for example, may garner a punishment of life in prison, while the theft of a car might only require a year in prison. In this case, there are clear, appreciable differences in how much worse we think murder is than theft, and the difference in punishments *express* the difference in our moral condemnation.

4.2 Moral Education Conclusion

The moral education theory of justifies punishment when it is used for two purposes: first, for the wrongdoer's own good; and second, to express the wrongness of the action. In the first way, a punishment is meant to force the wrongdoer to reflect on her actions and determine how she was wrong. Ideally, she will come to understand that she acted immorally and realize that she actually deserved the punishment that she received. Moreover, if the punishment was given in the correct amount, she should also realize the severity of her actions; that is, if the punishment was quite harsh, for example, then she will realize the gravity of the action and that it was truly immoral and unacceptable. The punishment also works to shape society by expressing the moral imperatives that others are meant to live by. If everyone knows that stealing is wrong because people have been punished for stealing in the past, then fewer will consider stealing in the future.

Whether or not these motivations for punishment succeed is not entirely clear. Indeed, in the next chapter I will argue that there are numerous reasons for thinking that they do not. However, the moral education theory provides a powerful description of why we may think that a punishment is justified. It gives an intuitive description of why we *want* to punish wrongdoers (i.e., to express our disapproval of the action), and also an optimistic reason to think that we might actually be helping wrongdoers by punishing them if it successfully causes them to consider the moral reasons for not acting in a particular way.

5. Conclusion

I have now introduced three theories of punishment: the retributivist theory, deterrence theory, and moral education theory. In describing these theories, I have noted that there are different

motivations for punishing wrongdoers that are either backward-looking (desert-based) or forward-looking (consequentialist) which inform *why* we punish wrongdoers and *when* we think it is justified to do so. Whether our motivations are forward-looking or backward-looking, however, provide very different recommendations for how much, or even *if*, it is permissible to punish. The retributivist theory, for example, may recommend an extremely harmful punishment like capital punishment, even though it arguably will never provide a moral education for the wrongdoer (indeed, how could it?). On the other hand, the moral education theorist would vehemently deny that capital punishment could ever be justified. Since these punishments conflict with each other (i.e., the moral education theorist thinks the retributivist punishment is unjustified, while the retributivist thinks the moral education punishment is unjustified) they cannot both be correct. For this reason, then, it is important to discuss these different theories of punishment to determine which ones are justified and for what reasons, or even if they can be justified at all. In the next chapter I will take a further look at the three theories of punishment discussed here, where I will argue that all of them have serious moral issues which indicates that none of them provide a complete justification of punishment.

Chapter 2

1. Introduction

Chapter one discussed several different motivations for punishment, utilizing conceptions of desert, deterrence, and moral education. I described how theorists of punishment argue for one or more of these conceptions as a justification of punishment, which provides us with different reasons for punishing wrongdoers. The different theories also provide us with conflicting information about when it is justified to punish a wrongdoer, the amount the wrongdoer ought to be punished, and the *kind* of punishment that ought to be endured. The retributivist and deterrence theorist, for example, may be able to endorse capital punishment in principle, while the moral education theorist seems less likely to accept it as a justifiable kind of punishment. And while each theory of punishment might be internally consistent, I have done very little to compare and contrast the different theories to show which one (if any) might be more justified than the others.

In this chapter, I will be criticizing each of the different theories of punishment discussed in chapter 1. In each of the proceeding sections I will identify one distinct objection to punishment which provides strong reason for rejecting at least some of the competing justifications of punishment. I then compare and contrast the theories together, identifying which are able to resist the objection and which would need to be modified. Ultimately, I will show that each of the three theories discussed in chapter 1 has at least one strong reason for believing that it does not satisfactorily justify punishment.

2. Free Will and Moral Responsibility Scepticism

One useful way to talk about free will and moral responsibility is a way in which you cannot have one without the other.[6] If free will is understood to mean the ability to act on one's decisions without the interference of outside influences, and to have moral responsibility is to have *complete* ownership over one's actions, then this is a way in which one must also have free will in order to have moral responsibility. That is, if nothing other than one's own intentions and actions bring about a desired event, then it would be reasonable to say that she has full responsibility for that outcome. And if it is only possible to have complete ownership over an outcome if one's intentions and desires had not themselves been manipulated or influenced by factors outside of her control, then if she acted with full moral responsibility for her action, she must have acted with free will.

There are two positions we can take on this. Either we *do* have complete ownership over our actions, or we do not. In order to be held ultimately morally responsible for our actions, an individual must have acted with free will. This generally means at least one of two things must be true: (1) if time were reversed to a point just before an individual made a decision, that individual could, in principle, make a decision different from the one she had made before time had been reversed; (2) the individual must be able to make decisions without outside influence (i.e., the individual must be *causa sui* – "cause of oneself"; her decisions cannot be externally caused). The free will libertarian believes that we *could* make a different decision if we went back in time to a point just before choosing to do something[7] because we are not (always)

[6] For further discussion of different views of free will and moral responsibility, see Vargas, M. (2013) "How to Solve the Problem of Free Will". *The Philosophy of Free Will*.

[7] This does not mean that we *would* make another decision, just that we had to ability to do so in principle.

causally determined to act. That is, we have the ability to make decisions without being influenced by factors that are beyond our control. The hard determinist, on the other hand, believes that we are causally determined such that both (1) and (2) must be false. That is, our actions are always the result of past events, and if events occurred in the exact same way, we would *always* make the same decisions. We also cannot be the ultimate source of our actions, since causal events which ultimately result in our actions have been happening since long before we had any ability to control them. Lastly, the compatibilist denies (1) and (2) are required for free will altogether because she defines free will so that we do not require *complete* freedom from outside influence in order to still have free will.

The compatibilist understanding of free will is somewhat less strict in what constitutes an exercise of free will. To the compatibilist, an action that was caused with one's free will merely means an action one intended to carry out while not under duress (e.g. coercion). Similarly, a compatibilist might say that one has moral responsibility for an action just in case what she intended to happen (without being coerced) is also what *actually* happened. For example, if I intend to throw a baseball to a catcher and – barring some unforeseeable and unavoidable event – I succeed, then this would be an action of free will that I would have moral responsibility for. If, however, I only threw the ball out of anger because my child had recently been hurt by the catcher, the compatibilist would still say I acted with free will and am therefore morally responsible, while a hard determinist would say that I did not have free will and therefore was not *ultimately* morally responsibility for the action because it was in some sense *caused* by the actions of the catcher.

In addition to the three views of free will (i.e., compatibilism, libertarian free will, and determinism) which describe what free will is and what it would take to have it, there is also the

sceptical position. As opposed to the other positions which claim that we can have free will only if certain conditions hold true, the free will sceptic, on the other hand, maintains that no how we describe it, there is no way that people have *enough* free will to bear *ultimate* morally responsible for our actions. That is to say, no matter where one stands on the free will debate, the free will sceptic argues that it is impossible to possess the kind of control that is necessary in order to be blamed or praised for our actions. This is because, ultimately, our actions were not caused by *us*, but by events over which we had no control, which does not leave room for ultimate moral responsibility.

In this section I will argue that, in order to deserve punishment in the basic desert sense, we must have the kind of *ultimate* moral responsibility and free will that is only acquired when our actions are not causally determined such that we could choose otherwise or that we are the ultimate cause of our actions (i.e., when either (1) or (2) is true). Moreover, I will show that we do *not* have free will or moral responsibility in this way, regardless of whether determinism or libertarianism is true. To do this I will argue that we always, in every circumstance, lack the kind of control that is necessary for moral responsibility. I will then show that without this sort of moral responsibility, we cannot deserve to be punished, which ultimately undermines any justification of punishment that is ultimately concerned with the harm that a wrongdoer deserves in the basic desert sense.

Suppose that determinism is true. In that case, literally all of our actions are causally determined, including the brain state that makes you think you want to do something, so that when you act, it is only because of those preceding causal events. For example, suppose an individual wants to rob a bank. The desire to do so, if determinism is true, is entirely caused by factors outside of her control so that, even if she were to go back in time a hundred times, those

factors would still influence her to want to rob a bank. In this way, she cannot possibly desire anything else, since those factors will always cause her to want the same thing. As a result of this, it seemingly does not make sense to blame her for wanting to rob the bank, since, in any meaningful way, it is not her fault that she wants to do so. That is to say, she is not the ultimate cause of her own desires because she cannot control the factors which influence them, and thus, she cannot be held ultimately morally responsible for having them.

On the other hand, if indeterminism is true and we have libertarian free will, then our wants and desires are not caused by previous events and our actions amount to nothing more than random chance. One common argument in favour of libertarian free will is to suggest that quantum mechanics shows that not everything is causally determined. Instead, they say that all events have fundamentally stochastic outcomes that are probabilistic instead of certain. Without committing them to mind-body dualism, then, the libertarian can claim that this leaves room for an individual to act on decisions that are not causally determined. This, too, is insufficient control (or *ownership*) of the outcome of our actions in order to have moral responsibility. Even if there is an element of randomness in our actions, it does not leave room for our wills to act independently of outside factors; since, if it is the case that we are able to form desires or the will to act in isolation of outside factors, those desires would amount to nothing more than a random (or probabilistic) outcome that are not based on anything that we have actual ownership over. In this way, even if the world is *not* causally determined, there is no point in which free will is possible.

Still other free will sceptics claim that we lack free will and moral responsibility because of the pervasiveness of luck (Levy 2011). That is, we cannot have free will because, no matter if everything is causally determined or not, the vast majority of the circumstances we find

ourselves in, our ability to act, and our desires are the result of either good or bad luck. And since almost everything (or, indeed, everything) is the result of luck, then we cannot say that an individual deserves blame or praise for her actions in the basic desert sense – i.e., to morally deserve a punishment in and of itself, regardless of the consequences or positive outcomes of the punishment – since she cannot bear full or ultimate moral responsibility for her actions.

To briefly discuss these implications for the justification of punishment, let me return to the compatibilist position. As I discussed above, the compatibilist is comfortable with saying that the world is causally determined, but claims that we have free will just in case we are able to act in the way that we desire. In this way, even if determinism is true, we can still say that we have "free will" and "moral responsibility" for our actions, but only because we have moved the goalposts to include *degrees* of responsibility where we never have *complete* or *ultimate* control over our actions, but enough to say that we have *some* responsibility for our actions. The obvious question, then, is to what degree can an individual be morally responsible for her actions? If she is able to exert, say, control over 90% of the factors which ultimately cause her to act in a particular way, then we might say that she *deserves* most of the responsibility for her actions. However, we have good reason to believe that she controls *very little* (or, more likely, none) of the factors which eventually cause her to act in particular ways. For example, the when I threw a ball at the catcher, I had no control over the fact that he hurt my child, nor the fact that him doing so made me angry. So how much control did I have over my actions? I would say none at all, once literally everything that went into determining the action had been accounted for.

Now, in terms of moral responsibility, it seems that in every case there is very little reason to think that we ever have ultimate control over our actions. For that reason, then, we

should also have very little reason to think that we have ultimate moral responsibility for our actions. Therefore, any theory of punishment which requires moral responsibility for an action in order to justify the punishment will necessarily fail to accomplish this goal if the free will sceptic's position is correct.

The retributivist theory of punishment, of course, holds that a punishment is only deserved when a wrongdoer bears moral responsibility for her immoral action. And, since she can only bear moral responsibility if she were able to use free will, but – as we already discussed – since there is very good reason to believe that she *does not* have free will in the relevant way, then she cannot have the requisite moral responsibility for a deserved punishment. Therefore, if the free will sceptic is correct, individuals cannot deserve punishment in the basic desert sense and retributivist punishment cannot be justified.

The deterrence theory, on the other hand, does not require the moral responsibility of a wrongdoer in order to justify punishment, so it is compatible with the free will sceptic's position. The moral education theory, too, must be compatible with this view, since punishments are only "deserved" in the sense that wrongdoers *require* the punishment in order to teach them moral behaviour and to express the wrongness of their actions to others. Thus, the free will sceptical argument, if convincing, harms the justification of the retributivist theory, but does not affect either the deterrence theory or the moral education theory.

3. Retributivism

In chapter 1, I briefly described how both Moore and Zaibert's arguments claim that a punishment need not have good consequences in order for it to be morally good. Each of these arguments presents a view of retributivism which is intuitively appealing; it provides an intuitive

case for why it is morally acceptable for state-sanctioned, retributive punishment. In the next few pages, though, I will argue that we have strong reasons for believing there is no intrinsic value in harming wrongdoers by providing a counterexample to the Fraternity Boys example from chapter 1.

The bare bones of Zaibert's justification for punishment is this: The punisher must believe that a person is deserving of blame for her action.[8] The punisher then does something which she believes will be painful[9] to the blameworthy person in an act of (retributive) revenge. When the harm inflicted is correctly administered in the amount deserved, an organic whole which is intrinsically good is produced. When the organic whole is intrinsically good, the punishment is therefore justified – provided other confounding factors do not exist that would make the punishment unjustified/inappropriate.

One way to attack this justification of punishment – which I think ultimately fails, but is worth mentioning – is simply an epistemic scepticism. There does not seem to be any principled manner in which a punisher could be sure that an organic whole with positive intrinsic value had been produced by punishing a wrongdoer, either because it is impossible to know the wrongdoer was really responsible for her actions, or that it is impossible to know if the correct (retributive, proportional, appropriate) amount of punishment had been administered. At best, we might accept that a good had been produced in the rare case where both the punisher and the wrongdoer felt that the punishment was justified. Or, perhaps less satisfyingly, in the event where the punisher is *pretty sure* that the wrongdoer was justly punished.

[8] i.e., the punisher believes that the individual has done something wrong and that she ought to feel guilty for her action – even if she does not feel guilty.

[9] "Painful" could also be substituted for harmful, unpleasant, etc. It does not necessitate the type of harm involved either; this could include mental, physical, emotional, or financial pain, among others.

It is entirely possible, though, that the retributive theorist might accept that *in practice* we could never be certain that an intrinsically good organic whole had been produced, but that does not undermine the fact that there would, in theory, be justified instances of retributive punishment. More interesting, in my opinion, is how we might argue that that retributive punishment fails to guarantee the intrinsic value of the punishment, and therefore cannot be a necessary or sufficient condition in justifying the punishment.

To do this, let us first consider why we might think that retributively punishing an individual will necessarily produce an organic whole that is intrinsically good. Zaibert thinks, and most other retributivists who hold similar views would agree, that this is the case because getting what one *deserves* is intrinsically good. On the face of it, this seems intuitively plausible simply because of our understanding of the word "desert." If I get what I deserve then the world seems to be functioning as it should since I received exactly what I was owed. We seem perfectly satisfied when a regular transaction between individuals works this way: if the owner of an object is paid that object's value in order to part with it, then both parties should surely be satisfied with the exchange. It seems, then, that the intrinsic value of being fairly paid your due is good overall.

But this intuition does not seem to get us very far when we ask *why* it is intrinsically good, and if it is necessarily good in all cases. For example, is it necessarily good even in cases where getting what one deserves is, on its own (i.e., in isolation – an *organic part*), an intrinsically bad thing, such as the harm that is caused to an individual from a punishment? We might assert that it is a good thing, but explaining its goodness by appealing to the intrinsic goodness of getting what one deserves is simply question begging. Justifying a punishment on the basis that the wrongdoer is "getting what she deserves" is placing all of the normative force

on the intrinsic goodness of desert, while the justification for the intrinsic goodness of desert is merely a language trick in which the word "desert" sneaks in its own intrinsic value.

Consider, then, the intrinsic value of desert and the *kind* of desert the retributivist is talking about. The kind of desert which can be justified by appealing to other things, such as the consequences of an action or the institutions in place which provide rules that must be followed, cannot be what we are talking about. If, for example, we claimed that the fraternity boys who party outside of a prison where executions occur deserve to be punished because it will bring about desired consequences such as the deterrence of future partying, then we are not talking about the right kind of desert. In those cases, we could simply discuss desert as a utility maximizing function, ignoring the intrinsic value of the thing itself in favour of utilitarian or consequentialist considerations. Rather, we should be more interested in the *basic desert* described by Pereboom: "the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations" (Pereboom 2014, p.2). This kind of desert is what the retributivist has in mind for justifying punishment, so it must be shown that when this kind of desert is fulfilled it is always intrinsically good.

The intrinsic value in having a basic desert fulfilled (that is, to receive what is deserved), is supposed to be good in virtue of the fact that it is deserved. It seems, though, that there exist confounding circumstances where it does not make sense to consider "getting what one deserves" an intrinsic good, even as an organic whole. To show this, consider the following example:

> A young mother in a war-torn country must acquire enough food for herself and her small child to survive. There are no jobs available where she can work for food or wages, there are no options for hunting, gathering, or scavenging food. She has two (and only two) available options:

1. She can steal enough food for her family to survive (but not so much as to cause anyone else to starve), or
2. She can allow herself and her family to starve.

If we look at the mother's options in isolation (purely considering the actions themselves, regardless of context), we can see that both of her options are intrinsically bad. We ordinarily think that stealing is wrong (in and of itself), and also that allowing your family to starve is wrong (again, in and of itself), especially when another available option would prevent anyone from starving. It seems, then, that no matter which action she decides to take, she will be deserving of some kind of punishment due to whatever immoral action she chooses. We might say that the two deserve different punishments, but a punishment would be deserved nonetheless.

According to the theory, then, Zaibert will be forced to conclude that punishing the mother will produce an intrinsically good organic whole (i.e., the punishment is *good*). And while Zaibert uses the Fraternity Boys example to produce the intuition that punishing the deserving is good, it seems as though this example does the very opposite. It seems as though in the case of the mother, punishing her should have no place in our consideration of how we should treat her actions. Indeed, I think it would be very strange if we were to assert that the organic whole produced by punishing her, given the context, could be anything but intrinsically *bad*. We might say that the only thing the mother reasonably *deserves* is some food, but certainly not punishment. Nevertheless, Zaibert's description of the theory will commit him to asserting that a punishment in this case is intrinsically good.

Now, several things can be said in Zaibert's defense. First, arguing for the intrinsic goodness a deserved punishment, one might contest that the mother does not deserve, in the basic desert sense, to be punished because she is not culpable for her actions. One might claim that, since the mother did not have the ability to choose an option that was intrinsically good, she

cannot be blamed for choosing an option that was intrinsically bad.[10] Essentially, because she did not have the ability to choose a good option, she cannot be held accountable for doing something wrong. An essential part of deserving a punishment, after all, is the freedom to have wilfully chosen to do the immoral action. And since she did not have the freedom to choose the immoral action because it was the only *real* available option available to her, then she cannot deserve to be punished.

On Zaibert's view, however, this is not the correct assessment of the situation. In order for someone to have done something *intrinsically wrong*, she need only have had the ability to freely[11] choose to act immorally. Supposing the mother decided to (rightly, in my view) steal enough food for her and her family to survive, she would have been culpable for the intrinsically wrong action. Zaibert's view of organic parts and wholes views wrong actions in isolation of the context of the situation. As long as the individual acted with the relevant moral understanding of the intrinsic value of the action,[12] and was not coerced into choosing the action, then she is blameworthy for the action, and – since it is intrinsically bad – thus deserves to be punished for it.[13] And, ultimately, if the punishment is in the amount deserved, Zaibert will insist that an intrinsically good organic whole has been produced.

Since the mother was aware of the intrinsic badness of stealing, and she freely chose to steal regardless of that fact, then that is all that is required to say that she acted intrinsically

[10] We might ostensibly blame her *more* if she does not choose the least intrinsically bad action, but as long as there are no good options, the least bad option will still be bad, in and of itself.
[11] Here I mean "free" only in the thin sense that it is not coerced and the agent had the regular mental faculties to understand what she was choosing to do. If Zaibert required a thicker sense of freedom for desert, then it seems that he might quickly lose the ability to say anyone could *ever* deserve punishment given that everyone is constrained and influenced by their environment (this is the free will scepticism argument that is discussed in the next section).
[12] Zaibert also claims that it is possible for individuals to deserve punishment if they were not aware of the wrongness of their actions. For example, if they acted negligently. For my purposes, however, this is not relevant.
[13] Zaibert discusses blame on pages 31 and 32 of his book.

badly. A punishment, in the amount deserved, is then supposed to produce an organic whole that is intrinsically good. We can see, then, that this first defense of Zaibert's view does not succeed because Zaibert himself must deny it. Meanwhile I think we would still like to claim that it would *not* be good if the mother were punished for her action, despite what a basic desert (retributivist) theory suggests.

The second way that a retributivist might argue for the intrinsic goodness of punishing the mother, is that while punishing the individual for her wrongs in the amount deserved will produce an organic whole that is intrinsically *good*, it may still not be *right* to punish her due to other normative principles which we also hold (Zaibert p. 199). Those other principles, presumably, would include the fact that the mother has no better option available to her, making it wrong to punish her in that circumstance. In this way, though, Zaibert admits of a gap between the *good* and the *right* (Zaibert p. 215) which ensures that a retributive punishment can *never* justify a punishment by appeal *only* to the basic desert claim. Instead, it could only ever make the negative claim that innocent people do not deserve to be punished (which, as far as I can tell, is not a terribly interesting conclusion).

By claiming that it is always *good* to punish wrong actions, but not always *right* due to contextual considerations, punishment can never be justified without appeal to the positive consequences which would make it "right". In my example from above, then, the retributivist will claim that it is intrinsically good to punish the mother, but also *wrong* to do so (for hopefully obvious reasons). If the act of punishing is wrong, as it appears to be in this case, then the punishment is unjustified. This shows that there are clear instances where an individual might *deserve* to be punished in the basic desert sense, but punishing that individual for her actions would be unjustified.

This may not seem to be a particularly interesting conclusion, considering Zaibert is only attempting to justify a negative retributivist position: "to be a retributivist is not merely to claim that desert is a necessary condition for the infliction of just punishment, **nor quite to claim that it is a sufficient condition either**. To be a retributivist is to recognize that deserved punishment is an intrinsic good" (ibid., 214). However, it seems as though Zaibert provides us with an evaluation of desert that is unverifiable and unusable except to be able to apply the normative term "good" to, such that we can say that a punishment is "good" even when it is "wrong". In terms of how we ought to act in instances where the action is both good and wrong, Zaibert seems to recommend that we refrain from the action on the seemingly contradictory grounds that it is wrong to act in a way that is good.

While the intrinsic goodness of deserved punishment appears to be internally consistent, it does not seem to capture the sort of justification of punishment that retributivists are looking for when they say that a punishment is deserved. For instance, it does not justify punishment or tell us what to do in instances where all of the available options are intrinsically bad, or in situations where some necessary evil must be done for the greater good. Instead, it merely preserves the notion that some actions are morally wrong, while others are good, and supposes that it is good to punish individuals when their actions are wrong. But this amounts to mostly a language trick, since the action can be wrong while a punishment for that wrong may still be unjustified. For this reason, the theory seems to be an *argumentum ad absurdum* which warrants simply abandoning it in favour of some other justification.

The non-intrinsic goodness of deserved punishment, then, is another strong objection to the retributivist theory of punishment. Since there are instances where "getting what one deserves" is not inherently good, then it is not clear that punishing wrongdoers retributively will

also be *good*. Therefore, the retributivist cannot use the goodness of a deserved punishment as a justification for the retributive theory of punishment.

On the other hand, I must note that this criticism of a basic desert justification of punishment does not harm Dimock's justification of punishing retributively in order to restore trust in society. This is because her justification does not rely on the wrongdoer deserving punishment in the basic desert sense. Instead, her argument attempts to justify retributive punishment with the "by relation",[14] whereby the justifying state of affairs (restoring trust in society) happen *by* retributively punishing wrongdoers. The claim is that "the punishment and its justifying event/state of affairs begin simultaneously and are non-causally related" (Dimock 1997, 40). This view claims to be retributivist not because the punishment is deserved, but the punishment is required in order to *non-causally* bring about the desired trust in society. It still fails to be justified, however, for the reasons I discussed in chapter 1. That is, we may have very good reason to think that there are empirically better methods of restoring trust in society than by punishing wrongdoers. Indeed, in chapter 3 and 4 I argue that making prisons resemble life on the outside as much as possible, and harming prisoners as little as possible, has provided compelling evidence in being able to restore trust in society better than our current retributive penal systems.

4. Deterrence Theory

Briefly, I will mention three limitations that exist in the deterrence theory of punishment, but which are argued against by Rawls in his rule utilitarian justification of punishment. If

[14] For further details see Mark A. Michael, "Utilitarianism and Retributivism: What's the Difference?", *American Philosophical Quarterly* 29 (1992): 2.

punishment is purely meant to deter individuals, then these limitations will show instances where deterrence will not be an effective method of preventing crimes (or of providing a safe society). That is, some or all of the mechanics at play in deterring individuals will not be present, and thus, the potential wrongdoers will not be deterred by the punishment that is threatened for the wrongful act. While Rawls provides reasons for thinking that deterrence cannot provide the *only* justification of punishment, the limitations of a purely deterrent-based theory of punishment are useful in showing why we might think that deterrence cannot justify punishment on its own.

4.1 Limitations of the Deterrence Theory

First, if the punishment is insignificant compared to the crime, then it cannot rationally deter an individual. For example, if the punishment for stealing a car is a $5 fine, then a cost-benefit analysis will show that stealing the car will almost always be the rational choice. Similarly, if the punishment appears exceptionally unlikely to happen, again it would be rational to commit the crime. If I can steal a million dollars from a bank with only a one or two percent chance of being caught, in many situations I ought to choose to rob the bank (assuming my priority is my own self-interest without regard to any other moral considerations).

Those two limitations can, in principle, be corrected in order to deter effectively: increase the severity of the punishment appropriately so that the negative consequences outweigh the positive consequences of committing the crime, and ensure there is a strong public perception that if an individual were to commit a crime, it is extremely likely that they will be caught and punished.[15] Much harder to correct, however, is when the agent does not rationally consider the

[15] There is a relatively recent paper which suggests that the severity of the punishment actually has little to do with the effectiveness of the punishment as a deterrent; rather, only the certainty of whether or not the individual will be punished has an effect on deterrence (Wright 2010). I imagine, however, that the severity of the

negative consequences of committing a crime. In this situation, no amount of punishment will be able to rationally deter the individual since she is not rationally considering the consequences of her desired action or its alternatives. Crimes of passion – when individuals act in the heat of the moment, spontaneously, and without consideration for the consequences – for instance, may be instances where the agents involved do not have the ability to correctly identify alternative courses of actions, resulting in the agent "choosing" the only action which appears to be available to her. In these instances, if the agent has chosen rationally, then it must have been from a very limited list of possible actions. An outsider looking in on the situation would have been able to provide a longer list, but this is merely because the outsider was not constrained by the circumstances of the situation.

Since the deterrence theory is strictly utilitarian or consequentialist, to be able to justify punishing individuals who do not have a significant amount of control over their actions, the punishment must be able to deter others from committing similar crimes (or the same individual from committing the same crime again). When a crime is committed without any regard to the consequences, however, the threat of a punishment cannot hope to deter that crime. In those situations, then, a punishment cannot be justified under the deterrence theory of punishment. This leaves us with the question of when, or even *if*, punishing those individuals will be effective for deterring future crimes of the same sort or when punishment should not be used as the source of deterrence.

While it may seem as though the above limitations do very little in limiting the scope of where punishing for the purpose of deterrence can be justified, it is not clear that we can ever guarantee that an agent will rationally determine that she should not want to commit a crime.

---

punishment must at least provide a base amount of deterrence (*more than a $5 fine for car theft, for example*) in order to be effective at all.

Indeed, it becomes an empirical claim about when punishment is the most effective deterrent compared to any other deterrent. There *might* be more effective methods of deterring future crimes than punishing wrongdoers. For example, making systemic changes to society which prevent the circumstances from arising that would result in those crimes being committed, might be a much more effective method at reducing the rate of crimes in a society than simply punishing everyone who commits a crime. Any time an alternative to punishment is found to be more effective than the punishment, it would, in principle, undermine the justification of punishment for all crimes committed under those circumstances. If the effectiveness of punishment as a deterrent becomes a purely empirical question, we might discover that punishment is *never* justified.

4.2 Mere Means Objection

Rawls shows that a rule utilitarian is able to argue in favour of the deterrence theory of punishment while taking care to show that a justification of the theory does not necessarily entail justifying the punishment of the innocent people, or punishment that is disproportionate to the crime. He shows that a deterrence theory of punishment is only concerned with punishing wrongdoers because it is useful as a deterrent of crime; but that individual instances of punishment require additional justification beyond merely deterring. The objection raised to this justification by retributivists and others, however, is that this treats wrongdoers as a means to accomplishing the larger goal of less crime (or whatever the goal happens to be) and neglects to respect the autonomy of the individual.

The concern stems from the claim that we have a moral duty to treat others with respect. Retributivists assert that this is a foundational part of morality, because, if we demand moral

respect for ourselves (which we do when we consider ourselves moral agents), then we must also owe that same respect to other individuals since they are no morally different from us in any relevant sense.[16] If an individual commits some wrong, with full knowledge of the relevant moral imperatives (i.e., she knows her actions were morally wrong), then the best way to respect her autonomy is to attribute her the moral responsibility which entails that she deserves the punishment for her actions. Yet the deterrence theory of punishment is not directly concerned with the wrongdoer's autonomy when justifying punishment. It is only interested in punishing in order to keep society safe, and to send a message to would-be criminals so that they do not commit any crimes. "Sending a message", in this way, is using the moral agent as a tool for accomplishing the goal of a safer society. This, as any good Kantian would say,[17] is an unacceptable way of treating another person because it does not respect their autonomy.

Since the deterrence theory does not treat persons with the moral respect that we claim they deserve, the mere means objection – if true – is quite damaging to the justification of the deterrence theory of punishment. It means that the deterrence theory is not able to express our disapproval of crimes *in the right way* because it treats people as a mere means to an end. Although using the rule utilitarian justification of deterrence from Rawls ensures that punishments are "fair" because they do not punish too harshly or punish the innocent, there appears to be something profoundly lacking in punishing purely for the greater good. That is, it still appears as though we are justifying a punishment merely so that the rest of us are better off without consideration for the autonomy of the individual. In this way, it appears that there might be a distinction between doing the *right* thing, and doing the *best* thing. Doing the *best* thing

[16] The wrongdoer is morally different from an innocent individual in the sense that she has done something morally wrong while the innocent person has not, but not in the sense that she is any less of a moral agent because of her actions than anyone else. For further discussion on this, see chapter 3.1.

[17] Kant says exactly this in *The Metaphysics of Morals* (1996, p. 105, 6:331 by Akademie pagination).

(even if it is done fairly) by producing the best outcomes, does not necessarily entail that we have done the *right* thing. And, for this reason, we might still claim that the deterrence theory of punishment has not been sufficiently justified.

5. Moral Education Theory

The moral education theory is also susceptible to the mere means objection. Since it is primarily concerned with punishing in order to teach the wrongdoer a moral lesson, or to express the wrongness of an action, it effectively *uses* the wrongdoer in order to accomplish this goal. While that goal may be for the wrongdoer's "own good", it does not consult the wrongdoer on whether or not they *want* something that will allegedly help them. Even if a punishment could help me better myself and society in numerous ways, I might still rationally prefer not to receive the punishment since it will harm me. For example, even if punishing me by restricting my access to the internet would ultimately cause me to realize that stealing was wrong,[18] I may still rightly prefer not to have my liberty restricted in this way. In this way, the moral education theory disrespects the autonomy of the individual in order to produce a desired result (i.e., uses wrongdoers as a means to an end).

Aside from the mere means objection to punishment for the purpose of moral education, the theory can be objected to in two other ways: it is not clear that punishments will *actually* teach moral education to wrongdoers, and in relevantly similar situations we do not believe that it is permissible to harm someone merely because it is for her own good.

[18] Presumably the restriction of my liberties would not itself cause the moral education, but rather the time during which I did not have access to the internet would allow me to reflect on the wrongness of my actions and therefore come to the conclusion that stealing was wrong.

In order to justify the punishment of wrongdoers for the purpose of teaching them moral imperatives, Jean Hampton makes an analogy of punishing children in order to instruct them in the difference between right and wrong. That is, a punishment is not given to exact retribution on a child, but in order to show her that the action was morally wrong. And since moral education is generally a worthy goal, we might think that a punishment which was successful in accomplishing this goal might be justified. There are, however, relevant differences between punishing a child for the purpose of moral education and punishing an adult for the same purpose. Namely, adults often already have a sufficient understanding of the moral norms of their society. It is not clear, then, that punishing them would produce the desired result of providing a moral education. Furthermore, if the adult *did not* have the relevant moral knowledge of the situation, then we are more likely to say that she is not morally responsible for the action and does not deserve to be punished than to say that she *must* be punished. For example, a person who drives drunk and kills a child who was playing on the side of the road already *knew* that driving drunk was immoral. A punishment does not seem any more likely to teach that moral lesson than already having to live with the fact that her actions caused the child's death. Or, if she truly had been morally incompetent to the point that she acted innocently, it seems as though a punishment would have little hope of producing the desired result. In any case, it is an empirical question whether such a punishment could hope to succeed.

To the second point, the moral education theory seems to recommend punishments that we would otherwise not think are justified. While a punishment might be for a drunk driver's own good to teach her that her actions were wrong, we might similarly force obese people to adopt a sustainable calorie deficit to improve their health. Being morbidly obese, surely, is bad for one's health, so forcing such an individual to diet would be better than not, since it would

improve the person's life overall. Yet there is no difference between forcing an individual to become more morally healthy than there is in forcing an overweight person to diet (Boonin 2008, 191). But we would not consider the state-sanctioned harm of an overweight person to be justified merely because it caused her to become healthier, so we should similarly not consider a moral education punishment to be justified. In either situation, the punishment is paternalistic in that it does not treat the individual with the autonomy and respect that she deserves.[19]

To conclude, the moral education theory seems to have similar objections as the deterrence theory. While it differs in the justification of punishment by ensuring that the punishment is for the individual's own good (which prevents punishments that are intuitively too harsh), it similarly fails to respect the autonomy of the individual. For this reason, it seems that we have strong reason to suggest that the moral education theory cannot sufficiently justify punishment on its own merits.

6. Conclusion

In this chapter I have discussed some of the reasons we have for thinking that the retributivist, deterrence, and moral education theories cannot justify punishment. The final tally for the three theories is as follows. The retributivist argument cannot be justified if we lack the moral responsibility required for basic desert. If the free will sceptical position is true, then we must lack this moral responsibility and retributive punishment cannot be justified. Positive retribution also cannot be justified (without utilizing additional justifications) if it can be shown that harming a wrongdoer is not always good in and of itself. Of the three theories, retribution best

[19] In this way, the paternalism argument is similar to the mere means objection; that is, in both cases the individual's autonomy is not respected because they are merely treated as a means to an end.

avoids the mere means objection, although there may be some reason for thinking it can punish too harshly. The deterrence theory, on the other hand, is perfectly capable of avoiding the free will scepticism and intrinsic goodness objections. It is very susceptible, however, to the mere means objection which allows for punishments that are too harsh, punishing the innocent, and the paternalism objection. Lastly, the moral education theory is also able to avoid the free will scepticism objection, and with some charity we can say the intrinsic goodness objection as well. Like the deterrence theory, though, it is susceptible to the mere means objection and seems to provide a paternalistic justification for punishment that we would object to under similar circumstances. As we can see, then, there are strong reasons for rejecting any or all of the theories of punishment from chapter 1. In the following chapter, I will introduce a fourth theory of punishment, the public health-quarantine model, and argue that it is not subject to any of the objections described here.

Chapter 3

1. Introduction

In the previous chapter I discussed three different justifications of punishment, arguing that we have strong reasons for concluding that none of them is able to convincingly justify punishment. In this chapter, I discuss a new theory of punishment, the Public Health-Quarantine (PHQ) model, and argue that it succeeds in justifying punishment where the others have failed. This relatively recent addition to the literature on punishment is an expansion of Derk Pereboom's model of punishment based on a quarantine analogy (2014) to place it "within a broader justificatory framework drawn from public health ethics" by Gregg Caruso (2016).

The key focus, or the main goal, of the public health-quarantine model of punishment, as Caruso describes it, is to motivate or prioritize steps which prevent individuals from committing crimes. That is, rather than assume society is functioning correctly and that there are some no-good wrongdoers trying to ruin the way we all get along, it instead asserts that there are currently systems in society which cause or greatly influence individuals to commit crimes. And, rather than blame the individuals who find themselves in circumstances where they must commit a crime, the PHQ model emphasizes that society should instead be working to remove those underlying causes. The PHQ model, therefore, says that any justified system of punishment for wrongdoers must focus its efforts on the public health, whereby it promotes the rehabilitation of wrongdoers as well as ensuring preventative measures are taken such as education and other social programs that remove the underlying systemic causes of crime.

In the following few pages I will outline Gregg Caruso and Derk Pereboom's arguments in favour of the public health-quarantine model of punishment. To do this, I will provide a brief overview of an important underlying assumption of the PHQ model: that we have strong reasons

to believe that we lack free will in the morally relevant sense that we cannot possess ultimate moral responsibility for our actions. From there, I will discuss in more detail the key aspects of the PHQ model of punishment, laying out its justification and inherent limitations. Then, in sections 3, I will argue that that the PHQ model provides a more appealing system of punishment than any of the forward-, backward-, or mixed theories of punishment that I discussed in Chapters 1 and 2.

2. Public Health-Quarantine Model of Punishment

2.1 Free Will Scepticism

In Chapter 2, I argued that one reason we may have for rejecting a retributivist justification of punishment, is that it seems as though our actions may be caused or influenced by factors outside of our control to such an extent as to preclude ultimate moral responsibility for our actions. That is, if it is the case that our actions are determined to the extent that we cannot act of our own free will, then we cannot deserve blame or praise for the consequences of those actions. And if we cannot deserve blame or praise for our actions, then we cannot deserve to be punished when we commit some morally wrong action. It would therefore be impossible to justify punishment in the basic desert sense (i.e., in the way that retributivists justify punishment), since "to hold them responsible in a non-consequentialist desert based sense would be to hold them responsible for the results of the morally arbitrary, for what is ultimately beyond their control, which is fundamentally unfair and unjust" (Caruso 2016, 26).

If the free will sceptic's position is to be accepted, a justification of punishment must not appeal to the notion of basic desert. The deterrence and moral education theories succeed in this regard, but as I argued in Chapter 2, there are independent moral reasons for rejecting those

justifications.[20] This seems to leave very little room for the justification of punishment. Indeed, Pereboom and Caruso argue that the standard alternatives to retributivism, deterrence and moral education, are not appealing on these grounds (Caruso 2016, 28-29), and that if we could hope to justify punishing wrongdoers, it must be for different reasons than any of those discussed in Chapter 1.

Pereboom and Caruso believe that, despite these arguments against, we can hope to justify punishment by using a model of punishment called the Public Health-Quarantine model. Both Pereboom and Caruso have argued independently (Pereboom 2001; 2014, Caruso 2016; forthcoming) and jointly (forthcoming) that we ought to deny basic desert claims on the grounds that we lack moral responsibility for our actions. They also agree, however, that it is neither pragmatic nor morally acceptable for the state to allow dangerous and harmful individuals to remain among society where they will actively make the lives of (other) innocent individuals unsafe. For this reason, then, they offer a different justification of punishment based on the right to self-defence, which I will discuss in the proceeding section.

2.2 Quarantine Model of Punishment

As I have described, both Pereboom and Caruso assume that, no matter if determinism or libertarianism is true, the world does not allow for the kind of free will that is required for moral responsibility. That is, they are hard incompatibilists, claiming that the stuff we normally associate with free will[21] is incompatible with both determinism and libertarianism, and

[20] i.e., the "using people as mere means" objection against the deterrence theory, and the argument that punishing criminals cannot be expected to teach moral values in the same way as punishing children for the moral education theory

[21] That is, we associate free will with the ability to have chosen otherwise in certain scenarios, or else to be the ultimate source of our actions.

therefore, punishment cannot be justified with an appeal to the individual's moral responsibility or to basic desert claims. Even if an individual commits a serious crime, that individual does not *deserve* to be harmed by punishment; indeed, the individual *deserves* to be treated in the exact same way as everyone else. However, Pereboom argues that we are still justified in punishing those who pose a serious threat to us because we have the right to self-defence, even in cases where this requires harming someone else. That is, we have the right to self-defence or the defense of others by threatening or harming individuals who pose a threat to us, but only in the minimum amount required to effectively deter the threat (Pereboom 2014, 166). It is not permissible to harm individuals beyond this right to self-defence because the one causing harm does not deserve to be harmed in the basic desert sense. Pereboom provides the following example:

> Suppose that someone clearly aims to kill you, and that to prevent his doing so you may knock him out with a baseball bat. You may then threaten him with this amount of harm. Suppose he does attempt to kill you, but in the process he trips over the toys on the floor, and this allows you to pin him to the ground and tie him up. At this point is it still legitimate for you to knock him out with the bat? To do so would not be justified by the right to harm in self-defence. (Pereboom 2014, 168)

While Pereboom's example gives a scenario for the amount of harm that is justified in protecting oneself from the immediate threat of an aggressor, the same is not true for an individual who is already in the custody of the law (Pereboom 2014, 169). Here, Pereboom argues that only incarceration of the aggressor, and no further harm, can be justified after the aggressor has been detained. To show this, Pereboom draws an analogy between the right to self-defence and the quarantine of individuals with dangerously infectious diseases such as tuberculosis. In those instances, he says, we have the right to separate those who are carrying communicable diseases so that they do not come into contact with others and spread the disease throughout a community. Similarly, if someone poses a serious threat to society's safety by

threatening to commit a murder (or other dangerous and harmful crime), then we have the right to incapacitate the individual such that the threat no longer exists. The way to do this *and only this*, without causing additional harm, is to incarcerate the individual. In this way, the least amount of harm is caused to the individual while still removing the possible threat to society.

The free will sceptic cannot, however, endorse the incarceration of anyone who commits non-violent or "victimless" crimes. Since the individual does not pose a threat to the safety of society, it would not be justified to harm the individual. Therefore, crimes that are currently punished quite harshly around the world that are not harmful to individuals, such as drug possession, would not merit quarantine of the individual. At most, these sorts of crimes might call for monitoring, counselling, or some other form of behaviour correction that in no way takes away the individual's liberties. In this way, the quarantine model of punishment that Pereboom suggests would provide a proportional system of punishment where the harshest punishment that could be justified would be incarceration, while the weakest would be no punishment at all.

It is also important to note, however, that while the right to self-defence cannot justify harming wrongdoers more than strictly necessary, and cannot justify harming the innocent, one *is* able to provide a threat of harm in order to deter would-be wrongdoers. For example, the state can explicitly state (i.e., threaten) that the punishment for bank robbery is incarceration. In regards to quarantine, the "threat" from the state is also made explicit: if you contract a dangerously infectious disease, you will be quarantined until such a time as it can be determined that you are no longer a threat to society. This threat does not apply to anyone who does not pose a threat of spreading the disease, but can be acted upon in order to protect the safety of others.

The last thing I would like to discuss before I describe Caruso's extension of the quarantine model of punishment to include the public health, is one seeming way in which we can justify punishing the innocent by using the quarantine analogy. That is, when we *suspect* that an individual might be a carrier of an infectious disease, we still think it is justified to quarantine her until we can be sure that she no longer poses a threat even though she is effectively "innocent". Similarly, then, it might seem as though we could incarcerate any individuals who might potentially cause harm to society, even if they have not yet done so. For example, if an individual were to seem exceptionally *likely* to commit a crime, we might think that it was absolutely imperative to incarcerate the individual (even if we were mistaken). Notably, this would seem to expose the model to serious concerns such as racial or ethnic discrimination, where certain people of a particular race or ethnicity are deemed to be threatening and are "justifiably" incarcerated, even when the threat is purely imagined due to racist tendencies.

To answer this concern, I think we can bite the bullet – albeit very softly. First, the quarantine analogy does suggest that we can be justified in quarantining those who are very likely a threat, even if it is the case that they would not have *actually* caused any harm. For example, an individual who speaks openly about inciting violence on a particular group of people could reasonably be detained even if she never intended to harm anyone. In that case, it would seem reasonable and justifiable to incarcerate the individual, even if it is only for the purpose of establishing the individual's intentions, mental state, etc. If nothing else, it would help to make society feel more safe, knowing that a potential threat to their safety had been eliminated. However – and, I think, this should be emphasized – someone who obviously appears to be a threat to everyone is not an edge case, and does not deal with the concern for

discrimination and prejudice, especially among law enforcement. For those cases where the methods for predicting whether or not one is likely to be a violent criminal are objectionable (e.g. by racial profiling), detainment should not be justified. Pereboom suggests that:

> To avoid this problem, it seems that invasive preventative measures should be restricted to those who have committed crimes.[22] The right to liberty should count heavily here. This right would yield strong reason not to detain someone even if there were some reason to believe that he is likely to commit a crime. (Pereboom 2001, 176-7)

That is, and individual's right to liberty should overrule any reason to believe an individual might be a threat unless that threat is immediately obvious and imminent. In this way, reasons such as skin colour or choice of clothing could never pass as a legitimate justification for detainment of an individual who has not yet committed a crime. This means that the analogy with dangerously infectious disease may not be perfect, but is still similar. While it may be justifiable to issue a blanket quarantine to everyone, indiscriminately, who may have come in contact with an infectious disease, it would not be justifiable to quarantine everyone who has expressed a dissenting view, disagreeable opinion, or *heard* an incitation to violence. For the most part, though, when quarantining people who are a risk for spreading a communicable disease, only individuals who pose the highest risk for carrying the disease are detained, and a similar (although perhaps not exactly parallel) assessment could be used for the incarceration of dangerous individuals.

[22] It is not clear to me if Pereboom suggests detaining individuals at the first sign of threat if they have a past history of being incarcerated for committing crimes, or if he is suggesting only detaining people after they have attempted to cause harm (in which case I'm not sure if it could still be considered *preventative*). If it is the former, some caution would have to be taken to ensure a criminal record did not immediately target that individual for being detained at the slightest provocation; while if it is the latter, then some kind of measures might have to be put in place so that people who are disposed to causing harm (and have a history of causing harm), could not easily cause harm without being stopped in some way. The details of this have not been laid out, and additional discussion would likely be helpful in determining what kind of preventative measures are necessary and justifiable. It seems that, as a start, always erring on the side of caution (i.e., scepticism about the likelihood of harm) seems to be the morally correct route.

2.3 Public Health

Pereboom's view is narrowly focused on justifying criminal punishment, while Gregg Caruso expands this view to place it within a broader framework of public health ethics. Like Pereboom, Caruso is a free will sceptic and therefore rejects retributivism and basic desert on the grounds that we do not have ultimate moral responsibility for our actions. He also agrees that the quarantine analogy provides the best justification for incapacitation. In order to expand on the quarantine model, however, Caruso argues that by placing the quarantine model in a framework of public health ethics, it "will not only provide a justification for the incapacitation of dangerous criminals but it will also provide a broader and more comprehensive approach to criminal behaviour generally" (Caruso 2016, 31). That is, while the quarantine model provides us with an account of punishment which justifies causing the least amount of harm necessary for self-defence, Caruso expands on this to prioritize preventative measures and provide us with "a more detailed set of principles for resolving the conflict between individual liberty and public safety" (Caruso 2016, 31).

As well as justifying the punishment of individuals, Caruso draws from the traditional medical ethical approach in order to emphasize autonomy, beneficence, nonmaleficence, and justice (see Beauchamp and Childdress 1989). The goal of this model is to provide a method of dealing with dangerous criminals which promotes the health of society as a whole (i.e., the public health). Caruso describes public health as containing four unique characteristics (Caruso 2016, 34):

(1) It is a public or collective good;

(2) Its promotion involves a particular focus on prevention;

(3) Its promotion often entails government action; and,

(4) It involves an intrinsic outcome-orientation

First, In regards to punishment, the public or collective good is simply ensuring that the goal of the criminal justice system (e.g., safety, security, justice, etc.) is maximized across the entire public. The needs of the individual, then, are outweighed by the needs of society such that safety, security, and justice are maximized for everyone, and not just particular individuals.

Second, A public health model of punishment would also focus on preventing crimes and criminal behaviour before it happens. This is preferable to merely incarcerating individuals after they have committed a crime, not only because it means less crimes are being committed (making society safer), but also because it reduces the burden on society. For example, less taxes would be required to house and feed criminals if there are fewer criminals, thus providing economic incentive. The public health model therefore recommends eliminating (or alleviating) systemic disadvantages which are typically causes of crime. It therefore makes the prevention of crime the primary concern for the criminal justice system. Essentially, punishment (i.e., incarceration) is only required as a last resort when preventative measures have failed. If, on the other hand, preventative measures are a success, individuals will not commit crimes because they have no reason to; the preventative measures would have removed systemic disadvantages, provided help for mental illnesses, and other underlying causes of crime such that there are no longer incentives, causes, or desires to commit crimes.

The third point exists to note that those preventative measures do not yet exist; that governments must take action to provide the help that is required to remove the underlying issues which cause crimes. And, furthermore, it must provide the help that is necessary to those who have committed crimes so that they are no longer a threat to society and are able to rejoin society

to continue their lives. Since the quarantine model does not permit harming the individual any more than absolutely necessary, this also means that wrongdoers should not be treated in the way we treat criminals today (badly), but instead they should live as close to their regular lives as possible except that they are unable to physically be part of society.

Lastly, a public health model of punishment must have a focus on considerations of social justice and fairness. That is, the public health model of punishment must promote autonomy, beneficence, nonmaleficence, and justice among the public in order to ensure that the criminal justice system is maximizing the public health. To do this, it must ensure the autonomy of the individual is infringed upon as little as possible – only when the individual poses an immediate threat of harm; that it benefits everyone as much as possible; that it harms individuals as little as possible (i.e., incapacitation is only used minimally in order to ensure the safety of society); and treats everyone equally and fairly.

These sections have given a brief overview of how the public health-quarantine model recommends the state ought to deal with the criminal behaviour of members of society. In the next sections I will discuss why the PHQ model should be preferred to the ones discussed in chapter 1 and 2. Briefly, I will discuss how the PHQ model provides an alternative to retributivism that does not rely on basic desert, but still provides a proportional system of punishment similar to what the retributivist position hopes to achieve. I will then discuss how Caruso's model of punishment prevents the mere means objection that we saw in a purely consequentialist deterrence theory, as well as some optimism for the success of the PHQ model in comparison to a moral education theory.

3 How PHQ Improves on the Other Theories

3.1 Retributivism

In my discussion of free will scepticism, I noted that the PHQ model of punishment must necessarily deny the retributivist justification of punishment since it denies the ultimate moral responsibility needed for basic desert. Any justification that hopes to succeed, then, must not appeal to a notion of basic desert. Possible alternatives include the moral education theory, deterrence theory, the right to harm in self-defence, and an incapacitation theory (Caruso 2016, 28). In the following two sections I will discuss why the PHQ model of punishment (which justifies punishment through the right to harm in self-dense and as an incapacitation theory) should be preferred to both the moral education theory and deterrence theory. For this section, though, I will discuss why the PHQ model is able to maintain the main appeal to the retributivist justification of punishment: namely, the respect for autonomy.

The retributivist argues that the best way to respect the autonomy of individuals is to treat them as though they are indeed responsible for their actions. This means that if an individual commits some wrong, with full knowledge of the relevant moral imperatives (i.e., she knows her actions were morally wrong), then the best way to respect her autonomy is to attribute her the moral responsibility which entails that she deserves the punishment for her actions. The retributivist argues that if we want to deserve praise for our good actions, then we require moral responsibility in the basic desert sense, which necessarily entails that we *also* deserve blame for our bad actions.

By denying moral responsibility, we seemingly undermine the justification for praise or blame and therefore also deny the autonomy which we desire. To a certain extent, the free will

sceptic must accept that autonomy in the sense described is not a priority, since it denies that autonomy can exist in this way. However, the *dignity* of the individual can still be respected in much the same way.[23] First, we can still have proportional punishments so that wrongdoers are treated more harshly for more harmful crimes, while more leniently for less harmful crimes. Second, we still treat individuals with respect inasmuch as they are individuals who cannot be used as a mere means to an end.

Even though the PHQ model takes a free will sceptical position on moral responsibility, it does not mean that individuals are merely being used as a means for deterring further crime or for the purpose of keeping society safe. That *is* a goal of the PHQ model, but it also provides the guarantee that individuals are treated with respect and allowed to have as much autonomy over their own lives as possible (with the obvious limitation that they are not able to physically engage with society). For example, the PHQ model would explicitly exclude any punishment which restricted the individual's ability to pursue their own goals, such as handling their own finances, participating in discussions, voting, etc. In this way, the PHQ model succeeds in respecting the autonomy of the individual as much as possible, while allowing for the regrettable but necessary infringement on their liberty to ensure the safety of society.

The proportionality of punishments, as I discussed in the quarantine section (section 2) of this chapter, is also able to provide a principle of proportionality which recommends punishments that are proportionate to the danger being posed by the individual. While the minimum punishment that could be justified might be something like a small fine for harmless crimes (for example, running a stop sign, or possessing a small amount of an illegal drug), the maximum punishment, of course, would be life imprisonment for the wrongdoer. The way it

[23] Caruso argues that this respects the individual's dignity even more than the retributivist would agree to (forthcoming, 5-12)

respects the individual, however, is to ensure that the punishment is only given in the amount that is required to rehabilitate the individual so that recidivism did not occur. Thus, the individual is shown respect by only punishing until the individual is rehabilitated, at which point we respect the individual enough to trust that she will be a functioning member of society. For example, Caruso (forthcoming, 12) provides a possible scenario:

> Consider again the hypothetical scenario used in the Shariff et al. study. The fictional case involved an offender who beat a man to death but after serving two years in prison was nearly 100% effectively rehabilitated. The case further stipulated that "the prosecution and defense had agreed that the rehabilitation would prevent recidivism and that any further detention after rehabilitation would offer no additional deterrence of other potential criminals" (Shariff et al. 2014, 4). On [the public health-quarantine] model, it would be unjust to continue to incapacitate this individual.

In this scenario, the PHQ model maintains that individuals who no longer pose a threat to society should not be incarcerated. This treatment of the individual greatly respects her dignity by showing that the state trusts her to be a functioning member of society. In this way, while it is not the case that she is punished in the amount *deserved*, but exactly in the amount that her dignity demands by allowing her to rejoin society once she no longer poses a threat of harm to others. Therefore, the PHQ model is still able to provide a "deserved" punishment, although not in the basic desert sense.

3.2 Consequentialism / Deterrence Theories of Punishment

The issues for the deterrence theory of punishment that I discussed in chapter 2 were these: deterrence may not be effective in certain circumstances, especially when people are less likely to rationally weigh the consequences of their actions; deterrence can, in principle, justify

punishing the innocent or punishing unfairly;[24] and using people as mere means to an end without respect for their autonomy. In my discussion of the public health-quarantine model of punishment above, I hinted at the way in which it can avoid these issues. In this section, however, I will provide a more detailed account of the issues and how the PHQ model is preferable to the deterrence theory.

One clear difference between the PHQ model and the deterrence theory, is that punishment is not justified solely on the basis of whether it deters future crimes. While the primary goal of the criminal justice system is to prevent future crimes, the actual punishment of individuals is not *because* it provides a tangible threat to others who are considering committing a crime; instead, its purpose is to ensure the safety of society. The punishment may deter future crimes because of the threat of incarceration, but it is not the primary goal of punishment, nor what provides its justification. In this way, then, one of the concerns for the deterrence theory is overcome by the PHQ model. While we may have legitimate concerns for the efficacy of using punishment as a deterrent (e.g., when individuals commit crimes of passion), the PHQ model is not concerned with when the punishment is an effective deterrent – it is primarily concerned with removing the threat to individual's safety. If, in a fit of rage or passion, an individual attacked someone, the correct response would always be to remove the threat from society, without any regard to whether or not that would effectively reduce future crime.

In my discussion of the quarantine model above, I briefly discussed how the use of punishment cannot be justified when it is in excess (i.e., goes beyond incapacitation) or used against the innocent in order to deter. That discussion is directly related to this concern for the deterrence theory where it seems as though, in order to deter most effectively, we might be

[24] i.e., punishing some criminals more than others simply to make an example of them in order to deter future crimes.

justified in punishing the innocent or punishing harshly so that others are aware of the negative consequences for committing a crime. In the PHQ model, however, there are very strict upper bounds for the amount of punishment that can be justified: only as much as necessary to remove an immediate threat to society. This means that, while the death penalty is obviously effective at removing a threat to society, it cannot be justified because it does not follow the principle of least infringement. Instead, after the aggressor has been detained, the threat has been removed as long as the individual is no longer able to cause further harm to society.

It might be the case that the only way to stop an immediate threat to someone's life is to take lethal force, so the PHQ model is able to justify a use of force in this way, but after the treat has been neutralized by whatever means are necessary at the time, then no further punishment can be justified. As I discussed in the example above, if it is possible to remove a threat of harm by tying someone up, then after this has been done it can no longer be justified to beat that person with a baseball bat. In this way, the worry for the deterrence theory that excessive force can be justified does not exist with the PHQ model since only the minimum amount of harm is ever allowed in order to ensure the safety of society. While the deterrence theory deters wrongdoers by any means necessary including the use of force to intimidate, the PHQ model can only use the *threat* of incapacitation as a deterrent from committing the crime. The actual harm itself cannot be justified unless it is preventing further harm, and since the effectiveness of the harm as a deterrent does not play into it its justification, it can never be used excessively or on the innocent.

Lastly, the PHQ model provides a way to address the "mere means" objection that retributivists argue makes the consequentialist deterrence theory unjustified. That is, the deterrence theory of punishment sometimes justifies harming individuals merely as a means to

deterring others or "to provide credibility for a system of threats" (Pereboom 2014, 169). The retributivist claims that "using" people in this fashion is unjustified because it does not respect the autonomy of the individual. That is, since people are not merely objects to be used as tools for keeping society safe, then we should respect the principle of autonomy[25] and not treat them as such.

Two things can be said about the principle of autonomy in regards to the PHQ model of punishment. First, that the principle is not the primary concern and is therefore sometimes infringed upon, but second, that considerable weight is given to the principle of autonomy so that it is infringed upon as little as possible and only when it can be justified by a more general principle of justice that we think to be true. To the first point, it is obviously the case that there are regrettable circumstances where the state is forced to infringe on an individual's right to freedom and liberty in order to protect the safety of society. For example, a serial killer who outspokenly plans to kill again should not be allowed to remain at liberty to do so by the state. In this sense, the PHQ model allows for the infringement of the individual's autonomy, however, it only does so in the interest of others, and only when the threat to others' safety is serious enough to warrant it. It would not, for example, be justified in incarcerating individuals who commit victimless or harmless crimes such as running a stop sign when there are no other cars on the road. In situations where there is no threat to anyone's safety, "punishments" that did not infringe on the individual's autonomy would be more appropriate, for example, a small fine or other non-invasive punishment. In that way, the individual's autonomy is respected as much as possible, whenever possible, provided that individuals' autonomy does not threaten the

[25] As Caruso describes the principle of autonomy in terms of public health ethics: "places primary emphasis on the liberty, privacy, and informed consent of individual persons in the face of a health intervention carried out by other parties. It acknowledges a person's right to make choices, to hold views, and to take actions based on personal beliefs" (Caruso 2016, 37).

wellbeing of anyone else. In chapter 4, I further discuss the requirements for respecting the individual's autonomy and how the PHQ model can ensure that it is infringing upon it as little as possible.

3.3 Moral Education Theories of Punishment

In chapter 1 I explained that the moral education theory of punishment justifies causing harm to individuals based on an analogy of punishing children. That is, we do not punish children in order to give them what they deserve, but instead to teach them a moral lesson. For example, we may censure a child's behaviour when she steals from a sibling in order to show her that such behaviour is morally wrong. Since this justification of punishment does not rely on the moral responsibility or the basic desert of the wrongdoer, a free will sceptic may potentially be able to accept this justification (Caruso 2016, 28). As I discussed in chapter 2, however, it is not clear that punishment is as effective at teaching morals to adults as it is to children. For this reason, there appears to be empirical concerns for the justification of punishment in the moral education theory.

The PHQ model of punishment does not justify punishment only if it is able to teach moral knowledge to the wrongdoer – it is only justified when it minimally harms an individual to ensure the safety of society – so, in that case this worry is not a concern for Caruso and Pereboom's justification of punishment. Taken more broadly, however, is the more general concern that the PHQ model of punishment can be reasonably expected to actually succeed in providing preventative measures, preventing harm to society, and rehabilitating individuals effectively without harming them more than necessary.

The success of the PHQ model, of course, relies heavily on the structure of the criminal justice system under a well-justified theory. The benefit of this model over other theories (including the moral education theory) is that it prioritizes the preventative measures to reduce crime *before* they occur, rather than merely waiting for individuals to commit crime and then punishing them harshly. Caruso provides the following comparison of the PHQ model of punishment with public health ethics:

> The primary function of [public health] agencies is to *prevent* disease, food borne illnesses, environmental destruction, injuries, and the like. A non-retributivist approach to criminal justice modeled on public health ethics would similarly focus on prevention … Instead of focusing on punishing criminals and building more supermax prisons, the public health model would advocate addressing the systemic causes of crime, such as social injustice, poverty, systemic disadvantage, mental health issues, and addiction. (Caruso 2016, 33-34)

While the moral education theory asserts that punishment can be justified when it teaches moral behaviour, the PHQ model judges success on crime rate and recidivism rate. So, while the analogy between punishing children and punishing adults seems to fall apart in theory (as discussed in Chapter 2), the PHQ model can succeed in theory and merely requires a correctly structured and focused criminal justice system in order to succeed in practice. There is good reason to believe, therefore, that the PHQ model is capable of providing a theory of punishment that is both justified and likely to work in practice. In the next chapter, I will discuss further reasons to believe the PHQ model will succeed in practice, where I provide several examples to help showcase this.

4. Conclusion

In this chapter, I have introduced the public health-quarantine model as an alternative to the theories that were discussed in chapter 1 and 2. I have shown that the PHQ model is able to

justify the punishment of harmful and dangerous individuals when it is for the explicit purpose of protecting society, and only by causing the least harm necessary in order to achieve that goal. I argued that, by justifying punishment only under those specific conditions, it avoids worries about unjustly harming innocent people or excessively harming certain wrongdoers. It also provides us with a reasonable expectation of ensuring the safety of society, while respecting the autonomy of individuals and providing proportional punishments. In the next chapter I will further argue that the PHQ model is able to provide an appealing response to the "mere means" objection. I will also show that the PHQ model ought to be interpreted in a way that takes the health of the wrongdoer seriously, and that there exists a moral obligation for the state to compensate wrongdoers for unfairly incarcerating them. I will argue that this obligation can be met and I will provide several suggestions for how this can be done while showing that the obligation does not make incarceration so lenient as to be inviting.

Chapter 4

1. Introduction

In chapter 3, I outlined Gregg Caruso's free will sceptical model of punishment based on Derk Pereboom's quarantine analogy within a broader justificatory framework of public health ethics (Caruso 2016, 25). He defends a view of punishment from a position of free will scepticism; denying that we have ultimate moral responsibility for our actions and therefore cannot justify punishment from basic desert.[26] Instead, Caruso proposes a model of punishment which focuses on the public health, while still being justifiable (i.e., does not punish too harshly or in another way that cannot be morally justified). To do this, Caruso provides a model of punishment which justifies removing dangerous and harmful individuals from a society for the overall benefit of that society. This model of punishment focuses on an overall, public health, while taking care to consider that individuals do not have ultimate moral responsibility for their actions.

I have argued that the public health-quarantine model of punishment gives us the best justification of punishment so far; or at least that it is preferable to the alternatives I discussed in Chapter 1. For that reason, the aim of this chapter is not to criticize the PHQ model of punishment, but instead to explain and clarify the obligation that exists to improving the wrongdoer's health in order to justify a punishment in the PHQ model. The way Caruso has currently presented the PHQ model of punishment, the primary concern may be seen to be the public or collective good at the regrettable but necessary expense of the individual wrongdoer. This interpretation of the model would open it to the "mere means" objection raised against the deterrence theory in chapter 2 in which individuals are treated as a means to an end instead of

[26] As Pereboom describes basic desert: "the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations" (Pereboom 2014, 2).

respecting their individual autonomy. I believe that such an interpretation of the model would be incorrect; rather that an individual's rights must be weighed heavily when she does not hold moral responsibility for her actions.[27] It is the aim of this chapter, then, to argue that there exists a moral obligation such that any state-sanctioned punishment must also compensate a wrongdoer by improving her circumstances after the period of incarceration.[28]

In order to show why Caruso's PHQ model ought to be interpreted in the way I am proposing, I will first provide a brief review of the PHQ model. Then, I will show that Saul Smilansky's objection (Smilansky 2011; 2016) to Derk Pereboom's quarantine analogy produces a tricky problem for the PHQ model of punishment that, I think, has not sufficiently been met by either Pereboom or Caruso.[29] I will argue that this problem can be met by interpreting Caruso's model in the way I have proposed above. More specifically, I will argue that Smilansky correctly points out that we owe a wrongdoer compensation for her incarceration, and that this compensation can be achieved with special attention to that individual's health. Following this, I will show that compensating wrongdoers in this way does not provide us reason to believe that any justifiable punishment must necessarily result in luxurious accommodations for a wrongdoer such that there is an incentive to commit crime, rather than refrain from it. To conclude, I will argue that the PHQ model is able to respond to the "mere means" objection by providing a focus

[27] Caruso's model takes as a starting point the assumption that we cannot assume an individual has free will in any meaningful sense, which ultimately results in the individual losing moral responsibility for her actions (Caruso 2016, 25-28).

[28] Here I say "improving her circumstances," but do not mean to imply just any sort of compensation will suffice. I might replace "circumstances" with "public health," except that this seems to conflate the public health of society as a whole with the health of the individual *within* that society, as a functioning member of that society. In any case, I go into more detail about what exactly is owed to a wrongdoer later in the chapter, so if the term seems slightly confusing now, it should become more clear shortly.

[29] Although both have provided responses (see: Pereboom 2014; Caruso, forthcoming)

on the health of the individual such that wrongdoers are not merely being treated as an object for improving overall public health, but also as autonomous individuals.

2. Overview of the Public Health-Quarantine Model

The Public Health-Quarantine model ensures that quarantine is only used minimally, in a way that is least harmful to the individual while still guaranteeing the safety of the public. At the same time, additional focus is placed on the prevention of future crimes; ensuring society is educated and aware of laws and moral norms, and removing or preventing social barriers and circumstances which conspire to place people into situations where their only option is to commit a crime. This focus places considerable attention on the needs of society. As Caruso puts it, public health is a "public or collective good" (Caruso 2016, 33). To a certain extent, this means that the needs of society trumps the needs of the individual (for example, when an individual is too dangerous to remain free in society because they will cause grievous harm to others).

The individual, however, never *deserves* this treatment in the basic desert sense. It must be made clear, then, that this cannot be pressed to extremes: it cannot be acceptable to take away an individual's liberty merely because that is what is best for society; it should not be acceptable, for example, to simply kill all wrongdoers, even if it would maximally alleviate the burden of wrongdoers from society (since, for example, it is no longer necessary to feed, educate, or rehabilitate a wrongdoer at the cost of tax dollars and economic resources). To ensure that this does not happen, it must be made clear that this is a morally unacceptable way to deal with wrongdoers. In the following few pages I will argue that it is possible to justifiably quarantine

wrongdoers in a way that respects the individual's autonomy while also taking seriously the need for health and rehabilitation of the wrongdoer.

3. Obligation for Wrongdoers Who Are Incarcerated/Quarantined

Caruso goes to some lengths to ensure that the quarantine and rehabilitation process of wrongdoers is only as harmful as strictly necessary to ensure the safety of society, and describes his position such that it weighs the principle of autonomy carefully against society's need to restrict the wrongdoer's liberty for its own safety (Caruso *forthcoming*, 10). All of this works to define the boundaries and limitations to the PHQ model so that whatever course of action is taken can be justified. However, it does not go very far toward explaining what moral obligations we have *to* the wrongdoer. That is, Caruso has spent a good deal of time describing what we are *not* allowed to do to the wrongdoer, but has not described what we must do *for* the wrongdoer. This latter part will be the focus of this section.

Given the distinct lack of basic desert from the free will sceptic's position, Saul Smilansky rightly points out an objection to the PHQ model in the following way. An individual ought to be treated by the state as innocent (since individuals cannot be morally responsible in the basic desert sense, they must be treated equally regardless of their actions). However, since the state treats some people substantively worse than others by quarantining those that it deems a threat to the safety of society, the state "needs to offer such compensation as will right the balance" (Smilansky 2016, 11). I think that this needs to be taken very seriously and cannot simply be dismissed by supposing that the actions of the state are justified for the "greater good" of society (i.e., for purely utilitarian concerns).

Smilansky objects to the PHQ model of punishment, suggesting that since the state harms wrongdoers by quarantining them, any compensation that would balance the harm caused to them would necessarily make the quarantine so attractive that prison would look more like a five-star hotel than a correction facility (Smilansky 2011, 173). This, he believes, results in "funishment," where individuals are better off in prison than they ever could be in their daily lives, resulting in an *incentive* to commit crime rather than punishment being a deterrent for crime. Rather than create a method for rehabilitating wrongdoers, the kind of compensation that would be necessary for unfairly incarcerating an individual who is not morally responsible for her actions would likely result in an increase in crime, as many would choose the luxurious accommodation the state would be required to provide for merely taking away one's freedom. Smilansky's claim, then, is that the obligation that is owed to individuals for unfairly harming them is so high that a punishment is not "punishment" at all, but instead something resembling an all-inclusive vacation.

I think that Smilansky is correct in asserting that any individual whose liberty is limited by quarantine must be compensated, but I think that there are alternatives to fulfilling the state's obligation to compensate those it harms, while also avoiding the concerns of funishment. Namely, a simple improvement to the individual's conditions before she was incarcerated would go a long way to providing a necessary and sufficient compensation for the infringement of liberty. This could be done in two parts: first, while being quarantined, ensuring that every effort is being taken to "heal" the wrongdoer, and second, enabling the individual to have a better life after being rehabilitated than they had before, without being required to make prison into a 5-star resort. I will briefly describe each of those in more detail before discussing why we should be

satisfied with these two recommendations for fulfilling our obligation to compensate the individuals.

As we have seen, the quarantine model of punishment draws on the analogy of the necessity for quarantining those with dangerously infectious diseases, such as Tuberculosis. We do not, however, think that we have achieved justice and fulfilled our obligation to society after successfully isolating the infectious person from society. Society as a whole might be safe, but we also feel that we owe this person every reasonable effort to cure their infectious disease so that they can rejoin society; that is, we do not think that we can be morally justified in merely protecting society from a threat; we are also morally obligated to provide the individual with quick and effective health care. This too serves to promote the public health: alleviating the burden to medical caregivers and adding a functioning member into society. Similarly, I think, we have a moral obligation to "heal" or "cure" a wrongdoer when they are not guilty in the basic desert sense; and, since the PHQ model denies basic desert, individuals are *never* guilty in this sense and the state always has this moral obligation. While the state has a moral obligation to society to keep the individual quarantined while they pose a risk, it also has a moral obligation to individuals to ensure that they are only a risk for the minimum possible amount of time, and that they are only quarantined for just as long as – and no longer than – they are a threat to others' wellbeing.

To show that this analogy holds for the obligation to help those with infectious diseases as well as to help those who are a harm to society through their actions, let me first discuss how luck plays a significant role in which actions are available to us, and how this often causes us to *not* be guilty in the basic desert sense. This is because we are not responsible for our actions in the relevant and important way that we *could have done otherwise*, or that we are the *ultimate*

source of our actions. For example, we would not blame an individual who became paralyzed after being pushed down some stairs for not being able to run a marathon in under four hours; it would be ridiculous to expect such a thing to be within her abilities. No amount of resolve or strength of will could cause running a marathon to be within her abilities and she could not simply choose to run the marathon instead of remaining paralyzed. Similarly, we do not blame a child who is born into poverty or praise one who is born into wealth. It is completely out of their control which family they are born into, yet it plays an enormous role in the opportunities available to the individual throughout her life.[30] These kinds of unfair advantages and disadvantages seem unjust to many egalitarians which motivates and justifies redistributive policies which aim to negate these inequalities in one's life.

The circumstances which caused the individuals in those two examples to be able to act only in certain ways is not because of their willful refusal to act otherwise, it is simply the result of bad brute luck. Ronald Dworkin describes brute luck as being different from option luck in the following way: "Option luck is a matter of how deliberate and calculated gambles turn out – whether someone games or loses through accepting an isolated risk he or she should have anticipated and might have declined… [whereas brute luck is] a matter of how risks fall out that are not in that sense deliberate gambles" (Dworkin 2000, 73). That is, when situations arise purely out of events that are beyond one's control (such as an accident which requires one's legs to be amputated, or having abusive parents, etc.), then that is the result of brute luck – and, when those circumstances result in a disadvantageous situation for the individual, then it is a case of *bad* brute luck.

[30]For example, the richest 1 percent of men live an average of 14.6 years longer than the poorest 1 percent in the US. (Stepner, M. et al. 2016. The Association Between Income and Life Expectancy in the United States, 2001-2014 *JAMA.* 2016;315(16):1750-1766. doi:10.1001/jama.2016.4226)

There are two things to be said about brute luck, and one thing to be said about the free will sceptic's position on brute luck: first, it is inherently "unfair" when an individual is subject to either bad brute luck or good brute luck since they have done nothing to *deserve* their circumstances; second, that a fair society ought to try to compensate for bad brute luck; and third, that the free will sceptic will deny option luck entirely and say that every circumstance is the result of brute luck. Once I have explained these in a little more detail, it will be clear why the analogy between dangerously infectious individuals and dangerous criminals holds true.

First, the individuals did nothing to *deserve* the unlucky events which resulted in their unfortunate inability to act in certain ways; the paraplegic did not ask to be pushed down some stairs,[31] just as the child did not ask to be born into a family of negligent parents. In no way, then, is it their fault that they find themselves in those situations and unable to act in certain ways. That being the case, it is unfair that their circumstances have limited them in certain ways while other individuals have been fortunate enough not to be limited in the same ways: most people have not been involved in a horrific accident which left them paralyzed, and certainly some children are born into affluent families with loving, attentive parents that afforded their child with every opportunity. Because of this, all instances of brute luck are inherently unfair since there is nothing anyone can do to change whether they are the recipient of either good or bad brute luck; that is, brute luck is entirely outside of the control of the individual.

Since brute luck cannot be influenced by one's choices, "most egalitarians believe that justice requires the nullification of all differential effects of brute luck … feeling that it cannot be just that some people are worse off than others simply because they have been unfortunate, say,

[31] For the sake of the argument, assume the individual did not want to become a paraplegic and did not take actions which would knowingly result in that outcome.

to have been born with bad genes" (Lippert-Rasmussen 2014).[32] In terms of distributive justice, it is important that everyone is placed on roughly equal footing in order to have a fair chance at succeeding in one's pursuit of the good life.[33] Therefore, it is necessary to eliminate the effects of bad brute luck so that no one is disadvantaged to the point where they are unable to pursue their conception of the good life as effectively as anyone else.

An important aspect of the luck egalitarian position is that we should compensate for bad brute luck, but that we are still responsible for our own option luck (the calculated risks and gambles that we knowingly take in our lives). For example, the luck egalitarian may argue that we are responsible when we buy a lottery ticket and knowingly accept the risk of losing the amount spent on the ticket for the exceedingly small chance of winning the jackpot. In this case, there is no obligation to nullify the bad option luck when we inevitably do not win the jackpot because it was not the result of bad brute luck that we were disadvantaged. However, while distributive justice does not require compensation for bad option luck, the pervasiveness of brute luck makes it so that even when there exists some option luck in a decision, there is always some underlying brute luck which has an influence on the decisions being made or the gambles being taken.

The free will sceptic argues that *all* instances of an individual acting are instances that are ultimately caused by brute luck.[34] By denying free will (or assuming that it cannot exist), there can be no point in time where an individual makes a decision that is not the result of factors outside of one's control, which undermines one's ultimate moral responsibility for the action.

---

[32] For examples of egalitarians who believe this, see Cohen (2011, 5, 29); Dworkin (2000); Rakowski (1991); for a recent critique, see Elford (2013).

[33] That is, people are able to pursue their conception of the good life, whatever it may be, on equal footing as everyone else.

[34] Neil Levy (2011) employs an argument to this effect which describes the pervasiveness of luck as ultimately undermining free will entirely.

Even if it is not the case that the pervasiveness of brute luck completely undermines free will, however, we can still see that the existence of brute luck causes the PHQ model to have an obligation to nullify the bad brute luck that results in individuals being incarcerated. Take, for example, an individual with exceptionally bad constitutive luck (i.e., the bad luck involved with having bad personality traits). Through no fault of her own (perhaps as a result of growing up in an underprivileged family), the individual is more likely to cause harm than other people. If that person then commits a crime, it is in large part due to her bad constitutive luck. She was only in the situation in which she was more likely to commit the crime than other people would have been.

Take a more concrete example: the Nazi prison guard at a concentration camp. All things being considered, it is exceptionally unlucky that had been alive in Nazi Germany and indoctrinated with Nazi rhetoric and ideals. Even more unlucky, she was in a situation where someone was able to choose *her,* out of all possible candidates, to be a guard in a concentration camp where she was part of unspeakably horrible crimes. Today, we are lucky enough not to Nazi prison guard job positions, so no amount of bad brute luck could place us in that position. However, had we gone through the same experiences as the Nazi prison guard, it is entirely possible that we would have committed the same crimes as her.

This bad brute luck in the form of bad constitutive and circumstantial luck largely outweighs the option luck that is present in the Nazi guard case. Although she may have been able to weigh the risks of being hired by the Nazis and required to morally atrocious acts, it is largely due to factors outside of her control that she was able and willing to do those things in the first place. The same can be said of any wrongdoer; while we are lucky enough not to be disposed to commit wrongs, or to find ourselves in situations where we are able commit serious

crimes, wrongdoers *are* placed in those situations. Because of this they cannot hold ultimate moral responsibility for their actions. Thus, when a wrongdoer commits a crime, she cannot be held morally responsible for her actions and therefore cannot *deserve* to be incarcerated in the basic desert sense. This suggests that when an individual is incarcerated, the state owes her some form of compensation for treating her in a way that she does not morally deserve.

After describing the inescapable amount of brute luck that exists in influencing an individual's actions, it should become clear how the PHQ model of punishment requires compensation for the incarceration of dangerous criminals. Since it is a matter of bad brute luck that the state is required to incarcerate or quarantine a dangerous individual for the safety of society, it therefore has an obligation to nullify this bad brute luck by way of compensation. That is, when individuals are not ultimately responsible for their actions, the PHQ model must offer some compensation for treating some individuals considerably worse than others in order to justify the incarceration of those individuals. In relation to the analogy of our obligation to those with infectious diseases, just as we can only fully[35] justify quarantining individuals with infectious diseases by curing them, we must provide every reasonable effort in helping the wrongdoer as well. Thus, we also have an obligation to compensate wrongdoers for unfairly harming them by way of incarceration. In the next section, I will discuss how the obligation for compensating individuals can be fulfilled.

[35] Part of the justification comes from the right to harm in self-defense, while the remainder must come from the nullification of the bad brute luck which caused them to be in that situation to begin with.

4. Two Ways to Fulfil the Moral Obligation of Compensation

Fulfilling this moral obligation is certainly not a simple matter. It is also important to stress that this should not leave room to neglect the autonomy of wrongdoers on the grounds that it is for their own good. The harm principle should still hold weight, so it would not be acceptable to attempt to use methods of rehabilitation that undermined the individual's autonomy. For example, it is possible that we might be able to prevent all wrongdoers or would-be wrongdoers from committing another crime by forcing them to undergo lobotomies or electro-shock therapy or some other ridiculous "treatment." Those, of course, would *clearly* be unacceptable under both the harm principle and the principle of least infringement, given that those means are more harmful than necessary and clearly do not respect the autonomy of the individual. Balancing the principle of least infringement and the harm principle *would* allow for the detention or incarceration of harmful and dangerous individuals, but it would not permit the sorts of punishment that exist in North America today: the duration of sentences and the living conditions that are routinely prescribed to nonviolent criminals are certainly more harmful than strictly necessary to ensure public safety.

On the other hand, methods that *can* be used must prioritize the principle of autonomy. That is, it must place "primary emphasis on the liberty, privacy, and informed consent of individual persons in the face of a health intervention carried out by other parties" (Caruso 2016, 37). The methods employed, then, would always be with the consent and willing participation of those being rehabilitated. The methods themselves would, presumably, include whatever rehabilitation techniques are supported by trained professionals whose goal it is to successfully rehabilitate wrongdoers into society. I will not try to give an exhaustive list of methods, not least

of all because I am not a trained professional in the area, but also because any list that I might provide would be subject to current scientific information that might become outdated or ineffectual in different contexts. It is not the goal of this chapter to provide a method of rehabilitation, but instead only to express the limitations of what sorts of rehabilitation can justly be provided. The sorts of rehabilitation methods that I have in mind, however, might include therapy or jobs meant to expose wrongdoers to ways in which they can meaningfully contribute to society.

The second method of compensation for quarantine is closely related to the first, and can be supported by what has already been said about the PHQ model. That is, since I think we are morally required to compensate an innocent (in the basic desert sense) individual for unfairly but justifiably limiting their liberty, I will argue that it would be reasonable to provide compensation in the form of *better* opportunities in life after enduring quarantine than were available to that individual before. This does not mean that individuals are paid handsomely for committing a crime, nor any other form of compensation that incentivizes committing a crime. Rather, it should incentivize individuals to want to be a part of the rehabilitation process, thereby making society better off overall, and fulfilling the moral obligations of the government for quarantining an individual who does not *deserve* to be quarantined.

I think there are three key areas in which rehabilitators can make a wrongdoer "better off" than when they had committed the crime in the first place. Providing education, work/employment opportunities or other programs that contribute to society, and preventing and eliminating contributing factors that led the individual to act wrongly are all areas that can make the individual better off, while also providing compensation for their undeserved quarantine. The specifics of these will probably depend on the severity of the punishment, and it may be the

case that the rehabilitation process could provide the morally required compensation for individuals. I will discuss the specifics of these types of compensation when I argue that these types of compensation are sufficient to fulfil our moral obligation.

There are two problems for the sort of compensation I am suggesting that must be overcome. First, why is this compensation sufficient to nullify the harm caused by incarcerating an individual who is not guilty in the basic desert sense? Second, how can we ensure the compensation does not act as an incentive to commit further crime? The first is a moral question, so we will need to determine what moral obligation is owed and how much; while the second is an empirical question that may simply require empirical data before it can be determined to actually succeed. For the former, I will argue that this compensation is morally sufficient, and for the latter, I will provide several empirical examples to suggest that this does not produce the latter problem.

Although we can morally justify the incarceration of harmful wrongdoers with the PHQ model, we still have a moral obligation to nullify or compensate their unfortunate circumstances since their circumstances are the result of bad brute luck. One method of doing so is to provide equality in initial prospects, such that the effects of bad brute luck do not place an individual at serious disadvantage for succeeding in society (Vallentyne 2002, 543).[36] Currently, for example, individuals who are convicted of a crime are routinely marked as criminals by society (or labelled with criminal records that must be disclosed), drastically disadvantaging those individuals in their ability to succeed in society. For example, individuals with a criminal record have much more difficulty finding a job or housing; in some cases, they lose the right to vote. If

[36] Vallentyne provides an argument against the nullification of bad brute luck, but requires equality in initial prospects. I think this distinction disappears, however, since it seems as though "equality in initial prospects" just means *pre*-morally responsible actions, and when moral responsibility is denied, all points in time are pre-moral responsibility.

instead, the rehabilitation process was to include removing previously existing systemic or personal disadvantages for the individual to succeed, this would be one way of negating the bad brute luck that caused the individual to be incarcerated in the first place.

In this regard, the rehabilitation provided during incarceration would certainly go a long way towards making individuals better off than they were before: for one thing, by being properly rehabilitated they are no longer going to be a threat to society. In a very trivial sense, that will make them better off than they were before, merely in virtue of the fact that they no longer risk being incarcerated. This alone *may* be enough to compensate certain individuals, if the harm caused by the state is very minimal and rehabilitation could occur relatively swiftly.

On the other hand, different cases may require different levels of compensation. For instance, especially lengthy periods of incarceration, or cases where there were pre-existing grievous inequalities that led to the wrongdoer being incarcerated, may require additional compensation. It may be the case that the rehabilitation process would necessarily always provide enough compensation to cancel the negative effects of incarceration; if the process is carried out in such a way as to always make the individual better off by the correct amount. Caruso has already provided some arguments for the proportionality of the punishment, arguing that the PHQ model, although denying moral responsibility, is able to recommend harsher (i.e., longer) or softer (i.e., something like close monitoring instead of incapacitation) punishments depending on the danger that the individual poses to society (Caruso 2016, 40). Similarly, these punishments would require more or less compensation depending on the type of punishment. An especially long punishment, then, might require that the rehabilitation process ensures that efforts be taken to reduce or eliminate the underlying systemic problems which led to the wrongdoing in

the first place, while certain smaller crimes may simply involve a brief one-on-one therapy session with the wrongdoer (for example).

5. Guarding Against Funishment

Saul Smilansky has proposed a potentially powerful objection against approaches to punishment such as the PHQ model. Smilansky argues that we have a moral obligation to compensate individuals for harming them since they do not *deserve* to be harmed. He believes this obligation would necessarily require us to compensate offenders who are quarantined to such a degree that our prisons "would need to resemble five-star hotels" (Smilansky 2016, 3), thereby incentivising crime since the living conditions of being incarcerated would be so preferable to the regular struggles of society. According to the so-called funishment objection, any approach to punishment that aimed to compensate the punished would be self-defeating, since it would have the effect of incentivizing crime. In other words, if the PHQ included fair compensation, it would not achieve the aim of promoting public health, since it would *increase* crime. And while we may reasonably say that the PHQ model of punishment has other methods of providing compensation to wrongdoers that does not resemble a five-star hotel, provided it takes seriously the health of the individual; it has not been shown that these efforts at rehabilitation will not be so convenient and luxurious, that it would inadvertently act as an incentive to commit crime.

Let me first address a worry that, in any given society, the least well-off in society would happily give up their liberty in order to secure room and board in a rehabilitative prison. I can conceivably think of a society in which this is the case. Indeed, there are certainly some countries today where, if they implemented an ideal punishment system, it would be more appealing to be in prison than living free in society. I do not, however, believe that this is a

failing of the prison systems for not being harsh enough, but rather that the systems in place in society are not sufficient to satisfy the needs of its citizens. That is, if it is ever the case that being incarcerated and unable to pursue your conception of the good is preferable to being a functioning member of society, it should be taken as a clear indication that the society is not structured correctly. It *may* even be the case that there exist no well-structured societies today that are able to satisfy this claim, but I do not think this is a worry for the PHQ model of punishment so much as it is a worry for the structure of our societies.

It is also not the case that the steps to compensate individuals must be taken *after* the wrongdoer's action, meaning an individual must commit some wrong in order to reap the rewards of the compensation, so to speak. The PHQ model of punishment is required to both incarcerate and rehabilitate harmful individuals *and to prevent crimes from happening wherever possible* to promote public health (i.e., a safe, healthy, functioning society). When the overall aim of the PHQ model is to maximize public health, doing so by providing *preventative* measures (e.g. education, health care, etc.) will always be the first and most important step in achieving that goal. Making social programs available which work to prevent the circumstances from arising that result in individuals committing crimes *before* the crime has been committed is key in not only reducing crime, but also in preventing individuals from being incentivised to cause a crime. Since the PHQ model would not be offering any significant improvement to one's circumstances that were not already available prior to becoming a danger to society, there is no incentive for the individual to commit a crime merely to gain the advantage of whatever compensation is being provided.

With that being said, the kind of punishment being suggested by the PHQ model (i.e., the smallest restriction of liberty possible while being able to ensure the safety of citizens) does not

guarantee that all people would find prison less desirable than freedom. There is nothing, after all, about the rehabilitative process or the PHQ model in general that requires that a punishment should make the individual's life worse off than it was previously, except that wrongdoers no longer have the liberty to pursue their own lives in the same way as they had before. It must merely provide an environment in which the rehabilitation process is most likely to occur. Richard Wortley suggests that such an environment would consist of (Wortley 2005):

> (a) setting positive expectations through domestic furnishings that confer trust;
>
> (b) reducing anonymity through small prison size;
>
> (c) personalizing victims through humane conditions;
>
> (d) enabling a positive sense of community through ownership and personalization of the space;
>
> (e) reducing provocation and stress by designing in the capacity for inmates to enact control over environmental conditions and personal space.

But while none of these recommendations require that the individual should be worse off than it was previously, but it also notably does not require that the individual be any *better* off either. Indeed, depending on the individual's circumstances and personal preferences, such a prison environment might be much worse off than they were previously without also making the prison system more morally culpable for incarcerating the individual than it already was for restricting the individual's liberty. This is because the individual is not being harmed more than is absolutely necessary in order to ensure the safety of society. On the other hand, it *is* doing everything possible to rehabilitate individuals as quickly and effectively as possible, thereby taking steps to fulfil its obligation of compensation to the individual.

As I have argued, there is no need to make prisons any *more* accommodating than this because there is no further moral obligation to compensate individuals than what is already being provided. This provides some reason to believe that individuals would not prefer to be in prison because it would not be making the individuals life any better; ideally their lives would be nearly identical to before they were incarcerated, except that they no longer had the liberty engage with society directly. One may still, however, argue that those who are worse off than the average person would still prefer whatever accommodation the prison could offer over their current lives. This may be true, but I think there are three things to say in response to this. First, individuals are incentivized to actively participate in the rehabilitation process, such that they are more quickly able to pursue their own goals. Second, it seems that we are actively making people better members of society, so even if the occasional person does something harmful just to get a stay at the so-called prison resort, then it seems to be a good thing overall anyway. And third, there exist several compelling real-life examples in Norway and Denmark which seem to indicate that people would not choose this option, even if it were available.

To the first reason, individuals should want to participate in becoming rehabilitated since it will enable them to continue their lives where they left off before being incarcerated; hopefully in a way that will be better for society overall. This process, whatever it is, is meant to reduce a wrongdoer's stay at the correct facility as much as possible, such that individuals cannot stay and leech[37] off of society. It is also meant to *actually* help the individual. The focus, throughout the entire process, is to take the health and mental wellbeing of the individual as a serious concern.

The second reason, I think, clearly shows the absurdity of demanding that we not make prison so inviting to wrongdoers. The incapacitation of potentially dangerous individuals is

[37] I strongly dislike this term since I think that it implies that people are lazy by nature, which I think is untrue, but I see the term used frequently enough that I used it only to explicitly say that it would not happen.

helping society *and* the individuals to become better members of society. So, if an individual felt that they would be better off in prison and therefore causes some harm in order to be incarcerated, it is entirely possible that such an individual ought to have been in prison anyway. The majority of individuals will not want this, as is evidenced by several real-world examples. The Halden Prison in Norway is a real-life example of a prison that has been designed to specifically reduce crime. It is a high security prison which closely resembles Smilansky's concern for a prison of funishment. Here, the normality principle has been implemented which aims to ensure that during a prison sentence life inside will resemble life outside as much as possible.[38] The best part, however, is the empirical data that comes with implementing such a prison. Here's how Caruso has summarized the data (Caruso forthcoming, p. 31):[39]

> …when criminals in Norway leave prison, they tend to stay out. Norway's recidivism rate of 20% is one of the lowest in the world. The recidivism rate of Bastoy Prison, for example, is about 16%. By contrast, in the U.S. more than 76% of prisoners are rearrested within five years. The recidivism rate in the U.K. is lower, about 45%, but still more than double that of Norway. These statistics reinforce what researchers are finally beginning to realize, that prison has at best a negligible—and at worst a damaging—impact on the likelihood a person will re-offend (see Weatherburn 2010).

These things indicate that we don't really have to worry about funishment being a damning objection to the PHQ model of punishment, even though we really do have a moral requirement to compensate individuals for their time at the prison. While today's prisons regularly produce repeat offenders, the kinds of prisons that would most closely resemble Smilansky's funishment have been shown to have the greatest effects at reducing the recidivism rate. I have argued that by taking the restorative and rehabilitative processes seriously, we are

[38] For more details, see the Norwegian Correctional Service's full document: http://www.kriminalomsorgen.no/information-in-english.265199.no.html

[39] See the Norway 2015 Crime and Safety Report. https://www.osac.gov/pages/ContentReportDetails.aspx?cid=16970

able to adequately meet this moral obligation, while also ensuring that we are doing everything in our power to actually help the wrongdoers who require it in order to remain or become functioning members of society.

6. Conclusion

In this chapter, I first exposited Caruso's argument for a public health-quarantine model of punishment, given free will scepticism. I showed that, starting from the right to protect ourselves from harm, we can justifiably incarcerate harmful or dangerous individuals in order to protect the safety of society. This justification does not, however, preclude a moral obligation to compensate wrongdoers for treating them worse than we would otherwise, as Smilansky points out. In order to meet this obligation, I have argued that we need only compensate for the bad brute luck which caused the individual to act wrongly (or to be a danger to society), and that by sufficiently negating this bad luck, the state's moral obligation for compensation is fulfilled. I have suggested that the only way in which one's bad brute luck may be negated in this way, is to take seriously the individual's health and autonomy, so that serious effort is focussed on *helping* that individual and ensuring that the help she receives is both wanted *and* needed in order to become a functioning member of society. Both aspects of this help are required to ensure that the autonomy of the individual is respected while also being effective in rehabilitating and reducing recidivism. Lastly, I have shown that compensation for bad brute luck need not result in funishment since the moral obligation does not necessitate living conditions well above the average and still requires a good deal of effort on the part of the wrongdoer to acquire the help that is needed.

By taking seriously my recommendation of ensuring the health of the wrongdoer, I have provided an interpretation of the PHQ model which avoids the "mere means" objection to punishment. That is, it ensures that wrongdoers are not merely used as a means of promoting the general wellbeing of society, but are treated as autonomous individuals who deserve the respect that is required of all people. Most importantly, it ensures that dangerous wrongdoers who are incarcerated for the safety of society are helped as quickly and efficiently as possible, similar to the way in which we are obligated to treat those with dangerously infectious diseases who are quarantined to ensure they cannot unintentionally harm others. By ensuring wrongdoers are treated this way, the PHQ model of punishment cannot lose sight of the needs of the individual in favour of the greater good for society. In that way, my argument that state-sanctioned punishment entails an obligation to the wrongdoers, actually strengthens the PHQ model from some common (and often damning) objections to the justification of punishment.

Bibliography

Beauchamp, T. and Childress, J. (1989) *Principles of Biomedical Ethics*, 3rd edition. New York: Oxford University Press.

Bentham, J. [1789] (1907). *An Introduction to the Principles of Morals and Legislation.* Oxford: Clarendon Press.

Berman, Mitchell N. (2008) Punishment and Justification. *Ethics* 18: 258-290

Boonin, D. (2008) *The Problem of Punishment*. Cambridge: Cambridge University Press

Carlsmith, K. (2008). On justifying punishment: The discrepancy between words and actions. *Social Justice Research*, 21, pp. 119–37.

Carlsmith, K., Darley, J., and Robinson, P. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology*, 83, pp. 284–99.

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

——. (forthcoming) *Unjust Deserts: Free Will, Moral Responsibility, and Criminal Punishment*.

Caruso, G. and Pereboom, D. (forthcoming) "Hard-Incompatibilist Existentialism: Neuroscience, Punishment, and Meaning in Life," for *Neuroexistentialism: Meaning, Morals, and Purpose in the Age of Neuroscience,* Gregg. D. Caruso and Owen Flanagan, eds., New York: Oxford University Press.

Cohen, G. A. (2011) *On the Currency of Egalitarian Justice and Other Essays in Political Philosophy*, Princeton, NJ: Princeton University Press

Dimock, S. (1997) Retributivism and Trust. *Law and Philosophy*. 16:37-62.

Dworkin, R. (2000) *Sovereign Virtue*. Cambridge MA: Harvard University Press.

Elford, G. (2013) Equality of Opportunity and Other-Affecting Choice: Why Luck Egalitarianism Does Not Require Brute Luck Equality. *Ethical Theory and Moral Practice*. 16:39-49.

Feinberg, J. (1991) The Classic Debate. *The Philosophy of Law* 4th Edition, eds, Joel Feinberg and Hyman Gross. Belmont, California: Wadsworth Publishing Co. 624-629

Hampton, J. (1983) The Moral Education Theory of Punishment. *Philosophy and Public Affairs.* Vol 13, No. 3: 208-238.

Hart, H. L. A. (1959). The Presidential Address: Prolegomenon to the Principles of Punishment. *Proceedings of the Aristotelian Society* 60:1 - 26.Hart, H. L. A. (1961). *The Concept of Law*. Oxford University Press.

Kahan, Dan M. (1996) What Do Alternative Sanctions Mean? *Faculty Scholarship Series.* Paper 114 http://digitalcommons.law.yale.edu/fss_papers/114.

Kant. (1996) *The Metaphysics of Morals.* p. 105, 6:331 by Akademie pagination.

Kant, Immanuel [1788] (1909). *Critique of Practical Reason*. Dover Publications.Levy, Neil. (2011) *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. New York: Oxford University Press.

Lippert-Rasmussen, Kasper, "Justice and Bad Luck", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2014/entries/justice-bad-luck/>.Mackie 1982

Michael, M. A. (1992) Utilitarianism and Retributivism: What's the Difference?. *American Philosophical Quarterly* 29: 2.

Mill, J.S. (1843) *A System of Logic, Ratiocinative and Inductive*. University of Toronto Press.

——. (1859) *On Liberty.* In Ian Carter, Matthew H. Kramer & Hillel Steiner (eds.), Freedom: A Philosophical Anthology. Blackwell. pp. 129.

——. (1861) *Utilitarianism*. Tuttle. (1993).

Moore, G.E. (1903) *Principia Ethica*. Cambridge: Cambridge University Press.

Moore, M. (1997) *Placing Blame.* Oxford: Clarendon Press.

Morris, H. (1968) Persons and Punishment. *The Monist*. 52: 475-501.

Pereboom, D. (2001) *Living Without Free Will.* New York: Cambridge University Press.

——. (2014) *Free Will, Agency, and Meaning in Life.* Oxford: Oxford University Press.

Rakowski, E. (1991) *Equal Justice*. Oxford University Press.

Rawls, J. (1955) Two Concepts of Rules. *Philosophical Review.* 64: 3-32.

Sharif, A.F., Greene, J.D., Karremans, J.C., Luguri, J., Clark, C.J., Schooler, J.W., Baumesiter, R.F., and K.D. Vohs. (2014) Free will and Punishment: A mechanistic view of human nature reduces retribution. *Psychological Science* published online June 10: 1-8.

Smilansky, S. (2011) Hard determinism and punishment: A practical reduction. *Law and Philosophy* 30: 353-367.

——. (2016) Pereboom on punishment: Funishment, innocence, motivation, and other difficulties. *Criminal Law and Philosophy.* doi:10.1007/s11572-016-9396-3

Stepner, M. et al. 2016. The Association Between Income and Life Expectancy in the United States, 2001-2014 *JAMA.* 2016;315(16):1750-1766. doi:10.1001/jama.2016.4226

Vallentyne, P., (2002) Brute Luck, Option Luck, and Equality of Initial Opportunities. *Ethics*. 112: 529–557.

Vargas, M. (2013) "How to Solve the Problem of Free Will". *The Philosophy of Free Will*.

Wortley, R. (2005). *Situational Prison Control*. Cambridge University Press.

Wright, V. (2010) *Deterrence in Criminal Justice, Evaluating Certainty Versus Severity of Punishment.* The Sentencing Project.

Zaibert, L. (2006) *Punishment and Retribution.* Ashgate Publishing.