CHALLENGING RETRIBUTIVIST INTUITIONS

A thesis submitted in partial fulfilment of the requirements for the degree of

MASTERS IN PHILOSOPHY

of

RHODES UNIVERSITY

by

JONATHAN HAWKES

February 2009

Abstract

Can punishment, a practice which involves the deliberate infliction of suffering, be justified? Retributivists and consequentialists argue that punishment can be justified, whereas abolitionists argue that it cannot. Retributivists argue that punishment is justified because wrongdoers *deserve* it, whereas punishment is justified for consequentialists because it is beneficial for society. A popular form of abolitionism is restorative justice, which is the view that all those affected by crime (perpetrators, victims and members of society) should be reconciled. In this thesis I argue that retributivist justifications for punishment are mistaken, and argue in favour of a consequentialist view. I also argue that consequentialism can accommodate the valuable features of restorative justice while avoiding the challenges faced by it. My arguments against retributivism will turn on a thought experiment. The experiment is designed to draw out the fundamental retributivist intuition that people who cause harm deserve to suffer harm in return, yet excludes most of the principles retributivists would use to justify the intuition. I will go on to argue that, even if the retributivist considerations did apply to the experiment, they would still not justify the claim that wrongdoers deserve to be punished. Most of the retributivist considerations are, therefore, not necessary for the intuition, and none of the considerations are sufficient for it. The retributivist considerations are, I contend, rationalisations, as the claim that wrongdoers deserve to suffer is based, not on good reasons, but on an unreliable intuition. I shall argue that the consequentialist considerations, while not being necessary, are sufficient for the claim that wrongdoers should be punished, and they should be punished, I maintain, in the interests of preventing greater harm from occurring.

Acknowledgments

To all who were involved, you have my heartfelt gratitude. I want to thank the following in particular for all they have done:

My supervisors, past and present, Dr. Pedro Tabensky and Prof. Marius Vermaak, for your time, unwavering dedication, sound advice and the tangential conversations that had nothing to do with this thesis.

My family, for the sacrifices you have made so that I could have this opportunity, for your emotional and financial support and your dogged, sometimes dogmatic and often frustrating belief in my abilities.

Thandi, for your companionship, your seemingly endless patience and for clouting me and telling me to stop being stupid when I needed it most.

My friends and digsmates, for being there through the whole thing, for putting up with my rubbish and listening to me talk about nothing but philosophy and my thesis for two years (that should stop now, for a while at least).

TABLE OF CONTENTS

Challenging Retributivist Intuitions

Chapter One – The Problem: Punishment and Justification

The main question I address in this thesis is: can punishment for crimes be justified? Retributivism is a widely held view of punishment that is based on the deep-rooted intuition that wrongdoers *deserve* to suffer by being punished for their crimes. I maintain that this intuition is unreliable, and argue against retributivism in favour of a consequentialist view of punishment. In this introductory first chapter I firstly explain why punishment is in need of justification, and then introduce the theories which are purported to provide it. The chapter has four sections, in the first section (*1.1 –The Need for Justification*) I explain why punishment needs to be justified. The second section (*1.2 – Purported Justifications for Punishment*) details the attempted justifications offered by retributivists and consequentialists, and in the third section (*1.3 – The Abolitionist Position*) I mention the view that punishment cannot be justified. I briefly discuss my own position in the fourth section (*1.4 - My Position*), and end the chapter with a short summary.

1.1 –The Need for Justification

The reason why punishment requires justification can be stated succinctly. Punishment involves inflicting something that is unpleasant or burdensome on those who commit crime, and such a practice, which *aims* to cause suffering to those it is inflicted upon, is in need of justification (Scarre, 2004). It is in virtue of the deliberate imposition of burdens on criminal wrongdoers that punishment needs to be justified "in a way that most other political institutions do not" (Burgh, 1982: 193), that is, it needs *moral* justification. It is important to point out that there is an assumption at work in most legal and philosophical discussions of punishment, namely that punishment *can* be justified. Those who discuss the issue of punishment fall into two camps, those who argue it *can* be justified and those who argue it *cannot*, and should therefore be abolished.

1.2 – Purported Justifications for Punishment

Attempted justifications for punishment are offered by retributivists (who argue punishment is *deserved*), consequentialists (who argue punishment is *socially beneficial*)

and those who offer 'mixed' views which combine consequentialist and retributive concerns (Duff, 2008). Retributivism is a very influential theory of punishment, and provides the "major justification" (Burchell, 2005: 81) for punishment, and, as such, is the theory I focus on in this thesis.

In order to understand the key features of retributivism, it is useful to contrast it with consequentialism. Retributivism is based on the basic idea that those who cause undue harm *deserve* to be harmed themselves, an idea which is deeply entrenched in popular opinion (Burchell, 2005). Retributivists hold the view that wrongdoers, who are morally responsible, *deserve* to be punished in a way that is proportionate to the suffering that their crimes caused their victims: the punishment has to fit the crime (Tabensky, 2006). In order for a person to be responsible for an unlawful act, that act has to be significantly under that person's "control" (Metz, 2006: 227). By control, Metz means that the act was a result of the agent's deliberation, and was not the product of "external factors such as duress, brainwashing and trickery, nor by internal factors such as psychosis, sleepwalking and being a minor" (Metz, 2006: 227). So the agents are responsible so long as they *chose* to perform the crime, and were not compelled by internal or external factors over which they had no power[1]. By contrast, consequentialists hold that punishment should be inflicted on criminals in order to promote beneficial consequences for society, such as deterrence, protection of society and correction or reform of the offender (Graham, Kreider and Svatos, 2002).

The principle of proportionality is an integral feature of retributivism as it protects against what could be considered excessively harsh forms of punishment. Consequentialists, so the story goes, would employ excessively harsh punishments, such as keeping dangerous, unreformed offenders in prison after they have completed their sentences (Burchell, 2005), if it served the goal of crime prevention (Duff, 2008). According to retributivists, imposing the death penalty for traffic fines, for instance, may make roads a safer place, as drivers would be more careful, but many would think that this excessive and disproportionate punishment is wrong because it is "not deserved" (Allais, 2008a: 10). The principle of proportionality also supposedly sets retributive

[1] I shall not deal with the question of whether or not people can be responsible for their actions, and hence their crimes, but shall assume for the sake of argument that they can be responsible. I explain my reasons for avoiding this question in *4.2.3 – The Kantian Theory*.

punishment apart from revenge, as it sets a limit on the amount of harm that can be inflicted on offenders. Punishment should not inflict on offenders more harm than the offenders inflicted on their victims, it "must bear some relationship to the harm" (Burchell, 2005: 69) caused by the offenders. Revenge, on the other hand, is not governed by any such principle, and this is why retributive punishment (supposedly) differs from it[2].

Punishment, for retributivists, is justified intrinsically, whereas it is justified instrumentally for consequentialists (Allais, 2008a)[3]. Retributivists hold that punishment is justified because it is an end in itself, that is, we punish for the sake of punishing and not for the sake of some other good. Consequentialists, on the other hand, maintain that punishment is justified because it is a means to promoting good consequences. Consequentialism is therefore 'forward-looking', as the rationale for punishment is based on facts about the future, on states of affairs that punishment is supposed to bring about, such as crime control. Retributivism, on the other hand, is 'backwards-looking' insofar as it bases the justification for punishment solely on events that occurred in the past (Korman, 2003). Retributivists advocate that offenders should be punished independently of whether or not future benefits will come about as a result of punishing. So punishment is justified either "because it is deserved [retributivism], or because it is socially beneficial [consequentialism]" (Burchell, 2005: 68).

From the above we can see that the main features of retributivism are the notions of *desert* and *proportionality*, and the ideas that punishment is *justified intrinsically* as an end-in-itself, and that it is *backward-looking*. There are many different forms of retributivism (Cottingham, 1979)[4]. All these views hold the 'backwards-looking' justification for punishment (it is justified because offenders who are responsible for their crimes deserve to be punished), and attempt to answer the question: why do the guilty

[2] It is important to note that not all retributivists seek to separate revenge from punishment. J. F. Stephen (cited in Scarre, 2004: 115), for example, argued that "the criminal law stands to the passion of revenge in much the same relation as marriage to the sexual appetite". That is, punishment is the appropriate outlet and safeguard for the passion of revenge. Retributivists in general, however, do not follow Stephen's thinking, and attempt to separate, as far as possible, the notions of revenge and justice (Scarre, 2004). In any case, whether punishment expresses the desire for revenge or not, retributivists would hold that that punishment should still be proportionate to the offence in order to be considered just.

[3] I shall argue in Chapter Six that punishment is instrumentally, and not intrinsically, valuable.

[4] Cited in (Duff, 2008).

"deserve to suffer" (Duff, 2008: 11)? I consider three prominent versions of retributivism:

1. First, there is the "fairness theory" (Metz, 2006: 224), the view that crime upsets the social order, as criminals, by breaking the law, gain an advantage over law abiding members of society, and punishment is required to restore that order by depriving criminals of the advantage they have gained (Burchell, 2005).
2. The expressive theory is the second theory I discuss; it is the view that punishment expresses attitudes such as anger and resentment and judgments of condemnation and disapproval (Feinberg, 1994).
3. Finally, I examine the Kantian version of retributivism. Kant (2002) states that we should treat people as ends in themselves, which involves respecting their rationality and "holding them responsible for their actions (Rachels, 2007: 139). So we hold wrongdoers accountable by punishing them for their wrongful deeds[5].

1.3 – The Abolitionist Position

Retributivists and consequentialists argue that punishment can be justified, abolitionists, on the other hand, argue that it cannot be justified, and that we should find alternative means of responding to criminal wrongdoing (Duff, 2008). Burgh (1982), for instance, argues that neither the retributive nor the consequentialist views of punishment can be justified. He focuses predominantly on retributivism "due to a growing realization that a [consequentialist] justification… is ultimately inadequate" (Burgh, 1982: 194), consequentialists would supposedly treat people merely as a means to the welfare of others. In other words, he takes issue with the fundamental consequentialist idea that the suffering of the offender serves to safeguard law-abiding members of society by deterring potential offenders[6]. After dismissing consequentialism, Burgh turns to retributivism, and finds fault with the core retributivist intuition that wrongdoers, in virtue of committing crimes, *deserve* to be punished. Burgh examines what is, in his opinion, the strongest version of retributivism, namely the fairness theory[7], and argues that it fails to provide a

[5] See Appendix A.
[6] Burgh (1982) is not alone here; many argue that the consequentialist justification for punishment is insufficient. See, for instance, Allais (2008a); Burchell (2005), Metz (2006).
[7] Falls (1987: 26) also states that the fairness theory, or what she calls the 'theory of reciprocity', is "*the*

justification for the intuition that wrongdoers deserve to suffer[8]. To conclude, Burgh (1982: 210) suggests that we should cease acting on this unjustifiable intuition and seek an alternative "paradigm upon which to base our concept of the criminal law"[9].

One alternative to punishment which has been growing in favour of late is restorative justice (Duff, 2008; Burchell, 2005). On this view the appropriate response to wrongdoing involves a process of reconciling victims, offenders and all interested parties from the community (Duff, 2008). So the key theme of restorative justice is reconciliation, and the aims of the reconciliation process include repairing the harms done to victims instead of punishing perpetrators, and restoring or establishing relationships between all affected parties (Allais, 2008a). The process involves all those who have a stake in the wrongdoing coming together to discuss the harm done, its possible repercussions and how best to deal with the situation (Braithwaite, 1997)[10].

1.4 - My Position

I maintain that punishment *can* be justified, and argue in favour of the consequentialist view. I differ from Burgh (1982), then, in that I do not think that the consequentialist justification for punishment is inadequate. Burgh (1982) claims that the consequentialist justification is inadequate because we would treat offenders as mere means. I will argue (*contra* Burgh, 1982) that this criticism is based on the mistaken assumption that autonomy is intrinsically valuable. I do, however, agree that the fundamental retributivist intuition, that offenders deserve to be punished, cannot be justified. My argument in support of this claim will turn on a thought experiment (which I present in Chapter Three). The experiment is designed to draw out the retributivist intuition, yet it excludes most of the retributivist considerations which are supposed to justify that intuition. Given that the intuition remains even though most of the retributivist considerations are absent,

---

interpretation with which to deal, whether one argues for or against retributivism". I do not agree with Burgh (1982) and Falls (1987) on this point. My view is that the Kantian theory is the one to deal with, as most "contemporary retributivists situate their account within a roughly Kantian moral theory" (Allais, 2008a: 10), and emphasise the duty to respect people as ends in themselves (Allais, 2008a). It is for this reason that I focus predominantly on the principle of respect when I discuss Kant's retributivism.

[8] I agree that the fairness theory does not provide adequate justification for the intuition that wrongdoers deserve to be punished, and, as such, I shall say more about the theory's failings in Chapter Four, which is devoted to the shortcomings of all three of the retributive theories I mentioned earlier.

[9] Whether we should, or even can, cease acting on this intuition is the topic of Chapter Five.

[10] Cited in (Llewellyn & Howse, 1999).

it follows that those considerations are not necessary for the intuition. I will go on to argue that the considerations are not sufficient for the intuition either, as they would not justify it even if they were present in the thought experiment. Consequently, I maintain that the justifications for punishment put forward by retributivists are mere rationalisations, and that retributivists are merely attempting to hold onto the unfounded intuition that wrongdoers deserve to be punished. Rationalisations disguise the real reasons individuals have for their beliefs or actions, and, even if they are plausible, they are not true. For example, a government might rationalise its intervention in a civil war in a foreign country, claiming the intervention is based on humanitarian concerns when it is actually based on a desire to acquire the country's natural resources (Warburton, 2001). The retributivist justifications, I contend, are mere rationalisations because the driving force behind the retributivist theories is the intuition that wrongdoers deserve to suffer by being punished, yet the retributivist considerations are neither necessary nor sufficient for this intuition. This intuition, I maintain, is founded, not on good reasons, but on emotional responses to crime, such as anger and resentment, which typically involve the desire to strike back at the offender (Duff, 2008). Many, such as Mill (1907) and Strawson (1974), have emphasised that feelings such as anger and resentment are natural and commonplace. Mill (1907), for instance, claims that the desire to strike back is natural and common to all animals. I argue, however, that retributivists are misidentifying and misrepresenting this desire. The desire is misidentified and misrepresented because animals strike back to protect themselves or their young, and protection (of ourselves or others) is a consequentialist concern, not a standard retributivist one.

To conclude, I shall argue that consequentialism could provide a better justification for punishment, and that the charges against consequentialism (that it is unjust and treats people merely as a means) can be met. I also contend that consequentialism can accommodate the goals of restorative justice, but only insofar as reconciliation will indeed be the best thing for all involved. Advocates of restorative justice, for instance, claim that perpetrators and victims should be reconciled, and that perpetrators should be reintegrated into society (Llewellyn & Howse, 1999). There are times, however, when it seems misguided to try to reconcile victims and offenders, consider rape cases or domestic abuse as examples. Confronting victims with those who

have harmed them so severely, and claiming that the relationship between them should be restored seems to add insult to injury. An attempt to reconcile victims and perpetrators in these cases "would be a betrayal both of the victim and of the values to which we are supposedly committed" (Duff, 2008: 6). Furthermore, in order for reconciliation to take place, wrongdoers, and community members, need to recognise and admit that wrong has been done (Duff, 2008). Reconciliation, however, seems unobtainable where wrongdoers are recalcitrant, obdurate or unrepentant. In instances such as these, and in cases involving rape or domestic violence, punishment seems to be a more appropriate response to criminal wrongdoing, or so I will argue.

Chapter Summary

In this chapter I explained the central issue I address in this thesis, namely that punishment requires moral justification because it involves the deliberate infliction of suffering on those who break the law (Burgh, 1982). I then presented the cases for and against punishment, describing the views of those who argue that punishment *can* be justified (retributivists and consequentialists), and those who argue that it *cannot* be justified (abolitionists). I sit in the camp that argues that punishment can be justified, specifically in the consequentialist camp, and I concluded the chapter by briefly introducing the arguments I shall employ in this thesis, namely the arguments against retributivism and in favour of consequentialism.

Chapter Two – Retributivist Justifications

In this chapter I discuss the three prominent retributive theories I mentioned in Chapter One, so the chapter has three sections: *2.1 – The Fairness Theory*, *2.2 – The Expressive Theory* and *2.1 – The Kantian Theory*. In each section I shall explain the respective theory, and then show how that theory incorporates the main features of retributivism, namely the notions of desert and proportionality, and the ideas that punishment is justified intrinsically as an end-in-itself, and that it is backward-looking.

2.1 – The Fairness Theory

The first version of retributivism I discuss is the fairness theory. The fairness theory is a modern version of retributivism which is based on the uncontroversial observation that the law, just by existing, constitutes a benefit and a burden to all citizens (Falls, 1987). The law benefits people by prohibiting others from interfering with their individual rights, and imposes burdens on those who do interfere with those rights. Crime, according to the fairness theory, upsets the social order, as criminals, by breaking the law, gain an advantage over law abiding members of society. Two of the foremost fairness theorists are Morris (1976) and Murphy (1978). Morris (1976: 33)[11] says:

> If a person fails to exercise self-restraint even though he might have and gives in to such inclinations, he renounces a burden which others have voluntarily assumed and thus gains an advantage which others, who have restrained themselves, do not possess.

Murphy (1978: 100)[12] makes a similar point:

> If the law is to remain just, it is important to guarantee that those who disobey it will not gain an unfair advantage over those who do obey voluntarily. It is important that no man profit from his own wrongdoing, and a certain kind of "profit" (i.e., not bearing the burden of self-restraint) is intrinsic to criminal wrongdoing

The advantage that Morris and Murphy mention does not come in the form of "ill-gotten material goods" (Falls, 1987: 28). Criminals benefit from crime because they acquire a freedom that is not rightly theirs (Scarre, 2004). The freedom criminals acquire is a freedom from the constraints of the law. Law-abiding citizens restrain themselves whenever they desire to break the law. Criminals, on the other hand, do not. Criminals

[11] Cited in Falls (1987: 28).
[12] Cited in *Ibid*. at 28.

thus benefit from other people's obedience to the law when they do not constrain their own behaviour (von Hirsch: 1976)[13]. So criminals, to use Metz's (2006: 229) words, do not undergo their "fair share of the burden of obedience" when they commit crime. Criminals who go unpunished would benefit from the law, as they would be protected from others interfering with their lives, but would not experience the burden of the law to the same extent as law-abiding citizens, as they forgo their burden of obedience and act on their unlawful desires (Falls, 1987).

Advocates of the fairness theory state that, since criminals gain an unfair advantage, punishment is required to restore the social order by depriving criminals of the advantage they have gained (Burchell, 2005). Punishment restores the balance upset by crime (the balance of burdens and benefits) by imposing a burden on criminals. The balance, however, is not restored merely by returning whatever ill-gotten material goods offenders may have gained. Criminals have to suffer a loss of freedom in order to cancel the freedom they gained by committing crime (Scarre, 2004). Punishment, therefore, restores the balance by returning suffering for suffering, thus ensuring an "equal distribution of the burdens and benefits of law, a distribution which is demanded by fairness" (Falls, 1987: 30).

In response to the question 'Why do the guilty deserve to suffer?' the fairness theory provides the following answer: Criminals *deserve* to be punished because it is unfair that they should gain an advantage over law-abiding citizens. So criminals deserve to suffer the loss of the unfair advantage they gained through crime (Duff, 2008). Furthermore, in order to restore the balance, the punishment has to be *proportionate* to the crime. So the worse the crime, the harsher the punishment has to be to restore the balance (Metz, 2006). The fairness theory therefore includes the notions of desert and proportionality. The fairness theory is also backwards-looking and justified intrinsically because the purpose of punishment on this view is to restore a legal equilibrium that has been upset by crimes which occurred in the *past.* Once that equilibrium has been restored, there is no more work for punishment to do. So punishment is not future directed, but is viewed as the appropriate way to rectify what has happened in the *past*. The fairness

[13] Cited in Scarre (2004).

theory therefore also includes the ideas that punishment is backwards looking and is justified intrinsically and, as such, contains all of the main features of retributivism.

2.2 – The Expressive Theory

I turn now to the expressive theory, which is the second version of retributivism I discuss. Feinberg (1970: 98), a leading proponent of the expressive theory, claims that punishment expresses "attitudes of resentment and indignation, and of judgements of disapproval and reprobation". There are therefore two central components of this view, firstly the expression of attitudes such as resentment, and secondly the communication of judgements such as condemnation. These two components are sometimes separated, and discussed as different accounts of retributivism (see Duff, 2008). For the sake of simplicity, I shall follow Feinberg (1970) and discuss the two components together under one view (the expressive view), rather than as two separate views.

The first component of the expressive view appeals to our emotional responses to crime. Crime may arouse emotions such as resentment and anger, and these often include the desire to make wrongdoers suffer because of the harm they have caused. Punishment is therefore required to express or satisfy these attitudes (Duff, 2008). So punishment, rather than merely communicating judgements of condemnation, becomes a symbolic way to get even with criminals, as it expresses the desire to strike back at them. Punishment, therefore, is a kind of "legitimised vengefulness" (Feinberg, 1970: 100).

The expressive theory incorporates the notions of desert and proportionality in the following ways. Criminals, according to the expressive theory, should be punished because they *deserve* censure, and punishment expresses our emotional responses to crime. The expression of condemnation needs to be *proportionate* to the offence, so the "worse the offence… the stronger the expression of disapproval must be and hence the harsher the punishment should be" (Metz, 2006: 224). Censuring criminals through punishment may (or may not) deter potential future offenders, but such deterrence is not the goal of expressive punishment. Censure, according to the expressive theory, is an *end in itself*. To illustrate this point I cite Feinberg (1970: 101), who says that

> [symbolic] public condemnation... may help or hinder deterrence, reform, and rehabilitation – the evidence is not clear. On the other hand, there are other functions of punishment, often lost sight

> of in the preoccupation with deterrence and reform, that presuppose the expressive function and would be difficult or impossible without it.

These functions include, for instance, the disavowal of criminal wrongdoing by state authorities. By punishing criminals the state demonstrates that it does not condone crime, and that criminal activity is not the sort of behaviour it accepts, or expects, from its citizens (Feinberg, 1970). On the expressive view, punishment is a practice that society uses to characterize its standards and values and affirm its dedication to them (Scarre, 2004). From this one can see that the expressive theory is backwards-looking, not forward-looking, as the point of punishment is to condemn *past* actions, and not to prevent *future* ones. The expressive theory therefore contains all of the essential features of retributivism.

Finally, it is important to note that expressive theorists only attempt to modify offenders' behaviour by reminding them of the "good moral reasons" (Duff, 2008: 13) they have to shun a life of crime. The expressive theory can therefore avoid the criticism often levelled against consequentialism, namely, that it attempts to force wrongdoers to comply with the law (Duff, 2008).

2.3 – The Kantian Theory

The Kantian version of retributivism is the final theory I discuss. Kant (2002) states, repeatedly and forcefully, that people are to be treated as ends in themselves, and not as mere means to some other end. Treating people as ends involves respecting the fact that they are rational beings, and holding them responsible for actions they chose to perform (Rachels, 2007). The Kantian principle of respect for persons not only prevents people from being treated merely as a means to some other end, but, according to Metz (2006), requires "condemning people proportionately to their responsible wrongdoing" (Metz, 2006: 222). In other words, according to the Kantian retributivist, we punish people in order to treat them as ends in themselves.

Punishment, for the Kantian retributivist, should be enacted irrespective of whether or not it will bring about any future social benefits. Kant claimed that punishment "can never be inflicted merely as a means to promote some other good for the criminal himself or for civil society. It must always be inflicted upon him only

because he has committed a crime" (Kant, 2003: 105)[14]. Kant stated that only the principle of equality (*ius talionis*) can determine the type and amount of punishment a criminal should receive (Kant, 2003). The *ius talionis* states that those who cause unnecessary harm deserve to suffer "like" (Falls, 1987: 25) harm in return. Criminals deserve like suffering because they have the capacity for autonomy (the ability to consider whether certain goals are worth pursuing) and act in light of this consideration. Human agents, according to Metz, have the "highest intrinsic value in the world" (Metz, 2006: 226) because they have the capacity for autonomy, and such intrinsic value merits respect. Such respect involves individuals and institutions "tracking" (Metz, 2006: 227) the choices people make. By 'tracking' people's choices, Metz means that we respond in kind to those choices, and impose burdens on those who make wrong choices, and refrain from doing so (and possibly offer benefits) when people choose correctly (Metz, 2006). Why, though, should we respond 'in kind'? The first formulation of Kant's Categorical Imperative provides the answer. The first formulation is as follows: "*act as if the maxim of your action were to become by your will a* **universal law of nature**" (Kant, 2002: 689)[15]. That is, to put it crudely, act in a way that you think every rational being should act. So if I am sympathetic, I demonstrate that I think people (including myself) should be treated with sympathy. Conversely, if I am callous, then I am effectively announcing that people (including myself) should be treated in the same way. Rachels (2007: 139) sums the issue up well. He writes:

> when a rational being decides to treat people in a certain way, he decrees that in his judgment *this is the way people are to be treated*. Thus, if we treat him the same way in return, we are doing nothing more than treating him *as he has decided* people are to be treated… and so we are, in a perfectly clear sense, respecting his judgment, by allowing it to control our treatment of him.

Consequently, we respond in kind to criminals by punishing them because, by treating others badly, they have made it known that, in their view, that is how people should be treated.

On the Kantian view, criminals *deserve* to suffer because they *chose* to commit crime, and as they are ends in themselves, they should be held accountable for their misdeeds. Punishment for the Kantian is *proportionate* because it has to follow the

[14] Italics in original.
[15] Italics and bold in original.

principle of equality (*ius talionis*); wrongdoers have to suffer similar harm to that which they inflicted on their victims. Finally, punishment for the Kantian is intrinsically justified and backwards-looking in that it can never be inflicted on wrongdoers to produce socially beneficial ends. Punishment is inflicted *just because* the wrongdoer has committed a crime. So the Kantian theory contains the four main elements of retributivism.

Chapter Summary

In this chapter I have described the three versions of retributivism that I focus on in this thesis, and demonstrated how each incorporates the main features of retributivism, namely the notions of desert and proportionality, and the idea that punishment is both intrinsically justified and backwards-looking. The first theory I mentioned was the fairness theory, which is the view that wrongdoers, in virtue of committing crime, gain an unfair legal advantage over law abiding citizens, and punishment is required to remove that advantage. The expressive theory was the second view I described, it is the view that punishment expresses our emotional responses to crime, and communicates censure. Finally, I explained the Kantian theory, which is the view that rational beings should be held accountable for their actions, and the way to hold offenders accountable is to punish them.

Chapter Three – The Thought Experiment

In the previous chapter I described three prominent retributive theories, the fairness theory, the expressive theory, and the Kantian theory. In this chapter I present the thought experiment that I will use to test the justification for punishment that each of these theories provides. The point of the experiment is that most of the retributivist considerations are not necessary, and none are sufficient, for the intuition that offenders deserve to be punished. The chapter has three sections, I explain thought experiments in the first section (*3.1 - The Nature of Thought Experiments*), in section two I present my own thought experiment (*3.2 – My Thought Experiment*). In the final section I discuss two possible criticisms of my thought experiment, and respond to them *(3.3 – Criticisms and Responses*).

3.1 – The Nature of Thought Experiments

Thought experiments are imaginary situations that are designed to explore questions about the nature of our concepts and theories (Brown, 2007), and are intended to clarify issues surrounding them (Warburton, 2001). In order to better understand thought experiments, it is useful to contrast them with scientific experiments, as they mimic the scientific process (Baggini & Fosl, 2003). Traditional thought experiments, like scientific experiments, are ways of exploring questions about "the nature of things" (Brown, 2007: 1). The main difference between the two types of experiment is that scientific experiments are generally conducted in laboratories and involve empirical investigation, whereas traditional thought experiments are generally conducted "in thought alone" (Baggini & Fosl, 2003: 58). In order to conduct a thought experiment, we attempt to visualize a given scenario, carefully reflect on this situation, and then see what the outcome of the exercise will be. The key point of thought experiments is that we are able to grasp the nature of concepts or theories simply by thinking about them (Brown, 2007)[16].

Considering examples of both thought and scientific experiments will further illustrate the important differences between the two. First imagine a scientific experiment

[16] It must be noted, though, that some more recent thought experiments involve presenting an audience with a situation, and then documenting and analysing their responses to it. I will say more about these experiments later.

designed to determine how soap powder bleaches. Normally, many factors could be responsible for the bleaching power of the soap, such as the soap's "active ingredients, the type and temperature of the water in which the ingredients are dissolved, the materials being cleaned, and the machinery – if any – used to do the cleaning" (Baggini & Fosl, 2003: 58). In order to discover which of these factors causes bleaching, the experiment has to make sure that the extraneous variables are isolated from the essential factors. If the scientist's hypothesis is that chlorine causes bleaching, the experiment "needs to show that *if all other factors remain the same* the presence or absence of the chlorine will determine whether the soap powder bleaches" (Baggini & Fosl, 2003: 58)[17]. Put more plainly, the aim of scientific experiments is isolate the essential variables, variables which cause a particular effect if they are present, one which would not occur if those variables were absent. Thought experiments operate in a similar fashion, as they also test certain variables and separate crucial variables from extraneous ones. The difference is that the variables are changed in one's imagination alone (Baggini & Fosl, 2003: 59).

One of the most famous philosophical thought experiments is Robert Nozick's 'experience machine'. The experience machine is "a type of virtual reality machine which gives you the illusion of actually living your life but with the added twist that everything that you do or happens to you is intensely pleasurable" (Warburton, 2001: 131). The machine can mimic any real life scenario, but replicate that scenario in its most pleasurable form. Moreover, once plugged into this machine, you will actually believe that what is happening to you is real. Nozick asks if you would plug into the experience machine for the rest of your life. Most people, claims Nozick, would say 'no' to this question, and this suggests that those who do answer 'no' value something more than pleasure. Nozick's hypothesis, it could be argued, is that people value some things more than unlimited pleasure. To test this hypothesis, the thought experiment isolates pleasure from other things in life that we value, and draws out our attitude towards it. Nozick's thought experiment of the experience machine therefore provides us with a useful way of testing our intuitions about pleasure.

It is important to point out that some, more recent, thought experiments, however, are not conducted in thought alone, but involve presenting an audience with a scenario,

[17] Italics in original.

and then recording and analysing their responses to it. One of the most important modern thought experiments, introduced by Greene *et al*. (2001)[18], is the famous trolley problem. A concise version of this experiment can be found in Greene & Haidt (2002: 519), it is as follows:

> *Suppose a runaway trolley is about to run over and kill five people. Suppose further that you can hit a switch that will divert the trolley onto a different set of tracks where it will kill only one person instead of five. Is it okay to hit the switch? Now, what if the only way to save the five people were to push a large person (larger than yourself) in front of the trolley; killing him but saving the others? Would that be okay?*

So there are two possible scenarios, in the first, the only way to save the five people from the runaway trolley is to flick a switch, and kill another person on a different track. In the second scenario, the only way to save the five people on the track is to kill one person by pushing him in the path of the trolley. Greene *et al*. (2001)[19] found that most people say 'yes' to the first scenario, but 'no' to the second, even though both cases involve sacrificing one person in order to save five others. Greene and his colleagues have thus drawn a distinction between 'personal' and 'impersonal' moral violations and judgements. A personal moral violation has three components, it is "i) likely to cause serious bodily harm, ii) to a particular person, iii) in such a way that the harm does not result from the deflection of an existing threat onto a different party" (Greene & Haidt, 2002: 519). An impersonal moral violation, on the other hand, does not meet these criteria. The second trolley scenario does meet all of these criteria and it is therefore a *personal moral violation*. Pushing someone in front of the trolley is going to cause that particular person grievous bodily harm (the person dies as a result), and the death of that person is not a consequence of diverting some greater harm, it is a means to preventing that harm. The first trolley scenario, by comparison, is an *impersonal moral violation* as "diverting a trolley involves merely deflecting an existing threat" (Greene & Haidt, 2002: 519).

[18] Cited in Greene & Haidt (2002).
[19] Cited in *Ibid*.

Green *et al*. (2001)[20] scanned participants' brains with fMRI (functional magnetic resonance imaging) when they presented the participants with the two trolley scenarios. They found that the cognitive areas of the brain showed increased activity when participants were faced with *impersonal* moral dilemmas, whereas both cognitive *and* emotional areas showed increased activity when they were presented with *personal* moral dilemmas. Green *et al*. concluded that, while *reasoning* plays an important role in the production of moral judgments, specifically in impersonal moral judgments, that role is limited, and *emotions* feature strongly in many of those judgments, specifically in personal moral judgments. The work of Green *et al*. is significant in that it demonstrates that both reasoning and emotions play a role in moral judgment, "but automatic emotional processes tend to dominate" (Greene & Haidt, 2002: 517).

Another recent thought experiment, which of particular importance to my own work, is presented by Haidt (2001). The experiment elicits a particular response from the reader, yet excludes the standard reasons used to justify that response. Haidt concludes that the initial response is emotional, rather than rational, as there are no good reasons one can appeal to in order to justify that response. Haidt's experiment is important because my own thought experiment is similar to it. I intend to elicit a particular response with my thought experiment, yet exclude most of the standard considerations offered as justifications for that response in an attempt to show that the response is emotional, rather than rational. I shall discuss the similarities between the two experiments (Haidt's and my own) in greater depth in the next section of this chapter (where I present my experiment). For now I explore Haidt's (2001: 814) experiment in more detail, it is as follows:

> *Julie and Mark are brother and sister. They are travelling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love?*

[20] Cited in *Ibid*.

Haidt claims that most people who hear the above story first say that is was wrong for Mark and Julie to make love, and then try to find reasons for their statement. Haidt's example specifically excludes the reasons that most people would use to support their statement that what the siblings did was wrong. There is no danger that a child with genetic deficiencies would be born as a result of their act, as two forms of birth control were used. Julie and Mark are in no way psychologically harmed by their actions, so one cannot say that what they did was wrong because they will be hurt by their actions. In the end, according to Haidt, many people say something like "I don't know, I can't explain it, I just know it's wrong." (Haidt, 2001: 814). Haidt's explanation for this is that people feel "a quick flash of revulsion at the thought of incest" (Haidt, 2001: 814) and then go about searching for reasons to justify their statement. In other words, it is the 'flash of revulsion' that leads people to say that the siblings' actions are wrong, and not some well thought out argument. In this regard, when people search for reasons after they have felt the flash of revulsion, they "construct post-hoc justifications" (Haidt, 2001: 815). Even though the 'flash' is what guides the responses offered by those who hear this story, it is not acknowledged by those searching for reasons, who, according to Haidt, "experience the illusion of objective reasoning" (Haidt, 2001: 815). That is, they do not acknowledge that the driving force behind their convictions is an emotional response and not a rational one.

3.2 – My Thought Experiment

The thought experiment I present is designed to undermine the retributive theories of punishment by showing that the theories are based on an unreliable intuition, namely the intuition that those who cause harm deserve to suffer harm in return. The experiment is as follows:

> *Jeff lives in a small, close-knit community where there is no crime. Jeff has prostate cancer, and he is going to have an operation that will leave him incapable of having sexual intercourse. Late one night he goes to the corner café to grab a bite to eat. He sees an attractive woman, Alice, leave the café alone. Alice is a stranger to the town, and has no family or friends. Jeff follows Alice, and rapes her. In an attempt to keep Alice from crying out, he smothers her with*

his jacket, and, in the process, suffocates her. These actions are uncharacteristic of Jeff. He acknowledges that he has done a terrible deed, and genuinely feels bad because of what he has done. Jeff's community do not feel they have been disadvantaged by Jeff's crime, and do not believe that he gained anything beneficial from it either. In their view, what happened to Alice was tragic, but they are not incensed by Jeff's actions. Rather, they view him as a person who has gone horribly wrong and is in need of support and guidance. Jeff is consequently not punished. Did the community do the right thing?

According to Appiah, a large portion of moral philosophy has been devoted to thinking about cases such as the one above, "forming an intuition about the right answer, and then trying to discover principles that will explain why it's the right answer" (Appiah, 2008: 84). The particular intuition I seek to draw out with the thought experiment is the retributivist intuition that those who commit crime deserve to be punished. I am not claiming that everyone will say that Jeff deserves to be punished, and I do not intend to find out just how many would respond in this way. Rather, my goal is to show that *if* one has the intuition that Jeff should be punished, *then* that intuition is unreliable in this situation. The intuition is unreliable because it is unfounded, and it is unfounded because the thought experiment excludes most of the principles that retributivists rely on to explain why punishing Jeff is the right thing to do. Those who claim that Jeff should be punished will therefore have no resources to back up their assertion, and the assertion will be an empty one.

I have good reason to think that many, retributivists included, will think that Jeff should be punished. I, for one, find the idea that Jeff is not punished deeply unsettling. If someone close to me was raped and murdered, and the person responsible was not punished I would be incensed as well as distraught. The thought alone is distressing enough. I would want the person responsible to suffer for what he had done. In the same vein, I would want Jeff to be punished for his actions, and I do not think I would be alone in this response. Most people, claims Burchell (2005), would agree that criminals deserve to be punished. Many hold that offenders should suffer a penalty that is "equal in its severity to the pain or injury they originally inflicted on their victim(s)" (Scarre, 2004: 113). Burgh (1982: 195) takes it to be an "unquestionable fact" that the intuition that

wrongdoers deserve to be punished because they have committed crimes is shared by most people. He bases his view on the "empirical observation that *generally* the response of most people to wrongdoing is that the wrongdoer deserves to suffer and that this is especially true of the sufferer of wrongdoing" (Burgh, 1982: 195, footnote 4). Moreover, retributivists assert that the belief that offenders should receive their just deserts "is a central plank of most people's moral outlook" (Scarre, 2004: 114). I do not have any evidence to show that retributivists are correct in their assertion that many hold the view that offenders deserve to suffer, but I have already shown that the notion of desert is deeply imbedded in the retributive theories, as it is the foundation of these theories (Burchell, 2005). Allais (2008a: 10) for example, claims that "the guilty… deserve to be punished, and they deserve punishment the severity of which is proportionate to the seriousness of the offence". Retributivists are therefore theoretically committed to saying that Jeff should be punished, as he is guilty of committing crimes, and this is enough to support my claim that the example will draw out the intuition that Jeff deserves to suffer for his crimes.

In the previous section I mentioned that my thought experiment was similar to Haidt's (2001) experiment (which involves an example of incest). My example of Jeff and Alice is similar to Haidt's example of Julie and Mark in that it draws out a certain intuitive response from the reader, yet excludes the standard reasons used to justify that response. The idea of incest elicits a particular type of emotion, or "affect" (Appiah, 2008: 103), namely that of disgust, or "moral repugnance" (Appiah, 2008: 103). Crime also elicits emotional responses from people, such as indignation and anger, which often involve "a desire to make the wrongdoer suffer" (Duff, 2008: 11). We can, it seems, safely say that crime too educes the affect of moral repugnance. In Haidt's (2001) example, the flash of moral repugnance leads people to say that what Julie and Mark did was wrong. Haidt, however, has removed the elements of incest which would explain why their actions were wrong. Those who try provide reasons to explain why Mark and Julie's actions are wrong are, in Haidt's opinion, merely offering post-hoc justifications for their intuitions. My example of Jeff is akin to Haidt's as a flash of moral repugnance leads people to say that Jeff should be punished, even though most of the elements of crime which would explain why he should be punished have been removed. Those who

attempt to provide reasons for punishing Jeff are, I maintain, merely offering post-hoc justifications in attempt to support their intuitive response, justifications which are, in the end, mere rationalisations.

3.3 – Criticisms and Responses

In this section I discuss two criticisms which could be brought against my thought experiment. The first criticism is that thought experiments do little more that draw out our intuitions, which, according to some, is not a good way of doing philosophy (Baggini & Fosl, 2003), and the second is that my experiment is too far-fetched and contrived. I argue that neither of these objections undermines my thought experiment.

3.3.1 – The First Criticism

Thought experiments have been criticized for doing little more than drawing out our intuitions, and some have argued that this is "an unreliable method of doing philosophy" (Baggini & Fosl, 2003: 60). The worry behind this criticism is that our intuitions are not trustworthy, especially in the bizarre or far-fetched scenarios provided by highly contrived though experiments (Brown, 2007). Consequently, any view which is based on such unreliable intuitions will itself be unreliable. It is important to note that there are two main types of thought experiments, constructive and destructive. Constructive experiments are used in support of a theory, whereas destructive ones are used to undermine a theory (Brown, 2007). The worry that our intuitions are unreliable seems to be more a concern for thought experiments which are designed to bolster theories, that is, for theories that are based on certain intuitions. Destructive thought experiments that are intended to refute a theory may do so by showing that the theory is based upon an unreliable intuition. Nozick's experience machine is an example of a destructive thought experiment, as it undermines the intuition that the best life to live is the most pleasurable one. The thought experiment I put forward is also a destructive one, as I seek to undermine the retributive intuition that undue harm warrants similar harm in return, and it is therefore important that my experiment effectively draws out this intuition. The first criticism, then, does not seem to apply to my thought experiment.

3.3.2 – The Second Criticism

In order for a thought experiment to work, it has to be plausible and coherent. One could argue that my example is too contrived and far-fetched, and should therefore be rejected. I have two responses to this objection. Firstly, it does not matter if the experiment is far-fetched, since the point is to test our intuition and draw out a particular response, which, as I argued in the previous section, it does. Secondly, I argue that my thought experiment is conceivable *and* possible.

My first response to this objection is that dismissing the example because it is far-fetched and contrived misses its point. Earlier I mentioned Nozick's famous thought experiment of the experience machine. The thought experiment is obviously far-fetched, as the chances of such a machine being built are slim to none. That it is far-fetched, however, does not matter, as the point of the experiment is to "pick out our fundamental attitude to pleasure, and it is good at making clear our intuitions on this" (Warburton, 2001: 131). In a similar vein, it is does not matter that the example of Jeff is far-fetched. The purpose of the experiment is to test our intuitions on crime, and to draw out a particular response to it.

The second response I have to the objection that my example of Jeff is too far-fetched and contrived is that the scenario is not just conceivable, but possible as well. One might say that there are a few elements of the thought experiment that are highly unlikely. First, that a community could exist that had no crime prior to Jeff's actions; second, that someone (in this case, Alice) could have no family or friends; and third, that the community would not call for Jeff's punishment. Firstly, it should be noted that imagining this scenario is not like attempting to imagine a square with only three sides, which would be impossible. It is not the case that we just cannot imagine Jeff's scenario because the scenario has contradictory elements in the same way that we cannot imagine a square that only has three sides. The scenario is *difficult* to imagine, but not *impossible.* We might struggle to imagine a crimeless community because crime is such a prevalent feature of societies in general. Humans are social creatures, and families and friends are integral parts of our lives. So the fact that Alice has no family or friends might strike us as quite odd. That the community does not call for Jeff's punishment might also appear

unlikely because we know how most communities respond to heinous crimes like rape and murder, that is, with indignation and horror.

My task now is to show that the scenario presented in my thought experiment is possible. I discuss the three elements of the scenario I mentioned in the previous paragraph, namely the crime rate in the community, Alice's lack of family and friends and the community's response to Jeff's crimes. The first element of the scenario I discuss is the crime rate. There are many people in the world who do not commit crime, so it does not seem impossible for there to be a community full of law-abiding citizens. Secondly, Alice has no family or friends, but even this should not be inconceivable, as there are some solitary people in the world who do not have these types of relationships, such as hermits. Finally, the community view Jeff as a person in need of guidance and support and do not punish him for his crimes. Such a benevolent attitude towards wrongdoers is an uncommon one, but not as uncommon as one might think. There are many examples of people who do not adopt a condemnatory attitude towards those who commit crimes against them or those closest to them, "some family members of victims have even gone so far as to try to help the perpetrators regain their lost dignity" (Tabensky, 2006: 144). One need only think of the South African Truth and Reconciliation Commission in order to imagine a scenario where a community does not punish offenders. Many have criticized the commission, though, on the grounds that not punishing the perpetrators has made "some who suffered under apartheid feel that their wrongs have not been taken seriously enough" (Scarre, 2004: 100). Similarly, one could criticize the members of Jeff's community for not taking Jeff's crime seriously enough. This criticism, however, assumes firstly that punishing Jeff is justified, and secondly that punishment is the most appropriate way to take Jeff's crime seriously. The onus is on the critic to provide a story that can validate these assumptions. Claiming that Jeff should be punished just because he deserves it does not tell us why he should suffer, as Montague puts it, "this purported explanation is no explanation at all without some justification of the move from propositions about deserved treatments to propositions about requirements to accord such treatments" (Montague, 2002: 21). That is, the notion of 'just deserts' is not a viable justification for punishment, as the idea itself calls for further explanation. In the previous chapter I mentioned three prominent retributivist theories (the fairness,

expressive, and Kantian theories), advocates of each of these views would say that Jeff should have been punished. The task of the next chapter, then, is to examine the reasons each theorist would offer in support of the claim that Jeff should have been punished and see if they are viable.

Chapter Summary

The point of this chapter was to introduce the central thought experiment which serves as the foundation of my argument against retributivism. I first explained what thought experiments are and how they work, and discussed two examples of modern thought experiments, namely Greene *et al.'s* (2001)[21] trolley problem and Haidt's (2001) example of incest. I then presented my thought experiment, which involves Jeff raping and accidentally killing Alice, and ended the chapter by defending the experiment against two possible criticisms, namely that the experiment only draws out our intuitions, and that it is too far-fetched and contrived.

[21] Cited in Greene & Haidt (2002).

Chapter Four – The Failings of Retributivism

The purpose of this chapter is to explain, firstly, why most of the retributivist theoretical considerations do not apply to the thought experiment I presented in the last chapter, and, secondly, why they would all fail as justifications for punishment even if they did apply to the experiment. The chapter, then, has two main sections (*4.1 – Why Retributivist Considerations do not apply to the Thought Experiment*) and (*4.2 – Why Retributivist Considerations are Insufficient Justifications for Punishment*). Each of these sections has three parts; each part is dedicated to one of the three prominent theories of retributivism I discuss in Chapter Two. So section *4.1* will have the following sub-sections: *4.1.1 – The Fairness Theory*, *4.1.2 – The Expressive Theory* and *4.1.3 – The Kantian Theory*. Section *4.2* will cover the retributivist theories in the same order as *4.1*.

4.1 - Why Retributivist Considerations do not apply to the Thought Experiment

In this section I argue that the retributivist considerations, except for the Kantian consideration of autonomous choice[22], do not apply to the thought experiment of Jeff and Alice, and are therefore not *necessary* for the intuitive response that Jeff deserves to be punished. The reason the considerations are not necessary is that the desire to punish Jeff arises even when the considerations have been excluded from the experiment. So why do most of the retributivist considerations not apply to the experiment?

4.1.1 – The Fairness Theory

The considerations of the fairness theory do not apply to the experiment because it does not seem as if Jeff has gained an advantage over the members of his community.

An advocate of the fairness theory would claim that Jeff has gained an unfair advantage over law-abiding citizens as he exercised desires that law-abiding citizens would constrain. This advantage is the freedom to exercise his desire to rape Alice. Punishment is therefore required to remove the unfair advantage that Jeff has gained. Punishment would restore the balance of burdens and benefits by restricting Jeff's freedom and imposing something unpleasant on him in return for the advantage he gained. No-one in Jeff's community, however, harbours the desire to rape anyone, and so

[22] More on this in *4.1.3 – The Kantian Theory*.

they have not been disadvantaged because Jeff has acted on inclinations that they have constrained. They do not think the opportunity to exercise the desire to rape is advantageous or desirable, and none of them wish they were in Jeff's situation, or in a position to gain a similar 'advantage'. Consequently, Jeff has not gained an unfair advantage, so punishment is not required to remove it.

One could reply that, even though Jeff did not gain an advantage by raping Alice, he benefited from the law in general, as he broke the law, whereas the members of his community did not, so they are still legally disadvantaged. The problem with this sort of response, as Burgh (1982) points out, is that *all* wrongdoers benefit from the law in general, as they break the law, whereas law-abiding citizens do not. If punishment is required to remove the unfair advantage gained from crime, and all wrongdoers gain the same unfair advantage from breaking the law in general, then it follows that all wrongdoers would deserve the same punishment, no matter what crime they committed. Rapists would therefore deserve the same amount of punishment as traffic offenders, and equating rape and traffic violations seems absurd. Furthermore, claiming that Jeff has gained an unfair legal advantage seems to misrepresent both what it is about Jeff's actions that make them wrong, and what disturbs us about the example. The view that crime upsets the social order distracts our attention from the wrongs that have been done to the victims of crime "when it is those wrongs that should be our central concern" (Duff, 2008: 5). What makes Jeff's actions wrong, and what offends us about them, is surely the undue harm they caused Alice, and not the supposed unfair advantage he gained from his crime. Attempting to restore some obscure legal balance seems inadequate as well as inappropriate in the face of the harm Jeff caused.

We should, one could plausibly argue, be trying to mend the harm that Jeff caused instead of trying re-establish an abstract equilibrium of benefits and burdens. How, though, could we mend the harm that Jeff has caused? Ten (2000)[23] argues that punishment should, in some way, recompense victims, as they suffered directly as a result of crime. The most appropriate form of punishment then, according to Ten, involves the offender paying a fine or offering some other service to the victim. If the victim is somehow compensated by punishment, then it appears as if punishment can repair the

[23] Cited in Scarre (2004).

harm done. Ten's (2000)[24] reformulated version of the fairness theory, however, does not appear to be useful in Jeff's scenario. There is no victim to compensate through services or monetary offerings, nor are there any family members or friends who could claim compensation. Other forms of punishment, such as incarceration, would, it seems to me, do little to repair the harm suffered by Alice. Incarcerating Jeff would obviously not bring Alice back to life. Even if Alice had survived the assault, it seems odd to say that punishing Jeff would heal Alice, and annul or undo the damage done to her, as Montague (2002: 4) says "[clearly]… no kind or amount of punishment is likely to have this restorative capability for the woman who is raped".

The justification for punishment offered by fairness theorists, therefore, does not apply to the case of Jeff, as the principle of unfair advantage does not capture the wrong making features of Jeff's actions. Ten's (2000)[25] idea that the balance is restored when the harm suffered by the victims is mended does not apply either, as there is seemingly no way to mend the harm.

4.1.2 – The Expressive Theory

The considerations of the expressive theory do not apply to the experiment because the members of Jeff's community do not feel anger or resentment, so punishment is not required to express them, nor is it required to communicate condemnation as all members of the community (including Jeff) acknowledge that wrong was done and should not be repeated.

An advocate of the expressive theory, or an 'expressivist', would hold that Jeff deserves to be censured because he has committed terrible deeds. An expressivist would also maintain that our emotional responses to his behaviour, those of disgust, resentment and anger, need to be expressed through the appropriate means. So an expressivist would say that Jeff should be punished as punishment is the proper way to communicate the censure he deserves, and adequately expresses our emotional responses to Jeff's crime. Jeff's community, however, do not feel anger (or anything like it) towards him. Rather, they view him as someone who is in need of support and guidance. The community, therefore, does not need to express emotions such as anger and indignation through

[24] Cited in *Ibid*.
[25] Cited in Scarre (2004).

punishment, as they do not experience them. Moreover, punishment is not required to express Alice's emotional responses to what was done to her, as she is dead, and she had no family or friends who would be affronted by Jeff's actions, so punishment is not required to express their emotions either. One may, however, claim that it is implausible to think that the members of Jeff's community did not feel anger or resentment, whereas others who hear the example of Jeff's atrocious behaviour will experience these emotions. I have, however, argued in *3.3.2* (*The Second Criticism*) that this scenario is not implausible. In addition to what I argued in *3.3.2*, consider the following example, cited in Tabensky (2006: 143-144):

> 'Westerners', recently remarked Marcos Sandoval of the Triqui people of Oaxaca, 'represent justice with a blindfolded woman. We want her with her eyes well open, to fully appreciate what is happening. Instead of neutrality or impartiality, we want compassion. The person committing a crime needs to be understood, rather than submitted to trial.'
>
> These open eyes of their justice do not, for example, look for punishment when a person violates a shared custom. He or she is perceived as someone in trouble, who needs understanding and help; including the opportunity to offer compensation to the victim of his or her misdemeanour … Rather than confine wrongdoers in jail, many of these communities tie them to trees or confine them to places for a few hours or days with the express hope of allowing their passions to calm down; or for the safe return from their delirious condition. These practices are not conceived as forms of punishment. Instead, they offer communal support: according opportunities for the soul to heed the wisdom and advice of elders, when they come to converse and reflect with those who have wronged others.

The Triqui, assuming that Sandoval is reporting accurately, are an example of a group of people who, like the people in Jeff's community, respond with compassion and understanding, rather than with anger and resentment. The reactions of the members of Jeff's community are therefore not implausible.

The other component of the expressive theory is that punishment should communicate the appropriate censure that criminals deserve for their offences. Punishment as communication seeks to modify future behaviour only by reminding the agent of the good moral reasons for refraining from crime (Duff, 2008). Jeff, however, already appreciates the gravity of the situation, and does not need to be provided with good reasons to refrain from committing crime, as he cannot commit the same crime in

the future due to his operation. What is the point of censuring Jeff when he already knows that what he has done is wrong and feels genuinely bad because of what he has done? Moreover, the members of Jeff's community are all law-abiding citizens, so they do not need to be convinced that Jeff's actions are wrong, and do not need to be persuaded against committing similar crimes. Punishment, therefore, is not needed to perform any expressive function, such as the expression of feelings of indignation or the communication of censure. The reasons expressivists offer to justify punishment are consequently not applicable in Jeff's situation.

4.1.3 – The Kantian Theory

The Kantian theory is different from the other theories, as the main Kantian consideration, namely autonomous choice, *is* present in the thought experiment, as Jeff chose to rape Alice. According to Kant (2002) we have to treat people as ends-in-themselves, which involves respecting their rational nature and holding them accountable for their actions (Rachels, 2007). By punishing wrongdoers we hold them responsible for the wrongful deeds that they have "freely chosen" (Rachels, 2007: 139). So a Kantian would state that we should punish Jeff because we need to treat him as an end-in-himself, and respect the fact that he *chose* to commit crime. The Kantian principle of respect for persons calls for individuals and institutions to track actions that people *chose* to perform (actions that are not influenced by internal or external factors that are beyond the agent's control) and respond in kind to those actions (Metz, 2006). In Jeff's case, he chose badly, and consequently, he should be punished. Punishing Jeff would treat him as an end-in-himself, as an autonomous moral agent who can be held accountable for his actions, and not punishing him is disrespectful (Scarre, 2004). The members of the community therefore did not act appropriately when they did not punish Jeff, as they did not treat him as an end-in-himself.

I could not remove the consideration of autonomous choice from the experiment, as, without the presence of criminal intent, Jeff's actions would not be classed as a crime. The standard view is that an offence needs to have both *mens rea* (guilty mind) and *actus reus* (guilty act) in order to be categorised as a crime (Scarre, 2004). In other words, for an offence to be considered a crime both the intent to cause unlawful harm and the act of

causing unlawful harm have to present[26]. So if Bill kills Tom in a "fit of insanity" (Scarre, 2004: 140), then Bill cannot be convicted of murder, as there is no *mens rea*. That is, the intent to end Tom's life is absent.

It must be noted that, even though the Kantian consideration of autonomous choice applies to the original thought experiment (as Jeff did chose to rape Alice), it only does so because Jeff's intention was carried through to fruition, and unlawful harm was done to Alice. It seems to me that it is mainly this undue harm, over and above the fact that Jeff chose to rape Alice, which elicits the desire to make Jeff suffer by punishing him. Furthermore, focusing solely on the fact that Jeff chose to rape Alice, claiming that we need to treat Jeff as an end-in-himself and respect his autonomy distracts our attention from the unwarranted suffering that has been inflicted on Alice, suffering which, as I mentioned in *4.1.1 – The Fairness Theory*, should be our primary concern (Duff, 2008). Talk of respecting Jeff seems inappropriate when faced with the harm suffered by Alice. A Kantian might respond to this point by claiming that we *should* respect Jeff because he is intrinsically valuable in virtue of possessing the capacity for autonomy, and this value warrants respect. But why is autonomy intrinsically valuable? I take up this issue in *4.2.3*.

Section Summary

I have just argued that the considerations retributivists would use to justify punishing Jeff do not apply to Jeff's situation (except the Kantian consideration of autonomous choice). The considerations of the fairness theory do not apply because Jeff has not gained an advantage over others in his community, and claiming that he has gained an advantage misses the point, Jeff's actions are wrong because he caused undue harm, and not because he gained an obscure legal advantage. The considerations of the expressive theory do not apply because punishment is not required to express resentment or condemnation. Finally, the Kantian consideration of autonomous choice *does* apply, but it does not seem to fully capture what it is about Jeff's actions that elicit a desire to punish Jeff, as it seems to be the presence of *both* criminal intent and unlawful harm that elicits the intuitive response that Jeff should be punished. Consequently, the considerations of the fairness

[26] There are exceptions to this rule, such as manslaughter, which is the "act of killing a person unlawfully but not intentionally or by negligence" (Graham, Kreider and Svatos, 2002: 783)

and expressive theories are not necessary for the intuitive response that Jeff should be punished, as the desire to punish Jeff arises even when those considerations do not apply to the experiment. Even though the Kantian consideration of autonomous choice does apply to the experiment, it appears to be only partially applicable.

4.2 - Why Retributivists Considerations are Insufficient Justifications for Punishment

In the previous section I argued that most of the retributivist considerations are not necessary for the intuitive response that Jeff should be punished. In this section I shall argue that none of the retributivist considerations are sufficient justifications for punishment either, as, even if these considerations did apply to the experiment, they would still not provide good reasons for punishing Jeff.

4.2.1 – The Fairness Theory

The main reason I think that the considerations of the fairness theory do not provide sufficient justification for punishment is that the notion of proportionality is obscure[27]. I begin this sub-section by discussing the problems associated with the principle of proportionality. I then discuss a possible way for the fairness theorist to respond to this criticism, and argue that this response is inadequate and does not circumvent the initial criticism of the fairness theory.

A supporter of the fairness theory would say that Jeff gained an unfair advantage over the members of his community, as he broke the law and they did not, but how do we establish how great that advantage is (Duff, 2008)? Moreover, how do we determine what type and amount of punishment is proportionate to the offences Jeff has committed (Scarre, 2004)? In other words, what is the appropriate sentence for rape and manslaughter in this situation? The fairness theory does not provide us with any criterion which would help us answer these questions, as Scarre (2004: 116) states, the theory does not offer any "clear criteria for determining when punishments are proportionate to offences".

[27] This criticism of the principle of proportionality applies to all three versions of retributivism that I discuss, as the principle is an integral feature of each of them.

The problems with the principle of proportionality become even more noticeable when there are multiple offenders. In the example of Jeff, there is only one offender, yet it is still difficult to determine what punishment would fit his crimes. What would happen if Jeff had been a member of a gang, and the gang had raped and murdered Alice? Does each gang member bear an equal portion of punishment in order for that punishment to be proportionate to Alice's suffering? Or do we say that they each must suffer as Alice did? So, for the sake of argument, imagine that the court decided that fifty years in prison would be proportionate to Alice's suffering. Are those years to be divided equally amongst the gang members, so that each member serves a portion of the total sentence, or does each member have to serve fifty years? The fairness theory does not seem to offer us any way to answer questions such as these.

A supporter of the fairness theory could reply that the above criticism misses the point, as the criticism concerns what counts as proportionate and not whether punishment should be proportionate. To put it another way, the criticism is not directed towards the justification the fairness theory offers, which is that criminals have gained an unfair legal advantage. In order to criticise this justification, one has to show why the idea of unfair advantage does not provide good grounds for punishing criminals.

I have two responses to the fairness theorist here. My first response is that it is not clear why punishment has to be proportionate in the first place. Proportionality, as I mentioned in *1.2 – Purported Justifications for Punishment*, is supposed to protect against excessive punishment, such as putting traffic offenders to death in order to deter further traffic offences. Furthermore, it is supposed to set punishment apart from revenge, as it sets a limit on the amount of suffering that can be meted out on wrongdoers. Nozick (*Philosophical Explanations*: 366–368)[28] writes, "[retribution] sets an internal limit to the amount of the punishment, according to the seriousness of the wrong, whereas revenge internally need set no limit to what is inflicted"[29]. It is difficult, though, to see how a difference in the *degree* of suffering changes one *kind* of phenomenon (revenge) into another (punishment). Both revenge and punishment are retaliatory acts which involve

[28] Cited in Zaibert (2006: 87). Zaibert (2006) does not reference a year for Nozick's text.
[29] Nozick mentions four other features of retributivism which supposedly set it apart from revenge (see Zaibert, 2006: p 87), but, since I am focusing on the problems associated with the principle of proportionality, I shall not discuss them.

harming individuals for the (perceived) wrongs they have committed, and, as Zaibert (2006) points out, the harm caused by revenge, like punishment, can be equivalent to the initial harm caused by the wrongdoer. Imagine that a person is shot once, and retaliates by shooting the attacker once, is this an act of revenge or an act of punishment? The principle of proportionality supposedly distinguishes punishment from revenge, but it seems that we cannot use this principle to determine whether the act of shooting the attacker is revenge or punishment, as the retaliatory harm inflicted is equivalent to the initial harm caused by the attacker. If revenge and punishment really are two different kinds of phenomenon, as many claim[30], then it should follow that equivalent revenge is still revenge, and, conversely, disproportionate punishment should still be punishment. Consequently, we cannot answer the question "Why should punishment be proportionate?" by claiming that it has to be proportionate to distinguish it from revenge, as the *degree* of suffering inflicted on wrongdoers cannot be used to separate one *kind* of activity (punishment) from another (revenge).

My second response to the fairness theorist is that he has missed the deeper point behind the initial criticism (the criticism that it is difficult to determine what counts as proportionate). The deeper point is that it is difficult to fathom what counts as proportionate because the whole idea of criminals gaining an unfair advantage is obscure, and it misrepresents both what it is about Jeff's actions that make them wrong, and what disturbs us about the example. It is the undue harm suffered by Alice that is upsetting, and not the fact that Jeff has gained some obscure legal advantage. We should, following Ten (2000)[31], be trying to mend the harm that Jeff caused instead of trying to balance the equilibrium of benefits and burdens. Nevertheless, Ten's reformulated version of the fairness theory does not escape the problems associated with the notion of proportionality. If Alice had survived the ordeal, or she had friends or family that could claim compensation on her behalf, we would still face the problem of deciding, for example, how large a fine Jeff should pay if there was someone to pay it to.

[30] Zaibert (2006: 82) mentions a handful of philosophers who maintain that punishment and revenge are different activities. The list includes Elster (1990), Feinberg (1970), Flew (1954), Honderich (1967), Kleinig (1973), Lacey (1988) and Ten (1987).
[31] Cited in Scarre (2004).

The fairness theory, then, does not offer good reasons to punish Jeff, as the theory is plagued by the problem of explaining the notion of proportionality, and the practical difficulties of establishing when punishment fits the crime.

4.2.2 – The Expressive Theory

The expressive theory does not provide sufficient justification for punishment because it is a *non-sequitur* (Warburton, 2001), as it does not follow from the premise that resentment and condemnation need to be communicated to the conclusion that *punishment* is the most appropriate method of communication. Furthermore, the message that advocates of the expressive theory want to communicate is too open-ended. There are two components of the expressive theory, first, expressing resentment, and second, communicating censure. I discuss my objections to the expressive function first, and follow with my objections to the communicative function.

The expressive theory states that punishment is required to express our feelings of resentment. Jeff's community, however, was not incensed by Jeff's actions. One might say that they should have gotten angry, and that by not getting angry they did not show that they value Alice, or take the wrongs done to her seriously. Say, then, that the members of Jeff's community did experience anger and indignation, one could agree that these feelings are appropriate emotional responses to crime, yet still argue that we should resist the desire to hit back typically involved in these emotions (Duff, 2008). Klimchuk (2001: 81) puts it nicely, he says "[t]he desire for revenge, one might argue, is something to be overcome rather than satiated". The expressivist could reply that it seems odd to say that it is wrong to express emotions, such as anger and resentment, if they are appropriate responses to crime, and it is right to feel them (Scarre, 2004). There are, however, many different ways for people to express their anger or indignation, and the expressive theory does not give us a reason why our emotional responses to crime are best expressed through punishment. In other words, the expressive theory needs to tell us more about "why the infliction of suffering should be an appropriate way to express such proper emotions" (Duff, 2008: 12).

To the charge that we need to take the suffering of the victims seriously, and that, by allowing him to escape punishment, Jeff's community did not take Alice's suffering

seriously enough, one can pose the following question: "Why does taking crime seriously necessarily have to involve punishment?" To use Tabensky's (2006: 140) words,

> there are alternative means of one's understanding of the seriousness of a given transgression... One way is to acknowledge in deed and belief that the perpetrator is, in some sense, sick, and to try as hard as we can to cure him or her, or to change the circumstances leading to recalcitrant forms of criminal behaviour.

In other words, we can show that we take crime seriously by attempting to address the conditions that lead to crime, such as the lack of education or training (Rachels, 2007), social inequality and the lack of job opportunities (Tabensky, 2006). Attempting to prevent crime from occurring again by addressing these issues would demonstrate an appreciation of the seriousness of crime. So, once again, the onus is on the expressivist to show why punishment is the most appropriate way to show that we take crime seriously[32].

The second component of the expressive theory is that punishment is required to communicate the condemnation that criminals deserve. There are two questions that can be asked of this component of the expressive theory. First, why should the communication of censure involve 'hard treatment'[33]? Second, why is the point of communicating censure merely to *remind* wrongdoers that there are good moral reasons not to commit crime, and not to dissuade them?

To the first question then, why does communicating censure have to involve hard treatment? Censure can be communicated through a number of different means, such as an official sentence in a criminal court, by an additional formal reproof issued by a judge (or representative of the law), or by a process of wholly symbolic punishments that are not actually burdensome or painful (Duff, 2008). Censure can also be communicated by hard treatment, that is, through punishments that are actually burdensome or painful, such as imprisonment, fines or mandatory community service. Why should we choose to communicate censure through harsh punishments rather than the alternative methods mentioned by Duff?

[32] I will discuss the claim that we have to express attitudes such as resentment and indignation through punishment in more detail in Chapter Five, specifically *5.1 – Ineradicable Reactive Attitudes*.
[33] The term "hard treatment" comes from Feinberg (1970: 100)

One answer to the question of why we should communicate censure through hard treatment is that respecting persons requires holding them fully accountable for their actions, and consequently burdening those who behave badly. An integral feature of communicating censure is respecting wrongdoers as morally responsible, rational agents (Duff, 2008). Respecting the rational nature of individuals requires holding those individuals accountable for their actions, and this involves individuals and institutions "tracking" (Metz, 2006: 227) peoples' actions, and responding in kind to them. So we burden those who choose wrongly, and refrain from burdening (possibly even praise) those who choose appropriately (Metz, 2006). We respond "in kind" because the wrongdoers' own actions have invited us to respond in this way. According to Kant's (2002) first formulation of the Categorical Imperative, when we act in a certain way we show others how we think people should be treated, as we "proclaim our wish that our conduct be made into a 'universal law'" (Rachels, 2007: 139). When wrongdoers treat other people badly, they are inviting us to treat them badly, and by doing so we act in accordance with their own choices. By treating wrongdoers badly, we show that we respect their decisions, and allow their behaviour to determine how we respond to them (Rachels, 2007). So, in order to respect wrongdoers' rationality, we have to burden them because they have chosen to behave badly, and hard treatment is, of course, burdensome by nature. It follows, then, that we punish wrongdoers in order to burden them, as Kant (1965)[34] writes, the wrongdoer's "own evil deed draws the punishment upon himself".

Falls (1987) uses the example of being lied to by a friend to illustrate the importance of returning harm with "like" harm (in accordance with the *ius talionis*, the principle of proportionality/equality). She claims that, in order to hold wrongdoers accountable, we have to unequivocally communicate our disapproval of their actions, and stress why they were wrong. Calmly stating that the friend should not have lied does not, in Falls' (1987: 42) opinion, effectively communicate the pain the lie caused, nor the "unqualified insistence that we not be so treated". The appropriate way to communicate this pain and insistence would be to temporarily shun the friend who lied. In the same vein a verbal reprimand issued by the state would not adequately convey to wrongdoers the pain caused by their actions or the same categorical insistence that we should not be

[34] Cited in Rachels (2007: 139). Rachels does not reference a page for Kant's (1965) quote.

treated in this way. Only the temporary separation from society (through incarceration) effectively communicates the censure wrongdoers deserve. Holding wrongdoers accountable therefore necessarily involves hard treatment (punishment that will cause wrongdoers similar physical and/or psychological pain to the pain they inflicted on their victims). Furthermore, holding wrongdoers accountable does not involve reforming them, even though we hope that they will be motivated to refrain from committing crime in future. She writes "[we] hold him accountable for the sake of respecting him as a moral agent, not for the sake of improving him, even though we hope and insist upon this" (Falls, 1987: 43). Falls' view of punishment, and the expressive theory in general, avoids the charge levelled against consequentialism that it seeks to manipulate wrongdoers into conforming with the law. Most expressivists only attempt to modify offenders' behaviour by reminding them of the "good moral reasons" (Duff, 2008: 13) they have to refrain from committing crime.

I have three responses to the view that communicating censure has to involve hard treatment because we have to respect wrongdoers' rationality, and this involves holding them accountable for their misdeeds and responding in kind by burdening them. My first response is that the example Falls (1987) uses to demonstrate the importance of returning harm with "like" harm is flawed. Falls claims that we should shun a friend who has lied in order to convey the pain the lie caused, and the insistence that we should not be lied to, Calmly telling her she should not have lied will not suffice. This example, though, seems to be a false dichotomy, which is occurs when someone "sets up a dichotomy in such a way that it appears there are only two possible conclusions when in fact there are further alternatives not mentioned" (Warburton, 2001: 63). Falls' (1987) example only involves two possible options, namely calmly telling the friend she should not have lied, or shunning her. There are other ways of communicating pain and displeasure to the friend, such as adopting an affronted demeanour, and telling her in an offended and disapproving tone of voice she should not have lied, and the exclusion of these options undermines the example. Falls' claimed that, just as shunning effectively communicates pain and displeasure, incarceration effectively communicates censure, but since the example of the lying friend is flawed, the analogy between shunning the friend and incarcerating perpetrators fails. If there are other ways of communicating pain and displeasure (which

could prove to be more effective than shunning) to the friend who lied, then it follows that there are other ways of communicating censure to offenders (which could prove to be more effective than incarceration). Consequently, Falls' account is weakened by her exclusion of these other means of communication[35].

My second response is that responding in kind, and burdening wrongdoers, may obscure the message we are trying to communicate. Falls (1987: 46) claims that punishment is an act that demands that wrongdoers respond to their punishment as "moral agents", and reflect on the censure that has been communicated. She also claims that hard treatment effectively communicates this censure. It is questionable, though, whether hard treatment is conducive to moral reflection. Consider Nietzsche's (1989: 81) statement that "punishment makes men hard and cold; it sharpens the feeling of alienation; it strengthens the power of resistance". A central component of communicating censure is treating wrongdoers as *rational* agents. So, if we want to elicit a reasoned moral response from wrongdoers, then surely appealing to their *rational* nature by expressing disapproval and denouncing their actions through rational discussion would be a more effective way of firstly, respecting their rationality and secondly, eliciting a moral response? People generally respond emotionally when they are harmed. I have already mentioned that victims of crime respond emotionally with feelings of anger or resentment (Duff, 2008), and it stands to reason that causing wrongdoers physical and/or psychological pain will mostly lead them to respond *emotionally* rather than *rationally*, and experience feelings of distress, anger or frustration for instance. So it appears unlikely that communicating censure through hard treatment will cause wrongdoers to respond as rational moral agents.

My third response to the view that punishment has to involve hard treatment is that, if we want to take crime seriously, then our focus should not be on merely communicating censure. The point of communicating censure is *not* to dissuade wrongdoers from committing crime, but to hold them accountable for their actions (Falls, 1987). Censuring wrongdoers respects them as rational moral agents, and only seeks to dissuade wrongdoers from committing crime by reminding them of the good reasons they

[35] I mention other possible means of communicating censure earlier in this section and in *5.1.1 – My First Response to the Strawsonian*.

have to not commit crime (Duff, 2008). Respecting wrongdoers as rational, responsible agents involves allowing them to determine their own actions, and leaving them "free to remain unpersuaded" by the censure that has been communicated (Duff, 2008: 14). If, however, we are serious about lowering crime rates, and preventing crimes (especially atrocious ones) from occurring again in future, it does not make sense to run the risk that those who have committed crimes will not be persuaded by punishment, and only encourage them to refrain from crime by expressing our disapproval of their actions. Are we really happy to leave wrongdoers to decide for themselves whether or not they want to commit crimes in future? The alternative to letting wrongdoers determine their own behaviour is to try rehabilitating or reforming them with "psychological therapy, educational opportunities, or job training, as appropriate" (Rachels, 2007: 135). Rehabilitating wrongdoers, though, violates the central component of communicating censure, which is respecting wrongdoers as "rational and responsible" agents (Duff, 2008: 13). In particular, rehabilitation violates the right wrongdoers have, as autonomous beings, to be self-governing people who get to determine how they will behave (Rachels, 2007). Why, though, should we respect wrongdoers' rationality (autonomy)? The view that we should treat offenders as autonomous moral agents is, of course, a Kantian principle, and one I shall explore further when I discuss the Kantian theory of retributivism, which I do next.

4.2.3 – The Kantian Theory

In this section I want to focus predominantly on the Kantian principle of respecting people as ends in themselves, but, before I do, I need to say more about the fact that consideration of autonomous choice *does* apply to my original thought experiment. I mentioned in *4.1.3 – The Kantian Theory* that I could not remove the consideration of autonomous choice from the experiment, as Jeff's heinous actions would not be classified as a crime without the presence of criminal intent. I argue that, even though this consideration applies the original experiment, it is not sufficient for the response that Jeff should be punished. My reason for saying this is that criminal intent (*mens rea*) alone does not constitute a crime. Let me change the original thought experiment slightly to illustrate my point. Imagine that Jeff intends to rape Alice and follows her after she

leaves the café, but, before he can attack her, a couple of policemen walk past. Jeff does not attack while the policemen are there, and Alice has time to get to her car and drive away. In this situation, *mens rea* is present, as Jeff intended to rape Alice, but *actus reus* is absent, as no harm came to Alice. No crime has been committed and so there is no reason to arrest Jeff and charge him with rape. I doubt many would say that Jeff should be punished in this situation just because *mens rea* was present, so why should we say that the presence of *mens rea* alone justifies punishing Jeff in the original thought experiment when both *mens rea* and *actus reus* are present? I maintain that the consideration of autonomous choice only applies to the original thought experiment because Jeff *actually* raped Alice, thus causing her undue harm (*actus reus*). I mentioned in *4.1.1 - The Fairness Theory* that this unlawful harm should be our primary concern (Duff, 2008). Kantians might respond that we have a duty to punish Jeff, as by doing so we hold him accountable for his actions and thereby respect him as an end in himself. Why, though, should we respect people as ends in themselves? I have two main criticisms of this principle. First, there seems to be a contradiction between the principle of respect for persons and the principle of proportionality. Second, we are supposed to treat people as ends in themselves because they have intrinsic worth in virtue of possessing the capacity for autonomy (the ability to set goals and pursue them). I argue (*contra* Kant) that autonomy is *not* intrinsically valuable. Before I discuss these criticisms, I need to say more about what it means to treat people as ends in themselves.

Kant (2002) states, repeatedly and emphatically, that human beings should always be *treated as an end, and never a mere means*. What does it mean to treat persons as ends in themselves, and not merely as a means? Treating people as ends in themselves involves respecting their rationality (specifically their capacity for autonomy), and acknowledging that they have their own projects that they wish to pursue (O'Neill, 2002). Treating someone as a *mere means*, on the other hand, "is to involve them in a scheme of action *to which they could not in principle consent*" (O'Neill, 2002: 717)[36]. So treating people as a mere means involves using them as instruments or objects to achieve our own goals or ends without their approval or consideration of the fact that they too have their own ends that they wish to pursue. None of this, however, implies that people cannot be

[36] Italics in original.

treated as means. In fact, many of our day to day interactions require treating people as means, such as interactions which involve the exchange of goods or services. If my tap is leaking and I call a plumber, I use the plumber as a means to fixing the leaky tap, and he uses me as a means to earn an income. What is important in this scenario is that each party *consents* to the transaction, and appreciates that the other has his own goals and is not just a tool or an object "to be manipulated" (O'Neill, 2002: 717). Kant (2002: 693) is explicit on this point, he states that "a human being… is not a thing and hence not something that can be used *merely* as a means"[37]. If I forced the plumber to fix the leaky taps (say I threatened to shoot him if he did not) then I would be using him as *mere* means, as he cannot consent to be treating in this way, "for consent precludes… coercion" (O'Neill, 2002: 717). In other words, if I did not threaten the plumber, he would not have fixed my taps free of charge. So we have to respect persons because they are autonomous (they can set ends for themselves), and this capacity is intrinsically valuable. Respect entails treating persons as ends in themselves, and not as mere means to other ends.

Furthermore, we are, according to Kant (2002), supposed to treat people as ends-in-themselves, and not as mere means, because they have *intrinsic value*, or dignity. Humans have this dignity "because they are *rational agents*, that is, free agents capable of making their own decisions, setting their own goals, and guiding their conduct by reason" (Rachels, 2007: 131)[38]. In other words, they are intrinsically valuable because they are autonomous; they can set themselves goals and pursue them, and such value merits respect (Metz, 2006). It is important to note that we do not respect people just because they are human beings, we respect their rational nature, which, according to Kant (2002: 693) "*exists as an end in itself*"[39]. So we do not respect specific individuals, but rather the humanity embodied in each individual. The ability to establish worthwhile goals sets persons apart from beings that cannot determine whether their desires are worth acting on, beings whose actions are based on instinct or conditioning (Metz, 2006).

Having explained the principle of treating people as end-in-themselves, I want to start my discussion of the objections to this principle by getting a popular challenge to

[37] Italics in original.
[38] Italics in original.
[39] Italics in original. I shall use "end in itself" and "intrinsically valuable" interchangeably.

Kantianism (and to retributivism in general) out of the way. One way of attacking the principle of respect (one I will *not* pursue), is to argue that people (wrongdoers included) are not autonomous, and hence not morally responsible agents. Wrongdoers only deserve to be punished if they are responsible for their actions, if they are not responsible, then they cannot be punished[40]. I seek to avoid questions about moral responsibility, however, as, even if we could provide a plausible account of moral responsibility, we can still ask "why a morally autonomous being of intrinsic worth who voluntarily causes the undue suffering of another deserves a return of proportionate suffering or even any suffering at all" (Falls, 1987: 26). In other words, even if we assume for the sake of argument that people *are* autonomous, we still need an argument for why autonomous people *deserve* to suffer just because they caused harm. The versions of retributivism that I discuss in this thesis attempt to provide arguments for why wrongdoers *deserve* to suffer, which is why I focus on them and not on questions of moral responsibility.

My first criticism of the principle of respect is that there seems to be a contradiction between it and the principle of proportionality (or equality). Kant (2003) was adamant that only the *law of retribution* (*ius talionis*), otherwise known as the principle of equality, could specify the kind and amount of punishment to be meted out to criminals. He also held that retribution "must still be freed from any mistreatment that could make the humanity in the person suffering it into something abominable" (Kant, 2003: 106). The 'humanity' Kant is referring to here is the capacity for rational thought, which allows agents to choose worthwhile goals, as, for Kant, it is "the purely human element in man" (Kant, 1965: 101-102 and 132-133)[41]. That is, rational thought is the capacity that sets humans apart from other creatures. In order to respect the humanity of a person, punishment should not impair or eradicate that person's ability to think rationally. Accordingly, savage punishments, such as torture, which would ruin the agent's capacity for rational thought, are prohibited. The prohibition of certain punishments is, however, contrary to the principle of proportionality, which demands like punishment for any undue suffering caused (Falls, 1987). In order to preserve the principle of proportionality,

[40] For arguments on this debate (whether people are autonomous) see the collection *Judging and Understanding* edited by Tabensky (2006), specifically 'Part II – Free Will, Determinism and Moral Responsibility: Challenging Retributive Judgement', and Metz's paper in the same collection.
[41] Cited in Falls (1987: 37).

a Kantian retributivist would have to maintain that cruel murderers who torture their victims before killing them deserve to be treated in exactly the same fashion. If the Kantian is unwilling to punish cruel murders "with the same degree of cruelty" (Scarre, 2004: 118) they inflicted on their victims, then he is not fully committed to the principle of proportionality. So while torture is prohibited by the principle of respect for persons, it is allowed by the principle of proportionality. There is, therefore, a contradiction between the principle of respect for persons and the principle of proportionality (Falls, 1987).

Let me relate the above discussion of the tension between the principle of respect for persons and the principle of proportionality back to the example of Jeff. The principle of proportionality requires that Jeff suffer to the same extent Alice suffered. So, according to the principle of proportionality, Jeff should be raped and murdered. The principle of respect for persons however prohibits such punishments, as they compromise the human dignity (conceived of as the capacity for rational thought) of the wrongdoer (Scarre, 2004). A strict application of the *ius talionis* could be one way of dealing with the aforementioned contradiction. That is, the contradiction can be dealt with by allowing punishments that are considered to be cruel or degrading to be meted out to criminals who treated their victims in a cruel and demeaning manner. The problem with this move, however, is that few today would find such punishments to be morally acceptable (Scarre, 2004). The Kantian is, it seems, in a quandary. If he holds onto the principle of proportionality, then he cannot be true to the principle of respect for persons. If he is true to the principle of respect for persons, then he is not fully committed to the principle of proportionality.

My second criticism of the principle of respect is that it relies on the mistaken view that autonomy is intrinsically valuable. Kant (2002) insists that we treat people as ends in themselves because they are intrinsically valuable in virtue of being autonomous agents who can set goals for themselves and pursue them. What is meant by intrinsically valuable here? A more detailed discussion of the notion of intrinsic value will prove useful.

Traditionally, something is intrinsically valuable if it has that value in itself, and does not derive that value from any relations, connections it has, or consequences it leads to (Nozick, 1981). Intrinsic value is therefore non-derivative (Zimmerman, 2007).

Similarly, McLeod (2003: 11) says, "[the] *intrinsic* value of a thing is the value it has simply in virtue of what it is, rather that the value it has in virtue of what it leads to, signifies, entails, purchases, and so on"[42]. Traditionally, intrinsic value is seen as a non-relational concept, that is, if something is intrinsically valuable, it is not valuable *for* someone or something else (Lemos, 1994). Extrinsic value, on the other hand, is the value something has in virtue of its connections, relations or consequences. If something is extrinsically valuable, then that value is based "upon the occurrence of something else" (Chisholm, 1978: 121). Extrinsic value is therefore derivative in that it is valuable because it is related in some way to something else which is considered good or valuable (Zimmerman, 2007). So, traditionally, when something has *intrinsic* value, it is valued for its own sake, and its value is non-relational and non-derivative. Conversely, when something has *extrinsic* value, it is valued for the sake of something else, and its value is relational and derivative.

Zimmerman (2007) suggests that, when we consider the claim that rational beings are ends in themselves that possess intrinsic value and dignity, we understand Kant, and others who follow him, as being concerned, not with the traditional concept of intrinsic value, but with the question of how we should treat such beings. Kant and his followers, as I have mentioned, are indeed concerned with the way rational beings should be treated, as we are, according to Kant (2002) supposed to treat people as ends in themselves, and not merely as means, in order to respect their rationality (more specifically their autonomy). We should not violate their autonomy by lying to them, manipulating them or using them for our own purposes. In the case of punishment, wrongdoers are not to be used as a mere means to the end of social welfare. Even though Zimmerman (2007) is right to say that Kantians are concerned with the way we should treat rational beings, the reason we are supposed to treat them as ends in themselves is *because* they are autonomous, and when one considers the descriptions of autonomy provided by Kantians, it seems that they are, arguably, employing the traditional notion of intrinsic value.

The main features of the traditional concept of intrinsic value are that it is non-derivative (it does not derive its value from anything else) and non-relational (it is not valuable *for* anyone or anything). To say that rationality is an end-in-itself and is not to

[42] Italics in original.

be used as a mere means to any other end, as Kant (2002) does, implies that the capacity is to be valued for its own sake, not for the sake of any other ends or consequences, and hence is not to be considered as instrumentally valuable[43]. In addition, consider the following statement from Falls (1987: 40):

> The principle [of respect] maintains that individuals, simply in their role as persons - and hence independently of their behavior, social status, usefulness, or desirability - have intrinsic worth meriting them a certain kind of treatment. The judgment that they have this worth is based upon nothing more than that they have a capacity for reasoning and autonomous moral decision-making.

So, according to Falls, humans are intrinsically valuable *just because* they possess the capacity for autonomy, and not for any other reason. Remembering that *autonomy* is an end in itself, and not the individual who possesses it *per se*, it stands to reason that autonomy is not valuable because it allows humans to (by choosing and acting appropriately) be useful or desirable, behave admirably or gain social status; it is valuable in its own right. Autonomy is therefore to be seen as non-derivatively (it does not derive its value from anything else), rather than derivatively valuable. Moreover, autonomy is also, arguably, presented as a non-relational value, as it is not valuable because of what it allows the autonomous person to do, and is therefore not valuable *for* any particular person *per se*. From the above, it appears, *prima facie* at least, that, contra Zimmerman (2007), the traditional notion of intrinsic value is being employed in the Kantian discussions of autonomy, as autonomy is presented as being non-derivatively and non-relationally valuable.

I have just argued that autonomy is presented as being intrinsically valuable in the traditional sense. I now want to argue that, even though autonomy is presented in this way, it is not, in fact, intrinsically valuable. Writers such as Moore and Aristotle suggest that in order for something to be intrinsically valuable it has to have that value "in isolation" (Chisholm, 1978: 121), in other words, it would be valuable even if it was the only thing that existed. This "ontological isolationism", as Lemos (1994: 10) calls it, is problematic as there are certain things, which are considered to be intrinsically valuable, that cannot exist in isolation. Lemos (1994) discusses the example of happiness. He asks

[43] Instrumental value is a form of extrinsic value, as whatever has instrumental value derives that value from something else it is related to, specifically to some other end that is considered to be good (Frankena, 1963).

the reader to consider "the fact of Smith's being happy" and points out that "Smith's being happy could not exist without Smith's existing" (Lemos, 1994: 11). So it is impossible for happiness to exist in isolation, as a certain type of being (a being that has the capacity to be happy) has to exist in order for happiness to exist. Furthermore, when we consider what it is about happiness that we value for its own sake, it appears that we value what it feels like to be happy, that is, the *experience* of being happy. In this regard happiness can be seen as inherently valuable. If something is inherently valuable, it is valuable because the contemplation or experience of it is valuable or "rewarding in itself" (Frankena, 1963: 66). It is important to note that inherent value is still a form of extrinsic value, as what has inherent value (in this case happiness) has it because it is in some way related to something else which is purportedly good (the experience of being happy) (Zimmerman, 2007). Happiness, then, appears to be extrinsically valuable, as it is both relationally and inherently valuable, as it is valuable *for* someone (the person who is happy) and what is of value for that person is the experience of being happy.

What of autonomy then, is it valuable in the same way that happiness is valuable? The first thing to notice about autonomy is that we cannot talk about it without talking about an individual who possesses the capacity, that is, without talking about a being who possesses, and can make use of, this capacity. It is difficult to see how autonomy can be valuable in isolation, as its value is dependent on the existence of beings that possess and use it. Moreover, it is hard to see how the capacity could have any value if there was nothing to deliberate about, or no options available to choose from. Imagine that only one autonomous being exists and that this being is incapable of performing any action. This being cannot set itself goals, as there are not options available to it, and so nothing for it to do. Even if any options existed (say it could decide which direction it wanted to go in, and whether it wanted to go there slowly or quickly), and it could determine which course of action it wanted to follow, it would still not be able to pursue that goal, as it cannot act in any way. In this scenario it seems as if autonomy has no value at all, as the capacity is impotent. Agents cannot be autonomous if they cannot act on decisions they have made, and they cannot make these decisions if there are no options to reflect on and choose from. Remembering that extrinsic value is "dependent… upon the occurrence of something else" (Chisholm, 1978: 121), it can be said that autonomy is extrinsically

valuable, as its value is dependent on the presence of viable options for agents to choose from, and the ability to act on those decisions.

I have mentioned two types of extrinsic value, namely inherent value (which is the value something has because experiencing or contemplating it is valuable or worthwhile in itself) and instrumental value (the value something has a means to an end that is considered good). Does autonomy have inherent or instrumental value? It is reasonable to say that the experience of making and acting on our own decisions is generally more rewarding and valuable than the experience of having somebody else make those decisions or direct our actions. One merely needs to consider the life of a slave, the experience of being manipulated by other people, or being forced to do something against one's will to appreciate this point. Governing one's own life, by setting one's own goals and pursuing what one wants to pursue, is indeed a rewarding experience, but we do not pursue goals merely for the sake of this experience, we pursue goals that we want to *achieve*. The experience of directing our own lives can, in fact, be quite frustrating when, after much deliberation and effort, we are unable to realise the projects we have set ourselves. So, even though it is rewarding to be able to say that we have achieved goals that we have selected for ourselves, and that we have achieved them through our own efforts, it seems that much (though not all) of autonomy's lustre is lost if we do not achieve these goals at all. To reiterate, autonomy does seem to have inherent value, but being autonomous *and* achieving one's objectives seems to be more rewarding, and hence more inherently valuable.

From the above, it seems that autonomy cannot be seen as an end-in-itself, as, traditionally, if something is an end in itself then we pursue it for its own sake, and not for the sake of anything else. Consider the example of happiness, when asked why we pursue happiness it is not bizarre to say that we do so *just for the sake of being happy*. While one could plausibly say that one desires autonomy *just for the sake of being autonomous* (one might have been a slave for instance), it is not contentious to say that we want to be autonomous so that we may do certain things and achieve certain goals. In other words, we can value autonomy, not just for the sake of being autonomous, but for the sake of the goals we can choose to pursue. With regard to others, we are not supposed to use people as a mere means to any other end, as doing violates their autonomy (Kant

2002). We are not supposed to deceive, manipulate or use people for the sake of furthering our own ends, but acknowledge that they have projects of their own, and, rather than interfere with them, occasionally aid them in the pursuit of these projects (O'Neill, 2002). Metz (2006: 227) states that "[respect] also requires… helping [people] on occasion to develop their capacity to choose goals or to attain the goals they have chosen". Notice, though, that we are not supposed to merely foster their *capacity* for autonomy, but help them *achieve their goals* as well. Furthermore, while it is rewarding firstly, to see others exercise their autonomy by pursuing and realising goals they have chosen for themselves, and, secondly, to help others in their endeavours, I doubt anyone would implore us to help others accomplish morally reprehensible goals. The achievement of such goals may be rewarding for the individuals who chose them, and thus inherently valuable, but ascribing any other value to those goals or to the autonomy of the agents who chose them seems inappropriate. In such cases, the Kantian cannot say that autonomy is still to be valued for its own sake, as the agents are not exercising their autonomy just for the sake of being autonomous, but for the sake of pursuing their nefarious ends. Given that autonomy is being exercised for the sake of these ends, and that morally reprehensible ends are not generally considered to be valuable or worth pursuing, it is hard to see how autonomy can derive any value from those ends. So even though we can pursue the autonomy for its own sake, we can also pursue the capacity for the sake of something else, namely the goals that we, and others, choose by exercising that capacity. Furthermore, as autonomy can be exercised for the sake of those goals, it seems that it can derive value from worthwhile goals, but not from malicious, morally reprehensible ones. What I have just said could imply that autonomy has instrumental value, but it is important to note that it is not the mere *capacity*, but rather autonomous *behaviour*, that is valuable as a means to the ends we want to achieve, as we could not achieve our goals if we never acted on our decisions.

In brief, I argued that, even though autonomy is presented as being intrinsically valuable in the traditional sense, it is extrinsically valuable. It is extrinsically valuable because it derives its value from the existence of possible options that we, as agents, can reflect on and choose from. The value of autonomy is also dependent firstly on the ability to act on the choices we make, and, secondly, on our choices being worthwhile and

morally acceptable. I then attempted to ascertain whether autonomy has either inherent or instrumental value. I concluded that autonomy has some inherent value, as the experience of being autonomous is indeed rewarding (and hence valuable *for* the person who possesses it), but not as rewarding as being autonomous and achieving the goals one has set. I also concluded that, while we can pursue autonomy merely for the sake of being autonomous, autonomous behaviour (of which autonomy is constitutive) is instrumentally valuable as a means to the goals we seek to accomplish, but, as I mentioned, only insofar as those goals are worthwhile and morally acceptable. If what I have said is right, then it follows that autonomy, because it is *not* intrinsically valuable, is not worthy of the utmost respect. If autonomy is not worthy of the utmost respect, then one cannot say that we should not try to rehabilitate offenders, or punish them as an example to potential offenders, because doing so would violate their autonomy, and that we should not violate autonomy because it is intrinsically valuable. I do not mean to imply, though, that autonomy has *no* value, and should not be respected at all, but I shall return to this issue, as well as the issues of rehabilitation and deterrence, in Chapter Six.

Section Summary

In this section I argued that the considerations retributivists would use to justify punishing Jeff would not be sufficient justifications even if they did apply to the thought experiment. The considerations of the fairness theory are not sufficient because: firstly, the principle of proportionality does not seem to be feasible, and, secondly, it is not clear why punishment should be proportionate. The considerations of the expressive theory are not sufficient because: firstly, the theory is a *non-sequitur*, secondly, punishment may obscure the message expressivists are trying to communicate, and, thirdly, we should, if we take crime seriously, be focused on more than merely communicating censure. Finally, the considerations of the Kantian theory are not sufficient because: firstly, criminal intent does not constitute a crime, secondly, there is a contradiction between the principles of respect and proportionality, and lastly because autonomy is not intrinsically valuable.

Chapter Summary

The point of this chapter was to argue that the retributivist considerations are neither necessary nor sufficient justifications for punishment, as those considerations do not apply to the thought experiment (except for the Kantian consideration of autonomous choice), and would not prove sufficient even if they did apply.

Chapter 5: Should We Abandon the Intuition?

I have argued that the retributivist intuition that wrongdoers deserve to suffer, and hence should be punished, is unreliable, and cannot serve as the basis of justification for the practice of punishment. I used the thought experiment of Jeff to illustrate the unreliability of this intuition, as the thought experiment excludes most considerations offered by retributive theorists to justify the intuition. Even though these considerations were removed, the intuition still remains. The fact that the intuition remains without the retributivist considerations is enough, in my opinion, to establish that those considerations are not necessary for the response that Jeff should be punished. The considerations are not necessary because the desire to see Jeff punished arises even when the considerations are not present, and they are not sufficient because those considerations are not satisfactory justifications for the response that Jeff should be punished. Previously I mentioned that crime often arouses emotional responses, such as "the resentment of 'retributive hatred'" (Duff, 2008: 11), that involve a desire to make wrongdoers suffer. Given that the reasons offered by retributivists to justify the intuition are not satisfactory ones, it appears that the intuition is motivated, not by these reasons, but by the desire to make the wrongdoer suffer. As such, the intuition seems to be an emotional, rather than an intellectual response (Crisp, 1997). I have mentioned (in Chapter Three), though, that this intuition is a deeply ingrained one. Burgh (1982), for one, claims that it is undeniable that most people have this intuition[44]. The main question of this chapter, then, is: given that this deeply rooted intuition is unreliable, should, or even can, we abandon it (and the related emotions and activities), or is the price too high?

This chapter has two main sections, in the first section (*5.1 – Ineradicable Reactive Attitudes*) I discuss a possible response to the above question (of whether or not to abandon the intuition) which could be offered by an advocate of Strawson's view of reactive attitudes (henceforth known as the Strawsonian). The Strawsonian could respond by claiming that we cannot abandon the intuition (which is based on retributive reactive attitudes) even if it is unreliable, as our commitment to reactive attitudes in general (and to the practices that express them) is ineradicable. I argue that this view is specious, and

[44] Burgh's (1982) view is based on empirical observation, and I tend to agree with him, as my own empirical observations are in line with his.

have two responses to it. First, the fact that we have reactive attitudes does not commit us to any particular way of expressing them. Even if these reactive attitudes are ineradicable (which is debatable), it does not follow that retributive punishment is a natural outgrowth of these attitudes. Second, the reactive attitudes of resentment and indignation are misidentified and misrepresented by retributivists. It is natural, claims Mill (1907), to resent and to repel those who harm us or those with whom we sympathise. This inclination is common to all animals, as they strike back at those who cause them, or their young, harm. The reason, then, that advocates of Strawson, as well as retributivists, misidentify this inclination to strike back is that animals strike back in order to *protect* themselves or their young, and protection is a distinctly consequentialist concern, not a standard retributivist one. The second section of this chapter (*5.2 – Criticisms of Mill*) concerns two possible criticisms of Mill's view, and my subsequent rejection of those criticisms. I devote the second section of this chapter to dealing with the objections to Mill (1907) because his view is especially important, as it provides the foundation for my own (consequentialist) account of punishment, an account which, I contend, can accommodate the intuition, though not as a justification.

5.1 – Ineradicable Reactive Attitudes

In this section I discuss the Strawsonian response to the question of whether or not we should abandon the central retributivist intuition that wrongdoers deserve to suffer. The response is that we cannot discard this intuition, or the practice of punishment, because they are based on ineradicable reactive attitudes. This response is based on Strawson's (1974) view that we have natural reactive attitudes which cannot be abandoned because of certain theoretical considerations. Before I examine the Strawsonian response, I first have to describe Strawson's view of reactive attitudes.

Reactive attitudes are feelings that we have towards other people, and arise from our interactions with them (Allais, 2008b). They include "gratitude, resentment, forgiveness, love and hurt feelings" (Strawson, 1974: 4). These attitudes presuppose a demand that participants involved in interpersonal relationships manifest a degree of respect and good will to other participants (Allais, 2008b), and are responses to the perceived good will, ill will or indifference of others (Eshleman, 2004). The reactive

attitudes that are of interest here are those of "offended parties" (Strawson, 1974: 4), viz. responses to manifestations of ill will, such as the attitude as resentment. For the sake of simplicity, I follow in Feinberg's footsteps, and use the term 'resentment' to refer to the "various vengeful attitudes" (Feinberg, 1970: 101) that crime arouses. Resentment, for Strawson (1974), is a personal reactive attitude, which is a natural reaction to the ill will or indifference shown to us by others. Strawson contrasts personal reactive attitudes with "the sympathetic or vicarious" attitudes (Strawson, 1974: 14), and the vicarious analogue of resentment is indignation. So while personal attitudes are responses to harms done to ourselves, vicarious attitudes are responses to harms done to other people (Allais, 2008b), and both "types of attitude involve, or express, a certain sort of demand for inter-personal regard" (Strawson, 1974: 16). That is, interpersonal relationships, for Strawson, are based on a requirement that those involved display a certain amount of respect and benevolence to other participants. It is also important to note that the reactive attitudes (personal and vicarious) are expressed by those involved in what are considered to be *normal* interpersonal relationships, as Downie (1966: 33) states "[the] reactive attitude (or range of attitudes) is that which characterises the inter-personal behaviour of normal adults".

Having described reactive attitudes, I turn to the Strawsonian response to the central question of this chapter, namely whether we should discard the retributive intuition and the practice that expresses it, namely punishment. According to Strawson (1974: 13), our commitment to interpersonal relationships is "part of the general framework of human life", and is

> too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical commitment might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand them precisely is being exposed to the range of reactive attitudes and feelings that is in question" (Strawson, 1974: 11).

In light of these comments, a Strawsonian could say that it does not matter that the intuition to punish wrongdoers is unreliable (as I have argued it is), as this intuition is based on the attitudes of resentment (when we have been harmed) or indignation (when others have been harmed), and these attitudes cannot be put aside because of any "theoretical commitment". Having these attitudes is an integral feature of being human, and, even if I am right to say they are unreliable, we are incapable of relinquishing them

completely. Moreover, we cannot do away with the practice of punishment, as punishment expresses these natural attitudes, as Strawson says "the preparedness to acquiesce in that infliction of suffering on the offender which is an essential part of punishment is all of a piece with this whole range of attitudes of which I have been speaking" (Strawson, 1974: 22). Punishment, therefore, goes hand in hand with the attitudes of resentment and indignation, as these attitudes involve a desire to make offenders suffer through punishment. Resentment and indignation can therefore be classed as "*retributive* reactive attitudes" (Allais, 2008b: 55), ones that include the view that the offender should be censured. "Retributive" here refers to the idea that proportionately condemning wrongdoing is the right thing to do (Allais, 2008b). The Strawsonian response, then, is this: we cannot forsake our retributive reactive attitudes (or the practice of punishment that expresses them), even if they are unreliable, as our commitment to reactive attitudes in general is "ineradicable" (Downie, 1966: 36).

Strawson (1974: 9) does discuss times when we do discard these reactive attitudes, and adopt what he terms an "objective attitude" towards other human beings. While participant reactive attitudes are responses to *normal* human beings, the objective attitude is adopted towards *abnormal* humans (Downie, 1966), those who cannot partake in normal interpersonal relationships because of psychological abnormalities or moral immaturity (Strawson, 1974). We adopt the objective attitude towards those who are "warped or deranged, neurotic or just a child" (Strawson, 1974: 8). One regards these individuals as ones that should be treated or controlled (Eshleman, 2004). While objective attitudes are generally reserved for abnormal behaviour (Downie, 1966), there are times when we can adopt the objective attitude towards *normal* humans, as "a refuge… from the strains of involvement; or as an aid to policy; or simply out of intellectual curiosity" (Strawson, 1974: 9-10). One can imagine a soldier at war as an example of one who adopts the objective attitude towards normal humans. Even though we can suspend our reactive attitudes to normal humans in certain situations, we cannot discard them completely (Strawson, 1974). We cannot maintain the objective attitude for too long, as the reactive attitudes are natural reactions, and we cannot do what "it is not in our nature to (be able to) do" (Strawson, 1974: 18). He goes on to claim that, even if we could adopt the objective attitude indefinitely (which we cannot, as we do not have a

choice in the matter), our lives would be impoverished as a result. There does seem to be some truth in this statement, as resentment seems to be a vital component of our moral makeup (Scarre, 2004). Butler (1953)[45], for instance, claims that resentment is both natural and essential for self-preservation. We need to get angry and indignant "against vice and wickedness" (Butler, 1953: Sermon VIII, para. 12: 130)[46] so that we do not tolerate unwarranted injury, and act to prevent it from happening again in future. In a similar vein, Scarre (2004: 101) claims that we "*ought* to" get angry and resent harms that we, and those close to us, suffer, being indifferent to those harms would leave one wondering whether we cared about anything at all. In light of these considerations, it seems to be true that our lives would be poorer if we abandoned our commitment to our reactive attitudes, and the practices that express them (Downie, 1966).

5.1.1 – My First Response to the Strawsonian

My first response to the Strawsonian response just discussed is that, even if we acknowledge that the attitudes resentment and indignation are natural and commonplace, it does not follow that we have to be committed to a particular means of expressing them, viz. the practice of retributive punishment. Tabensky (2006:144) claims that the "attitude of revengefulness [characterised by a desire for retribution] is laden with robust theoretical commitments which are not commonplace", specifically the commitment that wrongdoers *deserve* to suffer by being punished for what they have done. There are many examples of people who do have the attitude of revengefulness, and do not desire to punish wrongdoers. Reports of the South African Truth and Reconciliation Commission offer many examples[47]. Moreover, retributive punishment has not been the standard practice of judgment throughout history, as Tabensky (2006: 144) says "our practices of judgement have varied enormously across the span of time. It would be the height of absurdity to claim that our reactive attitudes have never changed in the light of the flux of history".

[45] Cited in Scarre (2004).
[46] Cited in *Ibid*., at 103.
[47] Allais (2008b: 39- 41) discusses two such examples. The first is of Babalwa Mhauli, who sought to forgive those who brutally murdered her father. The second example is of Pearl Faku and Doreen Mgoduka, whose husbands were murdered by Eugene de Kock. After a meeting with de Kock, a man nicknamed "Prime Evil", Faku said that she forgave him. Allais' (2008b) paper is devoted to making sense of the forgiveness given by these women.

To illustrate Tabensky's point that the attitude of revengefulness is 'theory laden', and that our practices have differed throughout history, consider J. F. Stephen's (1863: 99)[48] famous statement that "[the] criminal law stands to the passion of revenge in much the same relation as marriage to the sexual appetite", meaning that the law, like marriage, is an appropriate outlet for, and safeguard against, the natural dispositions of revenge and sexual desire (Scarre, 2004). Sexual yearning is indeed natural, but the activity of sex itself has taken on many different forms, and has been practised in different ways by different groups throughout the years. Marriage is not the universal conduit for sexual yearnings. Take Eskimo customs, which are very different from Western customs, as an example:

> The men often had more than one wife, and they would share their wives with guests, lending them out for the night as a sign of hospitality. Moreover, within a community, a dominant male might demand – and get – regular sexual access to other men's wives. The women, however, were free to break these arrangements simply by leaving their husbands and taking up with new partners – free, that is, so long as their former husbands chose not to make trouble. All in all, the Eskimo custom was a volatile practice that bore little resemblance to what we call marriage (Rachels, 2007: 17).

Even in the Western tradition, it has not always been the view that sex should only be practiced in marriage. During the first centuries of the Christian era, the Manichaean religion (which combined Christian and Zoroastrian teachings) condemned all "sex, even in marriage" (Russell, 1974: 326). According to Manichaeans, all matter was fundamentally bad, so sex and reproduction were bad (Midgley, 2001). Western views on sex and marriage have also been changing since Stephen's time (he wrote in the late 1800s). Consider the following statement from Singer (1999: 5):

> in the nineteenth century when data on the moral belies and practices of far-flung societies began pouring in. To the strict reign of Victorian prudery the knowledge that there were places where sexual relations between unmarried people were regarded as perfectly wholesome brought seeds of a revolution in sexual attitudes

Today many of our moral attitudes have changed dramatically, including our attitude towards extramarital sex, and, although this change in attitude might be considered controversial, sex outside of marriage is not generally seen as taboo.

[48] Cited in Feinberg (1970: 100-101).

From the above we can see that, while sexual desire is generally considered to be natural, this desire has been expressed in many different ways over the course of history. The same, I contend, can be said of the retributive reactive attitudes. Resentment and indignation may be natural, but our practices of judgment have indeed varied over the years. Braithwaite (1997: 3)[49] suggests that throughout history, restorative, and not retributive, justice has been the prevailing criminal justice system for most groups of people, including "ancient Arab, Greek, Roman, Indian, Hindu, Buddhist, Taoist, and Confucian traditions". Restorative justice, unlike retributive justice, is not solely focused on punishing perpetrators, but on repairing the harm done, and restoring relationships between perpetrators and victims (Allais, 2008a).

In modern, predominantly Western, cultures, punishment has become the conventional symbol of condemnation, which in Feinberg's (1970: 100) estimation, is

> neither more nor less paradoxical than to say that certain words have become conventional vehicles in our language for the expression of certain attitudes, or that champagne is the alcoholic beverage traditionally used in celebration of great events, or that black is the colour of mourning.

Feinberg's analogy between punishment expressing condemnation and champagne being used for celebratory purposes is not sufficient to justify employing punishment for expressive means. There is an important difference between punishments being employed to express condemnation, and, say, champagne being used for celebratory purposes, namely that punishment causes harm, whereas drinking champagne does not. Punishment therefore requires justification, while drinking champagne does not. Another reason why punishment requires justification is that it is not a natural phenomenon, but "a human institution" (Bedau, 2005: 7) that cannot be separated from human "purposes, intentions and acts" (Bedau, 2005: 7). Humans intentionally and deliberately organise and practice punishment, and it is not inevitable that every society should practice punishment (Bedau, 2005). One cannot assume that punishment is the appropriate method of expressing condemnation merely because it has become the conventional means of doing so.

Strawson admits that his own account of reactive attitudes may have been influenced by "local and temporary features of our own culture" (Strawson, 1974: 24). I contend that Strawson has downplayed the extent to which his own culture has influenced

[49] Cited in Llewellyn & Howse (1999: 372).

him. Strawson (and those who follow him), as well as most retributivists, appear to merely assume that punishment is a natural expression of the attitudes of resentment and indignation. The quote I cited earlier from Strawson (1974: 22), namely that "the preparedness to acquiesce in that infliction of suffering on the offender which is an essential part of punishment is all of a piece with this whole range of attitudes of which I have been speaking" illustrates this assumption quite clearly. Strawson (1974) also acknowledges the different forms reactive attitudes have taken for different peoples at different times in history, but he focuses on the attitudes, and not on the modes which express them. He says that "an awareness of variety of forms should not prevent us from acknowledging also that in the absence of *any* forms of these attitudes it is doubtful whether we should have anything that *we* could find intelligible as a system of human relationships, as human society" (Strawson, 1974: 24). What Strawson claims might indeed be true, but all that follows is that, for a society to be recognized as human, its members have to experience reactive attitudes. What *does not follow* is that we have to be committed to specific ways of expressing these attitudes. There are many ways to express these attitudes and condemn wrongdoing. Feinberg (1970: 115-116) discusses an intricate public ceremony as an example of a practice which expresses condemnation[50] equally as well as hard treatment (punishment), but in a less painful way. He says:

> One can imagine an elaborate public ritual, exploiting the most trustworthy devices of religion and mystery, music and drama, to express in the most solemn way the community's condemnation of a criminal for his dastardly deed. Such a rite might condemn so very emphatically that there could be no doubt of its genuineness, thus rendering symbolically superfluous any further hard physical treatment. Such a device would preserve the condemnatory function of punishment while dispensing with its usual physical media – incarceration and corporal mistreatment.

The Strawsonian can reply that, while all of these different methods certainly do *express* our reactive attitudes, they do not *satisfy* them, only punishment can. We do not desire some complex social ritual, what we desire is for the *one who caused the harm to suffer*, as Hershenov (1999: 90) says, we want those responsible to be "severely disadvantaged".

Hershenov (1999: 87) claims that punishing wrongdoers benefits victims as they feel "vindictive satisfaction" from "getting even" with those who have harmed them. On

[50] Feinberg uses the term condemnation to refer to attitudes of resentment and reprobation, where reprobation is "the stern judgment of disapproval" (Feinberg, 1970: 101).

Hershenov's view, an equality is restored between wrongdoers and their victims when wrongdoers are punished, as wrongdoers are reduced to a similar condition to that of their victims at the time of the crime, as both wrongdoer and victim have now been harmed. Moreover, the victims are (supposedly) elevated to their original pre-crime conditions because they experience vindictive satisfaction, as Hershenov (1999: 87) says, punishing wrongdoers "(re)creates an equality… by raising that of the…[victim] back to or near the status he enjoyed *prior* to being victimized"[51]. It is this equality, and the psychological benefits (in the form of vindictive satisfaction) gained by victim, that justifies punishing wrongdoers. So a Strawsonian, following Hershenov (1999), could claim that punishment is required to satisfy our retributive reactive attitudes (which involve the desire to "get even" with wrongdoers). In response I want to argue that there are times when we cannot, or should not, satisfy our desires in the ways that we want to. There are many situations where we have to sublimate our desires or impulses[52]. If, for instance, an employee is chastised at work for a minor infraction, and, as a result, wishes to punch his employer, it would be more socially acceptable for him to find a different outlet for his rage, such as punching a boxing bag, instead of actually assaulting his employer. Hershenov (1999: 90) seems to have something like sublimation in mind, as he says "since people do have vindictive feelings, it is prudent to channel them in a productive, civilized way". What Hershenov says here is fine, except that, on his view, the means of channelling these vindictive feelings is punishment, and punishment has to be justified. The mere satisfaction of vindictive feelings alone cannot provide this justification. Hershenov (1999) himself admits that it would be a mistake to view vindictive satisfaction as a justifying principle. If the mere satisfaction of a desire provided the justification for action, the chastised employee would be justified in assaulting his employer, *just because* it satisfied his desire, and that cannot be right. What supposedly justifies punishment on Hershenov's (1999) view is the restored equality between wrongdoers and victims that punishment brings, *not* the vindictive satisfaction felt by victims. The notion of equality (or equilibrium), however, is plagued by a number of serious problems (see Chapter Four). In Hershenov's case, it is difficult to see how punishment, and the subsequent

[51] Italics in original.
[52] Sublimation involves channelling desires or impulses into socially acceptable activities (English & English, 1958)

vindictive satisfaction, can restore a rape victim or an abused child to the conditions they enjoyed prior to being victimised. In Montague's (2002: 4) words, "[it] is unlikely that any punishment imposed on their victimizers could restore people harmed in these ways to conditions that even remotely resemble their 'pre-crime status'". The notion that we require punishment to satisfy our reactive attitudes is, therefore, not a viable justification for retributive punishment.

From the above it seems that the objection that we cannot theorise away our retributive practices, as they are based on common reactive attitudes which cannot be theorised away, is specious. It might indeed be true that humans naturally experience anger, frustration, resentment and indignation when they, or those close to them, are harmed, "but these trivial observations do not entail commitment to a specific narrowly-defined set of practices of judgement" (Tabensky, 2006: 144). There have been many different judgment practices over the years, and, as these practices are not natural, they need to be justified. Claiming that they are justified *just because* they express natural attitudes is not a satisfactory answer, as these attitudes can be expressed in a myriad of different ways.

5.1.2 – My Second Response to the Strawsonian

I have discussed the first response to the Strawsonian; I now discuss the second one, which is that the attitudes of resentment and indignation, while natural, are misidentified. Mill (1907: 76) claims that it "is natural to resent, and to repel or retaliate, any harm done or attempted against ourselves, or against those whom we sympathise". All animals, claims Mill, strike back, when they, or their young, are hurt. Humans differ only insofar as they have superior intellectual capacities, and can sympathise with beings other than just their young, with all "human, and even with all sentient, beings" (Mill, 1907: 76-77). The natural sentiments of resentment and sympathy are, according to Mill, the source of the desire to punish. Davis (1972: 140) is skeptical of Mill's ideas: he writes "[benevolence] might also be traced to 'animal desires', whatever they are; so what?" I think, however, that Davis (1972) is too dismissive of Mill. The important point here is not just that the desire to punish can be traced to the 'animal desires' to strike back, what is relevant is the explanation for that retaliation. Animals, human and non-human, strike

back when they are harmed to protect themselves, that is, in *self-defense*. So, according to Mill, in cases where others are harmed, humans, due to their extended intellectual capacities, desire to strike back, to punish, in order to protect those who are harmed. In Mill's (1907: 77) words:

> By virtue of his superior intelligence, even apart from his superior range of sympathy, a human being is capable of apprehending a community of interest between himself and the human society of which he forms a part, such that any conduct which threatens the security of the society generally, is threatening to his own, and calls forth his instinct (if instinct it be) of self-defence. The same superiority of intelligence, joined to the power of sympathising with human beings generally, enables him to attach himself to the collective idea of his tribe, his country, or mankind, in such a manner that any act hurtful to them rouses his instinct of sympathy, and urges him to resistance.

So advocates of Strawson (1974), as well as retributivists, misidentify the desire to strike back because animals retaliate to protect themselves or their young, and protection is a distinctly consequentialist concern, not a standard retributivist one.

5.2 – Criticisms of Mill

I mentioned at the beginning of this chapter that my second response to the Strawsonian (*5.1*.2 above) is of particular importance to my own view of punishment. Mill's (1907) view that the desire to punish is a combination of the natural inclination to strike back at those who harm us and our advanced intellectual capacities provides the foundation for my view (which I shall argue for in greater detail in Chapter Six). I will argue that punishment is a form of *protection*, and is justified if, and only if, it prevents greater harm from being done, and Mill's view is important because it focuses on the protection of self and others. It is for this reason that I have devoted a separate section for the criticisms of Mill's position. There are two ways one could criticise Mill's (1907) view (a view I agree with to a large extent). Firstly, the critic could say that while Mill is right that we resent and retaliate when we are in the process of being harmed, the desire to strike back is not the source of the desire to punish because we seek to punish *after* we have been harmed. Secondly, an implication of Mill's view is that every time we desire to punish wrongdoers, it is because of our attitudes, more specifically our *feelings*, of resentment or indignation coupled with sympathy for the victim. The critic may point out

that our *beliefs* are separate from our *feelings*, and there are times when the desire to punish is based, not on our *feelings*, but on our *beliefs*. In other words, we call for punishment because we believe that punishing is the right thing to do, and not in order to express the way we feel. I argue that neither of these criticisms knocks down Mill's view.

5.2.1 – The First Objection to Mill

The first objection to Mill's view is that his analysis of the animal desire to strike back is only appropriate when someone is in the process of being harmed, but does not apply when we seek to punish wrongdoers. In "The Wanderer and His Shadow" Nietzsche (1989: 180) states that when one is harmed, and one strikes back at the wrongdoer, one is doing so "merely in order to *get away* with life and limb"[53]. In other words, striking back when one is harmed is an act of "*self-preservation*" (Nietzsche, 1989: 180)[54]. Striking back in cases where one is being harmed is, according to Nietzsche (1989: 180), "almost an involuntary reflex". Punishment, however, is sought *after* one has been harmed. So how, the critic could ask, can a near involuntary reflex, which occurs *at the time* of being harmed, be the source of a deliberate, intentional judgment (that the offender should be punished) *after* the time when one is harmed? Bishop Butler (1953)[55] claims that there are two forms of resentment, the instinctive (or reflexive) resentment which occurs when we, or those close to us, are harmed, and deliberate resentment, which arises when we contemplate harm that has been done to us, or to others. It seems then, taking into account Butler's distinction between instinctive and deliberate resentment, that the desire to retaliate persists over time, and resurfaces when we reflect on harms done (to us and to others). At this point the critic could ask what exactly accounts for this persistence over time. When Mill (1907) discusses the source of the desire to punish, he refers to our extended intellectual capacities, specifically our extended capacity for sympathy. I hold that our advanced intellectual capacities, namely our capacities for memory and belief formation, account for the persistence of the desire to strike back at those who have caused harm. We know that memories can evoke strong emotional reactions, and memories of times when we, or others, have been harmed are no exception. Moreover,

[53] Italics in original.
[54] Italics in original.
[55] Cited in Scarre (2004).

we form beliefs about things that have been done to us, as well as to other humans. In cases where unwarranted harm has occurred, we, for example, form the belief that it should not have occurred. It is not inconceivable then, that our memories and beliefs of past harms can elicit the very same emotional responses we experienced *when we were harmed*. The fact that we have advanced intellectual capacities therefore explains how the natural inclination to retaliate is the source of the desire to punish.

5.2.2 – The Second Objection to Mill

Having dealt with the first objection to Mill (1907), I shall now deal with the second objection a critic could raise, which is that the desire to punish may be based solely on our beliefs, and not on our feelings of resentment of sympathy. An implication of Mill's view is that, whenever we desire to punish a wrongdoer, we do so because of our sentiments of resentment or indignation and sympathy for the victim. The critic could point out that this does not seem right, as there are times when we desire to punish, not because of how we feel, but just because we believe it is the right thing to do. The original thought experiment of Jeff is a prime example, as there is no victim to protect or sympathise with. Moreover, the critic could say that even though it might be true that humans *can* sympathise with all other humans, it is not apparent that they always *do* sympathise with them, and there are times when we desire to punish even if we do not sympathise with the victims.

Allais (2008b) discusses the influential view that the purpose of punishment is to express condemnation, and not resentment. An advocate of this view of punishment would maintain that we should punish because of the belief that wrongdoing warrants condemnation, and not because it will express our attitudes of resentment and indignation. There are two things to note here: first, claiming that we desire to punish merely because we *believe* that it is the right thing to do undermines the Strawsonian objection to my claim that the intuition that wrongdoers deserve to suffer is unreliable. To reiterate, the Strawsonian objection is that our practices of retributive judgment are based on ineradicable reactive attitudes, and hence cannot be abandoned. The reason this objection is undermined by the claim that we desire to punish purely because we believe it is right is that advocates of Strawson, and retributivists in general, cannot rely on the

retributive reactive attitudes to justify punishment. Consequently, retributivists and defenders of Strawson have to rely on their beliefs to justify punishment, and not on the intuition that wrongdoers deserve to suffer, an intuition which, at bottom, seems to be an emotional response than an intellectual one. I have, however, already argued that the retributivist justifications for punishment are erroneous. The view of punishment that Allais (2008b) discusses, namely that the purpose of punishment is to express censure rather than retributive reactive attitudes, is flawed because it has not been shown that condemnation necessarily has to be expressed via hard treatment (punishment).

The second point to note about Allais's (2008b) view that punishment can be based purely on beliefs, and not on emotional attitudes, is that neither Mill (1907), nor consequentialists in general, need disagree with this view, as long as the beliefs are justified. I have argued that retributivist justifications for punishment are erroneous. My task now is to show why I think the consequentialist justifications are not also mistaken, a task I will take up in the next chapter. I will argue that the consequentialist view of punishment is not justified (contra retributivism) because it is based on the desire to strike back at wrongdoers; they are justified because we strike back in the interests of *protection* (of ourselves or others). My account can, therefore, accommodate the desire to strike back and "get even" with those who cause harm, but the desire is not to be viewed as a justification. Punishment, I contend, is justified because we are concerned with protecting ourselves and others.

Chapter Summary

In this chapter I have dealt with the question of whether or not we should, or can, abandon the intuition that wrongdoers deserve to suffer because it is unreliable. I discussed a response which could be offered by a Strawsonian, namely that we cannot do away with this intuition (or the practice of punishment which expresses it) as it is based on ineradicable reactive attitudes. I rejected this view, and argued firstly that, even if our reactive attitudes are ineradicable, we need not be committed to any particular expression of them, and secondly that the retributive reactive attitudes of resentment and indignation are misidentified and misrepresented, as we strike back at those who cause harm in the interests of protection. These attitudes are misidentified and misrepresented because

protection is not a standard retributivist concern, but a consequentialist one. I then discussed two possible objections to my second response to the Strawsonian, namely that the desire to strike back cannot be the source of the desire to punish, and that punishment is based, not on emotions or inclinations, but on beliefs. I rejected both of these objections, but shall discuss my response to the second objection (that punishment is based on beliefs) in greater detail in the next chapter.

Chapter Six – The Consequentialist View of Punishment

I have argued that the retributivist view of punishment is not justified, as most of the considerations put forward by the retributivist theories are not necessary, and none are sufficient, for the claim that wrongdoers *deserve* to suffer by being punished for their transgressions. In this final chapter I argue that a consequentialist view of punishment can be justified, and offers a far more viable and rational alternative to retributivism. I do not have space to offer an in-depth defence of consequentialism, so I shall only outline why, in my view, consequentialism offers a better justification for punishment than retributivism. The chapter has four sections, in *6.1 – Consequentialism and the Thought Experiment* I restate the consequentialist position and revisit the thought experiment I presented in Chapter Three in order to test whether the consequentialist considerations apply to it. *6.2 – Criticisms of Consequentialism* deals with two broad criticisms brought against consequentialism: firstly, that it is unjust, and, secondly, that consequentialists do not treat wrongdoers as ends in themselves. I shall then respond to these criticisms, and reject them in turn. The fourth section *6.3 – Is Punishment Intrinsically or Instrumentally Valuable?* I deal with the question of whether punishment is good as an end in itself or as a means to some other end. In the final section, *6.4 – Consequentialism versus Abolitionism*, I argue that the consequentialist position is superior to the abolitionist one.

6.1 – Consequentialism and the Thought Experiment

The justification for punishment offered by consequentialists can be stated concisely and perspicuously: "it's right to punish criminals because doing so minimizes the net level of suffering" (Dolinko, 1997: 507). Consequentialists argue that punishment, which intentionally and deliberately causes harm, is only justified if it leads to beneficial consequences which outweigh that harm (Burgh, 1982). It is important to note, though, that the brand of consequentialism I am arguing for is concerned chiefly with preventing the harm caused by crime, and not with, say, maximising happiness or utility as some utilitarians would have us do. By preventing crime, we thereby prevent the harm caused by crime, and crime prevention is the "most plausible immediate good" (Duff, 2008: 7) that a system of punishment can provide. Punishment can prevent crime by discouraging prospective wrongdoers, removing wrongdoers from society and thereby preventing them

from committing additional crimes, and by correcting or reforming wrongdoers (Graham, Kreider and Svatos, 2002). So, the consequentialist justification for punishment I argue for is this: punishment is justified because it prevents the harm caused by crime.

I focus on the principle of preventing harm because it is well-endorsed, and advocated by both deontologists and consequentialists[56]. Butler (1993), for instance, begins his paper "The Moral Status of Smoking" with the assumption that people have some rights and that these rights impose duties on people to respect them. Butler, therefore, has deontological concerns, as he is concerned with people's rights. He goes on to claim that the "right to be free from harm is in some sense more basic than the rights one may have to perform certain activities. This 'harm principle' is perhaps the fundamental liberty-limiting principle" (Butler, 1993: 3)[57]. In other words, we should not act in a way that will violate another person's right to be free from harm[58]. Singer (1972), a well known utilitarian[59], assumes at the beginning of his paper "Famine, Affluence, and Morality" that, among other things, suffering and death (which are forms of harm) are bad, and he thinks that most would agree. After stating this assumption, Singer (1972: 231) introduces the main claim of his paper, which is this:

> if it is in our power to prevent something bad from happening, without sacrificing anything of comparable moral importance, we ought, morally, to do it. By 'without sacrificing of comparable moral importance' I mean without causing anything else comparably bad to happen, or doing something that is wrong in itself, or failing to promote some moral good, comparable in significance to the bad thing that we can prevent. This principle seems almost as uncontroversial as the last one. It requires us only to prevent what is bad, and not to promote what is good[60].

It should be noted that the views of Butler (1993) and Singer (1972) differ in that Butler (1993) tells us what we *should not* do (we should not act in ways that will harm others, as doing so violates their right to be free from harm), and Singer (1972) tells us what we *should* do (we should actively prevent bad things from happening whenever and

[56] Deontologists are principally concerned, not with the possible consequences of people's actions, but with the duties people have (Graham, Kreider and Svatos, 2002). Kantianism is one form of deontology. Other deontologists, such as Butler (1993), speak of the duties we have to respect the rights of others.
[57] Butler (1993) excludes harms such as sports injuries, self defence, just wars and similar harms from the harm principle.
[58] Butler focuses specifically on the harm second hand smoke causes to those around the smoker.
[59] Utilitarianism is, of course, a form of consequentialism.
[60] Singer's (1972) view is more relevant to my own work than Butler's (1993), and I will return to the claim that we should not forfeit anything of comparable moral significance when we try to prevent harm when I discuss the objections to consequentialism (*6.2 – Criticisms of Consequentialism*).

wherever we can, as long as we do not do something else which is equally as bad in the process). Nevertheless, both theorists can plausibly be said to be concerned with preventing harm, and we can thus see that the principle is plausible and spans the theoretical divide between deontologists and consequentialists.

In order to test the consequentialist justification for punishment (punishment is justified because it prevents the harm caused by crime), I revisit the original thought experiment of Jeff that I presented earlier. I set up the experiment so as to exclude almost all of the theoretical considerations offered by both retributivists and consequentialists[61]. I did this to ensure that the response to the experiment was an emotional, and not an intellectual, one. I then argued that, even if the thought experiment had been set up to accommodate the retributivist considerations, those considerations would still not provide adequate grounds for punishing Jeff. I now see whether the consequentialist considerations would be more viable justifications if the experiment is set up to accommodate them.

In the original thought experiment, a consequentialist would say that the main of goal of punishing Jeff should be to make sure he does not cause more harm in the future (here we see the forward looking element of consequentialism). In the original thought experiment, there is no reason for a consequentialist to say that Jeff should be punished. There is no need to incapacitate him in order to prevent him from repeating his offence, as he will be incapable of repeating it after his surgery. He already understands the wickedness of his actions, and so he does not need to be rehabilitated. Finally, the members of Jeff's community are all law-abiding citizens who also understand the gravity of Jeff's crimes, so Jeff does not need to be punished in order to deter potential offenders. So the consequentialist considerations do not apply to the original thought experiment, and are not necessary for the response that Jeff should be punished. The elements of the thought experiment which exclude the consequentialist concerns are Jeff's cancer and surgery, his acknowledgement of the wrongness of his actions, and the fact that he lives in law-abiding community. Let us see if these factors will be relevant if the

[61] The only consideration which could not be excluded was the Kantian consideration of autonomous choice, which was dealt with in *4.1.3 – The Kantian Theory*.

thought experiment is reworked to accommodate them. Imagine the original thought experiment differs in the following ways:

- Jeff is in good health, and does not need surgery (so he still rapes and accidentally smothers Alice).
- He does not acknowledge that what he did was wrong.
- Not all of the members of Jeff's community are law-abiding, and some have been convicted of crimes such as rape and murder.

When faced with the modified thought experiment, a consequentialist's main concern would be making sure that such harm does not occur in future. There are two ways of achieving this goal: making sure the *Jeff* does not repeat his offences, and making sure *others* do not commit similar crimes. In order to make sure that Jeff does not repeat his offences he should firstly be incarcerated to protect society. While he is imprisoned, rehabilitation should be attempted through suitable means, be it therapy, education or professional training (Rachels, 2007), and the wrongness of his actions should also be impressed upon him. Jeff's punishment can also serve as a warning to potential offenders in order to deter them. Punishing Jeff, according to the consequentialist, would therefore be justified because it would prevent harm from occurring in future.

In short, I have argued that the consequentialist considerations are not necessary for the response that we should punish Jeff, as they do not apply to the original thought experiment. It is my view though that the consequentialist, unlike the retributivist, considerations are sufficient for that response when the thought experiment is set up to accommodate them. My reason for saying this is that I think the criticisms brought against consequentialism can be met.

6.2 – Criticisms of Consequentialism

For the purpose of this argument I assume that most people, retributivists included, would agree that the aim of preventing harm is both a good and desirable one. The question I want to focus on now is whether or not we will sacrifice anything of equal or greater value in the pursuit of this goal. Critics of consequentialism claim that we will sacrifice things of greater value, specifically justice and the principle of treating others as ends in themselves, if our main goal is crime prevention. So there are two major criticisms of

consequentialism, firstly that it is unjust and secondly that we treat offenders as a mere means to the prevention of crime. I will now discuss each of these criticisms in greater detail.

6.2.1 – Consequentialism is Unjust

The first major criticism of consequentialism is that it is unjust. Consequentialists would, or so the story goes, employ "manifestly unjust punishments" (Duff, 2008: 8), to bring about beneficial ends, such as crime prevention. One allegedly unjust punishment is punishing an innocent person to deter potential offenders, as punishment will supposedly have the deterrent effect irrespective of whether the one punished is innocent or guilty (Burchell, 2005). Another example of an allegedly unjust punishment is the overly harsh treatment of the guilty (Duff, 2008). An example of this would be keeping dangerous, unreformed, criminals in prison after they have finished their sentences in order to protect society (Burchell, 2005)

Let me return to the thought experiment to further illustrate this first objection. In the original experiment Jeff intentionally raped Alice, so he cannot be classed as innocent. Imagine, then, that another person raped Alice, but did not kill her, and that Jeff found Alice after she had been raped and tried to help her. Police found him at the scene; the public are outraged and are calling for Jeff to be punished and threatening to take matters into their own hands if he is not. Consequentialists, according to critics, would punish Jeff, even though he is innocent, in order to appease the outraged community and send a message to potential rapists. Punishing the innocent, however, is considered to be wrong, as it is unjust (Duff, 2008). For an example of overly harsh treatment of the guilty, take the following situation: Jeff rapes Alice, is found guilty and sentenced to twenty five years in prison. When his sentence is finished, Jeff is still not rehabilitated. Consequentialists would purportedly claim that Jeff should be kept in prison in order to protect other members of society, but "disproportionate punishment is not deserved" (Allais, 2008a: 10) and so keeping Jeff in prison after his sentence is over is seen as unjust. Both of the above situations are considered unjust, as, in the first case Jeff is punished even though he is innocent, and, in the second, he receives a disproportionate degree of punishment. Consequentialism is therefore criticised for pursuing the goal of

crime prevention "at the expense of a conception of justice which most will refuse to abandon" (Burgh, 1982: 194).

The most common response to the first criticism (a response I do *not* endorse) is to adopt a side constrained version of consequentialism, which is a combination of retributivists and consequentialist concerns (Duff, 2008). Side constrained consequentialism incorporates the retributivist notions of desert and proportionality to safeguard the innocent from undeserved punishment and protect the guilty from overly harsh treatment. The beneficial consequences brought about by punishment still serve as the general justification and goal for punishment, but the pursuit of that goal must be constrained by the retributivist notions of desert and proportionality to guard against unjust punishments[62]. Side-constrained consequentialism, however, faces the same challenge as retributivism, viz. explaining the notions of desert and proportionality (Duff, 2008). I have argued that these notions are not feasible, and, as such, I cannot endorse a version of side constrained consequentialism.

I have three responses to the first criticism of consequentialism. My first response is that this criticism begs the question. Begging the question involves "assuming the very point that is at issue" (Warburton, 2001: 26), and while begging the question is not logically invalid, it is unconvincing and uninformative (Warburton, 2001). To say consequentialism is unjust presupposes the truth of a certain conception of justice, specifically the retributivist conception which includes the notions of desert and proportionality. Critics of consequentialism claim that punishing the innocent and disproportionate punishment of the guilty are unjust because such punishments are not deserved (Allais, 2008a). It is important to note, though, that these forms of punishment are unjust because they do not accord with *retributive* notions of justice. The first objection, therefore, begs the question, as the conclusion that consequentialism is unjust is based on the assumption that retributive justice is right. The notions of desert and proportionality, while deeply ingrained, are not unquestionably true, nor are they historically independent, that is, they have not been present in every society at every point in history (I mentioned earlier in Chapter Five that restorative, and not retributive, justice has been the dominant version of justice throughout history). One cannot assume

[62] For an example of side constrained consequentialism see Hart (1968), cited in Duff (2008).

that current conceptions of justice are justified just because they are deeply ingrained in most modern societies. I do not have room to argue for a consequentialist (or restorative), as opposed to a retributive, conception of justice; my point here is that the criticism that consequentialism is unjust is specious because it is based an unfounded assumption. In order for this objection to be convincing, critics of consequentialism have to show that the retributivist conception of justice is justified. This task would involve making the principles of desert and proportionality coherent, a challenging task at best given the preceding arguments against retributivism[63].

My second response to the first criticism has two parts. First, the criticism is based on the mistaken assumption that, for instance, punishing the innocent will, in general, prevent greater harm from occurring. Second, *even if* it can be shown that punishing an innocent will prevent greater harm, then it is the right thing to do. Let me return to the scenario where Jeff is innocent of raping Alice, but the community are vehemently insisting that he be punished to illustrate the first part of my response. Scarre (2004: 131) claims that such examples "are apt to dissolve under scrutiny", as punishing an innocent person is not the only, let alone the most effective, way to ward off an angry community[64]. If we are concerned that a riot will ensue, we can call the police or the army for riot control, try negotiating with the main agitators, or secretly remove the object of the community's ire, in this case Jeff, from the situation (Scarre, 2004). It is also important to note that the community is adamant that Jeff should be punished only because they think *he* raped Alice. They want the guilty to be punished, and they mistakenly believe that Jeff is guilty. In their minds, they would be punishing a rapist, not an innocent man. If they found out at a later stage that an innocent man was imprisoned while the man who actually committed the rape was still on the streets and capable of committing further crimes, then they would most likely lose confidence in the criminal justice system (Rachels, 2007)[65]. Punishing Jeff despite his innocence is consequently *not* the right thing to do in this situation.

[63] I do not mean to imply that this is an impossible task, or that my own arguments are foolproof, merely that it seems to be a difficult task.
[64] The particular example Scarre (2004) discusses is McCloskey's well-worn example of a man incriminating an innocent person in order to stop a riot.
[65] Rachels (2007) also discusses McCloskey's example.

Moreover, it does not, in my view, make sense for critics to say that consequentialists would, in general, punish the innocent in order to deter potential offenders, as, by that same logic, critics would be committed to saying that consequentialists would want to promote mediocre workers in order to motivate employees to work harder. It does not make sense for a company that wishes to encourage hard work through the incentive of promotions to promote people who have not worked hard. Promoting mediocre workers would, it seems to me, have the *opposite* effect of what was intended. Employees would not have a good example of a hard worker to emulate, so they would not know what improvements to make in their work routines. If one mediocre worker was promoted, then workers would have no reason to work harder, as it is possible to get rewarded for mediocre work. With this in mind, I doubt critics would accuse consequentialists of wanting to promote run of the mill employees, so why should they accuse them of wanting to punish innocent people in the interests of deterrence? The point of deterrence is to discourage certain types of behaviour, so what is the point of punishing someone who has not actually behaved in that way? Put simply, if we want to deter *rape*, then why punish someone who has not committed rape, surely we should punish *rapists*? If the goal is to deter certain behaviour, then it makes sense to use a good example of that behaviour, and innocent people are obviously not good examples of criminals. On the whole, then, the consequentialist's goals are not best served by punishing innocent people.

I come now to the second part of my response, which is that it would be right to punish an innocent if, and only if, it could be shown that it was *the only way* to prevent greater harm. Consequentialists cannot rule out the possibility that there are at least a few scenarios where punishing the innocent really will prevent greater harm from happening (Rachels, 2007). If it could somehow be determined that punishing the innocent will prevent greater harm, then that, for the consequentialist, is the right thing to do. To illustrate this point, I go back to the scenario where Jeff does not rape Alice, but is found at the scene of the crime. If, against all odds, we could know that punishing Jeff, even though he innocent, is going to be *the only way* to prevent greater harm from occurring, then the consequentialist would say that we should punish him. Such a situation would be tragic, as, either way, harm is done. If we punish Jeff then he suffers, but if we do not

punish him then other people suffer. In such situations, if the notion of preventing harm is correct, then we should try pick the lesser of two evils and inflict the least amount of harm possible. The suffering of one innocent person is tragic, but the suffering of many innocent is (generally) even more tragic (Scarre, 2004). The lesser evil in this scenario (and others like it) would be to punish Jeff, even if he is innocent. Moreover, if one really is concerned with protecting the innocent, then one should punish one innocent person in order to protect more innocent people, even though it is a very difficult decision to make.

The critic could reply that the problem with my response is that it is still not the case that the innocent are not punished *just because* they are innocent. Whether or not the innocent are punished is still contingent on the consequences of that punishment (Duff, 2008). If it is certain that punishing the innocent will prevent harm, then that, for the consequentialist, is the right thing to do, and this is what critics find repugnant. What reason, though, can critics give to support this response? They cannot fall back on the claim that punishing the innocent is wrong because it is unjust as they beg the question by assuming that the retributive conception of justice is right[66]. The critic's response seems to be based on a flash of moral revulsion at the thought of punishing the innocent, and not on good reasons[67]. One explanation for this flash of repugnance is that experience has taught us that, for the most part, harming innocent people does more harm than good, so we instinctively denounce all cases which would involve punishing the innocent (Rachels, 2007). When we denounce harming innocents for the sake of preventing greater harm, though, "our intuitive faculties are misfiring" (Rachels, 2007: 114), as the reason we say innocents should not be punished is that we think doing so will do more harm than good. If it can be established that punishing an innocent person will prevent greater harm from coming about, then *not* punishing that person will do more harm than good. The critic is therefore mistaken about what will actually cause the most harm, and consequently the critic's response, that punishing the innocent in the interests of minimising harm is wrong, is misguided.

---

[66] A Kantian critic could say that punishing the innocent people treats them as mere means, and not as ends in themselves, and therefore punishing the innocent is wrong. I shall deal with this objection in the next section (*6.2.2 – Consequentialism treats Offenders as Mere Means*), and shall argue that it relies on the false assumption that autonomy is intrinsically valuable. So the critic, it seems, cannot say that punishing the innocent is wrong because it treats them as mere means.

[67] The critic's response is therefore like the response to Haidt's (2002) thought experiment (in *3.1 – The Nature of Thought Experiments*) that it was wrong for Julie and Mark to make love.

My third response, which follows from the above discussion, is that the consequentialist's goal of minimising harm even if it means punishing the innocent, unlike the critic's aforementioned response, is a rational response, and is in-line with the intuition that preventing harm is a good thing. In *3.1 – The Nature of Thought Experiments* I mentioned the recent work of Greene *et al*. (2001)[68]. The scenario where a consequentialist would say that Jeff should be punished in order to prevent greater suffering would be classified as an *impersonal moral violation* by Greene and his colleagues. In this regard the scenario is akin to the first trolley problem, which involves pulling a switch to redirect the trolley onto another track where, instead of killing five people, it will only kill one person. Each of these scenarios lacks "the crucial sense of agency" (Greene & Haidt, 2002: 519) which would make them *personal moral violations*. In other words, a person faced with either scenario is not required to directly cause grievous bodily harm to a particular person. If the scenario called for one to punish Jeff by personally executing him in order to prevent some greater harm from happening, then this would be a *personal moral violation*, as one is asked to personally inflict grievous bodily harm on Jeff. Greene *et al*. (2001)[69] concluded that judgments concerning impersonal moral violations are predominantly rational ones, whereas judgments regarding personal moral violations are governed chiefly by emotions. The upshot of this conclusion is that the consequentialist decision to punish the innocent in order to evade some greater harm is a rational one, and is in-line with most people's moral intuitions about impersonal moral violations (most people said they would pull the switch in the first trolley problem). In this sense, consequentialism is unlike retributivism, which, I have argued, is based on an intuition that is fuelled by emotional responses to crime, namely the intuition that wrongdoers deserve to be punished.

The original thought experiment in *3.2 – My Thought Experiment* is designed to elicit a "flash of revulsion" (Haidt, 2001: 814) at the thought of Jeff's atrocious actions, yet excludes most principles retributivists would use to justify punishing Jeff. If retributivists maintain that Jeff should be punished, even though most of those principles do not apply, then they are offering "post-hoc justifications" (Haidt, 2001: 815). These

[68] Cited in Greene & Haidt (2002).
[69] Cited in *Ibid*.

justifications, I hold, are mere rationalisations, as they disguise the real reason retributivists have for saying Jeff should be punished. That is, retributivists claim he should be punished because they feel the flash of revulsion, yet they attempt to offer rational justifications for their claim.

I maintain that consequentialism offers a better, more rational, justification than retributivism, as the consequentialist justifications do not appear to be rationalisations. My first reason for saying this is that the fundamental consequentialist claim, that we should minimize the net level of suffering by causing offenders to suffer, is a rational one rather than an emotionally charged one, as was seen above in the discussion of impersonal moral violations. Secondly, even though consequentialists will most likely feel the same flash of revulsion that retributivists do when presented with the original thought experiment, there is no reason to say that Jeff should be punished, and so I do not think many consequentialists would say we ought to. Most retributivists, on the other hand, would, I assume, still say Jeff should be punished even though the majority of retributivist principles do not apply to the original experiment. It seems to me, then, that retributivists are attempting to hold onto the unfounded and unreliable intuition that wrongdoers deserve to be punished, whereas consequentialists are not. It must also be noted that consequentialists offer a fuller explanation of this intuition. I argued (in Chapter Five) that retributivists have left out an important facet of this intuition in their descriptions of it, namely that we respond emotionally, with feelings of anger, frustration and resentment, because we do not want people (ourselves or others) to suffer harm. Protecting ourselves or others from harm is *not* a standard retributivist concern, but a consequentialist one, but only consequentialists acknowledge this point. The intuition is therefore misidentified and misrepresented by retributivists. The intuition, once revised, does have a role to play in the consequentialist view of punishment, but as an *explanation*, and not as a *justification*. That is, consequentialists will refer to the revised intuition to explain why we desire to strike back at those who cause harm, namely because we want to protect those who are being harmed. It must be noted though that punishment is not justified for the consequentialist because we have this intuition; it is justified because we are concerned with the protection of others, and so with preventing greater harm from occurring. Furthermore, consequentialists can acknowledge that

punishment provides "comfort and gratification to victims and their families" (Rachels, 2007: 134), as well as "satisfaction to those who want to see wrongdoers suffer" (Duff, 2008: 7). So consequentialism does offer an outlet for the emotional responses evoked by crime. It is important to note, though, that the type of consequentialist I am discussing will not punish just to bring about feelings of satisfaction and comfort, but will do so only if the punishment really does prevent greater harm from coming about. Consequentialism, therefore, appears to be a far more viable and rational view of punishment.

6.2.2 – Consequentialism Treats Offenders as Mere Means

I have argued that the first criticism of consequentialism can be met. I now address the second criticism, which is that consequentialism does not treat offenders as ends in themselves, but as mere means to the end of crime prevention. Firstly, if wrongdoers are incarcerated in order to protect other members of society, then they are being used merely as a means to that end (Rachels, 2007). Retributivists claim that responsible agents are supposed to be left free to decide how they will behave in future. Imprisoned wrongdoers, however, are not afforded this opportunity as they are prevented from conducting their own behaviour (Duff, 2008). Secondly, imprisoning wrongdoers in an attempt to deter prospective wrongdoers is said to treat convicted wrongdoers merely as means to preventing crime (Allais, 2008a). Finally, rehabilitation violates the rights wrongdoers have as autonomous beings to determine the makeup of their own characters, and "we do not have the right to violate their integrity by trying to manipulate their personalities" (Rachels, 2007: 136). Retributivists, particularly Kantians, argue that, in order to treat wrongdoers as ends in themselves, we should only try to modify their behaviour by suggesting good reasons to turn away from crime (Duff, 2008). Attempting to rehabilitate offenders does not treat them as responsible, autonomous agents, but as "[patients]" (Allais, 2008a: 10) to be cured. Treating offenders merely as means violates their autonomy, and we are supposed to respect autonomy because it possesses the greatest intrinsic value in the world (Metz, 2006). We are forbidden from sacrificing autonomy for the sake of anything which is less valuable than it, and since autonomy is purported to have the greatest intrinsic value in the world, it follows that it cannot be sacrificed for

anything else at all. The pursuit of crime prevention, according to the critic, is therefore not justified because it would involve sacrificing the principle of respecting people as ends in themselves, and the price of sacrificing this principle is too high.

My response to the second criticism is that it relies on the mistaken view that autonomy has intrinsic value. I argued in *4.2.3 – The Kantian Theory* that autonomy is extrinsically valuable for the following reasons: firstly, it derives its value from the presence of possible goals to choose from. Secondly, the value of autonomy is contingent because we have to be able to act on our decisions and our choices have to be worthwhile and morally acceptable. Thirdly, autonomy has some inherent value, but not as much as being autonomous and achieving one's goals. Finally, autonomous behaviour (of which autonomy is constitutive) is instrumentally valuable as a means to achieving the goals we set for ourselves. If I am right, then autonomy is extrinsically valuable, and therefore does *not* possess the greatest intrinsic value in the world. What follows is that autonomy is not worthy of the utmost respect, and consequently retributivists are not at liberty to say that we should not punish wrongdoers with a view to rehabilitating them, or deterring potential wrongdoers, because doing so would violate their autonomy. The second charge against consequentialists (that pursuit of crime prevention would violate autonomy) is, therefore, spurious.

I do not mean to imply that autonomy has no value whatsoever, or that we should not respect the fact that people have their own agendas. There are times when consequentialists advocate principles that will have good consequences nearly every time they are followed, and the "principle of respect for autonomy would be a prime example of such a principle" (Singer, 1999: 100). A consequentialist would say that people should respect each others' autonomy because we would harm others if we disrespected them. A consequentialist would say that we should not, in general, deceive, manipulate or exploit others and for our own gains, but not because doing so violates their autonomy, but rather because we harm them. Harms "set back the interests, or are otherwise injurious to the welfare, of those who are harmed" (Butler, 1993: 3), so punishments that violate the autonomy of wrongdoers (incarceration, rehabilitation and deterrence) do indeed harm them, as they cannot further their own interests and their welfare is diminished. Crime, on the other hand, greatly harms many innocent people, as it diminishes the welfare of the

victims (drastically in cases of, for instance, rape and assault), and violates their autonomy by impeding their interests and pursuits. Respecting the autonomy of offenders seems to sacrifice the autonomy of law-abiding citizens. So, if respecting autonomy is our key concern, as the critic avers, then we should respect the autonomy of law-abiding citizens by taking steps to lower the crime rate, such as incarcerating, rehabilitating and deterring wrongdoers. Kantians would say that we cannot sacrifice the autonomy of wrongdoers for the sake of law abiding citizens, as it is intrinsically valuable. I have argued, however, that autonomy, as it derives its value from the goals we choose, cannot derive any moral value from morally reprehensible goals. So, while punishing wrongdoers harms them, we are not sacrificing anything of inestimable value by doing so, and we will cause greater harm to society if we *do not* punish them.

6.3 – Is Punishment Intrinsically or Instrumentally Valuable?

I have argued that the second criticism of consequentialism is spurious, as the criticism relies on the false assumption that autonomy is intrinsically valuable. It is important to note that a rejection of this criticism has implications for the claim that punishment is intrinsically valuable. The notion of respecting people as ends in themselves is closely linked to the idea that punishment is good in itself. If punishment was employed for instrumental reasons, for the purpose of crime prevention for instance, then those punished, according to Kantians, would be exploited as a means to those ends, and thus would not be respected as ends in themselves. Having argued that treating individuals as ends in themselves cannot be justified by claiming that they are intrinsically valuable, I now take a closer look at the idea that punishment is an end in itself. There are two reasons why punishment is thought to be an end in itself. Firstly, according to retributivists, punishment is an end in itself because the goal of punishing is not to bring about beneficial states of affairs such as crime prevention, we punish for the sake of punishing. Legal censure, according to Metz (2006: 225), "serves a morally sound function because it is in itself an appropriate response to what has happened in the past". If crime prevention was the goal of punishment, then punishment would be instrumentally valuable, that is, valuable because its "usefulness for some purpose" (Frankena, 1963: 66). Retributivists argue that it would be disrespectful to wrongdoers if

punishing them was instrumental to some other end. Retributivists argue that, by characterising punishment as an end in itself, their theory accords offenders the respect due to them as responsible moral agents (Allais, 2008a).

The second reason why retributivists argue that punishment is an end-in-itself is that "it is good that justice is done and bad that justice is denied" (Zimmerman, 2007: 1), and justice is "often thought to necessitate punishment" (Llewellyn & Howse, 1999: 356). Many think that offenders should get what they deserve, and what they deserve is to be punished (Burchell, 2005). Honderich (*Punishment*: 15)[70] criticizes this view, he writes:

> Sometimes… people say that a man deserves something and intend no more than it is right that he get it. To attempt to argue that a man's punishment is justified, by saying in this sense that he deserves it, is obviously pointless.

Davis (1972: 136) argues that Honderich's criticism misses the point, as he assumes that justifying punishment by saying offenders deserve it is "pointless" because "the supposed reason is identical with the supposed conclusion". He points out that what offenders expressly deserve is to suffer, and puts forward the proposition (R) "There is some intrinsic value in the suffering of the guilty" (Davis 1972: 136). There is nothing pointless or circular, he asserts, in justifying punishment by claiming that the suffering of the guilty is intrinsically valuable. Punishment is therefore good, and justified, because it involves making the guilty suffer. It is important to note that in Davis's view, it is not *punishment* that is intrinsically valuable, but rather the suffering of the guilty. He writes:

> Appeal to (R) does not involve the claim that the act of punishing has any intrinsic value. Similarly, one can justify scratching by appeal to the intrinsic value in the cessation of an itch; this would not be a claim that the act of scratching was intrinsically valuable (Davis. 1972: 131).

It must be noted that both of the reasons retributivists have for claiming that punishment is an end in itself are based on false assumptions. The first reason (people cannot be punished as a means to some other end) is based on the assumption that autonomy is intrinsically valuable (which I argued against in *4.2.3 – The Kantian Theory*). The second reason (it is good that justice is done, and justice involves punishing those who have committed crimes because it is what they deserve) is based on the assumption that people in fact do deserve to suffer, and consequently that retributive justice is justified as it

[70] Cited in Davis (1972: 136). Davis (1972) does not reference a year for Honderich's text.

metes out due punishment. The only justification Davis (1972: 139) offers for the proposition (R), that there is some intrinsic value in the suffering of the guilty, is that the "inclination to believe it seems very widespread among the people whose moral intuitions constitute the main data we have for settling questions of value". In other words, we are back where we started, namely with the widespread intuition that underlies retributivist theories, that "that persons who have caused harm should themselves suffer harm" (Burchell, 2005: 69). In an attempt to show that (R) is correct, Davis (1972: 139) puts forward the following scenario:

> Imagine an old-style Hollywood Western in which the villain, presented as irremediably wicked to the core, meets an unpleasant end in some natural disaster. Do you not feel that he has gotten what he deserved, that what happened was altogether fitting? If so, then you share in the intuition that (R) is correct.

I can agree that the villain meeting an unpleasant end is a good thing, but it does not necessarily follow from this that I think (R) is correct. I might think it is a good thing because it means he will not harm other people, not because he *deserved* it, and I have good reasons to believe that he would harm others given that he was 'irremediably wicked'. Moreover, Davis (1972) is, in essence, claiming that (R) is true merely because many share the same intuition. The mere fact that many agree that a proposition is true, however, is not good reason to think that it is; many people can share the same unreliable intuition. Appiah (2008) discusses a number of thought experiments which highlight common unreliable intuitions that many people have, including the famous trolley problem, which illustrates the squeamishness most people feel if they have to intentionally harm one person in order to save the lives of others. I have argued before that the intuition that those who cause gratuitous harm deserve to suffer harm in return is unreliable. The original thought experiment of Jeff that I put forward earlier in Chapter Three is designed to illustrate this very point. The experiment is supposed to elicit the intuition that Jeff deserves to suffer for his wrongs, yet removes almost of the all principles offered by retributivists which would show why this intuition is right. If the experiment is successful, then Davis's proposition (R) is unfounded, as there are no good reasons to think that the guilty deserve to suffer.

Finally, the criminal law exists to protect society from behaviour that threatens its welfare or security. Given that such protection is the chief purpose of the criminal law, it

seems implausible to say that punishment should not have a forward-looking, pragmatic role (Scarre, 2004). Punishment, according to consequentialists, can protect people by incarcerating dangerous criminals who are a threat to society, rehabilitate offenders by using (humane) means such as counselling, education and job training opportunities (Rachels, 2007), and deter potential wrongdoers. It follows then that the consequentialist view of punishment is more aligned with the chief purpose of the criminal law than the retributivist view, as consequentialism is forward-looking, while retributivism is backward-looking.

6.4 – Consequentialism versus Abolitionism

In this section I revisit the abolitionist view, and contrast it with the consequentialist view of punishment. I argue that consequentialism can accommodate the valuable components of abolitionism (specifically those of restorative justice), yet avoid the challenges faced by abolitionists[71].

Abolitionists argue that punishment cannot be justified, as neither retributivism nor consequentialism can offer viable justifications for it[72]. They argue that we should do away with punishment, and respond to criminal wrongdoing in other ways. Restorative justice is an alternative way of responding to wrongdoing that has been growing in popularity (Burchell, 2005; Duff, 2008), it is the view that we should reconcile all parties that have been affected by crime (Duff, 2008). Reconciliation involves repairing the harm that crime has caused and restoring, or building, relationships between victims, offenders and other interested parties (Allais, 2008a). The process of reconciliation involves all affected parties meeting in order to discuss the best way to deal with the harm that has been caused (Braithwaite, 1997)[73].

My intuition is that consequentialism is a more viable position because there are times when attempting to reconcile victims and offenders seems to be imprudent, as it might cause victims even more harm than they have already suffered at the hands of the offenders. It seems to me that suggesting to rape victims or victims of domestic violence

[71] I do not have space to go into great detail here, so I will only briefly say why I think consequentialism is a more viable view than abolitionism.
[72] I discussed Burgh's (1982) view as an example of abolitionism in Chapter One.
[73] Cited in Llewellyn & Howse (1999).

that they should seek to reconcile with those who harmed them so terribly would be very traumatic for the victims. Moreover, such a suggestion would not send the message that we value the victims and are repulsed by the way they have been treated, nor will it reaffirm our commitment to values we are supposed to have, namely that rape and domestic violence are abhorrent (Duff, 2008). Furthermore, reconciliation requires that wrongdoers acknowledge that they have done wrong (Duff, 2008). If, however, wrongdoers are intractable or unremorseful, then it seems that reconciliation will be unattainable. In addition, attempting to reintegrate these sorts of wrongdoers into society appears to be folly. In such cases, it seems that incarcerating these wrongdoers would be a more appropriate way to deal with them as the chances that they will commit crime again are high. If we are serious about reducing the crime rate, then we should remove individuals who are a threat to society. Consequentialists can agree with advocates of restorative justice, though, that punishing wrongdoers is not always going to be the most effective way of minimising suffering. Punishing wrongdoers, as I mentioned in *4.1.1 – The Fairness Theory*, will not generally be able to repair the harm suffered by victims of crime. Consequentialists, though, can endorse Ten's (2000)[74] view that, where possible, punishments should include compensation for victims (in the form of fines or services offered by offenders), as such compensation will repair some of the harm caused by the crime. Consequentialists can also endorse other responses to wrongdoing, such as mercy, forgiveness and reconciliation, but only as long as those responses really will prevent greater harm from happening, or reduce the amount of suffering already being experienced (Scarre, 2004). Consequentialism, therefore, seems to be a well-rounded view of punishment that can accommodate the attractive features of restorative justice, yet avoid the challenges faced by it.

Chapter Summary

In this final chapter I argued that consequentialism is superior to both retributivism and restorative justice, and is therefore the view of punishment that I endorse. I argued that the two major challenges faced by consequentialism can be met. The first criticism, that consequentialism is unjust, relies on two mistaken assumptions: firstly, that retributive

[74] Cited in Scarre (2004).

justice is right, and, secondly, that, for instance, punishing the innocent will, in general, minimise levels of suffering. I then argued that, even if it could be established that punishing an innocent would prevent greater harm from happening, it would be the right thing to do. This conclusion, I argued, is more rational and in-line with our intuitions than arguments put forward by critics. The second major criticism, that consequentialism treats offenders as mere means, is, in my opinion, based on the false assumption that autonomy is intrinsically valuable. In *6.3 – Is Punishment Intrinsically or Instrumentally Valuable?* I argued that punishment is instrumentally valuable because the view that it is intrinsically valuable is based on two mistaken assumptions (that retributive justice is right and that autonomy is intrinsically valuable). Finally, I argued that consequentialism can accommodate the valuable elements of restorative justice, and circumvent the problems associated with it.

References:

- Allais, Lucy, 'Punishment as Retribution', in ***The WISER Review***, **no. 3** (2008a), pp. 10.

  - Allais offers a succinct, accessible defence of retributivism, and good criticisms of both consequentialism and restorative justice.

- Allais, Lucy, 'Wiping the Slate Clean: The Heart of Forgiveness', in ***Philosophy & Public Affairs***, **36**, **no. 1** (Blackwell Publishing, Inc., 2008b).

  - In this paper Allais focuses on providing an account of forgiveness that makes sense of the idea of wiping the slate clean. What is of value for the purposes of this thesis is the clear discussion of Strawson's account of reactive attitudes she offers.

- Appiah, Kwame, ***Experiments in Ethics*** (Massachusetts: Harvard University Press, 2008).

  - An insightful discussion of recent thought experiments, specifically those that draw out unreliable moral intuitions, such as the well-known trolley problems.

- Baggini, Julian & Fosl, Peter S., ***The Philosopher's Toolkit, A Compendium of Philosophical Concepts and Methods*** (Oxford: Blackwell publishers Ltd, 2003).

  - Baggini and Fosl illustrate the key similarities and differences between thought experiments and scientific experiments, and provide a valuable discussion on the nature of thought experiments.

- Burchell, Jonathan, ***Principles of Criminal Law*** (edn), (Lansdowne: JUTA and Company Ltd, 2005).

  - A useful overview of the theories of punishment.

- Braithwaite, John, 'Restorative Justice: Assessing an Immodest and a Pessimistic Theory' (Review essay prepared for University of Toronto law course, '***Restorative Justice: Theory and Practice in Criminal Law and Business Regulation'***, 1997). Cited in Llewellyn & Howse (1999).

  - Braithwaite points out that restorative justice has been the dominant version of justice throughout history, a point which is valuable for my rejection of Strawson's view.

- Burgh, Richard W., 'Do the Guilty Deserve Punishment?' in ***The Journal of Philosophy***, **Vol. 79**, **No. 4** (1982), pp. 193-210.

  - A good example of an abolitionist, as he argues that neither the consequentialist nor the retributive theories of punishment can be justified. Burgh also presents a useful discussion of the intuition that wrongdoers deserve to suffer.

- Butler, Joseph, ***Fifteen Sermons Preached at the Rolls Chapel*** (London: G. Bell & Sons, 1953). Cited in Scarre (2004).

- Butler discusses the importance of resentment and points out that it comes in two forms: instinctive (or reflexive) and deliberate. His work is therefore useful for my discussion about the source of the desire to strike back.

- Butler, Keith, 'The Moral Status of Smoking' in ***Social Theory and Practice***, **Vol. 19. No. 1** (1993), pp. 1-26.
  - Butler argues that smokers should not smoke around others, as it causes them harm. His discussion of the 'Harm Principle' (do not act in a way that will violate another's right to be free from harm) is important for my defence of consequentialism.

- Chisholm, Roderick M., 'Intrinsic Value', in ***Values and Morals,*** Alvin I. Goldman & Jaegwon Kim (eds.), (Dordrecht: D. Reidel Publishing Company, 1978).
  - Chisholm offers a valuable explanation of what it means for something to have value in isolation, independent of any connections, consequences or relations.

- Cottingham, John, 'Varieties of Retribution', in ***Philosophical Quarterly***, **29** (1979), pp. 238-246. Cited in Duff (2008) and Montague (2002).
  - Cottingham identifies nine different types of retributivism.

- Crisp, Roger, ***Mill on Utilitarianism*** (London: Routledge, 1997).
  - Crisp provides an enlightening discussion of Mill's Chapter Five in ***Utilitarianism****, especially regarding Mill's position on the source of the desire to punish, namely the combination of our superior intellectual capacities and the natural inclination to strike back.*

- Davis, Lawrence H., 'They Deserve to Suffer', in ***Analysis***, **Vol. 32**, **No. 4** (1972), pp. 136-140.
  - A succinct account of the view that punishing wrongdoers is justified because the suffering of the guilty is intrinsically valuable. Davis also criticises Mill's views on the source of the desire to punish.

- Dolinko, David, 'Retributivism, Consequentialism, and the Intrinsic Goodness of Punishment', in ***Law and Philosophy***, **Vol. 16**, **No. 5** (1997), pp. 507-528.
  - Dolinko criticises Moore's view that retributivism can be framed as a consequentialist theory, and offers a striking criticism of the view that punishment is intrinsically valuable.

- Downie, R. S., 'Objective and Reactive Attitudes', in ***Analysis***, **Vol. 27**, **No. 2** (1966), pp. 33-39.
  - This paper is useful for understanding Strawson's account of commonplace reactive attitudes.

- English, Horace B. & English, Ava Champney, ***A Comprehensive Dictionary of Psychological and Psychoanalytical Terms*** (Toronto: Longmans, Green and Co., 1958)

- *English and English offer a succinct definition of the notion of sublimation.*

- Falls, Margaret M., 'Retribution, Reciprocity, and Respect for Persons', in ***Law and Philosophy***, **Vol. 6**, **No. 1** (1987), pp. 25-51.
  - *A good paper that discusses fairness theory, the communication of censure and Kantian elements of retributivism.*

- Feinberg, Joel, ***Doing and Deserving*** (Princeton, New Jersey: Princeton University Press, 1970).
  - *Feinberg is one of the leading proponents of the expressive version of retributivism.*

- Frankena, William, 'Intrinsic Value and the Good Life', in ***Ethics*** (Englewood Cliffs, NJ: Prentice-Hall, 1963).
  - *An insightful discussion on the different types of value (intrinsic, extrinsic, inherent, contributory and final), particularly with regard to inherent value.*

- Graham, Peter J, Kreider, Evan and Svatos, Michelle, 'Glossary' in ***Reason and Responsibility, Readings in Some Basic Problems of Philosophy*** (edn), Feinberg, Joel and Shafer-Landau, Russ (eds), (London: Wadsworth/Thomas Learning, 2002).
  - *Graham, Kreider and Svatos provide concise definitions of consequentialism, deontology and manslaughter.*

- Greene. J.D. *et al*., 'An fMRI investigation of emotional engagement in moral judgment', in ***Science*** **293** (2001), pp. 2105-2108. Cited in Greene & Haidt (2002).
  - *Greene and his colleagues present participants with the two trolley problems and they draw a distinction between 'personal' and 'impersonal' moral judgments. This distinction is important for my rejection of the criticism that consequentialism is unjust.*

- Greene, Joshua and Haidt, Jonathan, 'How (and where) does moral judgment work?', in ***Trends in Cognitive Sciences***, **Vol. 6**, **No. 12 (**2002), pp. 517-523.
  - *Greene and Haidt offer valuable discussions on the famous trolley problems.*

- Haidt, Jonathan, '*The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*', in ***Psychological Review, Vol. 108, No. 4*** (2001), pp. 814-834.
  - *A key paper, as my main thought experiment is similar to the one Haidt presents. Haidt also offers a provocative discussion on the role of emotions in our moral judgments*.

- Hart, H. L. A., ***Punishment and Responsibility*** (Oxford: Oxford Universtiy Press, 1968). Cited in Duff (2008).
  - *A good example of side-constrained consequentialism.*

- Hershenov, David B., 'Restitution and Revenge', in ***The Journal of Philosophy,*** **96** (1999), pp. 79-94.

  - Hershenov argues that punishing perpetrators benefits victims, as victims experience 'vindictive satisfaction' from 'getting even' with those who harmed them.

- Honderich, Ted, ***Punishment.*** Cited in Davis (1972). Davis does not reference the edition he quotes from. The latest version is: Honderich, Ted, ***Punishment: The Supposed Justifications Revisited*** (London: Pluto Publishing, 2005).

  - Honderich argues that claiming that wrongdoers should be punished because they deserve it is unhelpful.

- Kant, Immanuel, ***The Metaphysical Elements of Justice***, tr. J. Ladd, (Indianapolis, IN: Bobbs-Merrill Co., 1965). Cited in Falls (1987) and Rachels (2007).

  - Two important quotes come from this text. The first quote is that rationality is "the purely human element in man", and the second is that the offender's "own evil deed draws the punishment upon himself".

- Kant, Immanuel, 'The Good Will & The Categorical Imperative', in ***Reason and Responsibility, Readings in Some Basic Problems of Philosophy*** (edn), Feinberg, Joel and Shafer-Landau, Russ (eds), (London: Wadsworth/Thomas Learning, 2002).

  - In this canonical text Kant presents the two formulations of the famous Categorical Imperative.

- Kant, Immanuel, ***The Metaphysics of Morals***, Gregor, Mary (ed.), (Cambridge: Cambridge University Press, 2003).

  - Kant forcefully states that punishment can never be used as a means to another end, and argues that only the principle of proportionality can determine the requisite amount of punishment to be meted out.

- Klimchuk, Dennis, 'Retribution, Restitution and Revenge', in ***Law and Philosophy***, **20** (2001), pp. 81–101.

  - Klimchuk discusses Hershenov's view that punishment should compensate victims. The quote that the "desire for revenge…is something to be overcome rather than satiated" comes from this paper.

- Korman, 'The Failure of Trust-Based Retributivism' ***Law and Philosophy,*** **22** (2003), pp. 561–575.

  - Korman argues against Susan Dimock's view of trust-based retributivism and provides a useful way of distinguishing between Retributivist and Consequentialist theories.

- Lemos, Noah M., ***Intrinsic Value: Concept and Warrant*** (New York: Cambridge University Press, 1994).

- *A good introduction to the traditional notion of intrinsic value.*

- Llewellyn, Jennifer, J. & Howse, Robert, 'Institutions for Restorative Justice: The South African Truth and Reconciliation Commission' in ***The University of Toronto Law Journal***, **Vol. 49, No. 3**, (1999), pp. 355-388.
  - *Llewellyn and Howse provide a clear account of the notion of restorative justice.*

- Metz, Thaddeus. 'Judging Because Understanding: A Defence of Retributive Censure', in ***Judging and Understanding. Essays on Free Will, Narrative, Meaning and the Ethical Limits of Condemnation***. Pedro Tabensky (ed), (Aldershot: Ashgate Publishing Company, 2006).
  - *A good overview of the retributive theories, and a strong argument for retributivism.*

- Midgley, Mary, ***Wickedness*** (London: Routledge, 2001).
  - *Midgley's discussion of Manichaeism is useful as it highlights a view of sexual behaviour that is markedly different from modern views, namely the view that sex is bad. Her discussion is important for my rejection of Strawson.*

- Mill, John Stuart, ***Utilitarianism*** (edn), (London, Longmans, Green, and Co.: 1907).
  - *Mill's views on the source of the desire to punish (the combination of our superior intellectual capacities and the natural inclination to strike back) are integral to both my rejection of Strawson and my defence of consequentialism.*

- Montague, Phillip, 'Recent Approaches to Justifying Punishment', in **Philosophia**, **Vol. 29, No. 1-4** (2002), pp. 1-34.
  - *Montague criticises, among others, Hershenov's view that punishment should benefit victims in the form of vindictive satisfaction.*

- Morris, Herbert, 'Persons and Punishment', in ***On Guilt and Innocence*** (Berkley: University of California Press, 1976), pp. 31-88. Cited in Falls (1987).
  - *One of the foremost advocates of the fairness theory of punishment.*

- Murphy, Jeffrie, 'Marxism and Retribution', in ***Retribution, Justice, and Therapy*** (Boston: D. Reidel Publishing Co., 1987), pp. 93-115. Cited in Falls (1987).
  - *A leading proponent of the fairness theory of punishment.*

- Nietzsche, Friedrich, ***On the Genealogy of Morals***, tr. W. Kaufmann and R.J. Hollingdale, (New York: Vintage Books, 1989).
  - *Nietzsche provides a thought-provoking discussion on the view that punishment is supposed to stimulate feelings of remorse in wrongdoers.*

- Nietzsche, Friedrich, 'The Wanderer and His Shadow', in '*Appendix'* of ***On the Genealogy of Morals***, tr. W. Kaufmann and R.J. Hollingdale, (New York: Vintage Books, 1989).

- *An insightful discussion on the source of the inclination to strike back, and on two different types of revenge (reflexive and deliberate).*

- Nozick, Robert, 'Intrinsic Value' in ***Philosophical Investigations*** (Oxford: Clarendon Press, 1981), pp. 413-450.
  - *In this paper Nozick discusses the view that intrinsic value is best understood in terms of organic unity. He neatly explains the difference between intrinsic and instrumental value.*

- Nozick, Robert, ***Philosophical Explanations*** (Cambridge, Mass.: Harvard University Press), pp. 366–368. Cited in Zaibert (2006). Zaibert does not reference the year of this edition.
  - *Nozick discusses the differences between revenge and punishment, and offers a useful discussion on the principle of proportionality.*

- O'Neill, Onora, 'Kantian Approaches to Some Famine Problems', in ***Reason and Responsibility, Readings in Some Basic Problems of Philosophy*** (edn), Feinberg, Joel and Shafer-Landau, Russ (eds), (London: Wadsworth/Thomas Learning, 2002).
  - *While O'Neill focuses on famine relief in this article, her discussion on the principle of treating people as ends in themselves is extremely helpful.*

- Rachels, James, ***The Elements of Moral Philosophy*** (edn), Rachels, Stuart (ed.), (McGraw-Hill, New York: 2007).
  - *Rachels presents complicated theories and principles (such as Kant's two formulations of the Categorical Imperative) in a lucid manner.*

- Russell, Bertrand, ***History of Western Philosophy*** (edn), (London: George Allen & Unwin Ltd, 1974).
  - *Russell offers a discussion on the Manichaean view of sexual behaviour (the view that sex is bad) which is useful for my criticism of Strawson.*

- Scarre, Geoffrey. ***After Evil, Responding to Wrongdoing***, (Aldershot: Ashgate Publishing Company, 2004).
  - *An attack on retributive theories from a consequentialist perspective.*

- Singer, Peter, 'Famine, Affluence, and Morality', in ***Philosophy and Public Affairs***, **Vol. 1**, **No. 3** (1972).
  - *A well-known article where Singer argues that we should prevent harm no matter who is being harmed or where they are harmed. What is important for the purposes of this thesis is the claim that we should prevent harm only if we do not sacrifice something of comparable value in the process.*

- Singer, Peter, ***Practical Ethics*** (edn), (Cambridge: Cambridge University Press, 1999).
  - *Singer offers useful points on autonomy and the changing views on sexual behaviour.*

- Stephen, James, F., ***General View of the Criminal Law of England***, (London: Macmillan, 1863). Cited in Feinberg (1970) & in Scarre (2004).
  - *Stephen famously stated that punishment is to the desire for revenge what marriage is to sexual desire*.

- Strawson, Peter, 'Freedom and Resentment', in ***Freedom and Resentment and Other Essays***, (London: Methuen, 1974).
  - *In this well-known paper, Strawson argues that our commonplace reactive attitudes, such as anger and resentment, are natural and require expression through punishment.*

- Ten, C.L., 'Deserved Punishment and Benefits to Victims', in ***Utilitas***, **12** (2000). Cited in Scarre (2004).
  - *Ten argues that victims of crimes should be compensated through fines or other services which are offered by the perpetrators of those crimes.*

- Tabensky, Pedro, 'Moved Movers: Transfiguring Judgement Practices', in ***Judging and Understanding. Essays on Free Will, Narrative, Meaning and the Ethical Limits of Condemnation***. Pedro Tabensky (ed.), (Aldershot: Ashgate Publishing Company, 2006).
  - *A cogent argument against retributivism based on concerns about the nature of free will and responsibility.*

- von Hirsch, Andrew, ***Doing Justice: The Choice of Punishment*** (New York: Hill & Wang, 1976). Cited in Scarre (2004).
  - *von Hirsh is a proponent of the fairness view of punishment.*

- Warburton, Nigel, ***Thinking From A to Z*** (ed), (London: Routledge, 2001).
  - *Warburton defines thought experiments concisely*, and *offers a crisp definition for thought experiments, as well as a helpful discussion of Nozick's experience machine. Also provides a concise, but useful, explanation of rationalisations.*

- Zaibert, Leo, 'Punishment and Revenge', in ***Law and Philosophy***, **25** (2006), pp. 81-118.
  - *A detailed argument against the prevalent view that punishment and revenge are different activities. Zaibert's discussion of the principle of proportionality was particularly useful for my own work.*

Internet References:

- Bedau, Hugo A., 'Punishment, in ***The Stanford Encyclopedia of Philosophy***, Edward N. Zalta (ed.), (Fall 2005 Edition).
  URL = <http://plato.stanford.edu/archives/fall2005/entries/punishment/>.

- Brown, James Robert, 'Thought Experiments', in ***The Stanford Encyclopedia of***

Philosophy, Edward N. Zalta (ed.), (Summer 2007 Edition).
URL = <http://plato.stanford.edu/archives/sum2007/entries/thought-experiment/>.

- Duff, Antony, 'Legal Punishment', in ***The Stanford Encyclopedia of Philosophy***, Edward N. Zalta (ed.), (Summer 2008 Edition).
URL = <http://plato.stanford.edu/archives/sum2008/entries/legal-punishment/>.

- Eshleman, Andrew, 'Moral Responsibility', in ***The Stanford Encyclopedia of Philosophy***, Edward N. Zalta (ed.), (Fall 2004 Edition).
URL = <http://plato.stanford.edu/archives/fall2004/entries/moral-responsibility/>.

- McLeod, Owen, 'Desert', in ***The Stanford Encyclopedia of Philosophy***, Edward N. Zalta (ed.), (Fall 2003 Edition).
URL = <http://plato.stanford.edu/archives/fall2003/entries/desert/>.

- Zimmerman, Michael J., 'Intrinsic vs. Extrinsic Value', in ***The Stanford Encyclopedia of Philosophy***, Edward N. Zalta (ed.), (Spring 2007 Edition).
URL = <http://plato.stanford.edu/archives/spr2007/entries/value-intrinsic-extrinsic/>.

Appendix A

Thad Metz (Personal Communication, 17/11/2008) recently remarked that my mapping of the retributivist theoretical terrain is not completely accurate. Metz (2006) claims that there are three major versions of retributivism: the desert theory, the fairness theory and the (intrinsic) expressive theory (also known as the censure theory). I am aware that I do not discuss the desert theory as a main version of retributivism, and that the Kantian principle of respecting people as ends in themselves is a key feature of the censure theory. The censure theory, according to Metz (2006: 224), is the view that "the point of punishment should be to treat the offender as responsible for his behaviour, to affirm the value of his victim, or to disavow wrongful actions". I am also aware that, even though I discuss Metz (2006) and Falls (1987) when I discuss the Kantian theory, their views are examples of the censure theory. I would now like to clarify my reasons for not changing the structure of my thesis in light of these considerations. Firstly, a theory of retributivism has to provide an answer to the question: Why do wrongdoers deserve to be punished (Duff, 2008)? The answer provided by desert theory would be something like: "a person should be burdened for his wrongful behaviour because he deserves to be" (Metz, 2006: 224), and this answer does not answer the question at all. We still need to be told why perpetrators deserve to be punished. It is for this reason that I only discuss this theory briefly in *3.3.2 – The Second Criticism* and *6.3 - Is Punishment Intrinsically or Instrumentally Valuable?* Thirdly, Rachels (2007: 140) identifies a version of "Kantian retributivism", which combines the principles of desert, proportionality and the first and second formulations of the Categorical Imperative, and so talking about a distinct Kantian theory does not seem completely implausible. Finally, it seems to me that it is useful to have a separate section devoted specifically to the principle of treating people as ends in themselves, especially as it is such a ubiquitous and revered principle.