THE UNIVERSITY *of* EDINBURGH

Free Will, Punishment and Criminal Responsibility

By Elizabeth Shaw LLB (hons), LLM (dist.)

A thesis presented for the degree of PhD in Law.
Edinburgh University 2013

Table of Contents

Acknowledgments

I am very grateful to my supervisors, Professor Antony Duff and Professor James Chalmers, as well as to Professor Gerry Maher for his supervision near the end of my PhD programme. They bear no responsibility for any errors this thesis contains. I am grateful to the Arts and Humanities Research Council and the Clark Foundation for Legal Education for funding my PhD studies. I am also thankful for the support of my family and friends.

Abstract

Retributive attitudes are deeply held and widespread in the general population and most legal systems incorporate retributive elements. It is probably also the dominant theory of punishment among contemporary philosophers of criminal justice. However, retributivism relies on conceptions of free will and responsibility that have, for millennia, fundamentally divided those who have thought seriously about the subject.

Our legal system upholds the principle that the responsibility of the offender has to be proven *beyond reasonable doubt*, before the accused can be punished. In view of the intractable doubts surrounding the soundness of retributivism's very conception of responsibility, my thesis argues that it is ethically dubious to punish individuals for solely retributive reasons. Instead, my thesis proposes that a person should only be punished if the main theories of punishment agree that punishing that person is appropriate – I call this 'the convergence requirement'. This approach, I argue, is in accordance with the considerations underlying the beyond reasonable doubt standard.

In addition to considering the question of 'whom to punish' my thesis considers what methods of responding to criminal behaviour are acceptable. In particular, it attempts to explain, without appealing to the contested notions of free will or retributive desert, what is problematic about 'manipulative' methods of dealing with criminal offenders (focussing in particular on the possibility of modifying their behaviour through neurological interventions). The final part of this thesis also gives an overview of some of the practical implications for Scots criminal law of taking doubts about free will and retributivism seriously. Given the severe treatment that offenders undergo within the Scottish penal system (e.g. deprivation of liberty, stigma) and the high rate of recidivism, it is important to consider whether our

current penal practices are justified, what alternatives are available and what goals and values should guide attempts at reforming the system.

Declaration

In accordance with the regulation 26 of the University's regulations governing submission of theses I declare that:

a. This thesis was composed by me.
b. This thesis is my own work.
c. The work has not been submitted for any other degree or professional qualification except as specified.

(Elizabeth Shaw)

Note: During the course of my PhD I have published the following work based on Chapters Seven, Eight and Nine of this thesis:

Shaw E, 'Direct Brain interventions and Responsibility Enhancement' (2012) Criminal Law and Philosophy, DOI: 10.1007/s11572-012-9152-2 (online first).

E Shaw, 'Free Will, Punishment and Neurotechnologies' in van den Berg B and Klaming L (eds), *Technologies on the Stand: Legal and Ethical Questions in Neuroscience and Robotics* (Wolf Legal Publishers, Nijmegen 2011) 177-194.

E Shaw, 'Cognitive Enhancement and Criminal Behaviour' in E Hildt and A Franke (eds), *Cognitive Enhancement: An Interdisciplinary Perspective* (Springer, Dordrecht 2013).

Chapters One and Two of this thesis were based partly on material from my LLB honours dissertation (Aberdeen University), but in a significantly modified form (the most major modifications being the inclusion of a section on 'owning actions' and a section on 'the new dispositionalism').

Introduction

Retributive attitudes are deeply held and widespread in the general population and most legal systems incorporate retributive elements.[1] Victor Tadros observes that, 'Retributivism is probably the most popular theory of punishment amongst those people working on the range of issues within the philosophy of criminal justice, and amongst criminal justice academics more generally.'[2] However, this theory relies on conceptions of free will and responsibility that have, for millennia, fundamentally divided those who have thought seriously about the subject. [3]

Our legal system upholds the principle that the responsibility of the offender has to be proven *beyond reasonable doubt*, before the accused can be punished. In view of the intractable doubts surrounding the soundness of retributivism's very conception of responsibility, my thesis argues that it is ethically dubious to punish individuals for solely retributive reasons. Instead, my thesis proposes that a person should only be punished if the main theories of punishment agree that punishing that person is appropriate – I call this 'the convergence requirement'. This approach, it is argued, is in accordance with the considerations underlying the beyond reasonable doubt standard. In addition to considering the question of 'whom to punish' my thesis considers what methods of responding to criminal behaviour are acceptable. In particular, it attempts to explain, without appealing to the contested notions of 'free will' or 'retributive desert', what is problematic about 'manipulative' methods of dealing with criminal offenders (focussing in particular on the possibility of

[1] See, e.g., K Carlsmith, 'The Roles of Retribution and Utility in Determining Punishment' (2006) 42 Journal of Experimental Social Psychology 437.
[2] V Tadros, *The Ends of Harm: The Moral Foundations of the Criminal Law* (OUP, Oxford 2011), p44.
[3] Belief in this contested conception of free will is also widespread across cultures. See, e.g.: H Sarkissian et al, 'Is Belief in Free Will a Cultural Universal' (2010) 25(3) Mind and Language 346. Like retributivism, free will also figures prominently in religious discourse: Pereboom D, 'Free Will, Evil, and Divine Providence', in Chignell A and Dole A (eds) God and the Ethics of Belief: New Essays in Philosophy of Religion (CUP, Cambridge 2005) 77.

modifying their behaviour through neurological interventions). The final part of this thesis also gives an overview of some of the practical implications for Scots criminal law of taking doubts about free will and retributivism seriously. Given the severe treatment that offenders undergo within the Scottish penal system (e.g. deprivation of liberty, stigma) and the high rate of recidivism, it is important to consider whether our current penal practices are justified, what alternatives are available and what goals and values should guide attempts at reforming the system.[4]

Key Terms

I will now make some remarks on the terminology used in this thesis.

Retributivism

This thesis will focus on an influential version of retributivism, which holds that punishing the guilty is intrinsically good.[5] According to this version of retributivism, the state is not merely entitled to punish the guilty. Rather, the state has a moral duty to punish offenders, purely because they deserve to suffer, even if punishing them serves no further purpose.[6] For the retributivist, the judgement that someone is *morally responsible* for committing a criminal offence means that the criminal action belongs to the offender in such a way that she deserves to be punished for it, irrespective of the consequences of imposing punishment. (In this thesis, the term 'retributive responsibility' will be used to refer to this kind of moral responsibility.)[7]

---

[4] Reconviction figures in some UK prisons are over seventy per cent:
Ministry of Justice, Compendium of re-offending statistics and analysis (2010), available at http://www.justice.gov.uk/publications/docs/compendium-of-reoffending-statistics-and-analysis.pdf
Effective rehabilitation could also lead to considerable savings. The average cost of keeping one offender in prison for a single year is £40,000: Adebowale V, 'Diversion Not Detention' (2010) 17 (2) Public Policy Research 71.

[5] For a defence of this view see Michael Moore, *Placing Blame* (1997) (Henceforth: Moore, *Placing Blame*). For criticisms of other versions of retributivism see: John Mackie, 'Morality and the Retributive Emotions' 1982 Criminal Justice Ethics 3; Ted Honderich, *Punishment: The Supposed Justifications* (1984)

[6] Reform and deterrence are examples of purposes punishment might serve, which form no part of the retributive theory of justice.

[7] This definition of retributive responsibility is based on definitions given by Derk Pereboom in "Reasons-Responsiveness, Alternative Possibililities, and Manipulation Arguments Against Compatibilism: Reflections on John Martin Fischer's *My Way,"* (2006) 47 Philosophical Books 198,

The principle that the guilty should receive the punishment that they deserve is known as 'positive retributivism', since it is meant to provide a positive reason *in favour* of punishment. In contrast, the idea that those who are not guilty should be spared punishment (and that the guilty should receive no more punishment than they deserve) is known as 'negative retributivism'. Those who endorse 'negative retributivism' insist that this principle should *constrain* the state's power to punish. Unless otherwise indicated, the term 'retributivism' will refer to theories that include both the positive and the negative retributive principle.

Free Will

I will use the term 'free will' to refer to the ability to control one's actions in a way that could make one an appropriate candidate for judgements of retributive responsibility.

Determinism

The truth of determinism, as I will argue in Chapter Two, would pose a serious challenge for free will and thus for retributive responsibility. Determinism, as it applies to human behaviour, is the theory that the deliberations, choices and conduct of every individual are causally necessitated by factors that are ultimately beyond the individual's control. Determinism does not imply that our psychological states, such

---

at pp211-212 (henceforth: Pereboom, "Reasons-Responsiveness"); and in "Living Without Free Will: The Case for Hard Incompatibilism", in Kane (ed) *The Oxford Handbook of Free Will* (2002), 478, at p479, (henceforth: Pereboom, "Living Without Free Will"). Richard Double uses the term "retributive moral responsibility" in "Metaethics, Metaphilosophy and Free Will Subjectivism", in Kane (ed) *The Oxford Handbook of Free Will* (2002), 506, at p516. The connection between retributivism and responsibility is also discussed by Ted Honderich in *How Free Are You?,* 2nd edn. (2002) (Henceforth, Honderich, *How Free are You*). At p101-2 and p139, Honderich argues that retributive judgements depend on "holding people responsible in a certain way…a way that is inconsistent with determinism". See also H.L.A. Hart, *Punishment and Responsibility,* (1968). See also Galen Strawson, "The Bounds of Freedom", in Kane (ed) *The Oxford Handbook of Free Will* (2002), 442 (Henceforth, Strawson, "The Bounds of Freedom"). At p442 Strawson asks "Are [people] ever responsible for the their actions in such a way that they are, without any sort of qualification, morally deserving of …punishment…for them?". Honderich argues for an alternative, non-retributive conception of responsibility in *How Free Are You?,* chapters 8-11. Pereboom also advocates a non-retributive conception of moral responsibility in "Reasons-Responsiveness", p211-212. This thesis does *not* argue that that there is *no* sense of responsibility which is compatible with determinism.

as our intentions, desires and beliefs, make no difference to our actions. Rather, determinism implies that, if our actions are to be explained by reference to such psychological phenomena as mentioned above, then these phenomena were themselves produced by prior events that were causally sufficient for the occurrence of those psychological phenomena and that those prior events were themselves produced in the same manner by even earlier events etc. in an unbroken chain of cause and effect that can be traced back to before the person was even born. Nor does determinism imply that people will fail to modify their behaviour in response to good reasons for doing so. It merely implies that whether a person recognises and responds to one particular reason for action rather than another at any given time is determined by prior events in the manner described above. This thesis aims to cast doubt on the idea that determinism is compatible with retributive responsibility. However, it maintains that determinism is compatible with *rationality.*[8]

In this thesis, the term 'determinism' will be used to mean the theory that *human behaviour* is determined. The proposition that *all* events are caused has been contested by physicists. On one theory, indeterminism exists at the quantum level of subatomic phenomena. Some theorists have claimed that quantum physics may lend support to the idea that human actions are undetermined.[9] This claim will be discussed in Chapter One below, where it will be argued that, even if quantum events affect human behaviour, it is unlikely that this could provide a satisfactory basis for retributive responsibility.

Indeterminism

'Indeterminism' is the idea that determinism is false.

---

[8] See Chapter One, the section entitled "The Conceptual Coherence of Determinism".
[9] E.g. Richard Swinburne, *The Existence of God,* 2nd ed (2004) pp169-170.

Compatibilism

'Compatibilism' is the view that free will and retributive responsibility are compatible with determinism and that people can have free will. Most compatibilists believe that free will and retributive responsibility are also compatible with indeterminism.

Incompatibilism

This is the idea that determinism is incompatible with free will and retributive responsibility. There are two types of incompatibilism – 'libertarianism' and 'hard incompatibilism'.

Libertarianism

'Libertarianism' is the belief that free will and determinism are incompatible (thus libertarianism is a variety of 'incompatibilism') and that people can be free, because determinism is false.

Hard Incompatibilism

'Hard incompatibilists' believe that people lack free will and are not retributively responsible for their actions. Hard incompatibilists believe this either because they think determinism is true and that it is incompatible with free will and retributive responsibility, or because they think that free will and retributive responsibility are incompatible with both determinism and with indeterminism. Hard incompatibilism is also known as 'free will scepticism'.

Punishment

When I refer to 'theories of punishment' I intend this phrase to be construed in a broad sense. My use of the word 'punishment' is not restricted by definition to the notion of retributive punishment. Nor is it restricted to theories that inflict hardship on offenders *in order* to make offenders suffer (such a definition would include

retributivism as well as special and general deterrence theories that intend punishment to make offenders suffer, so that they or others will not offend/reoffend). Instead, when I refer to 'theories of punishment' I intend to include theories that recommend the use of state coercion in response to criminal behaviour. This broad definition includes incapacitation theories, such as that defended by Derk Pereboom, which attempt to justify subjecting dangerous offenders to various coercive measures to prevent them from being a threat to society, but do not require that these measures be imposed in order to make offenders suffer.[10] The reader should bear my stipulative definition of theories of punishment in mind when reading this text. I have opted for this interpretation of the term 'theories of punishment' partly for ease of exposition and partly because it encompasses theories which resemble central cases of theories of punishment in important respects. For instance, the incapacitation theories that it encompasses resemble central cases of punishment theories in that 1) they recommend significant interference with the offenders' liberty by the state and 2) this interference is imposed in response to a criminal offence.[11]

Structure of the Thesis

My Thesis has the following structure: In Part One I will argue that there is at least a reasonable doubt about the soundness of retributivism. This doubt arises from retributivism's reliance on a hotly contested conception of free will. Chapter One will provide reasons for doubting that retributivism could justifiably rely on the assumption of libertarian free will. Chapter Two will provide reasons for doubting the adequacy of compatibilist retributivism. Chapter Three will respond to an argument for retributivism, based on the implications of that theory for our practices: the argument that we need retributivism because it is the only theory that implies that

[10] Pereboom D, *Living without Free Will* (CUP, Cambridge 2001), pp174-186.
[11] But see chapter 6 for discussion of the possibility that incapacitation theories might allow for pre-emptive detention of those who have committed no crime. Readers who are reluctant to accept my stipulative definition of 'theories of punishment' should construe references to such theories as referring to 'coercive state responses to criminal behaviour'.

accused people and offenders should be protected by considerations of justice. I will argue that in fact there are good reasons for thinking that our practice of upholding these principles of justice could be defended on non-retributive grounds.

In Part Two I will relate the doubts about the soundness of retributivism that I raised in Part One to the literature on 'moral uncertainty' (i.e. uncertainty about which moral theory should guide our conduct). I will point out that, in the light of this literature, adopting a purely consequentialist theory of punishment would not be a rational response to the doubts about the soundness of retributivism since there is also uncertainty about the soundness of consequentialism. Chapter Four will present an overview of some of the main theories of moral uncertainty and will highlight some of their key shortcomings (at least in relation to dealing with uncertainty about theories of punishment). Chapters Five and Six will defend my own approach to moral uncertainty about theories of punishment. On my approach, the entire moral argument for punishing a person should be held to a high standard of credibility. I call this the 'cautious approach to punishment'. In Chapter Five I will argue that one reason for adopting a cautious approach to punishment stems from the underlying rationale for the beyond reasonable doubt standard in criminal trials – a standard that has widespread support. In Chapter Six, I will argue that in order to minimise the risk of punishing someone unjustifiably, we should only punish that person if the main punishment theories agree that doing so is justifiable. I will call this 'the convergence requirement'. If the convergence requirement is satisfied, then the state's argument for punishing a person has met the required standard of credibility. I will acknowledge in Chapter Six that, notwithstanding the arguments against retributivism presented in Part One, when we survey the state of the free will debate, there is a strong argument for 'free will agnosticism' rather than for *certainty that we*

lack free will in the sense required for retributive responsibility.[12] Given that the cautious approach recommends giving people who may be liable to punishment 'the benefit of the doubt', such people should only be punished if retributive theories (as well as forward-looking theories) would recommend this; i.e. in effect this means that punishment should be constrained by negative retributivism. The possibility that negative retributivism might be sound therefore provides a reason against such intuitively unjust practices such as punishment of the innocent – a reason in addition to those non-retributive reasons outlined in Chapter Three. Chapter Six will also give some reasons why theorists from different philosophical perspectives should endorse the convergence requirement and will defend this requirement against certain potential objections.

The convergence requirement assumes that punishment infringes the interests of the individuals who are punished – in particular their interest in not being seriously harmed - and therefore requires strong justification. Part Two focuses primarily on which individuals should be punished *at all*. Part Three will address the question of which method of responding to an individual's criminal behaviour should be preferred, once it is determined that some response is required, in situations where several possible alternatives are available. This question is complicated by the fact that offenders have a number of different interests that should be protected, but these interests can sometimes pull in different directions. In general, offenders have an interest in not being deprived of liberty. Therefore, where one mainstream theory recommends a sentence that involves less interference with liberty, that sentence should usually be preferred. However, I will also maintain that offenders have an interest in being treated as rational agents. Certain rehabilitative interventions might allow offenders to be released earlier into society, but may fail to respect the

[12] S. Kearns, 'Free Will Agnosticism' (2013) Nous (Online First). DOI: 10.1111/nous.12032

offender as a rational agent. This problem arises most acutely in connection with the possibility of using direct brain interventions to modify criminal behaviour. Therefore, Part Three will be devoted to examining this example in detail. It is also important to explore this topic, since it is of particular relevance to the arguments for free will scepticism which I advanced in Part One, when raising doubts about the justifiability of relying solely on a retributive theory punishment. It is tempting to explain the troubling nature of certain direct brain interventions by claiming that they threaten free will. If free will scepticism implied an acceptance of these troubling interventions then this could undermine free will scepticism and could possibly strengthen the case for a retributive system that stressed the importance of free will. However, in Chapter Seven, I will argue that, in fact, the objectionable nature of certain direct brain interventions has very little to do with free will. In Chapter Eight, I will argue that there are in fact non-retributive reasons for opposing the most intuitively-objectionable interventions. I will identify certain forms of biomedical intervention that are genuinely objectionable and that should not be used within the criminal justice system. In Chapter Nine, I will argue that, *in principle,* it would be morally permissible for the state to employ certain types of biomedical intervention (such as 'cognitive enhancements') in a limited way within the criminal justice system, provided that effective enhancements can be developed in the future that have minimal side-effects. Chapter Nine will then consider how we can distinguish interventions that enhance rational capacities from interventions that fundamentally change the person's character, and the extent to which this distinction matters.

In Part Four, I will discuss some of the practical implications of my approach for the criminal law. Chapter Ten will examine the criminal law's current position on the questions of free will and retributive responsibility. It will argue that the dominant view among criminal law theorists - that the criminal law is thoroughly compatibilist

- rests on dubious arguments. There is at least as much reason for thinking that principles of criminal law embody libertarian, incompatibilist assumptions as there is for thinking that they make only compatibilist assumptions. Chapter Eleven will examine what revisions to criminal law doctrines would enable these doctrines to be justifiable even if retributivism (and the notion of free will on which it depends) were regarded as unsound. It will suggest several changes to our *understanding* of the rationale for the provocation defence, self-defence and mental disorder defences; and for the overall *structure* of criminal defences. My intention in Part Four is to give a general overview of directions for future work in this area and not to give a comprehensive account of the criminal law doctrines discussed.

Part One: Free Will and Retributive Punishment

Part One: Overview

In Part One I will argue that the truth of determinism would seriously undermine the retributivist justification of punishment. Furthermore, I will argue that retributivists cannot successfully defend their view by dismissing determinism as empirically false or conceptually incoherent.

Retributivism can seem intuitively appealing. To take an example suggested by Ted Honderich, imagine that a man injures someone you care about, "or defrauds her in a financial transaction, or concocts evidence against her in a court… [you may well] have a *retributive desire*...[You may] want it to come about that [the man] suffers at least some unhappiness….. The desire may go a lot further than that."[13] People who have this kind of retributive attitude may well demand that wrongdoers are punished, without even considering the deterrent or reformatory effects such punishment may or may not have. Furthermore, such people may insist that their retributive attitudes are different from mere vengeance, because the former are based on *moral indignation,* involving a belief about the requirements of justice, whereas the latter involve purely personal feelings, such as hurt pride, anger, or vindictiveness.[14] The arguments presented in Part One aims to cast serious doubt on whether these retributive attitudes can be justified. If the institution of punishment is to be preserved, there are good reasons for seeking a non-retributive justification for it.

Chapter One will critique 'libertarian' defences of retributivism. Libertarian retributivists deny that human behaviour is determined and maintain that indeterminism enables us to be retributively responsible. Chapter One begins by citing evidence from neuroscience that supports the theory that the mental processes

[13] Honderich, *How Free Are You?,*p101.
[14] For an attempt to distinguish retributivism from vengeance see: Moore, *Placing Blame*, chapters 3 and 4.

that lead to action are determined. Libertarians have not shown that quantum events affect such processes in a way that could provide a basis for retributive responsibility. Furthermore, even if they could show this, it is extremely doubtful whether quantum indeterminacy could in any way enhance human freedom and responsibility. Finally, Chapter One will respond to a prominent libertarian argument against determinism – that it is conceptually incoherent and self-refuting.

Chapter Two will argue against the view that retributivism and determinism are compatible. Unlike libertarians, 'compatibilists' need make no claims about the actual truth or falsity of determinism. Hence 'compatibilist responsibility' is not a hostage to empirical fortune. This may partly explain why the majority of philosophers seem to favour compatibilism.[15] Since compatibilism is the dominant view, my critique of this theory will be longer than my critique of libertarianism. I will argue that the truth of determinism would seriously undermine retributivism. This is because determinism entails that human actions are inevitable, whereas retributive responsibility requires that the agent had the ability to avoid performing the wrongful action. Determinism also entails that the agent is not the ultimate source of her action; rather, ultimately, her action is a product of luck. This is also incompatible with retributive responsibility. Finally, Chapter Two will argue that attempts to rescue retributivism from these metaphysical difficulties by appealing to our 'moral experience taken as a whole' are unsuccessful.[16]

Chapter Three will then argue that a 'hard incompatibilist' approach to punishment (i.e. one that rejects libertarian and compatibilist retributivism) could be fair and humane. It is possible to have justice without retributive desert.

---

[15] Bourget D and Chalmers D (eds.) *The Philpapers Survey 2009*, available at http://philpapers.org/surveys/results.pl Accessed 31st May 2011.

[16] Moore, *Placing Blame,* p543.

Punishment is the deliberate infliction of suffering in the name of justice. The serious nature of legal sanctions is reflected in the requirement that the guilt of an accused person must be proven beyond reasonable doubt, before he can be punished. In Part Two, I will argue at length that the entire moral argument for punishing a person should be held to a similarly high standard of credibility, before it is fair to rely on it. Part One attempts to cast serious doubt on the idea that retributivism has met this standard.

Chapter One: Libertarian Retributivism

Introduction

This chapter will argue that the retributivist is unable successfully to defend his theory against the problem of determinism by simply dismissing determinism as empirically or conceptually flawed. Firstly, evidence from neuroscience provides some support for the idea that mental processes that lead to action are determined. Secondly, it will be argued that libertarians have not shown that quantum events affect such processes in a way that could provide a basis for retributive responsibility. Thirdly, this chapter will argue that determinism is conceptually coherent.

Empirical Support for Determinism

In order to avoid confusion, it is important at the outset to distinguish between two ways in which people use the word "cause". Firstly, "cause" is often used to indicate a *causal factor* which contributes to bringing about an effect. For example, fuel, kindling temperature and the presence of oxygen are all required in order to light a fire. Any one of these conditions could be described as a cause of the fire, in the sense of "causal factor".[17] In contrast, a *necessitating cause* (which Ted Honderich calls a "causal circumstance"[18]) is a set of conditions such that, if those conditions are present, a certain effect will *always* follow, whatever else is the case. For example, the presence of fuel, oxygen and kindling temperature together constitute a necessitating cause of combustion. It is the latter sense of "cause" which is central to

[17] Usually, the factor referred to as "the cause" is a new event e.g. the striking of the match, where fuel and oxygen are already present, or the application of fuel to a flame, where the oxygen is already present. Michael Moore, whose ideas will be discussed at greater length in chapter five, sometimes seems to imply that determinism relies solely on the concept of causal factor. See, for example *Placing Blame*, pp532-533 and p543.
[18] Honderich, *How Free Are You?*

the theory of determinism. (Throughout this thesis the term "cause" will be used to mean "necessitating cause", unless otherwise indicated. Sometimes, for emphasis, the full term "necessitating cause" will be used.)

Michael Moore accepts determinism because he believes that the alternative, indeterminism, is highly implausible. Moore asks, " is it not extraordinary to think that part of our most basic metaphysical picture of what the universe is like - in terms of causal relations – should have no application to persons? Is it not extraordinary to think that agents who clearly cause changes to occur in the world are themselves uncaused?"[19]

Indeed, scientists are generally agreed that neural processes in the human brain operate causally (e.g. in response to internal or to external (i.e. environmental) stimuli).[20] Furthermore, there is a large body of evidence to support the claim that conscious mental phenomena (including memories, emotions, imagination and, most importantly in this context, rational deliberation) necessarily go together with certain brain processes. For example, experiments have shown that when a particular area of a patient's brain is electrically stimulated, that patient will have the subjective experience of a particular "flashback" memory. Studies of stroke victims have established that specific areas of the brain relate to specific cognitive functions such that if those areas of the brain are damaged the associated cognitive functions will be lost or impaired.[21]

This evidence has been interpreted in various ways by determinists. One plausible view has been put forward by Ted Honderich. He argues that there is a nomic (law-

[19] Moore, *Placing Blame*, p504.
[20] For a lucid summary of some of the key evidence from neuroscience and its relevance to determinism see Honderich, *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes* (1988)
[21] See Wilder Penfield, *The Mystery of the Mind* (1975) N.R. Carlson, *Physiology of behaviour* (1994)

like) connection between mental events and neural events, such that "if or since one occurred, whatever else had been the case, the other would still have occurred".[22] In contrast, Libertarians, such as C.A. Campbell, believe that, in the case of genuinely "free" actions, the self brings about the action in a way that is not causally necessitated by brain processes, by nomic pairs of brain process and mental states, or by anything else.[23] Honderich makes a convincing case for thinking that the most plausible interpretation of the evidence of brain science does not support such an indeterminist view.[24] He argues that the findings of an intimate, necessary connection between neural phenomena and mental phenomena, *rule out* "what is fundamental to the [indeterminist's] free-floating self… Whatever, else was supposed to be true of the self…, it was supposed to be above and beyond the brain."[25]

Furthermore, one of the main "methods" which libertarians (such as C.A. Campbell[26] and John Searle[27]) have used to try to establish that mental processes are not necessitated does not seem capable of establishing any such thing. These theorists claim that we can discover such facts about the mind and brain through introspection on the activity of choosing. This involves examining what it is like to make a choice, from the perspective of the person actually making that choice, rather than from the viewpoint of an external observer. This introspection, libertarians claim, reveals "gaps" in the causation of human actions. However, introspecting on a particular experience of choosing cannot possibly establish whether that choice was determined. The Proposition "this choice was determined" refers to something

[22] Honderich, *How Free Are You?*. Some other interpretations of the mind-brain relationship which are consistent with determinism include Functionalism, Physicalism and Epiphenomenalism. For a critical discussion of some of these theories see John Searle, *The Rediscovery of the Mind* (1992)
[23] C.A. Campbell, *In Defence of Free Will* (1967).
[24] Honderich, *On Determinism and Freedom* (2005), chapter 5.
[25] Honderich, *How Free Are You?*, p69.
[26] C.A. Campbell, *In Defence of Free Will* (1967).
[27] John Searle, "Consciousness, Free Action and the Brain", Journal of Consciousness Studies, 2000, cited in Honderich, *On Determinism and Freedom* (2005), chapter 5.

outside the experience of choosing, presumably prior to it, i.e. its cause. Furthermore, even if the libertarian (after reflecting on the experience of choosing) somehow gets a powerful feeling that no necessitating cause preceded the choice, the libertarian cannot show that this powerful feeling is not an illusion, itself causally determined. Empirical support for the truth or otherwise of determinism can only be provided by wide-ranging studies of the causation of human behaviour, not by introspective reports about what the experience of choosing is like for particular agents.[28]

The Retributivist Cannot Rely on Quantum Indeterminism

On one interpretation of quantum theory, quantum events at the microlevel of atomic and subatomic particles are undetermined.[29] However, even if this theory is correct, the idea that quantum indeterminism may result in indeterminism at the macro-level (in particular at the level of neuron firings in the human brain) is speculative.[30] J.J.C. Smart argues that quantum effects are unlikely to be substantial enough to have any significant effect on deliberative processes. He says, "even a single neuron is a huge macroscopic object by the standards of quantum mechanics and furthermore, the

[28] For criticism of the "introspection method" along the lines given in this thesis, see Honderich, *On Determinism and Freedom* (2005), chapter 5.

[29] However, some writers have disputed this interpretation of quantum theory. According to Honderich, for example, a quantum "event" may not actually be an "event" in the ordinary sense of that term i.e. something that happens in space and time, such as an action, or the firing of a neuron. Rather, Honderich suggests, quantum "events" may be purely theoretical entities of the nature of propositions or numbers. For a full discussion see, Honderich, *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes* (1988) pp304-334. Robert Bishop, points out that some theorists believe quantum indeterminism is merely epistemic (i.e. the appearance of indeterminism is just due to our ignorance of the actual causes), see Bishop, "Chaos, Indeterminism and Free Will", in Kane (ed), *The Oxford Handbook of Free Will*, (2002), 110, at p118 and p120 (henceforth: Bishop, "Chaos, Indeterminism and Free Will").

[30] Even supporters of the view that quantum mechanics may provide a basis for retributive responsibility are extremely cautious in the way they express their views, e.g. Robert Kane writes, "It is *conceivable* that …indeterminacy could arise at macrolevels" [emphasis added], in "Some Neglected Pathways in the Free Will Labyrinth", in Kane (ed), *The Oxford Handbook of Free Will*, (2002), 406 at p434 (henceforth, Kane, "The Free Will Labyrinth").

failure to fire of an odd neuron is unlikely to affect behaviour….any thought or action almost certainly depends on a large number of neurons."[31]

Even supposing, contrary to fact, there were strong evidence suggesting that quantum events have noticeable effects on human thoughts and actions, the libertarian would have to show that they affect the causal chain leading to action at just the point required for retributive responsibility. He would have to show for example that quantum events occur during the deliberative process so that the decision the agent comes to is not determined. However, he would also have to explain why quantum events do not intervene between the decision stage and the action. (If they did intervene between decisions and actions people would constantly be performing actions they never intended and failing to perform ones they did intend!)[32] As Honderich says, "how can [the libertarian] consistently do this? Does quantum theory as interpreted have some clause, hitherto unheard of, that its random events occur only in such places as to make us morally responsible in a certain sense?"[33]

It also seems highly unlikely that quantum events could provide a basis for retributive responsibility even if they did occur in the "right place" in the causal chain and only in the right place. For it seems that this would just introduce randomness into our deliberative processes. This situation creates a dilemma for the retributivist. As A.J. Ayer argues, either it is a matter of chance "that [an agent chooses] to act as [he does] or it is not. If it is a matter of chance, then it is surely

---

[31] J.J.C. Smart, *Philosophy and Scientific Realism* (1963) p123. More recently there has been speculation that chaos may amplify quantum events and thereby introduce indeterminism in the workings of the brain. However, the presence of chaos in the brain is "currently hotly debated and inconclusive", and furthermore, as noted above fn 20, some theorists doubt whether quantum events are genuinely undetermined – Bishop, "Chaos, Indeterminism and Free Will", at p119

[32] Honderich, "Determinism as True, both Compatibilism and Incompatibilism as False, and the Real Problem", in Kane (ed), *The Oxford Handbook of Free Will*, (2002) 461, at p467.

[33] *Ibid.*

irrational to hold [him retributively] responsible for choosing as [he] did;…If it is not a matter of chance, then presumably there is some causal explanation of [his] choice: and in that case we are led back to determinism." [34]

The Libertarian may respond that an agent's choice is not a matter of chance but depends on something for which he *is* responsible, namely his character. He may argue that quantum events might provide a "window of opportunity" (i.e. a gap in the causal chain) for the agent's character to assert itself.[35] However, the claim that an agent is responsible for his character entails that he has made himself what he is. This seems to lead to an infinite regress for it raises the question: "who was the 'he' that made the character and how was he made?".[36]

To summarise the argument so far: evidence from neuroscience supports the theory that the mental processes that lead to action are determined. Libertarians have not shown that quantum events affect such processes in a way that could provide a basis for retributive responsibility. Therefore, the retributivist is unable successfully to defend his theory against the problem of determinism by simply dismissing determinism as empirically flawed.

The Conceptual Coherence of Determinism

Determinism is an empirical theory. Its truth depends on certain facts about the world. However, empirical theories must also be *conceptually* sound. Various philosophers have maintained that, whatever the facts may be, they can still show that determinism is necessarily unjustifiable. The objection that shall be discussed in this section was first formulated by the ancient Greek philosopher Epicurus and is

[34] A.J. Ayer, *Philosophical Essays* (1954) pp275-276. For a similar line of argument, see also: J.J.C. Smart, 'Free Will, Praise and Blame' 1961, *Mind*, 291.
[35] For a similar line of argument, see Kane, "The Free Will Labyrinth".
[36] For further elaboration of this last point, see G Strawson, "The Bounds of Freedom".

still influential in the contemporary debate about determinism and responsibility.[37] The objection states that determinism is self-defeating. Epicurus argued that, for determinism to be an intellectually respectable theory, it must be in accordance with reason, and reasoning is not a causal process. Therefore, the objection runs, a determinist cannot coherently argue for her position, because she admits that her own "arguments" are themselves merely a matter of cause and effect, rather than rationality. In this chapter, it will be argued that determinism does not imply that humans generally do not have the ability to engage in rational deliberation, nor does it imply that it is impossible to have a justified true belief (e.g. the belief that determinism is supported by scientific evidence.)

Stimulus-response versus rational deliberation

The statement that reasoning is not a causal process has seemed plausible to some critics of determinism because they take 'causal process' to indicate a simple, largely automatic stimulus-response mechanism. Indeed, some determinists have left themselves open to Epicurian-style objections by characterising the causal process in this way. For example, Hans Eysenck argued that the behaviour of human adults can be explained in terms of the conditioning they underwent in early childhood, as well as their biological make-up. This theory does not explain why people alter their behaviour (often rejecting the life-styles they were brought up with) in response to cogent reasons for doing so. Eysenck's account leaves no room for reasoned deliberation.[38]

[37] More recent discussions of versions of this objection include: A.C. MacIntyre, 'Determinism', in Bernard Berofsky (ed), *Free Will and Determinism* (1966), 240; Peter Westen, "Getting the Fly out of the Bottle: The False Problem of Free Will and Determinism" (2004) 8 Buffalo Criminal Law Review 599.

[38] For a critical discussion of Eysenck's view see Taylor, Walton and Young, *The New Criminology* (1973), pp 47-61.

However, M.C. Bradley has persuasively argued that there is a great difference between a simple, conditioned reflex "action" and a complex deliberative action, but insists that both actions can be explained in terms of causes. To illustrate his point, Bradley draws the following analogy: "There is a great difference between the machine which shoots out a chocolate when a sixpence is put in it, and the machine which calculates on the basis of range, direction and velocity of target, velocity of gun, wind-speed and direction, barometric pressure, etc., etc., where a shell must fall, fired from a moving gun to strike a moving target. But this vast difference in complexity has no bearing whatsoever on the question whether in each case the end-product is or is not strictly determined by a series of steps, themselves all strictly determined." [39]

It might be objected that human deliberation is not analogous to Bradley's "gun-laying machine", because no computer or robot could take into account the vast range of information which human beings are capable of taking into consideration. In reply to this point, it can be said that Bradley's analogy does not depend on the actual existence of such a complex machine or the practical possibility of making one. All that matters is that it is possible in principle to make one. The operation of some sophisticated machines is relevantly similar to deliberation in humans, because both these machines and individuals have the ability to *modify their behaviour/operations in the light of logically relevant data*. Bradley's example shows that the complexity of this data is irrelevant to the question of whether the mind/machine evaluating it is operating causally.

Flexibility – the ability to adapt one's behaviour in an appropriate way to changes in circumstances – is generally agreed to be a hallmark of rationality. A critic of

[39] M.C. Bradley, 'A Note on Mr MacIntyre's Determinism', in Bernard Berofsky (ed), *Free Will and Determinism* (1966), 256, at p260.

Bradley's example might concede that a being whose behaviour was determined could exhibit flexibility in response to a wide range of stimuli. Nevertheless, the critic might insist, if the being's behaviour is determined, then the being's flexibility will have limits, i.e. there will always be some stimuli to which the being will be unable to respond appropriately. For example, the sophisticated gun-laying machine might not be able to detect the difference between an enemy aircraft and a decoy, nor will the chess-playing machine (which is good enough to beat the human grandmaster) be able to play chequers or bridge and "input appropriate to these other games would reveal the system to be as non-rational and unresponsive as any stone".[40] Indeed, these facts should be acknowledged. However, it is submitted that they do not show that rationality and determinism are incompatible. For, it seems implausible to demand that a being must have *infinite* flexibility in its responses to stimuli, (i.e. perfect rationality) before one can say that its behaviour in general is rational. For what human being would satisfy such a demanding criterion? As Daniel Dennett puts it, "For every awe-inspiring stroke of genius…(of the Einstein-Shakespeare gambit), there are a thousand evidences of lapses, foibles, bumbling and bullheadedness to suggest to the contrary that man is only imperfectly rational."[41] If the critic were to stipulate that infinite flexibility is a necessary condition for rationality, it would seem that no-one is rational, irrespective of the truth of determinism.

Another critic of Bradley's example might point out that humans are different from machines in an important respect. Unlike a machine, a human can be consciously aware of taking on board and evaluating the relevant data and may have feelings and attitudes of pleasure/pain or moral approval/disapproval in connection with the data

[40] Daniel Dennett, "Mechanism and Responsibility", in Gary Watson (ed), *Free Will* (1982), 150, at p161.
[41] *ibid*, at p163.

being examined. Again this fact should be acknowledged. However, the fact that our mental processes have a subjective conscious dimension does not mean that those processes do not operate causally.

In the next chapter, it will be argued that retributivism is incompatible with determinism, because determinism rules out the ability to do otherwise, which is a necessary condition for retributive responsibility. However, it should be noted that a parallel argument cannot be constructed to show that rationality is incompatible with determinism. It is not necessary to be able to respond differently to a given stimulus in order for that response to be rational. The ability to reason soundly does not require the ability to reason otherwise, i.e. to reason fallaciously. As Dennett puts it: "the mere fact that [a certain] response *had* to follow, given its causal antecedents casts no more doubt on its rationality than the fact that the computer had to answer '108' casts doubts on the arithmetical correctness of its answer."[42]

In chapter four, it will be argued that a person cannot be morally responsible in the retributive sense if the fact that she made an immoral choice, or has a corrupt value-system is, as determinism implies, ultimately a matter of luck. However, again, it is submitted that a parallel argument cannot be made to show that rationality and determinism are incompatible. Ultimately, the fact that a person can respond appropriately to certain data (e.g. a theorist's arguments) may be explained by environmental factors (e.g. being taught by good teachers) and by genetic factors. If the presence of these factors were outwith the control of a given individual (and were therefore matters of luck), this would in no way detract from the rationality of the individual who would benefit from them.

[42] *Ibid,* at p164.

It might be objected that flexibility is not necessary for free will or rationality, because a person who is thoroughly committed to acting in accordance with a certain moral principle might not behave differently under any circumstances. For instance, a committed pacifist might never resort to violence, and yet would be considered free and rational. However, it is possible to distinguish the offender with an irresistible aversion to violence from a person with a firm moral commitment. If a person's non-violent conduct is genuinely a response to a moral reason, then one of the causal factors bringing about her behaviour is the perception that violence is morally wrong. She has the capacity to alter her behaviour if she revised her moral position in the light of new arguments or evidence. This capacity to alter one's behaviour in response to a change in one's values is an important kind of flexibility.[43]

Knowledge

Given that causally determined "deliberation" need not be of the crude stimulus-response variety, the Epicurian might still object that such deliberation could not lead to *knowledge.* He will urge that we could have no confidence in the truth of our beliefs if those beliefs are necessitated effects. Such beliefs *might* be true, but, the Epicurian will insist, we could have no confidence that they were true because, given that they were caused, we would have had them any way, even if they had been false. Ted Honderich makes the following persuasive reply to this kind of objection: The objection depends on "supposing that if my [belief] were false I would still have been caused to have it. But why should that be the case? If my [belief] were false, I might not have been caused to have it. I now think there is a keyboard in front of me. Would I still be caused to think that if there wasn't one?"[44] The Epicurian might

[43] Compatibilists differ over whether the flexibility possessed by rational agents in a deterministic world genuinely amounts to a capacity to behave differently from the way that one in fact behaves. The following theorists argue that it does: Fara, 2008; Vihvelin, 2004. The following theorists disagree, maintaining that the disposition to respond differently if different reasons were present is simply a feature of the way in which the agent *actually* behaves: Fischer and Ravizza, 1998.
[44] Honderich, *How Free Are You?*, pp 88-89.

concede the point in the case of Honderich's keyboard, but might still argue that causation cannot provide one with a comprehensive guarantee of the correctness of one's beliefs on *every* occasion. However, as Honderich points out "*Nothing* can give me all-encompassing confidence. In particular I couldn't get it from [indeterminism]." [45]

Conclusion

This chapter aimed to undermine indeterminist, libertarian defences of retributivism. I began by citing evidence from neuroscience that lends some support to the theory that human actions are determined. I then argued that quantum indeterminism has not been shown to affect mental processes in a way that could provide a basis for retributive responsibility. Next, I argued that determinism is compatible with reasoned deliberation and knowledge and is not self-refuting.

Any attempt to justify punishment must meet a fairly high standard. It cannot be based on wishful thinking or speculation (e.g. speculation about the *possibility* that someday science *might* show that mental processes are undetermined in a way that could leave room for retributive responsibility, assuming that this idea is even coherent). The retributivist would be on firmer ground, therefore, if she could show that her theory is compatible with determinism. In Chapter Two, it will be argued that she cannot satisfactorily show this.

[45] Honderich, *How Free Are You?*, pp 88-89.

Chapter Two: Compatibilist Retributivism

Introduction

In the previous chapter, it was argued that reasoning can be characterised as a causal process. A retributivist might then claim that if reasoning is compatible with determinism, it follows that retributive responsibility is also compatible with determinism. In other words, the retributivist might make the following claim: If a person who can reason, knowingly and voluntarily performs a wrongful action then she deserves to be punished even if her wrongful action were determined and irrespective of the consequences of punishing her. The chapter will argue that if a wrongdoer's action was determined, she is not morally responsible in the retributive sense for that action, *even if she has the ability to reason*. This is because determinism entails that human actions are inevitable, whereas retributive responsibility requires that the agent had the ability to avoid performing the wrongful action. Determinism also entails that the agent is not the ultimate source of her action; rather, ultimately, her action is a product of luck. This is also incompatible with retributive responsibility. Finally, this chapter will argue that attempts to rescue retributivism from these metaphysical difficulties by appealing to our 'moral experience taken as a whole[46]' are unsuccessful.

The Principle of Alternative Possibilities

Firstly, it is important to distinguish between two different kinds of reasoning – theoretical reasoning and practical reasoning. The former involves reasoning about *what is the case.* In the previous section, in the context of theoretical reasoning, it was argued that it is possible to have a justified true belief (e.g. the belief that determinism is strongly supported by scientific evidence) even if one is caused to

[46] Moore, *Placing Blame* (OUP 1997).

have that belief. This section will focus on practical reasoning. This kind of reasoning involves deliberating about *what to do*.

The arguments against retributivism that will be put forward in this section are based on "the principle of alternative possibilities" (henceforth: PAP).[47] This principle states that in order for someone to be morally responsible in the retributive sense for doing a wrongful action it must have been possible for her to have done otherwise.[48] It is *not*, however, being argued that the existence of alternative possibilities is a sufficient condition for retributive responsibility. There may well be other considerations which undermine retributive responsibility. Rather, it is submitted that PAP is a necessary condition for retributive responsibility.

In the first part of this chapter, it will be demonstrated that the principle of alternative possibilities is intuitively plausible. The second part of this chapter will focus on thought-experiments that are designed to refute the principle of alternative possibilities. It will be argued that these thought experiments are flawed. The final part of this chapter will criticise attempts to produce a version of PAP that is compatible with determinism.

[47] Carl Ginet, "In Defence of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing"(1996) 10 Philosophical Perspectives 404 (henceforth: Ginet, "In Defence of Alternative Possibilities").

[48] Ginet argues that a person needs to have the ability to do otherwise *right up until the moment the action is performed* in order to be responsible for that action. However, this is a debatable point. If the action flows inevitably from a choice the agent made and the agent had the ability to avoid making that choice, then, arguably, this is enough to render the agent responsible for the action. Both of these positions are incompatible with determinism since determinism renders all actions and choices inevitable. Therefore, the subtle differences between these positions will not be discussed in the text: In favour of the former position see Ginet, "In Defence of Alternative Possibilities". In favour of the position that one needs alternative possibilities when making a "self-forming choice" see Kane, "The Free Will Labyrinth".

The Intuitive Plausibility of PAP

Consider the following example of practical reasoning: Jane is choosing whether to give in to a temptation to steal John's bicycle. She is capable of grasping all the reasons in favour of refraining from stealing the bicycle and weighing them against the satisfaction she expects to derive from stealing it. In the end, Jane decides to steal the bicycle. Why does Jane decide to steal the bicycle? If determinism is true, the causes of Jane's decision can be traced back to factors ultimately beyond her control (such as the biological and environmental factors which formed her character). These determining causes would have rendered her decision to steal the bicycle inevitable. If determinism is true, she could not have done otherwise than steal it.

Imagine Jane is convicted of theft. Imagine the judge who heard Jane's case believes that determinism is true, but nevertheless believes that Jane is morally responsible in the retributive sense. After all, he thinks, Jane was not mad when she stole the bicycle. She knew what she was doing. Nor did anyone force her to steal it. Rather, she acted on reasons which seemed appropriate to her in the light of her character and value-system. The judge condemns Jane in the strongest terms. Consistent with his retributivism, he maintains that Jane deserves to be punished *purely* because of the nature of her deed, and irrespective of other considerations, eg whether punishment is likely to reform her, reduce the likelihood of her reoffending, deter other potential bicycle thieves etc. Yet, consistent with his determinism, in the next breath he acknowledges that, if he himself had the misfortune to be subject to the same determining causes which necessitated Jane's action, he would inevitably have done exactly the same thing as she did.

The judge's position is paradoxical to say the least. Surely, if judges were regularly to come out with statements like that, then respect for the law would be undermined.

Yet, the pronouncements made by the hypothetical judge are entailed by accepting that retributive responsibility is compatible with determinism. Jane could put the following challenge to anyone who endorses the position adopted by the judge: "since you wish to punish me purely because of the wrong that I have knowingly done, *please tell me what, in your view, I should have done instead*? You cannot claim that I should have decided not to steal the bicycle, because, given that you believe in determinism, that was not something I was able to do."[49] Jane's challenge is intuitively appealing as it is based on the widely accepted idea that ought implies can.

In real life, most people who express retributive sentiments do not endorse the kind of counter-intuitive position adopted by the imaginary judge. They simply do not accept determinism.[50] Imagine, for example, that Anne has retributive feelings towards Bryony, who has injured Anne. Clarence, in defence of Bryony, pleads the story of Bryony's disadvantaged background. A person who, like Anne, wishes to persist in her retributive feelings, typically responds by saying that Bryony could have resisted and overcome the pressures of her upbringing. Anne will claim that Bryony's upbringing did *not* necessitate her wrong-doing, that it did not make her wrong-doing inevitable. Anne may point to other cases of people who had disadvantaged backgrounds and yet remained morally upright citizens, who would not injure anybody. Furthermore, Anne does not believe that *anything* necessitated

[49] This kind of challenge was formulated by David Widerker and is known as the "W-defence". See Widerker, "Frankfurt's Attack on Alternative Possibilities" (2002) 14 Philosophical Perpectives 181, at p191.

[50] This is either because they have never thought about it or because, having thought about it, they reject it. In real life, judges typically do not accept determinism. For example, a Canadian court recently stated that "[The] criminal law rejects a determinist theory of crime.…The blunt fact is that a wide variety of societal ills ….are part of the causal soup that leads some individuals to commit crimes. If those ills are given prominence in assessing personal culpability, an individual's responsibility for his or her actions will be lost." *R v Hamilton*, [2004] 186 C.C.C. (3d) 129, per Justice David Doherty at 140.

Bryony's wrong-doing. Anne thinks that in some fundamental sense Bryony's wrong-doing was ultimately and absolutely up to Bryony.

Frankfurt-Style Cases

A retributivist might attempt to refute PAP by suggesting the following thought-experiment of the type originated by Harry Frankfurt[51]: Again, imagine that Jane is deciding whether or not to steal John's bicycle. Let us assume that Jane does not live in a deterministic universe, so in the normal course of events, she would be able to refrain from stealing the bicycle.[52] However, on this occasion, a demonic neuroscientist (who really wants Jane to steal) is monitoring Jane's thoughts. If he sees that Jane is about to refrain from stealing, the neuroscientist has the power to intervene to make her steal. However, in the event, Jane decides on her own to steal and so the scientist does not intervene. (In this scenario the neuroscientist can be referred to as a "counterfactual intervener", because he would only have intervened if the sequence of events which actually did occur had not occurred.)

This example is meant to illicit the intuition that Jane could be retributively responsible for stealing the bicycle, even though the presence of the neuroscientist meant that Jane could not have done otherwise. However, this example is flawed because, in a sense, Jane could have done otherwise. The scientist monitoring Jane's thoughts needs some sign which indicates that, unless he intervenes, Jane will not steal. It seems that this sign must be some mental event in Jane's brain such as the formation of an intention not to steal. Therefore, at some point in time prior to her theft, Jane did have an alternative possibility open to her. She could have *formed the*

[51] Harry Frankfurt, 'Alternate Possibilities and Moral Responsibility' (1969) 66 Journal of Philosophy, 829.
[52] In this example it is being assumed *for the sake of the argument* that, in the normal course of events (i.e. when neuroscientists are not interfering with Jane's actions) she is retributively responsible for her actions, although, as noted above, there are difficulties with the notion of retributive responsibility itself and how it could be reconciled with indeterminism.

intention to refrain from stealing. It is plausible to say that Jane is blameworthy because she did not form this intention.[53] The thought-experiment is presented in diagrammatic form on the next page. (The arrow indicates temporal succession i.e. events to the left of the arrow occur earlier than events to the right of the arrow.)

<u>Actual sequence of events:</u>

Jane forms an intention to steal ⟶ Jane steals

<u>Counterfactual sequence of events:</u>

Jane forms an intention not to steal ⟶ scientist intervenes ⟶ Jane steals[54]

John Martin Fischer, attempts to refute PAP by producing a modified version of Frankfurt's thought-experiment.[55] In Fischer's version, the neuroscientist knows that Jane will only form an intention to steal if she blushes beforehand. Jane's failure to blush by a certain time is the sign which would trigger the neuroscientist's intervention. However, in the actual sequence Jane blushes, the neuroscientist does not intervene and Jane forms an intention to steal which she then carries out. Fischer then concludes that it is plausible to hold Jane morally responsible in the retributive sense[56] even though the neuroscientist ensured that she could not have done

[53] Ted Honderich, *How Free Are You?,* p117

[54] It is implied that, in the actual sequence, after she forms the intention to steal, that intention persists throughout her action.

[55] John Martin Fischer, *The Metaphysics of Free Will* (1994), pp131-159.

[56] In this thesis it will be assumed that the sense of moral responsibility that Fischer is working with is *retributive* moral responsibility. He quotes Galen Strawson's conception of responsibility - "ultimately, truly and without qualification responsible and deserving of praise or blame or punishment or reward"- and comments, "I find it easier…simply to employ the term , 'morally responsible', where we keep it in mind that this involves genuine, robust moral responsibility (and not a revised or watered down version..)." Fischer, "The Cards that are Dealt You"(2006) 10 The Journal of Ethics 107, at p 111.

otherwise and could not have formed an intention to do otherwise. (We are also meant to assume that Jane was not in control of whether or not she blushed.) Fischer's thought-experiment is represented in diagrammatic form below.

Actual Sequence of events:

Blush ⟶ Jane forms an intention to steal ⟶ Jane steals

Counterfactual sequence of events:

Absence of blush ⟶ scientist intervenes ⟶ Jane steals

Fischer's example is flawed for the following reason. The blush which occurs in the actual sequence, indicating to the neuroscientist that his intervention is unnecessary, must causally determine Jane's decision to steal (or be associated with some other factor that causally determines her decision to steal). Otherwise, it would still be possible for Jane to form an intention not to steal, even after the blush. Yet, if Jane's action in the actual sequence is determined, then the thought-experiment is question-begging. One of the questions at issue in the debate over retributive responsibility is whether retributive responsibility and determinism are compatible. A person who is not already persuaded that they are compatible will not have the intuition Fischer wants us to have, i.e. such a person will not find it plausible that, given the actual sequence of events (in which Jane's actions are determined), Jane is morally responsible in the retributive sense.[57]

In response, Fischer argues that his thought-experiment can still show that alternative possibilities are unnecessary for retributive responsibility, even though some readers

[57] Ginet makes the objection that such modified Frankfurt cases beg the question. See, "In Defence of Alternative Possibilities".

may not have the intuition that Jane is responsible in the actual sequence.[58] According to Fischer, the thought-experiment defeats PAP as long as readers have the following intuition: If there is some factor in this thought-experiment that rules out retributive responsibility, that factor cannot *merely* be the absence of alternative possibilities.

Fischer, argues that if the thought-experiment is analysed correctly, the above intuition can be elicited. Fischer claims that there are *two* factors which make it inevitable that Jane will steal and that these factors can be analysed separately. One factor is causal determinism in the actual sequence of events. The other factor is the counterfactual intervener. If one focuses only on the counterfactual intervener, one has the intuition that his potential intervention is irrelevant to whether Jane is morally responsible in the retributive sense. Yet, the counterfactual intervener is one factor that rules out alternative possibilities. Therefore, Fischer argues, *if* some people have the intuition that determinism in the actual sequence means Jane is not responsible in the retributive sense, their intuition cannot be based solely on the fact that determinism rules out alternative possibilities. [59]

However, Fisher's way of defending his thought-experiment is questionable. It is possible to challenge his claim that the counterfactual intervener is one of the factors which removes alternative possibilities. The intuitive force of the Frankfurt example depends on the counterfactual intervener *actually* removing the ability to do otherwise. But, given that the blush occurs (indicating that Jane's action is determined), there is no ability to do otherwise and hence no ability for the intervener to remove. There is only the potential for the intervener to remove alternative possibilities if, *counterfactually*, the blush does not occur. The only

[58] John Martin Fischer, *My Way: Essays on Moral Responsibility* (2006), pp199-200.
[59] *Ibid.*

feature of the thought-experiment that was meant to be counterfactual was the intervention, not the removal of alternative possibilities by the intervener. For the thought-experiment to prove its point Jane's alternative must *actually* be removed by the intervener. Otherwise, our intuition that the counterfactual intervener is irrelevant could be explained by the fact that he does not actually exclude alternative possibilities, rather than it being explained by the irrelevance of alternative possibilities.[60]

Even if Fischer could modify his thought experiment yet again in response to these criticisms it is doubtful whether thought experiments of this kind can generate intuitions powerful enough to outweigh the intuitive force of PAP. In their most heavily defended versions these Frankfurt-style cases become so convoluted that they baffle our intuitions. In their simpler forms, thought-experiments designed to counter PAP do not succeed in excluding alternative possibilities, are question-begging, or just amount to a slightly more colourful way of asserting that PAP is irrelevant, which cannot hope to convince those who are do not already lean towards the view that PAP is irrelevant to retributive responsibility.[61]

Revising PAP

Another approach a retributivist could adopt, in order to defend the thesis that retributive responsibility is compatible with determinism, does not involve denying the relevance of a person's ability to do otherwise. Instead, on this approach, the retributivist claims that the phrase "could have done otherwise" should be given a particular meaning – a meaning which is compatible with determinism. G.E. Moore argued that when we say a person "could have done otherwise" we mean that she

[60] A related objection to Fischer's thought experiment is discussed in Pereboom, "Reasons-Responsiveness".
[61] This kind of objection to Frankfurt-style cases is made by David Copp, "Defending the Principle of Alternate Possibilities: Blameworthiness and Moral Responsibility" (1997) 31 Nous 441.

could have done otherwise *if she had chosen otherwise*.[62] If this analysis were correct, it would follow that a person could have done otherwise even if her choice to do what she did were causally necessitated.

Writers such as C.A. Campbell have persuasively criticised this view. Campbell gives an example where one could meaningfully and reasonably ask whether a person could have done otherwise, but where the meaning of that question could not be given in terms of Moore's analysis:

"Take lying, for example. Only in some very abnormal situation could it occur to one to doubt whether A, whose power of speech was evinced by his telling a lie, was in a position to tell what he took to be the truth *if he had chosen*. Of course he was. Yet it still makes good sense for one's moral thinking to ask whether *A,* when lying, 'could have acted otherwise'….It seems apparent, therefore, that in this class of cases at any rate one does not mean by '*A* could have done otherwise', ' *A* could have acted otherwise *if* he had so chosen'."[63]

Michael McKenna discusses another example.[64] Imagine that Danielle is psychologically incapable of wanting to touch a blonde-haired dog. Unaware of her condition, her father shows her two Labrador puppies on her birthday – one black and one blonde. He asks her to pick up the one that she wants to keep. She picks up the black lab. Now, Moore's account implies that she *could have picked up the blonde dog instead*, because she would have done this if she had wanted to. But, this seems wrong, because her psychological condition made her incapable of wanting to touch the blonde dog.

---

[62] G.E. Moore, *Ethics* (1912).
[63] C.A. Campbell, "Is Free Will a Pseudo-Problem", *Free Will and Determinism*, in Bernard Berofsky (ed), *Free Will and Determinism* (1966), p112.
[64]Michael McKenna, 'Compatibilism' in Stanford Encyclopaedia of Philosophy (2009), available at http://plato.stanford.edu/entries/compatilism/. (Henceforth: McKenna, 'Compatibilism').

Moore himself recognised that it is often legitimate to ask not only whether a person could have acted otherwise if she had chosen, but, further, whether that person could have chosen differently. In connection with the latter point, he contended that when we ask whether a person could have chosen differently, we are simply asking whether she would have chosen differently if she had made a different prior choice. However, Bernard Berofsky points out that "the latter strategy…falls prey to an infinite regress argument…[for] it is possible to raise the question 'But could the person have chosen differently?' at any level."[65]

In recent years, a number of theorists (sometimes called 'the new dispositionalists'[66]) have attempted to produce more sophisticated compatibilist accounts of the capacity to do otherwise.[67] For instance, consider the following definition of 'capacity': An individual has a capacity to perform an action, if she possesses certain intrinsic properties (including properties of her brain) which would be (non-deviantly) causally operative in her performing the action if she chose (tried, decided or intended) to exercise this capacity and if the circumstances were favourable to the exercise of this capacity.[68]

Like Moore's account, this is a conditional analysis of capacity. It defines capacity in terms of what *would* happen *if* certain conditions obtained. It states that the ability to

[65] Bernard Berofsky, "Ifs, Cans and Free Will: The Issues", in Kane (ed), *The Oxford Handbook on Free Will* (2002), 181, at p182.
[66] This term is used by, e.g. Michael McKenna, 'Compatibilism'; and by Randolph Clarke, 'Dispositions, Abilities to Act and Free Will: The New Dispositionalism' (2008) 118 Mind 323-351. (Henceforth: Clarke, 'Dispositions').
[67] New dispositionalist accounts include: Michael Smith, 'Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion', in Stroud S and Tappolet C (eds), Weakness of Will and Practical Irrationality (OUP, New York 2003) 17; Michael Fara, Masked Abilities and Compatibilism (2008) 117 Mind, 843; Kadri Vihvelin, 'Free Will Demystified: A Dispositional Account' (2004) 32 Philosophical Topics 427. I will refer to this last article as: Vihvelin 'Free Will Demystified'.
[68] The account of capacity in the text is based on the account given in Vihvelin 'Free Will Demystified'. I have added the qualification that 'the circumstances must be favourable' in an attempt to take into account an objection raised in Clark, 'Dispositions'.

do otherwise is a matter of having certain dispositions to act under certain circumstances. The new dispositionalism improves on Moore's account because it attends to the causal base or underlying structure of the agent's dispositions. Consider Danielle's psychological condition again. New dispositionalists could explain her inability to do otherwise than choose the black lab by pointing out that she lacks the intrinsic causal brain properties necessary to perform the action of picking up the blonde dog.[69]

However, new dispositionalist accounts of capacity do have some very counterintuitive implications. For instance, they seem to prove too much when applied to agents like Jane in Frankfurt-style cases (discussed above). I argued that, intuitively, Jane could have *tried* to behave differently, or could have formed *different intentions* from the ones she in fact formed. However, according to the new dispositionalists, Jane could actually have *behaved* differently to how the neuroscientist wished her to behave. In the actual scenario, Jane's brain still possesses the intrinsic properties that would be causally operative in her action of walking past John's bicycle, without touching it *if* she wanted to act in that way, and *if* the evil neuroscientist were not present. That, according to the new dispositionalists, shows that she can, in the actual scenario, walk by and ignore the bicycle. But given that the neuroscientist is present and will definitely intervene and ensure that she takes the bicycle, if he senses that she might wish to refrain, it seems odd to think that she still has the ability to leave the bicycle alone.[70]

However, there is an even more serious problem with new dispositionalism. Whether or not one's intrinsic properties will be causally operative in performing an action on a particular occasion depends on whether one wanted or intended to perform that

[69] McKenna, 'Compatibilism', section 5.1.5.
[70] See McKenna, 'Compatibilism', section 5.1.5.

action on that occasion. However, if determinism is true, the fact that one has or lacks that desire or intention on that occasion is the *inevitable* result of factors beyond one's control. Given the actual absence of such a desire or intention, it is impossible for the action to take place on that occasion. At most, the new dispositionalists succeed in describing what it means to have a general capacity to do something, e.g. the capacity to swim. They do not provide a convincing account of what it means to be able to exercise a capacity on a particular occasion.[71] But the retributive justification for punishing the wrong-doer for her specific wrongful action, is based on the idea that she could have refrained from that action *on the occasion.*

To summarise the argument that this chapter has presented so far: The principle that a person is not retributively responsible for a wrongful action unless she has the ability to do otherwise (PAP) has powerful intuitive force. Thought-experiments that are designed to refute this principle are unsuccessful and revised versions of PAP are inadequate. If determinism is true, no-one has the ability to do otherwise than they do. Therefore, if true, determinism would seriously undermine the retributive justification of punishment. Not only would determinism exclude alternative possibilities, but it also entails that the agent is not the ultimate source of her action. In the next section, it will be argued that this provides a further reason for thinking that retributive responsibility and determinism are incompatible.

Source Incompatibilism

Theorists like Frankfurt, and Victor Tadros invite us to focus on the desires that *actually* motivated the agent and to hold her responsible on that basis. In response to this approach, it is submitted that an agent's desires/values cannot provide an

[71] See Clark, 'Dispositions', pp338-339.

adequate basis for retributive responsibility unless the agent is the *ultimate* source of those values/desires.[72] If determinism is true, then, ultimately, factors beyond the agent's control fully account for her having the values/desires that she has. This "source argument" provides a further reason for holding that determinism and retributive moral responsibility are incompatible. The first part of this section on source incompatibilism will demonstrate the intuitive plausibility of the source argument. In the second part of this section, it will be argued that the fact that an agent whose actions are determined is not the ultimate source of her values, desires and choices means that, in an important sense, these values, desires and choices are products of luck and things for which she cannot be held retributively responsible.

A Manipulation Case

Like Frankfurt, Tadros argues that the ability to do otherwise is irrelevant to responsibility.[73] Tadros argues that, if an agent acts from a desire, she can be held responsible for her action as long as her desire to perform the action reflects on her *qua* agent. According to Tadros, in order for a desire to reflect on the agent *qua* agent, it must be accepted by her in the light of her value system. Tadros claims that it is not reasonable to ask whether this value system reflects on the agent, because a value system is constitutive of agency.

It is submitted that a retributivist could not rely on this kind of account. Imagine the following scenario: An artificial intelligence engineer of the future builds a conscious robot, called Professor Plum, with human-like powers of deliberation and action. The engineer programs this robot with a warped value system. Professor Plum perpetrates wrongful acts which he accepts in the light of his value system and

[72] Writers who accept the 'source argument' include Pereboom, *Living without Free Will* (2001); Saul Smilansky, "Free Will, Fundamental Dualism and the Centrality of Illusion" in Kane (ed), *The Oxford Handbook of Free Will* (2002) 489.
[73] Victor Tadros, *Criminal Responsibility* (2005), pp31-43, pp69-70.

which would therefore (according to Tadros) reflect on him *qua* agent. Given the full explanation of Professor Plum's behaviour, it seems highly counter-intuitive to suppose that it would be intrinsically good to inflict suffering on him only because of the nature of his actions (leaving aside consequentialist considerations). Yet, it is hard to see any difference between Professor Plum and an ordinary person (who is subject to causal determinism), which could exempt the robot from retributive responsibility, but would justify holding the ordinary person morally responsible in the retributive sense for her wrongful actions.[74]

It could be pointed out that, unlike ordinary people, the robot was programmed by a rational agent - the artificial intelligence engineer. It could be stipulated that, if a rational agent intentionally determines someone else's actions, the former agent is morally responsible for those actions, not the latter. In the normal case, it could be argued, responsibility lies with the rational agent who actually performs the action, as there is no other agent in the causal chain who could take the blame. But this would be an *ad hoc* manoevre. The example could easily be modified in response to this criticism. We could suppose that Professor Plum's creator is mad and so is not morally responsible, or that she created the Professor Plum by accident. It would still be counterintuitive to hold the robot morally responsible in the retributive sense.

Fischer responds to a similar thought-experiment by saying that the robot is indeed morally responsible, but is not blameworthy.[75] However, this kind of argument cannot help the defender of retributive responsibility. This is because the statement that X is morally responsible *in the retributive sense* for performing a wrongful action entails that X is blameworthy for it. As Pereboom puts it:

[74] This is based on Derk Pereboom's 'Professor Plum' thought-experiment in *Living Without Free Will* (2001), pp110-125.
[75] John Martin Fischer, *My Way: Essays on Moral Responsibility* (2006), pp230-234.

> '…for an agent to be morally responsible for an action in the sense at issue is for it to belong to him in such a way that he would deserve blame [or punishment] if he understood that it was morally wrong, and he would deserve credit or perhaps praise if he understood that it was morally exemplary, supposing that this desert is basic in the sense that the agent would deserve the [blame/punishment] or credit just because he has performed the action (given understanding of its moral status), and not by virtue of consequentialist considerations.'[76]

Fischer's response is especially puzzling as, in other passages of his work, he seems to assume that an agent who is morally responsible for a wrongful action is also blameworthy for it.[77] He does not give another example, apart from the case of the robot, where an agent is morally responsible for a wrongful action and yet is not blameworthy for it and he does not say precisely which features of the robot's situation should exempt him from blame/punishment. (In the remainder of this thesis it will be assumed that Fischer's main view is that an agent who commits a wrongful action is also blameworthy for it and that Fischer's position on the robot case is an aberration.)[78]

The Problem of 'Ultimate Luck'

Some retributivists might be willing to embrace the conclusion that it is intrinsically good to blame and punish the robot. Such retributivists might demand that the intuition that the robot is not responsible in the retributive sense stands in need of further justification. An answer to this objection is that the robot had the *misfortune* to be programmed with a warped value system, by an evil scientist. If he had been lucky enough to have been programmed with a virtuous value-system, by a good

[76] Pereboom, "Reasons-Responsiveness", p210.
[77] E.g., John Martin Fischer, "The Cards that are Dealt You"(2006) 10 The Journal of Ethics 107
[78] For further discussion of Fischer's response to the robot case, see Derk Pereboom, "Reasons-Responsiveness".

scientist, then he would not have committed the wrongful acts. It seems arbitrary for a moral judgement about the intrinsic goodness of punishing a particular offender to be based on something which is wholly fortuitous and outwith the offender's control. If determinism is true, ordinary human beings are relevantly analogous to the robot, because the fact that we have one value-system rather than another is also ultimately a matter of luck. If determinism is true, these arbitrarily given factors inevitably result in our behaviour, so that all our behaviour was "in the cards" before we were born.[79]

Although Fischer accepts that luck plays this fundamental role in human behaviour, he denies that this is relevant to moral responsibility. He points out that innumerable conditions which are *necessary* for an agent's behaviour are outwith the control of the agent and are therefore matters of luck. For example, it is necessary for the performance of any particular action that an asteroid does not hit earth, that the air remains breathable, that the agent is not struck by a bolt of lightning etc and the agent cannot control these things. Yet, Fischer argues, these facts are obviously completely irrelevant to the question of moral responsibility. He then asks: If it does not matter to us that there are an infinite number of necessary conditions for our behaviour that are wholly outwith our control, why should it matter that there is a set of *sufficient* conditions for our behaviour which is wholly outwith our control? He concludes: "Our behaviour may well be 'in the cards' in the sense that we simply have to play the cards that are dealt us… Yet we can still be responsible for playing the cards that are dealt us, even if we did not manufacture the cards, write the rules of the game and so forth."[80]

[79] Derk Pereboom, *Living without Free Will,* (2001) p6.
[80] Fischer, "The Cards that are Dealt You"(2006) 10 The Journal of Ethics 107, at p 128.

In reply, it should be acknowledged that the agent does not need to be in control of *all* the conditions which are necessary for the performance of an action in order for her to be morally responsible for that action; but it does not follow from this that the agent does not need to be in control of *any* of these conditions. The agent does not need to be in control of factors which merely enable the performance of her action, e.g. the presence of breathable air. After all, such factors enable the performance of a whole range of actions. However, the *act of will* that brings about the *particular* action at issue seems to be qualitatively different from background conditions such as the presence of breathable air. If the agent was not ultimately in control of this act of will (because it was causally determined by factors outwith her control), it seems unjustifiable to hold her retributively responsible for the action which is brought about by the act of will. To come back to the card-playing analogy, the fact that the agent did not manufacture the cards is indeed irrelevant to whether she is morally responsible for the way she plays them. Yet, it still seems reasonable to maintain that she must have been be ultimately in control of the *moves she made* before she can be held retributively responsible for these.

To summarise: for the retributivist, punishment is justified purely because of the nature of the agent's deed and not even partly because of factors outside the agent (such as the need to protect society). This kind of ultimate responsibility seems to require that the agent had ultimate control of the action. The retributivist focuses narrowly on the corrupt values that motivated the action (or on the wrongful choice of that action) and demands that the agent must "pay" for this. But if these values (or this choice) were determined by factors outwith the agent's control, the retributivist is basically demanding that the agent "pays" for her bad luck.[81]

[81] See Saul Smilansky, "Compatibilism: The Argument from Shallowness", (2003) 115 Philosophical Studies 257, at 268

‘Owning’ Actions: A Reply to the Manipulation Challenge

I will now attempt to rebut an influential compatibilist response to examples like the manipulation case discussed above. According to this compatibilist line of argument, manipulated beings appear to lack free will because they do not truly ‘own’ their actions in the way that ordinary human beings do.

Before discussing this issue further, it is important to note that examples similar to the manipulation thought-experiment can be found outside of science-fiction. There are real-life cases where people’s behaviour seems clearly to be caused by factors outwith their control, even though they are not ‘insane’ in the traditional sense. Such individuals seem to be unfit candidates for retributive responsibility and punishment. For instance, brain dysfunction due to head injury or disease can radically alter an individual’s character traits and motivations.[82]

Or consider the following case:[83] During the year two thousand, a forty year old school teacher from Virginia appeared to undergo a disturbing personality change. He began to engage in highly inappropriate sexual behaviour, culminating in making advances towards his twelve year old step daughter. He was convicted of child molestation. Given that this was his first offence he was sentenced to a diversion programme. He failed this programme because he repeatedly propositioned staff and

[82] See e.g., C Grady, ‘Neuroimaging and Activation of the Frontal Lobes’ in B Miller and J Cummings (eds), The Human Frontal Lobes: Functions and Disorders (The Guilford Press, London 1999), discussing the famous case of Phineas Gage. Frontal lobe damage can also be caused by vascular disease, see, e.g., H Chui and L Willis, ‘Vascular Diseases of the Frontal Lobes’ in Miller and Cummings (above). Antonio Damasio has done extensive work on the personality changes that can be caused by frontal lobe damage. See, e.g., A Damasio, *Descarte’s Error: Emotion, Reason and the Human Brain* (Putnam, New York 1994)

[83] The case was reported in J Burns et al, ‘Right Orbitofrontal Tumor with Pedophilia Symptom and Constructional Apraxia Sign’ 60 (2003) Archives of Neurology 437. For discussion of this case by legal commentators see, e.g., H Greely, ‘Law and the Revolution in Neuroscience: An Early Look at the Field’ 42 (2009) Akron Law Review 687; J Seiden, ‘The Criminal Brain: Frontal Lobe Dysfunction Evidence in Capital Proceedings’ 16 (2004) Capital Defense Journal 395.

other patients at the rehabilitation centre. He was then admitted to hospital after complaining of severe headaches. It was discovered that he had a brain tumour the size of an egg. It seems that the tumour was interfering with the functioning of his frontal lobes – an area of the brain associated with regulating socially appropriate behaviour. The tumour was removed and his deviant behaviour ceased. A year later, however, his headaches returned and he began once again to engage in inappropriate behaviour. The tumour had re-grown. As before, the tumour was removed and the deviant behaviour ceased. His doctors concluded that the tumour both caused his deviant sexual inclinations and caused his failure to refrain from acting on those inclinations.

Precisely how the tumour caused his failure to resist his impulses is still somewhat unclear. It seems it did not do so by altering his understanding of what he was doing or his knowledge of the wrongfulness of his conduct. Indeed he told doctors that he was aware that his actions were both illegal and immoral and he went to considerable lengths to try to conceal some of his unlawful behaviour. The neurologists who examined him concluded that the brain tumour made him incapable of resisting his urges.[84] However, it is far from clear that the impulses of all people with frontal lobe dysfunction are strictly speaking irresistible, in the sense that there is no possible incentive that would induce the individual to resist them. However, damage to the brain might *determine which incentives an individual will act upon*. It is certainly possible to imagine a case where all the psychological states that are relevant to bringing about the agent's action arise from a source (such as a brain tumour) that is outwith the agent's control – a source which is not a creation of the individual's agency, but which is 'alien' to her. Provided that these alien motivations cause the agent's behaviour in a deterministic fashion, i.e. they render it inevitable that the

[84] J Burns et al, ibid, p440.

agent behaves as she does, given the situation she is in, then it seems unfair to hold her retributively responsible for her behaviour. This presents a challenge to compatibilists. Either they must insist (counterintuitively) that such agents deserve retributive punishment, or alternatively they must point to some relevant difference between behaviour that is determined by ordinary factors (e.g. the agent's genes and environment) and determination by the kind of alien source considered above.

John Martin Fischer and Mark Ravizza argue that motivations arising from sources such as brain tumours do not belong to the agent because the agent has not *made them her own.* Their theory of responsibility has had a huge influence on the current free will literature.[85] According to Fischer and Ravizza, there are two prerequisites for retributive responsibility. The first prerequisite states that a person's action must result from a 'reasons-responsive mechanism'. The second prerequisite states that the agent must 'own' the relevant mechanisms. On Fischer and Ravizza's account, the particular mechanism (or thought/brain processes) that plays a role in producing any particular action need only be *moderately* responsive to reasons, in order for the agent to be held responsible for that specific action. There are two criteria for moderate reasons-responsiveness – moderate receptivity and weak reactivity. Moderate receptivity to reason does not require that an agent recognises *all* the reasons that exist for and against a particular action. It only requires that the agent recognises a pattern of reasons that is intuitively rational. The agent must also rank the relative strength of different reasons in an objectively rational way, given her value system and preferences. For instance if the agent considers a reward of £100 to be a sufficient incentive for performing a particular action then, intuitively, if she is rational she will also recognise that rewards of £200 and £300 are also sufficient incentives for performing that action. The agent must recognise the existence of

[85] Fischer and Ravizza, note 2 (above).

some *moral* reasons – the agent cannot be a psychopath. Weak reactivity to reason means that there is at least *one* possible incentive which, if it were available to the agent, would cause the agent's mechanism to react differently from the way in which the mechanism actually reacts. Fischer and Ravizza include the weak reactivity requirement to ensure that agents who suffer from irresistible impulses are not held responsible. If an impulse is truly irresistible, they claim, there will be no possible incentive that would induce the agent to act differently. They do not require that the agent reacts to every reason that she recognises as a sufficient reason for action, because they do not wish weak willed agents to be excused. Weak willed agents may be able to *recognise* a large number of reasons for action, but fail to *act* on them due to weakness of will. (Of course, it might be wondered why Fischer and Ravizza stipulate that the fact that a mechanism reacts to *one* reason for acting differently is sufficient for responsibility. Why not two, or three, or more? This is an area of their theory that has provoked much debate.)

If an agent fulfils this first requirement then, according to Fischer and Ravizza, her action is not the product of an irresistible impulse. The problem, however, is that a reasons-responsive mechanism could have arisen from a disease process or could have been implanted by a manipulator (the problem of 'alien mechanisms').[86] In response to this problem, Fischer and Ravizza developed their second prerequisite for responsibility - the agent must *take* responsibility for her mechanisms.[87] Taking responsibility, they claim, involves the agent adopting a certain subjective stance toward her own conduct. The agent must, they argue, view herself as an appropriate target for the reactive attitudes and must recognise that her actions have causal effects on the world. They claim that agents typically come to adopt this view of

[86] The terminology of 'alien mechanisms' is drawn from: M McKenna, 'Assessing Reasons - Responsive Compatibilism' 8 (1) (2000) Internatonal Journal of Philosophical Studies 89.
[87] Ibid, chapters 7 and 8.

themselves as they progress from childhood to adulthood. According to Fischer and Ravizza, coming to view one's actions in this way during this developmental stage makes one's subsequent motivations (and the actions that flow from them) one's own. However, (as discussed in detail below) they argue that agents do *not* make actions that flow from alien mechanisms their own.

Fischer and Ravizza claim that there are independent reasons for adopting their 'taking responsibility requirement', quite apart from its (supposed) advantages in dealing with the alien mechanisms problem. They claim that it captures important intuitions about the fundamental nature of responsibility. They write:

> 'A theory of moral responsibility is supposed to give expression to (and more concrete content to) our inchoate, intuitive conceptions of ourselves as active and in control; and it is highly plausible to think that our having a certain sort of *view of ourselves* is required in order for us to be active and in control.'[88]

They say that a person who fails to 'take responsibility' (in their sense of the term) is like a sailor on a boat with no rudder, simply tossed in different directions by the changing winds. [89]

This section will argue that Fischer and Ravizza's theory fails to provide sufficient conditions for retributive responsibility and should be rejected. Firstly, any intuitive appeal their idea of 'taking responsibility' has as a precondition for being responsible is lost when they try to refine it in order to deal with the alien mechanism challenge. Secondly, their theory when taken to its logical conclusion produces absurd results. Thirdly, even if their counterintuitive interpretation of 'taking responsibility' were accepted it would not be able to deal with all cases of alien mechanisms. Finally,

[88] Ibid, p223
[89] Ibid, chapter 8.

they fail to point to a relevant difference between actions which are determined by ordinary factors (e.g. our genes and environment) and actions that arise from sources such as brain tumours or manipulation.

What Does an Agent Take Responsibility For?

Fischer and Ravizza's first task is to explain what the agent takes responsibility for. They claim that agents only take responsibility for behaviour that comes about in a certain way – or, in their terminology, behaviour that arises from a certain kind of 'mechanism'. For instance, an epileptic may take responsibility for behaviour that results from his desires, beliefs and intentions etc. (his ordinary 'mechanisms' of practical reasoning) but not for behaviour produced by epileptic seizures.[90] The agent himself need not consciously think about these matters in exactly these terms, but Fischer and Ravizza's account (at this stage) captures something that seems to be *implicit* in ordinary thought – that when an agent 'takes responsibility', he is not 'accepting responsibility for all his actions *whatever* their source'.[91] According to Fischer and Ravizza, when an agent, at a particular time, comes to take responsibility for behaviour that flows from a certain type of mechanism, he thereby takes responsibility for his future behaviour that results from the *same kind* of source. They write:

> 'Having taken responsibility for behaviour that issues from a kind of mechanism, it is almost as if the agent has some sort of "standing policy" with respect to that kind of mechanism. Thus when the agent subsequently acts from a mechanism of that kind, that mechanism is *his own* insofar as he has already taken responsibility for acting from that kind of mechanism.'[92]

[90] Ibid, p215.
[91] Ibid.
[92] Ibid, p215.

The problem arises when Fischer and Ravizza try to explain what makes a mechanism belong to one 'kind' rather than another. They do not simply maintain that actions which flow from psychological states like 'desires', 'beliefs' and 'intentions' arise from one type of mechanism and actions that have nothing to do with such psychological states (such as epileptic seizures) belong in a different category. If they settled for this simple account then it would not help them to deal with the 'alien mechanisms' challenge. For instance, a neuroscientist could cause an agent to act in a certain way by implanting all the psychological states (desires, beliefs etc) that are sufficient to bring about the agent's behaviour. However, Fischer and Ravizza would not want to say that the agent's 'standing policy' of taking responsibility for his practical reasoning covers stretches of practical reasoning that were implanted by a neuroscientist. To meet this challenge, Fischer and Ravizza stipulate that when an agent takes responsibility for a mechanism underlying his action, he takes responsibility for the neurological details of that mechanism and the origins of those neurological details. Thus he does *not* take responsibility for subsequent acts that result from mechanisms with *different* neurological details (e.g. motivations that are entirely caused by brain tumours) or where the neurological details have a different causal history (e.g. they were implanted by a neuroscientist). They write:

'...in taking responsibility for this mechanism, we take responsibility for all its details (even if we are unaware of them): we take responsibility for the mechanism in its full reality. If causal determinism is true, our mechanisms of practical reasoning have always been deterministic; thus, in taking responsibility for ordinary practical reasoning, we take responsibility for its deterministic character (just as we take responsibility for its neural details.) We take responsibility for the entire iceberg, in virtue of seeing its tip. But in taking responsibility for ordinary practical reasoning, [the agent] does *not* thereby take responsibility for a *different kind of mechanism* –

one that involves direct stimulation of brain (or hypnosis, subliminal advertising, and so forth).'[93]

Fischer and Ravizza's interpretation of the notion of 'taking responsibility' certainly provides them with a response to the alien mechanisms challenge. But it does so at a very high price. Their interpretation has become so far removed from normal uses of the phrase 'taking responsibility' that it is hard to think of any reason why we should adopt their interpretation – any reason, that is, apart from a desire to rescue their theory. It is highly counterintuitive to insist that ordinary agents typically 'take responsibility' for complex neurological mechanisms that most of them *do not even know about*. Indeed, many of these agents may take responsibility for their motivations on the basis of assumptions about the fundamental 'reality' of these mechanisms that (if determinism is true) are *profoundly mistaken.* For instance, many agents may assume that their motivations are not the inevitable product of factors wholly outwith their control. Are we to think that when an agent takes responsibility for her motivations partly in virtue of the assumption that they are *not* entirely determined by factors outwith her control, she *thereby* 'takes responsibility' for mechanisms that *are* entirely determined by factors outwith her control? This seems to stretch the ordinary concept of 'taking responsibility' implausibly far; and the further Fischer and Ravizza depart from this ordinary concept, the less they are entitled to rely on the original intuitive picture of 'taking responsibility' which made it appealing to adopt this aspect of their theory in the first place.

Fischer and Ravizza might reply that although their interpretation of 'taking responsibility' is more technical than the ordinary way that notion is used, their account is still based on widespread intuitions, but just makes those intuitive ideas more 'precise'. They write: 'Of course, the non-philosopher would not characterize

[93] Ibid, p234.

his taking responsibility in quite [the way our theory characterizes it], but this theoretical characterization merely makes more precise the intuitive idea that one takes responsibility for actions that spring from certain sources (and not from others).'[94] In response, there is a very big difference between making an everyday intuitive idea 'more precise' and simply contradicting it. Fischer and Ravizza's account jars with the ordinary idea of taking responsibility. Furthermore, if they are to modify the notion of taking responsibility in the way they propose, then they need to justify this modification, by pointing to some *independent* reason for it (besides the fact that it helps them to meet a certain objection from their critics). Otherwise, their modification seems *ad hoc.*

As noted above, it is genuinely intuitive to think that agents take responsibility for behaviour that springs from certain sources and not from others; and it is not implausible to think that once an agent has come to take responsibility for some of his behaviour and the desires, beliefs etc that motivated it, his act of taking ownership also extends to *relevantly similar* motivations underlying his subsequent actions in such a way that those motivations also count as the agent's own. But which similarities should count as relevant? Fischer and Ravizza stipulate that it is crucially relevant whether the agent's psychological states have the same kind of neurological underpinnings. This is implausible. Consider the following example. An agent engages in some practical reasoning that feels subjectively exactly the same as deliberations that she engaged in previously (and for which she had previously taken responsibility). Her current deliberations were not the result of a mechanism implanted by a manipulator, nor were they caused by disease. However, it happens that this current stretch of practical reasoning involved totally different neural pathways from the pathways that were activated during her previous deliberations.

[94] Ibid, p215.

Fischer and Ravizza's theory seems to commit them to the view that the agent's act of taking responsibility for her earlier deliberations does not transfer to the later stretch of practical reasoning (in virtue of the difference in the underlying neural mechanisms). This seems arbitrary.

In What Sense of 'Responsibility' Do Agents 'Take Responsibility'?

According to Fischer and Ravizza, taking responsibility has the following features –

A: The agent must realise that her actions have a causal effect on the world.

B: The agent must see herself as an appropriate target for the reactive attitudes.

The first element of this account is clearly a necessary condition for taking responsibility (and for *being* responsible). Someone who genuinely cannot understand that her actions have practical consequences is so cut off from reality that she does not qualify as a moral agent. However, this element does not help Fischer and Ravizza deal with alien mechanism cases. An agent's behaviour can still have causal effects on the world (of which the agent can be aware) even if that behaviour is due to an alien mechanism.

Fischer and Ravizza rely on the second element of their account to deal with such cases. One problem with their approach stems from ambiguity surrounding the term 'responsibility'. Several theorists have argued that the term 'responsibility' can be used in retributive and non-retributive senses.[95] To say that someone is responsible in the retributive sense for a wrongful action implies that she deserves to be blamed/punished for it regardless of the consequences. In contrast, there are non-retributive (e.g. consequentialist) senses of responsibility. For instance, Derk Pereboom writes:

[95] E.g., T Honderich, *On Determinism and Freedom* (Edinburgh University Press, Edinburgh 2005); G Strawson, *Freedom and Belief* (Clarendon Press, Oxford 1986); D Pereboom, *Living without Free Will* (CUP, Cambridge 2001).

> '...[T]here are [non-retributive] senses of 'moral responsibility'. One might say that an agent is morally responsible when it is legitimate to demand of her that she explain how her decisions accord with morality...Making these demands of agents might be justified by its effectiveness in improving the agent morally—we humans are indeed susceptible to causal influence by challenge and counsel of this kind....[I]ncompatibilists would not find our being morally responsible in this sense to be even *prima facie* incompatible with determinism. The notion that incompatibilists do claim to be incompatible with determinism is rather the one defined in terms of basic desert [i.e. retributive responsibility].'[96]

Fischer and Ravizza's reference to the 'reactive attitudes' suggest that their notion of 'taking responsibility' involves the agent seeing herself as an appropriate candidate for attributions of *retributive* responsibility. PF Strawson, who first developed a form of compatibilism based on the reactive attitudes, advanced his view partly in response to non-retributive, forward-looking accounts of responsibility and punishment. Strawson observed that 'Some optimists about determinism point to the efficacy of the practices of punishment, and of moral condemnation and approval, in regulating behaviour in socially desirable ways.' He condemned this approach for 'over-intellectualizing the facts' in a manner characteristic of 'a one-eyed utilitarianism'.[97] He argued that our practices of responsibility and punishment are the expression of certain attitudes which themselves stand in need of no further justification. These attitudes seem to be essentially retributive. According to Strawson:

> 'Indignation and disapprobation, like resentment, tend to inhibit or at least to limit our goodwill towards the object of these attitudes, tend to promote an at least partial and

[96] D Pereboom, 'Reasons-Responsiveness, Alternative Possibilities and Manipulation Arguments Against Compatibilism: Reflections on John Martin Fischer's My Way' (2006) 47 Philosophical Books 198, p14.

[97] P Strawson, 'Freedom and Resentment' in G Watson (ed) *Free Will* (OUP, Oxford 2003), p73.

> temporary withdrawal of goodwill; they do so in proportion as they are strong; and their strength is in general proportioned to what is felt to be the magnitude of the injury and the degree to which the agent's will is identified with, or indifferent to, it.... The partial withdrawal of goodwill which these attitudes entail, the modification they entail of the general demand that another should, if possible, be spared suffering, is... the consequence of continuing to view [the wrongdoer] as a member of the moral community; only as one who has offended against its demands. So 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.'[98]

Now, it may be that many people do view themselves as appropriate targets for the reactive attitudes (interpreted as essentially retributive). In other words, many people may think that if they were to commit a crime it would be appropriate for the rest of the moral community to react by expressing attitudes of indignation and resentment and by inflicting the suffering which such attitudes 'entail' must be inflicted. They may also consider it appropriate for the offender to experience the first-person reactive attitude of guilt, together with the willingness to suffer which that emotion supposedly entails. Many people may feel that the appropriateness of all this is not contingent on whether expressing these attitudes and inflicting such suffering is likely to produce any further good consequences. However, many other people do *not* seem to view matters in quite this way. Those who take a more forward-looking, non-retributive approach to the issues of responsibility and punishment apparently do not 'take responsibility' for their behaviour in the sense of viewing themselves as appropriate targets for the (retributive) reactive attitudes. Taking responsibility in this retributive sense seems to be, according to Fischer and Ravizza, a precondition for being responsible. But surely Fischer and Ravizza cannot allow agents to evade retributive responsibility merely because the agents are not themselves retributivists.

[98] Ibid, p90.

Fischer and Ravizza's approach is flawed, because they fail to recognise, at this point in their argument, that there are different senses in which agents can see themselves as 'responsible'. They simply assume that there is only one relevant sense of responsibility, that either an agent views herself as responsible, or she does not, and that very few agents will fail to view themselves as responsible. It is plausible, they claim, that the vast majority of agents regard themselves as responsible, because the consequences of failing to do so are dire. True, agents who *view* themselves as non-responsible may escape being *held* responsible (at least if Fischer and Ravizza's ideas were put into practice). But, Fischer and Ravizza claim, such agents would be locked up on the basis that the way in which they view themselves makes them dangerous to society. They write:

> '...there are strong incentives not to opt out of moral responsibility. Agents who genuinely fail to take responsibility – and thus view themselves as lacking control – are legitimately sequestered from society, and are deprived of the opportunity to participate in the moral community.'[99]

Once we recognise that there are both retributive and non-retributive ways in which a person can view herself as responsible, it seems that Fischer and Ravizza face considerable difficulties. If they persist with the idea that a person must see herself as an appropriate target for the retributive reactive attitudes, then it can be objected that there may well be many people who do not view themselves in this way. Fischer and Ravizza cannot get round this problem by saying that such people should automatically be 'sequestered' from society. Otherwise a large number of philosophers (and some non-philosophers) with anti-retributive leanings would end up being sequestered!

[99] Fischer and Ravizza (fn2, above), P229.

Alfred Mele has raised a related point.[100] He describes a hypothetical philosopher, Phil, who is a committed hard determinist. Mele correctly notes that Fischer and Ravizza's theory implies that Phil is not responsible. Mele also notes that real-life hard incompatibilists (e.g. Derk Pereboom, Ted Honderich and Galen Strawson) will also, in virtue of their philosophical beliefs, be relieved from responsibility according to Fischer and Ravizza's theory (assuming of course that these hard incompatibilists are not secretly compatibilists in their personal lives). My argument differs from Mele's in two significant respects. Firstly, Mele's critique seems to assume a unitary concept of responsibility and, accordingly, states that philosophers like Phil do not consider themselves to be responsible. In contrast, I argue that such philosophers may in fact view themselves as responsible in a non-retributive sense. Secondly, my argument is not restricted to the minority of philosophers who are firmly committed to hard determinism. Rather, it extends to all those individuals who reject retributive attitudes for *whatever reason*. My argument therefore has much wider application. In a reply to Mele, Fischer has suggested that Mele's objection is not decisive because it only applies to a very small number of cases. It is harder to dismiss my argument in this way. Furthermore, Fischer's response to Mele's critique is itself dubious. Fischer writes:

> 'My general methodological disposition is to seek to capture the clear cases by appealing and intuitively natural principles, but to admit that these principles may well have jarring consequences in certain cases.... After all, the phenomena of moral responsibility are themselves messy around the edges, and it would be unreasonable to suppose that a largely successful and plausible approach would yield entirely comfortable results along all its perimeters.'[101]

[100] A Mele, 'Fischer and Ravizza on Moral Responsibility', 10 (2006) The Journal of Ethics 283; A Mele, 'Reactive Attitudes, Reactivity, and Omissions'61 (2) (2000) Philosophy and Phenomenological Research 447.
[101] J Fischer, 'The Free Will Revolution (Continued)' 10 (2006) The Journal of Ethics 315, p326.

Fischer raises a very important point here. Life is complicated and messy. We have good reason to be suspicious of any moral theory which claims that all these complications can be neatly resolved. Moral theories that attempt conclusively to 'tidy up' all the difficult cases that have been worrying people for centuries risk discarding important aspects of our moral experience. However, it is submitted that theorists should try to avoid generating *completely new* areas of mess. Prior to the development of Fischer and Ravizza's theory, people did not worry about whether hard incompatibilists like Galen Strawson should be completely relieved from responsibility (or, alternatively, sequestered to protect society) merely because of their philosophical views. People did not worry about this because it is obvious that the fact that someone subscribes to hard incompatibilism tells us nothing about whether that person is in fact responsible (or dangerous). A theory that implies the opposite has a serious flaw. It does not merely fail to yield 'entirely comfortable' results. It yields absurd results.

Fischer and Ravizza might try to avoid these problems by revising their position. They might allow that viewing oneself as responsible in a non-retributive (e.g. consequentialist) sense will suffice for being responsible. However, this manoeuvre would be extremely paradoxical. For Fischer and Ravizza ultimately aim to show that the kind of responsibility which can provide a basis for retributivism is compatible with determinism. It would be very odd to stipulate that viewing oneself as responsible in a non-retributive sense is a necessary condition for being responsible in the retributive sense. This seems like a *non sequitur*.

Furthermore, it should be remembered that one of the supposed merits of the notion of 'taking responsibility' was that it provided an answer to the 'alien mechanisms' challenge. Fischer and Ravizza claimed that agents do not take responsibility for

behaviour that is predetermined by 'alien' influences (e.g. hypnosis, brain manipulation etc). However, if they broadened their definition of 'taking responsibility' to cover consequentialist conceptions of responsibility, then certain agents with consequentialist leanings might judge themselves to be responsible even for behaviour that arose from alien mechanisms. This is because holding a wrong-doer responsible may help to produce good consequences, even if the wrong-doer's behaviour was due to an alien mechanism. Therefore, broadening the definition of 'taking responsibility' would not allow Fischer and Ravizza to deal with the alien mechanisms challenge (even if it would help them to overcome some of the other objections mentioned above). The alien mechanisms challenge is one that Fischer and Ravizza need to deal with because it is highly counterintuitive to say that an agent is morally deserving of retributive punishment if the agent's behaviour was entirely attributable to an alien mechanism; and Fischer and Ravizza wish ultimately to defend a retributive conception of moral desert.

Taking Responsibility for Alien Mechanisms

Imagine that the alien mechanism itself causes the agent to view himself as responsible (in the retributive sense) for actions that flow from that mechanism. Has this agent satisfied Fischer and Ravizza's requirements for being responsible (assuming that the alien mechanism is also reasons-responsive)? Fischer and Ravizza maintain that he has not. This is because a further element of their account has not been fulfilled - the agent's view of himself must, they maintain, be based 'on his evidence in an appropriate way'.[102] They assert that an agent whose view of himself is induced by the alien mechanism (e.g. it is caused by the brain tumour, or by hypnotic suggestion, or by a neuroscientist electrically stimulating his brain) does not form his beliefs in an 'appropriate' way. They state that 'the relevant notion of

[102] Fischer and Ravizza (fn 2 above), P236.

appropriateness must remain unanalysed'.[103] They tentatively suggest, however, that a belief is only formed in an appropriate way if it is produced by a mechanism that tracks truth. 'This implies (among other things) that, holding fixed the actual mechanism of belief production, the agent would not believe *P* if *P* were false'. [104]

One problem with this approach is that it is question-begging. Fischer and Ravizza maintain that the vast majority of agents who view themselves as retributively responsible and whose behaviour is determined by 'ordinary' factors (and not by alien mechanisms) are indeed retributively responsible. According to Fischer and Ravizza, such agents are retributively responsible partly because their view of themselves is based in an appropriate way on the evidence. But this assumes that their view of themselves tracks truth – i.e. it assumes that agents can in fact be retributively responsible for determined behaviour. This is the very question at issue in the free will debate.[105]

Another problem concerns an agent who believes (in a way not caused by the alien mechanism itself) that she would be retributively responsible even if her action is entirely determined by an alien mechanism. Imagine, for instance, that the agent is convinced by the arguments of Harry Frankfurt who holds precisely that position.[106] Fischer and Ravizza claim that such an agent might well be responsible on this basis. This seems highly counterintuitive. For it is plausible to think that this agent is simply mistaken about her responsibility status.

---

[103] Ibid.
[104] Ibid, p237.
[105] Fischer and Ravizza briefly mention another possible option (which they do not elaborate on). The truth-tracking requirement might be watered down. An agent, they claim, might still form a belief in an appropriate way even if she would still hold that belief in certain situations in which the belief was false. In response, it may be possible to envisage an alien mechanism that generates beliefs in this manner.
[106] H Frankfurt, 'Reply to John Martin Fischer,' in S Buss and L Overton (eds), *Contours of Agency: Essays on Themes from Harry Frankfurt*, (MIT Press, Massachusetts 2002).

What is Special about Alien Mechanisms?

It is useful at this point to recall exactly what the alien mechanism challenge amounts to. The argument goes as follows:

1) If the motivational states that bring about the agent's wrongful behaviour arise from an intuitively 'alien' source (e.g. they have been implanted by a neuroscientist) then the agent does not deserve retributive punishment, provided that the implanted motivations cause the agent's behaviour in a deterministic fashion, i.e. they render it inevitable that the agent behaves as she does.

2) There is no relevant difference between behaviour that is determined by ordinary factors (e.g. the agent's genes and environment) and determination by the kind of alien source considered above.

3) Therefore, people whose wrongful behaviour is determined by ordinary factors do not deserve retributive punishment.

Fischer and Ravizza maintain that people generally are responsible (in the retributive sense) for behaviour that is determined by ordinary factors. However, they think that there is something particularly troubling about behaviour that is determined by alien sources. One reason they give for this is that the agent is presumably *unaware* of the alien origins of her motivations. For instance, they say:

> 'Given that [the agent] does not know about the manipulation of the scientist, and has not explicitly considered such manipulation, it is plausible to say that [the agent] has *not* taken responsibility for the kind of mechanism that actually issues in the action…'[107]

However, it can be objected that people are also often unaware of (or have mistaken beliefs about) the causal origins of motivations that are determined by ordinary factors. Fischer and Ravizza seem to acknowledge this fact. As noted above, they

[107] Fischer and Ravizza (fn 2, above), P233.

stipulate that an agent whose behaviour is determined in the ordinary way can be said to have taken responsibility for her motivations even if she is unaware of (or mistaken about) the origins (or the deterministic nature) of those motivations. This renders their theory strikingly asymmetrical. The requirements that must be satisfied before agents can be said to 'take responsibility' for their mechanisms differ dramatically depending on the source of those mechanisms. When it comes to determinism by ordinary factors, it is not necessary for the typical agent to *know* that his mechanisms are determined, or which factors determined them. In contrast, according to Fischer and Ravizza, an agent cannot take responsibility for mechanisms that have been, e.g., implanted by a neuroscientist unless the agent is aware of the source of those mechanisms. What could justify this asymmetry?

Fischer and Ravizza appear to suggest that if an agent knows that her motivations have been implanted by a neuroscientist then the agent is in a position to refrain from doing what the neuroscientist wishes her to do, provided the implanted mechanism is reasons-responsive. (Of course, if the mechanism is not reasons-responsive then this case poses no difficulty for Fischer and Ravizza. This is because they argue that acting on a reasons-responsive mechanism is a necessary condition for responsibility. If the implanted mechanism is not reasons-responsive then there is clearly an important difference between the implanted mechanism and ordinary deterministic mechanisms, which can explain why agents can be held responsible for acting on the latter but not on the former.) They stress that 'if the scientist's manipulation of the brain induces a moderately reasons-responsive mechanism, then this mechanism has the general capacity to resist the reasons for the agent's actual behaviour'.[108] Fischer and Ravizza's suggestion is misleading. For one thing, it should be remembered that according to them a mechanism underlying an action is reasons-responsive as long as

[108] Ibid, p 239.

there is one possible incentive that would induce the agent to act differently. If that incentive is not actually present, then it is physically impossible for the agent to act differently. The neuroscientist can therefore ensure that the agent acts in one particular way despite having equipped the agent with a reasons-responsive mechanism. The neuroscientist can simply design the mechanism to respond differently to only one incentive (of the neuroscientist's choosing) and then ensure that this incentive is not actually available to the agent. The agent's knowledge that she had been manipulated could only lead to the agent acting differently if the agent could draw on some psychological state(s) (e.g. a desire to act differently, or a belief that she should act differently) which had not been implanted. However, to suppose that the agent could draw on psychological mechanisms that had not been implanted would be to suppose that the neuroscientist had *failed* to implant to all the psychological states that are involved in the occurrence of the agent's action. This simply does not address the alien mechanisms challenge, which states that the neuroscientist had succeeded in implanting all these states.

Fischer and Ravizza suggest (but do not fully develop) two other possible grounds for differentiating determination by alien influences from determination by ordinary factors. The first suggestion is that the appearance of an alien mechanism always results in an abrupt, apparently inexplicable change in the agent's motivations, whereas determination by ordinary factors does not involve this. This suggestion will be discussed in Chapter Seven.[109] The second suggestion is that it is just a basic fact that certain types of cause of human behaviour (i.e. all of those that are intuitively 'alien') are intrinsically troubling to most agents, whereas causation by one's genes and environment is not. However, it is far from clear whether this empirical claim

[109] See section entitled: "1) Similarity with previous mental states."

about ordinary people's attitudes is accurate.[110] Even if it were accurate, the question remains as to whether these attitudes have a rational basis – if they do not, it is not clear why they should play an important role in a compatibilist account of free will.

What is Special about Philosophers?

Finally, it is interesting to note one rather unusual feature of Fischer and Ravizza's account. The conditions that must be satisfied before an agent can fairly be held responsible differ depending on whether or not the agent is 'philosophically sophisticated'.[111] On their account, an agent who has not 'immersed himself in the debates about causal determinism, free will, and moral responsibility' is considered to have 'taken responsibility' for the mechanisms underlying his actions whatever his beliefs about the nature or origins of those mechanisms. (As argued above, this considerably stretches the notion of 'taking responsibility'.)[112] In contrast, an agent who is familiar with the philosophical literature on the free will/determinism problem is treated differently. Such an agent has *not* taken responsibility (and hence is not responsible) for his mechanisms, *unless* he finds it 'plausible' that the (retributive) reactive attitudes are compatible with determinism and he is willing to put aside any 'residual doubts' about this for all practical purposes. (As noted above, Fischer and Ravizza seem mistakenly to assume that the *only* reasons why someone

[110] See, e.g., T Sommers, 'Experimental Philosophy and Free Will' (2010) 5 Philosophy Compass 199–212. doi: 10.1111/j.1747-9991.2009.00273.x).

[111] Fischer and Ravizza (fn 2, above), p226.

[112] Imagine for instance that a person who is unfamiliar with the details of the academic free will debate nevertheless intuitively feels that a person is not responsible for behaviour that is the inevitable product of factors outwith her control. She assumes that her actions are not produced in this way. Someone might say that such views should be discounted when deciding whether the agent is responsible for behaviour that is in fact determined, because the agent is not sufficiently well-informed about all the relevant philosophical arguments. But it is quite another thing to insist that *the agent herself takes responsibility* for her deterministic mechanisms despite her (albeit ill-informed) views on the subject. If Fischer and Ravizza wish completely to ignore the agent' s own views in this way, it raises the question of why they think that the fact that the agent herself adopts the *subjective* stance of 'taking responsibility' for her motivational mechanisms should be a precondition for being responsible in the first place.

might have doubts about the appropriateness of the reactive attitudes are to do with determinism.) They continue:

> 'But why should a reflective individual view himself in the way suggested? Why should such an individual deem himself a *prima facie* plausible candidate for the reactive attitudes, and be willing to put aside metaphysical worries? We believe that the considerations developed thus far in this book can move a reflective individual in precisely this direction.'[113]

Now, Fischer and Ravizza's work is undoubtedly extremely impressive. Even one of their most persistent critics has praised Fischer's writings in the following terms:

> 'John Martin Fischer's theory of moral responsibility is one of the great compatibilisms in the history of philosophy, standing alongside those of Aristotle, David Hume and Harry Frankfurt, for example, and of these it is arguably the most thoroughly developed.'[114]

However, it would be somewhat premature to suggest that all philosophically informed individuals can be brought round to Fischer and Ravizza's way of thinking. Indeed, Fischer and Ravizza acknowledge this.

> 'We concede that some individuals will not be convinced.... Such individuals will not deem themselves apt targets for the reactive attitudes, and thus they will not take responsibility for the kinds of mechanisms that lead to their behaviour. Thus, on our account, they will *not* be morally responsible for their behavior. But we do not take this to be a defect of our theory.'[115]

[113] Fischer and Ravizza (fn 2, above), p227.
[114] D Pereboom, 'Reasons-Responsiveness, Alternative Possibilities and Manipulation Arguments Against Compatibilism: Reflections on John Martin Fischer's My Way' (2006) 47 Philosophical Books 198, p198.
[115] Fischer and Ravizza (fn 2, above), p228.

For the reasons stated earlier, it is submitted that a theory which relieves individuals of responsibility merely because they are not convinced by Fischer and Ravizza's arguments for compatibilism has a significant defect.[116]

Retributivism and 'Our Moral Life'[117]

One possible way of defending retributivism against the metaphysical objections raised in the last two chapters is by appealing to our "moral life" considered as a whole. Michael Moore believes that determinism and retributivism are compatible. He claims that the hypothesis that retributivism is "true" provides the best, most coherent explanation of the existence of many of our most powerfully felt (and "virtuous") moral attitudes and that we should therefore accept the retributive justification of punishment.[118] First, I will criticise Moore's account (and a somewhat similar view advanced by PF Strawson). Next, I will criticise another related way of defending retributivism. This latter defence states that we should not reject retributivism because doing so would undermine vitally important moral attitudes and social practices, which would be disastrous.

Moore's Moral Life Argument

Section A: The Coherentist Defence of the retributive principle

Moore attempts to defend retributivism by appealing to "our moral experience".[119] The moral experience he is referring to consists of emotional responses to the actions of ourselves and others (such as resentment, outrage and guilt) as well as more

[116] Furthermore, it should be noted that if an agent is not morally responsible for a crime, then according to retributivism, he should not be held criminally responsible. According to Fischer and Ravizza, philosophers who are convinced that compatibilism is false are not morally responsible. But it is hard to imagine any jurisdiction adopting a criminal defence specifically designed to cover a certain kind of philosopher.
[117] Moore, *Placing Blame,* p543.
[118] *Ibid,* chapters 2, 3, 4, 12.
[119] *Ibid*, p543.

cognitive responses (such as judging another to be blameworthy). Moore[120] argues that most ordinary people have retributive responses when they think about (at least a certain class of) criminals. He cites horrific examples of violent attacks, rapes, and premeditated killings, arguing that most readers would feel so outraged by such crimes that they would desire the suffering of the perpetrators *even if that suffering did not produce any good consequences*. He concludes that we have very good grounds for believing retributivism to be true, because it forms 'part of the most coherent [explanation] of our moral experience, considered as a whole.'[121]

Moore acknowledges that sometimes retributive responses come into conflict with other reactions people have towards criminals. For example, if we are persuaded that a person's criminal behaviour was caused by the deprivation and child abuse she suffered, then we may be moved to sympathy for her – a sympathy which can undermine our retributive impulses. According to Moore, however, a moral theory should aim to achieve maximum coherence among our intuitions. Therefore, he argues that sympathetic responses towards disadvantaged criminals should simply be discarded as 'moral hallucinations' because they only comprise a very 'small and isolated' class of responses and they do not 'fit' with the rest of our moral experience.[122]

Moore's coherentist approach is open to criticism. It is possible to have a completely coherent set of false or wicked principles. Furthermore, as Thomas Clark argues, 'conflict between retributive feelings… and sympathetic feelings…may simply reflect a real moral conflict and to discount one side of the conflict in order to preserve theoretical consistency might well compromise theoretical accuracy.' [123]

[120] Henceforth ,when the name "Moore" is used it refers to Michael Moore, not G.E. Moore.
[121] Michael Moore, *Placing Blame*, p542.
[122] *Ibid*, p543
[123] Thomas Clark, 'Against Retribution', *Human Nature review*, 2003, p 471.

Moore's aversion to moral conflict might result in a distorting over-simplification of our moral life.

Even if, as Moore suggests, our retributive responses are more numerous than our 'small and isolated' class of sympathetic responses towards disadvantaged offenders, this is not in itself a sufficient reason for thinking that our retributive responses are more worth preserving or are more likely to be justified. If determinism is true, and if, as was argued in previous chapters, retributive judgements depend on principles (e.g. the principle of alternative possibilities) which are incompatible with determinism, then these retributive judgements cannot be justified. Furthermore, Moore's claim about the internal coherence of his position is doubtful. Moore believes in both determinism and in retributivism, but if, as was argued in chapters three and four, these theories are incompatible, then taken together they do not form a coherent explanation of our moral life.

Even if Moore is right in saying that our sympathetic responses to disadvantaged criminals form 'a small and isolated' class of responses, it is doubtful that the best explanation for this phenomenon is that our sympathetic responses are 'moral hallucination[s]'. A better explanation is that we are usually wholly or largely *ignorant* of the causes of any particular criminal's behaviour. Certainly we are not usually aware of anything approaching a causal circumstance (i.e. a sufficient, necessitating cause) capable of explaining such behaviour. Occasionally, people become aware of a significant part of an agent's causal history (e.g. they may discover that the agent had a violent upbringing). Such people may be moved to adopt a kind of *partial* determinist position. Their retributive urges falter because they feel that, given the part of the agent's causal history that they know about, the agent *did not have much chance* of behaving differently. However, on most

occasions, people probably do not think that there exist any causes which wholly/largely remove the agent's ability to behave differently. It is for this reason that sympathetic responses to criminals are relatively few in number, which in no way suggests that such responses are hallucinatory.

Moore claims that our retributive responses persist in the face of knowledge of causes of behaviour and that this provides strong grounds for thinking that retributive responsibility is compatible with actions being caused. He says,

> 'we undeniably parcel out both praise and blame for actions and choices we know to be caused by factors external to the actor's free will. Our moral life is built on praising or blaming people ….[for their actions] – even though we know at least some of the factors that caused these actions'.[124]

However, it should be emphasised that the persistence of retributive responses despite knowledge of *some* of the causal factors *influencing* behaviour goes no way towards reconciling retributivism with determinism. The belief that an action was influenced by some causal factors does not amount to the belief that the action was determined. Moore might try to claim that people persist in their retributive judgements, despite being convinced that the behaviour in question was causally *necessitated* by factors outwith the agent's control. However, this claim is implausible. Unlike Moore, most people are probably not convinced of the truth of determinism (brain scientists, legal theorists, and philosophers aside, most people have probably never explicitly thought about determinism) and, as argued above, if they were convinced of the truth of determinism their retributive responses would not persist.

[124] Moore, *Placing Blame,* p543.

Section B: Virtuous Emotions

Another part of Moore's strategy for justifying retributivism appeals to the idea of 'virtuous emotions'.[125] He argues that our retributive inclinations are likely to reflect moral reality because these inclinations are based on virtuous emotions such as compassion for innocent victims of crime and outrage on their behalf. (Although he acknowledges that retributivist attitudes may sometimes be based on vicious emotions like vindictiveness and a desire for revenge). According to Moore, virtuous emotions are our 'main heuristic guide' to discovering 'moral truth'.[126]

Even if this part of his strategy were sound, it would still not overcome the objections raised above. If a virtuous emotion (e.g. outrage against a wrongdoer) is based on a false belief (e.g. the belief that there is no serious evidence for thinking that the wrongdoer's action was determined, or the belief that determinism and retributivism are compatible) then this 'virtuous' emotion cannot be a good heuristic guide to discovering moral truth. Furthermore, it seems implausible to suggest, as Moore does, that retributive responses are virtuous, while compassion for disadvantaged offenders is not. As Thomas Clark argues:

> 'Putting ourselves in the…shoes of an offender *should* inspire sympathy, for if it does not, then we are supposing that we would have been immune to the influences that shaped her. From a naturalistic perspective, which Moore shares, in which human beings are determined by environment (as well as heredity), such a supposition is clearly false and the lack of sympathy it generates is a clear moral defect.' [127]

However, it is submitted that, quite apart from the above objection, Moore's 'heuristic guide' strategy is flawed. His view raises the question of how we are to

[125] Moore, *Placing Blame,* chapter 3.
[126] *Ibid,* p135.
[127] Thomas Clark, 'Against Retribution', *Human Nature review*, 2003, p 471.

distinguish a virtuous emotion from one that is not virtuous. Moore claims that the theory which has been generated from virtuous emotions provides the criteria for distinguishing virtuous from non-virtuous emotions. This obviously seems circular. However, Moore claims that this is not a vicious circle. In defence of this claim, he draws an analogy between moral thinking and scientific enquiry,

> '…[S]urely in science we do not expect to have to come up with some prescientific test for the epistemic import of sensory experience *before* we meld those experiences into a scientific theory. Rather, we rely on the body of scientific theory itself to justify exclusions of experience from the data.'[128]

However, merely making this observation about scientific method is not sufficient to establish the soundness of Moore's strategy. Moore may have come up with an example of a virtuous circle, but he needs to demonstrate more than the fact that virtuous circles exist. He needs to show that his is one of them. In view of the numerous striking disanalogies between scientific thinking and moral thinking, this would be a formidable task, which Moore hardly begins to carry out.[129]

Section C: The Truth of Retributivism is Not the Best Explanation for our Retributive Inclinations

Moore claims that the best explanation for the fact that many people have retributive responses is that the "truth" of retributivism "causes" people to have these responses.[130] The many objections that have been raised above cast doubt on this claim. Still, given that retributive beliefs are very widely held, it might be wondered

[128] Moore, *Placing Blame,* p138.
[129] Giving a full account of the difference between scientific thinking and ethical thinking is beyond the scope of this thesis. Here is one key difference: Scientific theories, unlike moral theories, can generate predictions about the world, which can be tested by observation. Scientific theories are subject to repeatable experiments of this kind, which, in the case of a well-corroborated theory, deliver the same results for all competent enquirers. Moral theories are in no literal sense testable in this way by empirical observation.
[130] Moore, *Placing Blame,* p109.

how so many people could be mistaken. One response is to say that widespread error is not uncommon. Many people believed that the world was flat, that the sun moved around the earth, that the earth was at the centre of the universe, that base metals can be turned into gold etc. until scientific discoveries showed these beliefs were mistaken. In the moral domain, many people believed that slavery was justified, that only male property owners deserved the vote, that some races were inferior to others, that absolute monarchy was the ideal form of government etc. until moral argument persuaded people that these views were misguided. Furthermore, scientific investigation can show certain moral beliefs to be misguided. Scientific explanations of sudden deaths, and illnesses probably undermined the belief that people had a moral duty to burn witches, who had previously been held responsible for causing these deaths and illnesses. So too, it is possible for scientific discoveries supporting the theory that human behaviour is determined to undermine the belief that it is intrinsically good to make wrongdoers suffer irrespective of the consequences of doing so.[131]

Furthermore, there are several coherent explanations for why people have retributive responses, which could be adopted instead of Moore's explanation. A consequentialist might say that the retributive principle serves as a useful "rule of thumb" to indicate those occasions when inflicting suffering on someone will be necessary in order to prevent greater suffering. The consequentialist might say that it is the usefulness of the rule that explains why people to adhere to it (even if the rule's adherents do not fully realise that this is why they accept it).[132] Alternatively, it

[131] In one famous American case, a defence lawyer, Clarence Darrow, argued that the judge should have compassion for two murderers, because their criminal behaviour was ultimately the product of causes over which they had no control. The judge sentenced Darrow's clients to life imprisonment (although he was under great public pressure to sentence them to death). *People v Leopold and Loeb*, Cook County Crim Ct III [1924] *per* Robin West and Christopher Brown, "Opening Statements and Closing Arguments" (1978) 8 Maryland Law Forum 126

[132] On the role that a "rule of thumb" can play in moral thinking see, Jonathan Glover, *Causing Death and Saving Lives* (1977), p106.

could be argued that, for many people, retributive beliefs are simply based on a desire for revenge or on sadism (e.g. public executions once attracted huge crowds of spectators, who used the fact that the offender had committed some wrong as an excuse to enjoy watching him suffer). John Mackie presents a "biological explanation for the tendency to feel non-moral resentment of injuries and gratitude for benefits and a sociological explanation for their moral counterparts,…the retributive emotions."[133] Basically, Mackie argues that the tendency to retaliate against aggression directed against oneself or against one's offspring was a trait favoured by natural selection. He then argues that co-operation with others to protect the social group against aggression was also favoured by natural selection. Over time, Mackie maintains, certain *kinds* of behaviour came to be regarded as *generally* harmful and meriting a hostile response. The generality of these judgements give them the 'apparent impartiality' characteristic of moral judgements. This group morality was reinforced by social interaction and passed on to successive generations through education. Given the objections to retributivism that have been raised in this thesis, it is submitted that any one of the accounts just mentioned (or any combination of them) provides a better explanation of the prevalence of retributive attitudes than the supposition that retributivism is true.

PF Strawson's 'Reactive Attitudes'

Moore's discussion of our moral life draws to some extent on Peter Strawson's highly influential compatibilist theory based on the reactive attitudes, including resentment, (discussed at p62, above). Unlike Moore, Strawson believed that our reactive attitudes and practices of holding each other responsible are not merely heuristic guides to the truth about the nature of responsibility, but are constitutive of responsibility. Strawson seems to be too complacent about relying on our emotional

[133] John Mackie, "Morality and the Retributive Emotions" 1982 Criminal Justice Ethics 3, at pp8-9.

responses and current practices. As noted above, it is perfectly conceivable that these responses and practices are misguided. The fact that we currently hold criminals responsible in the retributive sense cannot by itself tell us that they *are* responsible in that sense. According to Strawson, retributive reactive attitudes are an inevitable part of human life and so it is pointless to argue about whether they are justified. In response, even if experiencing attitudes of resentment and disapprobation is inevitable, Strawson cites no evidence to suggest that their translation into institutions of retributive punishment is inevitable, so surely discussion of whether such institutions can be justified is not completely pointless.

Strawson also suggested that if, *per impossibile*, we could imagine giving up the retributive reactive attitudes, doing so would not be rational as life would be terrible without them. I will consider arguments of that kind below.

Retributivism, Illusionism and Criminal Behaviour

In defence of retributivism, it might be argued that attempting to remove the retributive elements from our attitudes and social institutions would be morally and socially disastrous. For example, the retributivist might claim that, in order for people to be *motivated to behave well*, they must feel that they are retributively responsible for their actions – i.e. they must feel that if they commit a crime, it will be appropriate for the state to punish them, regardless of the consequences of imposing such punishment. Also, the retributivist might claim, in order for criminals to be *reformed and rehabilitated*, they must feel retributive guilt – i.e. they must feel that they deserved to be punished purely because of what they did and irrespective of the consequences of such punishment being imposed. In Smilansky's words: "to put it bluntly: people as a rule ought not to be fully aware of the ultimate inevitability of what they have done, for this will affect the way in which they will hold themselves responsible….Humanity is fortunately deceived on the free will issue, and this seems

to be a condition of civilised society and personal value." [134] This illusionism, whatever its supposed merits, is no kind of retributivism, but a variety of consequentialism. It directs us to cultivate the illusion of retributivism because of the supposedly disastrous consequences of rejecting the illusion.

Furthermore, it is hard to take seriously such scare mongering about the corrupting psychological effect of rejecting retributive responsibility. So far we have not witnessed a crime-wave perpetrated by free-will-sceptical philosophers. Other writers have argued at length that rejecting retributive responsibility would not be psychologically demoralising and it is outwith the scope of this thesis to present a full discussion of these psychological claims.[135] It is submitted that even if these claims were correct, they would not establish the soundness of the retributive principle that punishing the guilty is an intrinsic good. A principle is not sound merely because we desperately want it to be, nor because we are afraid of what might happen if the principle were discovered to be unsound.

I will argue, however, in the next chapter that there is no *logical reason* why removing the retributive elements from our penal institutions would lead to an inhumane or unjust society. It would be unjust, however, to punish people solely on the basis of a penal theory, like retributivism, which is highly problematic. This thesis focuses specifically on punishment. I do not propose that people should try entirely to excise all retributive assumptions or sentiments from their mental lives and everyday dealings with others.

[134] S Smilansky, "Free Will, Fundamental Dualism and the Centrality of Illusion" in Kane (ed), *The Oxford Handbook of Free Will* (2002) 489, p500.
[135] See Honderich, *How Free Are You?,* chapter 10; Derk Pereboom, *Living without Free Will* (2001); Sommers T, 'The Objective Attitude' (2007) 57 (228) The Philosophical Quarterly 321.

A related objection is that my strategy of arguing against retributivism entails that the state should take no steps to deal with the criminally dangerous. However, this objection misconstrues my argument. The thought-experiments involving Jane and the robot, discussed in the last two chapters, were intended to demonstrate that it would be counter-intuitive to punish these determined agents *only because they had committed wrongful actions, irrespective of the consequences of punishing them.* These thought-experiments were not designed to show that, with respect to wrongdoers whose actions are determined, it is unjustifiable to punish them *simpliciter*.[136] On the contrary, there are important non-retributive reasons for punishing such wrongdoers, which are compatible with determinism.[137]

One theory of crime prevention which does not depend on the existence of retributive responsibility is suggested by Pereboom. He argues that, just as we have a right to quarantine carriers of severe infectious diseases in order to protect society, we also have the right to isolate dangerous criminals in order to protect society. The legitimacy of quarantine does not depend on the carrier being retributively responsible for having a dangerous disease. By analogy, society may have a right to isolate someone who has shown a sufficiently strong tendency to commit serious crimes, even though that person is not retributively responsible.[138] Alternatively, on the deterrence theory, punishment is justified if it prevents serious harm to society by deterring criminal behaviour. If, as argued in chapter two, determinism and

[136] The retributivist might claim that this would not be punishment in the fullest sense. If so, it is submitted, so much the worse for punishment in the so-called fullest (i.e. retributive) sense. The semantic manoeuvre of stipulating that only retributive punishment counts as punishment in the fullest sense does nothing to show that there is not some other meaningful and important sense of punishment which is compatible with determinism and which is justifiable.

[137] One might object that consequentialism is incompatible with determinism because consequentialism assumes that some things matter (e.g. suffering, preference satisfaction) and determinism entails nihilism i.e. that nothing matters. However, this objection is dismissed by most writers in the area of determinism and a full critique of this objection is outwith the scope of this thesis. For a discussion of determinism and nihilism see Daniel Dennett, *Elbow Room* (1984), pp153-156.

[138] Pereboom "Living Without Free Will: The Case for Hard Incompatibilism", in Kane (ed) *The Oxford Handbook of Free Will* (2002), 478, at p480.

rationality are compatible, then, even if determinism is true, rational would-be offenders can be deterred through their understanding of the consequences of law breaking. On the moral education theory, punishing wrongdoers may be justifiable if we have grounds for believing that punishment can (causally) bring about their reform and rehabilitation, through presenting them with good moral reasons for modifying their value systems.[139] There may be reasons why these non-retributive approaches are unacceptable, but my strategy for arguing against retributivism, does not in itself entail that these theories are unacceptable.[140] Subsequent sections of this thesis will focus on developing a defensible non-retributive approach to responding to criminal behaviour.

To summarise: Appealing to our 'Moral Life' cannot rescue retributivism from the metaphysical difficulties that have been raised in Chapters One and Two. Moore's arguments based on our moral experience and the supposedly virtuous nature of

---

[139] R.A. Duff has written extensively on the communicative function of the criminal law. See, e.g., *Trials and Punishments* (1986). However, it is not being suggested that his view of punishment as communication is necessarily consistent with determinism in all respects.

[140] It might be thought, for example, that a legal system which rejected retributive responsibility would have no rational basis for retaining the concepts of *mens rea* or excuses and that this would be unacceptable. HLA Hart argues that these features of our system should be preserved for consequentialist reasons, because, if they were abolished, people would be forced to live in a permanent state of fear. This is because, "if we are… to be liable if we [perform a prohibited act] by accident, by mistake, under coercion, etc., the chances that we shall incur the sanctions are immeasurably increased" and it will be difficult to predict when we will be subject to legal sanctions. H.L.A. Hart, *Punishment and Responsibility,* (1968), p47-48. The deterrence theory is regarded as unacceptable by many people because it treats criminals merely as a means. A view which stresses the importance of moral communication (perhaps in combination with other considerations such as the need to protect society) is not so susceptible to this criticism. If one of the goals punishment is intended to achieve is the moral betterment of the criminal, then the criminal is being treated partly as an end in herself. If the attempt to reform is through moral dialogue, addressed to the *rational faculties* of the agent, then arguably this also shows respect for the agent as an end in herself. Another possible objection to these non-retributive theories of punishment is based on the principle that it is intrinsically wrong to punish someone unless that person is retributively responsible for a wrongful action. One way of replying to this would be to argue against the soundness of that principle. Another, response is to argue that the principle is sound but can be outweighed by consequentialist considerations. The following writers believe that determinism and retributive responsibility are incompatible and have defended alternative theories of punishment: Honderich, *How Free Are You?,* Honderich, *Punishment: The Supposed Justifications* (1984) ; Derk Pereboom, *Living without Free Will* (2001).

retributive emotions are unconvincing. They cannot establish that the truth of retributivism is the best explanation for the existence of retributive attitudes and emotions. Furthermore, arguments based on the supposedly disastrous consequences of abandoning retributivism cannot show that that the suffering of the guilty is an intrinsic good.

Conclusion

The theory that human behaviour is determined is conceptually sound and seems to have empirical support. Quantum indeterminism has not been shown to affect mental processes in a way that could provide a basis for retributive responsibility. The truth of determinism would seriously undermine retributivism. A retributivist who tries to reconcile his view with determinism is in a highly paradoxical position. He must maintain that it is intrinsically good to punish an offender, just because that offender has voluntarily committed a wrongful action, irrespective of the consequences of imposing punishment. Yet, he must accept that had he, the retributivist, been subject to the same external factors that determined the offender's action, he would have committed the very same offence. He must maintain that it does not matter that the offender could not have done otherwise. Nor, on his view, does it matter that the offender's decision was ultimately the product of luck. He must advocate inflicting suffering, a thing we normally consider to be bad in itself, on the grounds that, despite all the above considerations, it can be intrinsically good in itself. The cumulative effect of this series of counterintuitive claims is to make the attempt to reconcile retributive responsibility and determinism seem almost incoherent. Any attempt to justify punishment must meet a high standard. Part One of this thesis has cast serious doubt on the idea that people can be held responsible in a way that could justify retributive punishment.

The arguments against retributivism that have been raised in this thesis have important practical implications for the legal system. Rejecting retributivism would free society from the obligation of ensuring that every guilty person is made to suffer in direct proportion to her moral guilt. Instead, it may be morally permissible to pursue alternative, more flexible responses to criminal behaviour. In view of the severe treatment that offenders undergo within the penal system (e.g. deprivation of liberty, loss of certain rights, stigma), over-crowding of prisons and the high rate of recidivism, it is important to consider whether our current penal practices can be justified on a non-retributive basis, what alternatives are available and what goals and values should guide attempts at reform.

As indicated above, some theorists worry that abandoning retributivism would lead to bad consequences for society, and specifically would result in an unjust and inhumane penal system. If this were correct, it would not show that the retributive principle is sound, but it would pose problems for the non-retributive approach to punishment that I advocate in this thesis. The next chapter will argue that a 'hard incompatibilist' approach to punishment (i.e. one that rejects libertarian and compatibilist retributivism) could be fair and humane. It is possible to have justice without retributive desert.

Chapter Three: Justice without Desert

Introduction

Ben, a sane adult, kills an old lady for her money. Martin, who is severely mentally ill, kills an old lady because he has the delusional belief that she is going to kill him. In both cases, the state has strong reasons to consider interfering with the killer's liberty. Retributivists claim, however, that the state's main grounds for interfering with Ben are completely different from its reasons for interfering with Martin.[141] According to traditional retributivism, the state is justified in interfering with Ben, because he is blameworthy and deserves to suffer for his crime; whereas Martin, who is blameless due to his insanity, may need restrictions placed on his liberty because he is dangerous. Non-retributivists, on the other hand, often cite the need to protect society as the main reason for interfering with both sane and insane offenders. Focusing on the forward-looking goal of social protection is claimed by the proponents of this approach to be more humane than the supposed vengefulness of traditional retributivism. Retributivists often respond that far from treating criminals humanely, this forward-looking approach actually *demeans* responsible offenders, by failing to distinguish them from insane lawbreakers on the basis of desert. For example, in an influential article, CS Lewis warned that once desert is abandoned in favour of harm prevention, responsible offenders are objectified and are no longer protected by considerations of justice*:*

> 'There is no sense in talking about a...'just cure'... We demand of a cure not whether it is just but whether it succeeds. Thus when we cease to consider what the criminal deserves and consider only what will cure him or deter others, we have tacitly removed him from the sphere of justice altogether;

[141] Or at least 'pure retributivists' claim this. Mixed theories are discussed briefly on below under the heading 'General Deterrence and the Mere Means Objection'.

> instead of a person, a subject of rights, we now have a mere object, a patient, a "case".'[142]

This chapter argues that discarding retributive desert, does not mean discarding justice, or treating offenders like objects.[143] By examining principles that apply equally to both sane and insane offenders, this chapter will show how a non-retributive response to crime can be fair to lawbreakers.[144] It is intuitive that a fair legal system must have the following features: 1) the state does not frame people who have not broken the law; 2) sentences are proportionate (or at least not grossly disproportionate); 3) rules of due process are upheld (e.g. the state bears a heavy burden of proof; there is independent judicial oversight; the suspected law-breaker is allowed to participate in the process); 4) the lawbreaker is not treated as a mere means. These principles are said to follow from retributivism. Philosophical discussions of punishment often remark that non-retributive (especially consequentialist) approaches sometimes recommend breaching these principles, or at least that these approaches fail to capture the real reason why the principles are so important.[145] It is also frequently assumed that these principles only govern the

[142] CS Lewis, 'The Humanitarian Theory of Punishment' (1953) 6 Res Judicatae 224. Lewis's article is cited approvingly in, e.g., S Morse, 'Thoroughly Modern: Sir James Fitzjames Stephen on Criminal Responsibility' (2008) 5 Ohio State Journal of Criminal Law 505, p511; and in N Vincent, in B van den Berg and L Klaming (eds), Capacitarianism, Responsibility and Restored Mental Capacities, *Technologies on the Stand. Legal and Ethical Questions in Neuroscience and Robotics* (2011 Wolf Legal Publishers, Nijmegen), p51. PF Strawson also advocated drawing a sharp distinction between the norms governing sane and insane law-breakers on similar grounds: P Strawson, 'Freedom and Resentment' (1962) 48 *Proceedings of the British Academy*, 187.

[143] Henceforth, the term 'desert' will refer to retributive desert. However, it should be noted that desert is sometimes used in an explicitly non-retributive sense. For instance, Benjamin Vilhauer uses the term 'personhood-based desert' to refer to claims that people have simply in virtue of their status as persons, rather than any action for which they are retributively responsible. He contrasts 'personhood-based desert' with 'action-based' desert. B Vilhauer, 'Persons, Punishment and Free Will Scepticism' (2011) *Philosophical Studies* (online first). DOI 10.1007/s11098-011-9752-z.

[144] The question of whether a non-retributive approach can be fair to victims and potential victims will be discussed in a subsequent chapter.

[145] See, e.g., A Duff, 'Crime and punishment' in E Craig (ed.), *Routledge Encyclopedia of Philosophy* (Routledge, London 1998). Retrieved October 29, 2011, from http://www.rep.routledge.com/article/T002; K Murtagh, 'Punishment', in J Fieser and B Dowden (eds.), *The Internet Encyclopedia of Philosophy*. Retrieved October 29, 2011 from *http://www.iep.utm.edu/punishme/;* H Bedau and E Kelly, 'Punishment', in E Zalta et al (eds.), *The Stanford Encyclopedia of Philosophy* (2010), Retrieved 29 October, 2011, from http://plato.stanford.edu/entries/punishment/.

treatment of responsible offenders. This chapter argues that these four principles (or analogous principles) also apply to insane offenders, whom no-one considers retributively responsible and that therefore these norms can be defended independently of the notion of 'desert'. It is argued that principles of justice can be endorsed by those who are sceptical about 'free will' and ultimate moral responsibility.

Consequentialist explanations of the above-mentioned principles depend on hard-to-verify empirical claims about the probable results of different penal policies. The norms of justice defended in this chapter do not depend on such contingencies. Nor does this chapter endorse a system of 'social hygiene' whereby no distinction at all is made between sane and insane offenders. But it does argue that such distinctions can be drawn on a non-retributive basis. Recognising the important similarities *as well as* the differences between these sane and insane offenders is necessary in order to treat both groups fairly and humanely.

Framing the Innocent

Here is an example that is frequently cited to demonstrate the short-comings of consequentialism:

> Framing a Moral Agent
> A horrible act of violence is committed and the culprit cannot be found. A riot will ensue that will harm many innocent people, unless the mob is persuaded that the wrongdoer has been apprehended and punished. So the authorities frame and punish an innocent man.[146]

[146] An example of this kind was used by H McCloskey, 'A Non-Utilitarian Approach to Punishment.' in G Ezorsky (ed), *Philosophical Perspectives on Punishment* Albany, (State University of New York Press, 1972), p127.

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

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

It seems intuitive to say that Timothy has been treated unjustly. Retributivists cannot explain this intuition with reference to retributive desert. Retributive desert does not come into it. The authorities do not claim that Timothy *deserves* to be locked up. Timothy is not a responsible agent and so would not deserve retributive blame even if he had been the attacker. The actual attacker (being insane) does not deserve retribution either.

The authorities' actions can be criticised for the following reasons. Firstly, they have wronged Timothy by lying about him. The lie is particularly objectionable because it denies important good qualities that Timothy actually has (e.g. gentleness and friendliness), and falsely asserts that he has serious negative qualities (a propensity to kill innocent people). True, Timothy is not responsible for having these good qualities, nor do the authorities claim that Timothy is responsible for his alleged negative qualities. Nevertheless, it seriously wrongs a person to tell this kind of lie about him. Furthermore, Timothy is detained on the basis of such outrageous lies. This also wrongs him, because the grounds of his detention are illegitimate. Timothy's detention also treats him merely as a means to avert a threat from elsewhere. It does not seek to eliminate a threat that he himself poses.[147]

The above-mentioned objections could also be raised against framing a sane person. There are therefore strong reasons against framing people who have not broken the law, which are neither based on retributivism, nor on consequentialism. Unlike consequentialist objections to framing people, the reasons given here seem to capture the intuitive idea that framing is unjust, because the *framed person* has been victimised. The consequentialist rationale refers to some calculation of the general welfare and this fails to capture our intuitions about the injustice done to the individual. Unlike retributivist objections, the reasons offered here help to explain why framing is unjust in both the agent and the non-agent examples. This is an advantage, since it seems intuitive that the injustice that occurs in both examples is relevantly similar.

[147] For more on the distinction between treating someone as a mere means and eliminating a threat they pose see below, under the heading 'General Deterrence and the Mere Means Objection'. The mere means argument was famously put forward by Kant. However, Kant seems to have tied this argument to the idea of rational agency, so it is not clear whether non-rational agents are protected by the duty of respect for persons as he originally formulated it. H Paton (tr), I Kant, *The Moral Law: Groundwork of the Metaphysic of Morals* (Routledge, London 1948), 85. However, it is submitted that the principle of respect for persons should extend to non-rational or partially rational people such as the mentally ill, learning disabled people and young children. See A Wood and O O'Neill, 'Kant on Duties Regarding Non-Rational Nature' in (1998) 72(1) *Proceedings of the Aristotelian Society* 211.

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

Benjamin Vilhauer has proposed a different kind of personhood-based, non-retributive argument against framing the innocent.[148] According to Vilhauer, respecting someone's personhood means treating them in a way that they would rationally consent to be treated. He does not rely on the person's actual consent, but on the notion of 'hypothetical consent'– i.e. they *would* consent to be treated this way *if* they were rational. He uses Rawls's idea of 'the original position' to model rational consent.[149] The original position is a thought experiment in which people choose the rules that will govern a society. The rules are chosen behind a 'veil of ignorance': the choosers are unaware of certain facts about what their own position will be in the society and what personal characteristics (e.g. race, gender, wealth, strength, intelligence and industriousness) they will have. They are aware of the fundamental interests that they all have in common (e.g. security and the freedom to pursue one's goals) and they have knowledge of relevant scientific and sociological

[148] Vilhauer, fn143, above.
[149] J Rawls, *A Theory of Justice* (2nd Ed, OUP, Oxford 1999).

theories. The veil of ignorance is designed to describe a situation of fairness among the social contractors, to ensure their impartiality and to filter out factors that are just down to luck. Each deliberator must also imagine that he or she is just as likely to be harmed by any principle that is chosen as to benefit from it. Vilhauer, unlike Rawls, includes knowledge of whether one will be a wrongdoer as a factor that is hidden from the social contractors. This is because Vilhauer is a free will sceptic and believes that one's moral character is, like race and gender, a product of the genetic and environmental lottery. Vilhauer claims that respecting someone's personhood means treating them as they would rationally consent to be treated, i.e. in accordance with a principle that would have been agreed to by deliberators in the original position. He claims that no rational deliberator could have chosen the principle that the authorities may, when it is expedient, frame innocent individuals. Such a regime would involve the authorities systematically deceiving the members of this society. Otherwise the scapegoating of innocent individuals would be ineffective. A deliberator in the original position must acknowledge that under this regime he could be one of those who are deceived about a basic principle governing that society. Consenting to systematic deception undermines one's status as a rational agent. Therefore, according to Vilhauer, the idea that a rational deliberator would choose to be systematically deceived about something so important is self-contradictory.[150]

Vilhauer's argument is intriguing and could be invoked to supplement the position defended in this chapter. However, it does not seem to capture the *main* reason why the authorities' actions are wrongful in the two framing cases. Intuitively, the main injustice in both cases is the wrong that has been done *to the framed individual*. However, Vilhauer's explanation focuses on the wrong of deceiving the general public. On Vilhauer's account, the wrong that is done to the framed individual

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

derives from the supposed logical problems with a principle that endorses deceiving the public. This seems too indirect.

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

Proportionality

If desert were abandoned, some fear that the state's response to law–breaking would no longer be governed by principles of proportionality. For instance, Lewis maintained that a medical model of punishment would permit the authorities to interfere with the liberty of citizens, whenever the authorities found this convenient.

They would simply label the citizens 'diseased'. He claimed that the authorities could impose on such unfortunate citizens any 'treatment', no matter how burdensome, and any period of confinement, no matter how lengthy. Ordinary people, he maintained, would have no basis for objecting to this on grounds of justice, since 'justice' is a retributive concept.[151]

This line of argument is based on a misconception of the principles that should apply to the mentally ill. It is unjust to confine someone or force her to undergo treatment against her will merely because she has a mental illness. She must pose a threat to the safety of herself or others.[152] Furthermore, certain treatments are so risky or so devastating to the individual that it would be unjust to impose them on her, even if she is mentally ill and dangerous. It would also be unfair to impose a particularly lengthy or onerous treatment/confinement on someone if her behaviour only had a relatively minor impact on the welfare of any particular individual.[153] For instance it would be grossly unfair to lock up a mentally ill person for life in a secure institution, just because she made loud noises in the street, causing only minor irritation. This is a consideration of proportionality (though clearly of a non-retributive kind). It is not merely a question of whether the intervention is *necessary* in order to prevent the objectionable behaviour. It is conceivable that for some people, a measure almost as drastic as confinement in an institution might be required in order to prevent them from causing a nuisance. Imposing such a drastic measure would still be unjust. This proportionality constraint is not merely the result of utilitarian calculation. Classical utilitarianism is aggregative. On an aggregative approach if enough people were each caused a tiny bit of distress by the nuisance,

[151] Lewis, fn142, above.
[152] See e.g. Mental Health (Care and Treatment) (Scotland) Act 2003, ASP 13.
[153] In the context of a discussion of the punishment of sane offenders, this principle is defended in T Honderich, *Punishment: The Supposed Justifications* (Penguin Books: Middlesex 1984), p 78.

then that could eventually outweigh the interests of mentally ill person and justify locking her up. In contrast the proportionality principle defended here states that the intervention must be proportionate to the impact that the harm to be prevented by the intervention would have on *any particular* victim. So a greater intervention, such as lengthy confinement, would be justified to prevent killing or a serious violent or sexual attack. Whereas a much more minor intervention, such as counselling, or supervision in the community would be justified to prevent nuisances. The proportionality principle is based on respect for the separateness of persons and on an ideal of equality – it is *prima facie* wrong to create a situation where people suffer grossly unequal levels of distress.[154]

If this principle of proportionality applies to insane law-breakers who are clearly not deserving of retribution, then an analogous principle of proportionality would also be available to sane offenders under a non-retributive system. It might be objected that the proportionality principle does not give very precise recommendations about the exact degree of burdensomeness that is appropriate in each case. However, this objection is equally applicable to retributive conceptions of proportionality. The most that can probably be said for both conceptions is that they serve as a safeguard against injustice, ruling out clear cases of *gross disproportionality*.

Due Process

Daniel Dennett, though far from being a traditional retributivist, has recently argued that there will be 'totalitarianism', unless we have a system of punishment based on desert.[155] However, this ignores the fact that important individual rights and rules of due process apply in contexts where desert is not an issue e.g. when the state wishes

[154] Like most of the principles of justice defended here, this is a strong presumption, but not necessarily an absolute prohibition in all cases. As noted earlier, retributivists themselves often admit that principles of justice can sometimes be outweighed if the consequences are serious enough.
[155] D Dennett, 'My Brain Made Me Do It' (2011), *Max Weber Lecture Series*. Retrieved 30 September, 2011 from http://cadmus.eui.eu/bitstream/handle/1814/16895/MWP_LS_2011_01.pdf?sequence=1.

to restrict the liberty of non-responsible, mentally ill offenders. Such individuals cannot be detained at the mere whim of a totalitarian dictator.

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

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

Thus it can be seen that several important principles of due process do not depend on desert and are applicable to sane and mentally ill individuals. To summarise, these

principles include the following: interventions are prescribed by law; the burden of proof is on those who wish to intervene; decisions are made by courts or tribunals that are independent of the executive; the person who may be subject to the intervention is entitled to participate in the process and to be fully informed and adequately represented; persons subject to interventions are entitled to initiate a review of the legitimacy of the interventions. Any non-retributive response to law-breaking should uphold these principles. However, there are further principles of due process that should apply specifically to sane offenders. These will be discussed in the final section.

General Deterrence and the 'Mere Means' Objection

According to consequentialist deterrence theories, the state may punish an offender in order to make an example of him. Consequentialists justify this with reference to the general welfare, rather than the culpability of the offender. However, some 'mixed theorists' argue that it is only permissible to make an example of the offender if he is *also* culpable. In this sort of case, the punishment 'kills two birds with one stone' by giving the offender what he deserves and also deterring potential offenders.[156] (Such theories are 'mixed' because they take account of retributive and forward-looking considerations.) It is often alleged that punishing someone for *purely* deterrent reasons instrumentalises that person. Mixed theorists, may claim, however, that making an example of a retributively culpable offender does not treat him *merely* as a means, as long as the state is also responding to his free choice.

[156] See, e.g. Lewis, fn142, above, p227. Andrew Von Hirsch defends a related line of argument. He maintains that criminals deserve censure for their wrongdoing because they are culpable. But he claims that the state is justified in providing prudential reasons for obeying the law, in the form of the harsh treatment aspect of punishment, in addition to moral reasons. The prudential reasons should help to strengthen people's resolve to act in accordance with the moral reasons. A Von Hirsch, *Censure and Sanctions* (OUP, Oxford 1993).

The analogy with insane law-breakers suggests that it is indeed wrong to punish criminals purely in order to make an example of them. The state can restrict the freedom of dangerous insane lawbreakers in order to incapacitate them. But it would be unjust to confine a non-dangerous person in order to 'make an example' of him, if that person broke the law due to a fit of temporary insanity that will never recur. The state would be using him as a mere means. It should be noted that the 'mere means' objection is not tied to retributivism (although its originator, Kant, was a retributivist). It is used by deontologists in a wide range of contexts, not just retributive punishment. The mixed theorist's defence against the mere means objection relies on the idea of retributive responsibility. If no one is retributively responsible then the mere means objection seems to apply equally to sane and insane offenders and to provide a strong reason against harming either group for *purely* deterrent reasons. (Although, as will be seen below, in some unusual contexts the prohibition on instrumentalisation can be outweighed by other considerations.)

This does not mean, however, that the state should try to ensure that its response to law-breakers does not have deterrent effects. A measure may be used as a deterrent, provided that it is *also* strictly necessary in order to incapacitate the dangerous person. It is possible that someone may be non-responsible, due to her mental condition, but also capable to a certain extent of being deterred. For instance, a person with severe learning disabilities may understand that some form of behaviour (e.g. running into the road, or being violent) will result in a negative consequence for her (e.g. she will have less freedom, and be subject to greater supervision). The thought of this negative consequence may help to restrain her from engaging in the dangerous behaviour. It is not wrong for her carers to explain to the person (in humane, non-inflammatory terms) that these negative consequences will occur as a result of such behaviour and have been imposed on others. They may explain this in the hope that this will affect the conduct of the person with learning disabilities. The

knowledge that mentally ill offenders will still be confined, if dangerous, may also deter some sane offenders from trying to fake an insanity defence. The state does not wrong mentally ill law-breakers by publically pointing out that such law-breakers need to be confined if dangerous. Any deterrent effect such statements may have is no bad thing, provided that the authorities do not use unduly stigmatising and inflammatory language. Therefore, the notion of retributive responsibility is not required in order to make this limited form of deterrence acceptable.

Victor Tadros has recently produced a strikingly original defence of general deterrence, which sharply differentiates between culpable and non-responsible people.[157] He maintains that his approach is neither consequentialist nor retributivist. This theory of punishment is based on an analogy with self-defence. It is impossible within the scope of this thesis to capture all of the details of his highly sophisticated and nuanced theory. This section will briefly summarise the significance of culpability in his account.

Tadros draws a distinction between 'manipulative' and 'eliminative' harm. Harming someone as a means to avert a threat from someone else is an instance of 'manipulative' harm. Whereas harming someone in order to eliminate a threat which that person is directly posing counts as 'eliminative' harm. He argues that both responsible and non-responsible (e.g. insane) people who pose a direct threat may be harmed in order to eliminate that threat. However, he claims that only people who are morally culpable (and had the opportunity to avoid being harmed) may be harmed manipulatively. He uses the following examples to support these claims:

> 'Hit Man
> I hire a hit man to kill you. The only way in which you can prevent the hit man from doing that is to pull me in front of x'

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

This is an instance of manipulative harm. The hit man's boss is no longer posing a direct threat to the intended victim. The intended victim uses the boss as a means to avert the threat from the hit man. It is intuitive to say that that this is permissible. Tadros uses this example to support the claim that it is permissible to subject a culpable offender to manipulative harm.

> 'Shield
> A maniac is attacking me. The only way to prevent myself from being killed is to pull you, an innocent bystander, in front of me, using you as a shield. If I do that you will be killed.'[158]

This also involves manipulative harm, since the innocent bystander poses no direct threat. It is intuitive to say that it is wrong to use the bystander as a shield. Tadros uses this example to support the claim that it is wrong to subject a non-culpable person to manipulative harm.

> 'Maniac
> I go temporarily insane and attack you. I will kill you unless you kill me.'[159]

This is an example of eliminative harm. The "maniac" is not killed as means to averting a further threat. The maniac is the threat. It is intuitive that it is permissible to harm the 'maniac'. Tadros uses this example to support the claim that it is permissible to inflict eliminative harm on people who pose direct threats, whether or not those people are culpable.

Punishing someone purely in order to deter others is an example of manipulative harm. According to Tadros, the above-mentioned self-defence cases help to show that it is permissible for the state to inflict this kind of manipulative harm on culpable offenders, but not on insane law-breakers.

---

[158] Tadros fn157, above, p241.
[159] Tadros fn157, above, p241.

However, it is far from clear that the self-defence cases genuinely support this sharp distinction between culpable and insane offenders when it comes to manipulative harm. Consider the following case:

> Hit man with an Insane Boss
>
> An insane person has paranoid delusions about a neighbour. He asks a hit man to kill the neighbour. The insane person accompanies the hit man to the neighbour's house to watch the killing. The neighbour knows the hit man's boss is insane. The only way the neighbour can prevent the hit man from killing her is to pull the insane boss in front of her, using the insane boss as a shield.

This case involves manipulative harm of a non-culpable insane person. It is far from obvious that it is wrong for the neighbour to use the insane boss as a shield. Yet Tadros would not want to conclude that it is all right for the state to inflict manipulative harm on insane law-breakers, e.g. by harming them purely in order to deter potential offenders. On balance, it seems that the prohibition on using someone as a mere means can be overridden in this unusual type of self defence case. But this does not imply that the state is entitled to harm the law-breaker as a means to some further end. The state has a strong duty of care to all its citizens, including law-breakers. In particular, it has an important duty to respect the personhood and basic equality of citizens. It is therefore especially problematic for the state to treat one citizen merely as a means, in order to benefit another group of citizens. In contrast, the neighbour in the self-defence case does not have that kind of special relationship to the insane boss, but she does have a strong right of self-preservation. Furthermore, the state's response to law-breaking is a public act that is performed after deliberation, in 'cold blood' and which expresses society's values. There is also a tremendous inequality of power between the state and the law-breaker. The state has

the offender at its mercy. Given this context, if the state were to inflict manipulative harm on a law-breaker this would do much to erode the prohibition on instrumentalisation. Harming someone as a means during a private act of self-defence does not send out such a strong message.

Furthermore, there seems to be a tension between Tadros's scepticism about retributive desert and the emphasis he places on culpability and opportunities. Tadros argues that the intuitions that support judgements about retribution are based on implicit ideas we have about free will. It is natural to think that the choices for which we deserve praise or blame were neither predetermined by factors beyond our control, nor a matter of chance. However, Tadros maintains that this implicit view of human choices is unlikely to reflect reality – it seems likely that our choices are in fact determined by our genes and environment. Therefore the intuitions in favour of retributive desert are probably misleading. He gives the following example to show that there are considerable difficulties in trying to reconcile free will and determinism. Imagine that you are contemplating a baby who will grow up to be a wrongdoer. (If determinism is true it would be possible in principle to predict the future with certainty if we had a complete knowledge of the facts of the past and the laws of nature.) This baby's genes and environment guarantee that he will become a wrongdoer. Tadros persuasively argues that a great misfortune has befallen the child. Furthermore, 'to think it a good thing that the badness of his life is compounded by making him suffer [for his wrongdoing] seems barbaric'.[160] Given these views about determinism, it seems odd that Tadros draws such a sharp distinction between the harm that it is fair to inflict on culpable versus non-culpable law-breakers. As we have seen, he maintains that only non-culpable individuals should be protected from manipulative harm, because culpable offenders had the opportunity to avoid being subjected to such harm. However, if determinism is true then it was predetermined

[160] Tadros, fn157, above, p63.

by factors outwith the offender's control that he would not take that opportunity. It seems that 'culpable' and 'non-culpable' law-breakers are both victims of bad luck.

Differences between Sane Law-Breakers and the Mentally Disordered

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

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

Different methods are appropriate for dealing with the behaviour of insane as opposed to sane law-breakers. Psychiatric counselling or treatment is typically the best approach for insane law-breakers. Sometimes it is justifiable to make such counselling or treatment compulsory, if the ability of the individual to make decisions about her own treatment is compromised by mental illness.

However, the behaviour of sane offenders may change for the better if they come to see the force of the moral reasons against wrongdoing. It is widely accepted that rationality is compatible with determinism, even if retributive desert is not. Presenting offenders with moral reasons for reforming themselves shows respect for the offender's ability to grasp such reasons. As we saw in the example involving Timothy at the beginning of this chapter, it is important for the state to acknowledge and not deny positive qualities that citizens may have, even if the citizen is not retributively responsible for having those qualities. Rationality is a quality that sane offenders possess and which the state must recognise. This point will be further developed in subsequent chapters.

Sane offenders might also benefit from certain limited kinds of psychological treatment or enhancement. However, such interventions should only be given to the offender if the offender consents. The subsequent chapters provide a much fuller account of the conditions under which such interventions are permissible.

The Trial Process

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

However, as noted above, moral reasoning, rather than medical help is typically the appropriate means of enabling sane offenders to reform themselves. The trial process can serve as a vivid form of moral communication, which can help the offender to appreciate more fully the impact of her conduct on others and to resolve to change her behaviour.[161] It also shows respect for the offender's rationality and membership of the moral community to allow her to give an account of her conduct in court, before other members of the community.[162] (This point is also developed in subsequent chapters.)

Actual Conduct and Standards of Proof

Before a sentence can be imposed on a sane offender, it must be proved beyond reasonable doubt that the person committed a crime. This principle can be justified on a non-retributive basis. It upholds the value of liberty by protecting the individual against the power of the state. The state also shows respect for citizens by having a very strong presumption that those citizens are non-dangerous. Past behaviour is one

[161] See e.g. R Duff, *Punishment, Communication and Community* (OUP, Oxford 2001); R Duff, RA Duff, *Trials and Punishments* (CUP Cambridge 1986).
[162] See Duff (2001) and Duff (1986), fn161, above.

of the best guides to future behaviour.[163] It is therefore appropriate that proof that the individual has actually engaged in dangerous conduct should be a necessary condition of interfering with the freedom of sane individuals. The state also shows respect for citizens by having a very strong presumption that their conduct is guided by the fundamental moral values embodied in the criminal law.

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

Nevertheless, some of the reasons behind the requirement of proof beyond reasonable doubt of actual prohibited conduct do not apply as strongly (if at all) to mentally disordered persons as to sane individuals. Consider the liberty-based rationale. A person who is so mentally disordered as to be non-responsible is likely to have limited liberty anyway. If her powers of critical reflection and practical

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

reasoning are impaired, she may not be able to form stable, authentic values and goals and may not be able to pursue these effectively. So placing her under psychiatric supervision would not be as great a deprivation for her as confinement would be for a sane individual who was quite capable of forming and pursuing his own goals. Indeed, compulsory treatment of the mentally disordered person might actually increase her liberty, if the treatment successfully restores her powers of practical reasoning. Furthermore, the presumption that citizens guide their conduct by good moral reasons, cannot apply to individuals who have been shown to be too mentally disordered to grasp and apply these reasons.

Conclusion

This chapter has argued that four of the most important principles of justice have analogues that govern our treatment of insane offenders. Given that insane offenders do not deserve retributive punishment, this suggests that these principles of justice need not depend on retributive desert. Traditionally, punishment theorists have often wanted to draw a very sharp distinction between sane and insane law-breakers. This may have been motivated by the poor treatment that people with mental health problems have historically received. These theorists did not want sane offenders to be treated equally badly. However, the treatment of both types of offender would be improved if we focussed on the need to respect personhood and the principles of fairness that apply to all law-breakers.

Part Two: Free Will, Punishment and Moral Uncertainty

Part Two: Overview

As a result of the problems faced by compatibilists and libertarians, hard incompatibilists have argued that (at the very least) there is enough uncertainty about retributive moral responsibility to mean that the kind of serious harm involved in punishment should not be inflicted for purely retributive reasons.[164] Unfortunately, many hard incompatibilists then rush to embrace some form of consequentialism (i.e. the view that punishment is justified solely because it produces the best overall consequences). However, there is also considerable uncertainty surrounding the soundness of consequentialism. For instance, in principle, consequentialism could sanction 'punishing' people who have not actually committed an offence, if this promoted the social good. My thesis will examine how we should respond to the uncertainty surrounding the soundness of both consequentialism and retributivism.

Different kinds of uncertainty can enter into our deliberations about what we should do. Firstly, there is uncertainty relating solely to the non-normative facts, e.g. whether or not the accused was the person who actually killed the victim. There is a vast literature concerning non-normative uncertainty. However, surprisingly little attention has been paid to the topic of *moral* uncertainty – doubt concerning which moral theory should guide one's actions.[165] This category includes uncertainty about which theory of punishment is morally preferable (the focus of this thesis). It should

[164] E.g. R Double 'The Moral Hardness of Libertarians' (2002) 5 (2) Philo 226; B Vilhauer, 'Free Will and Reasonable Doubt' (2009) 46 (2) American Philosophical Quarterly 131; D Pereboom, *Living without Free Will* (CUP, Cambridge 2001); *G Harrison***,** 'Hooray! We're Not Morally Responsible!' (2009) 8 Think 87.

[165] Recent publications that discuss this issue in detail include: J Hudson, 'Subjectivization in Ethics'(1989) 26 American Philosophical Quarterly 221; G Oddie 'Moral Uncertainty and Human Embryo Experimentation' in K Fulford et al (eds), *Medicine and Moral Reasoning* (CUP, Cambridge 1994); A Guerrero, 'Don't Know, Don't Kill: Moral Ignorance, Culpability and Caution' (2007) 136 (1) Philosophical Studies 59; T Lockhart, *Moral Uncertainty and its Consequences* (OUP, Oxford 2000); J Ross, 'Rejecting Ethical Deflationism' (2006) 116 Ethics 742; A Sepielli, 'Review of Ted Lockhart's Moral Uncertainty and its Consequences' (2006) 116 Ethics 601; A Sepielli, 'What to Do When You Don't Know What to Do' in R Shafer-Landau (ed.),
Oxford Studies in Metaethics, Volume Four (OUP, Oxford 2009); B Vilhauer, 'Free Will and Reasonable Doubt' (2009) 46 (2) American Philosophical Quarterly 131; M Zimmerman, *Living With Uncertainty* (CUP, Cambridge 2009).

be noted that uncertainty about the soundness of a moral theory can stem from doubt about the non-normative facts – e.g. one may be uncertain about the validity of libertarian retributivism because one does not know whether or not human actions are undetermined.

How certain must we be that an argument for punishing a person is sound before it is justifiable to rely on it? Chapter Four will present an overview of some of the main theories of moral uncertainty and will highlight some of their key shortcomings (at least in relation to dealing with uncertainty about theories of punishment). Some theorists have argued that theories of punishment must be held to the 'beyond reasonable doubt' standard.[166] However, they have not defended this claim in sufficient detail. In Chapter Five I will examine the underlying rationale for the beyond reasonable doubt standard in criminal trials and will argue that it implies that we should hold the entire moral argument for punishing someone to a high standard of credibility. In Chapter Six, I will argue that in order to minimise the risk of punishing someone unjustifiably, we should only punish that person if the main punishment theories agree that doing so is justifiable. I will call this 'the convergence requirement'. I will give some reasons why theorists from different philosophical perspectives should endorse this requirement.

[166] See references in fn 164, above.

Chapter Four: Approaches to Moral Uncertainty

Introduction

This Chapter will briefly describe some of the current literature on moral uncertainty, before defending my own approach to the problem of moral uncertainty about punishment.

A Simple Approach

The simplest approach to the problem of moral uncertainty would be to act on the moral theory in which one has the most credence. In situations where two or more theories seem equally plausible this approach is, of course, of no help at all. In other situations it would be possible but deeply irrational to adopt the simple approach. To see this, first consider an example involving uncertainty about the non-moral facts: Imagine someone is deciding whether to drink a cup of coffee and has just slightly more credence in the idea that the coffee is safe than that it is poisoned. It would obviously be crazy to go ahead and drink the coffee, because the consequences would be so serious if it were true that the coffee is poisoned.[167] This suggests that in the parallel moral case, it would be irrational to focus only on how much credence one has in a view which states that a particular action is morally wrong and to ignore how *seriously* wrong the action would be if that view turned out to be correct.

Maximising Expected Moral Value

In some cases it would be reasonably straightforward to take into account both one's degree of belief in each competing moral theory and also the risk of serious wrongdoing. For instance, imagine that a particular action is morally neutral on

[167] This example is taken from Sepielli (2009), fn 165 above.

theory A and morally heinous according to theory B. Imagine that one thinks that both A and B are reasonable, defensible positions, but on balance one finds A slightly more plausible than B. It nevertheless seems that one should, morally, refrain from performing the action. However, things can get much more complicated. Consider the following case: A judge is wondering which sentence she ought, morally, to impose on a particular offender. Her credence is divided between utilitarianism and retributivism. Giving the offender one particular sentence would maximise utility (and so would be obligatory according to act utilitarianism), but that sentence would be disproportionate to the offender's moral desert (and so would be unjust according to retributivism). One approach to this problem involves factoring in how much utility is at stake and how disproportionate the sentence would be.

Taking this approach would require a method of comparing the level of moral goodness/badness each theory accords to each course of action. One might imagine, on the one hand, a scale indicating the retributive moral worth of possible actions, with massive injustice at the bottom and perfect justice at the other extreme, and, on the other hand, a scale of utilitarian moral worth, with actions that would cause a huge amount of disutility at one end and actions that would produce a huge increase in utility at the other end. How could these two scales be made commensurable?

Ted Lockhart is one of the very few authors who have tried to tackle this question in detail in the context of moral uncertainty. He does not specifically consider the issue of moral uncertainty surrounding punishment theories (focussing instead on examples from medical ethics, and the ethics of charitable giving). However, his theory is meant to have general application to all cases of moral uncertainty. Lockhart defends what he calls the 'Principle of Equity among Moral Theories', which states that:

> 'The maximum degrees of moral rightness of all possible actions in a situation according to competing moral theories

> should be considered equal. The minimum degrees of moral rightness of possible actions in a situation according to competing theories should be considered equal unless all possible actions are equally right according to one of the theories (in which case all of the actions should be considered to be maximally right according to that theory).'[168]

In other words, if we rate the best possible action the person could do in that situation according to Theory A as scoring '10/10' on the scale of moral worth, then we must also take Theory B to accord 10/10 to whichever action Theory B views as the best action in that situation. If we rate the worst possible action in that situation according to Theory A as 0/10 then Theory B must be considered to give 0/10 to whichever action Theory B rates as the worst. If a theory says that all actions available in that situation are right and does not prefer one action over another then all actions are considered to score 10/10 on that theory.

Unfortunately, in many cases Lockhart's method of scoring degrees of moral rightness/wrongness completely defeats the purpose of trying to take the degree of rightness/wrongness into account in the first place. Consider applying Lockhart's principle to the following situation involving moral uncertainty about punishment theories. A judge has only two possible options – to acquit or to convict and is unsure what is the morally right thing to do. His credence is divided between a utilitarian theory (U) and a form of retributivism (R). According to R, the accused deserves to be acquitted, because he is morally blameless. According to U, the accused should be convicted in order to maximise utility. It seems that, for R, the best available option (the action which gets 10/10) is acquitting the accused and the worst (0/10) is convicting him. For U the situation is exactly reversed. This can be represented as follows:

[168] Lockhart (2000), fn 165 above, p84.

	U is True	**R is true**
Convicting:	10/10	0/10
Acquitting:	0/10	10/10

It seems that in this situation, according to Lockhart's principle, the judge must just try to work out which theory he thinks is more likely to be true – looking at the scores each theory gives each action cannot help him decide what to do. Yet this completely ignores the fact that certain situations will involve issues that are of tremendous moral significance according to one theory, but where according to a rival theory, nothing of very great importance is at stake. For it may be that, for one theory, the best option available *in this particular situation* is just about the best option conceivable in *any* situation and the worst possible option in this situation is just about the worst conceivable option in any situation.[169] In contrast, for another theory, it may be that the best option available in this situation is merely OK compared with the available options in other situations, and the worst available option in this situation is merely slightly bad, when compared with the options for wrong-doing available in other situations. To return to theories of punishment, imagine that, according to R, convicting the accused would involve a massive injustice. Acquittal is the best available option for R, so Lockhart would give an

[169] For related criticisms see: Ross (2006), fn 165 above and Sepielli (2009), fn165 above.

acquittal a score of 10/10. However, imagine that according to U convicting the accused is only slightly better than acquitting him. Lockhart would still insist that the conviction is given a score of 10/10, as long as this is the best option available in this situation according to U. This kind of scoring completely distorts things.

Andrew Sepielli attempts to deal with moral uncertainty by using an alternative method of comparing different theories' conceptions of moral value. He says that we can get some idea about the value that different theories accord to different courses of action by comparing each action to other actions that are generally agreed to have a certain value. He gives the following example. A woman is deliberating about whether or not to have an abortion. Her credence is divided between theory A and theory B. Theory A says that abortion is really bad – as bad as murder – and that *not* having an abortion is not wrong at all – it might be compared to any morally permissible action, e.g. using an innocuous form of birth control, such as the rhythm method. Theory B says that abortion is not wrong at all – it has the moral value of, say, the rhythm method - and that failing to have an abortion would be worse than not having one, but not as bad as, for instance, murder. So, Sepielli concludes, we can say that according to A the difference between having an abortion and not having one is as big as the difference between murder and using the rhythm method (the latter difference being uncontroversially very big indeed), whereas B treats the gap between having an abortion and not having one as being smaller than the gap between murder and the rhythm method.[170] Now, Sepielli's 'comparators' are fairly plausible because we can see more or less in what respects abortion might be thought by some to resemble contraception and by others to resemble murder. However, it is less clear how this approach would work for theories of punishment. It is not

[170] Of course, the woman's deliberations do not end there because she has to factor into the equation exactly how probable she considers each theory to be.

obvious, for instance, what action might be compared with the wrongness a retributivist would accord to a case of injustice.

The Presumption Approach

Both Sepielli's and Lockhart's approaches to the problem of uncertainty surrounding moral theories seek to work out the 'expected moral value' of an action. This involves determining the moral goodness/badness of the action if a particular theory is true, factoring in the probability of the theory actually being true and going through this process for each competing theory. These methods might be called 'maximization strategies' since they all recommend performing whichever action has the maximum expected moral value. A markedly different approach to the problem of uncertainty surrounding the soundness of punishment theories might be called the 'presumption approach'. Ben Vilhauer, for instance, proposes that there should be a presumption against punishment which the punishment theorist must rebut, by establishing her argument to an extremely high standard of credibility.[171] According to Vilhauer, the justifiability of this presumption is something about which there is a relatively high degree of moral certainty – most people, he claims, would agree that the deliberate infliction of serious harm on an individual by the state is something that is *prima facie* wrong and which should only be a last resort. He proposes that a moral argument for punishing a particular person must be established to be sound 'beyond reasonable doubt', before it is appropriate to punish the person on the basis of that argument. Vilhauer's approach could produce very different results from maximization strategies. For instance, a maximization theorist might claim that we should punish a person on the basis of a penal theory that is merely 'probably sound' as long as the potential moral benefits of punishing the person (if that theory were in fact sound) would be so substantial that they outweighed the potential wrongfulness

[171] Vilhauer (2009), fn165 above.

involved in punishing him if a competing theory that opposed punishing him turned out to be correct.

Presumption theories avoid the commensurability problem involved in trying to weigh the moral benefits of punishment on one theory against the moral wrongfulness of punishment on a different theory, where the theories have different conceptions of 'value'. Yet they face their own problems, including the challenge of explaining what precisely it means to say that a moral argument has been established 'beyond reasonable doubt'. Vilhauer argues that no retributive argument for punishing anyone can be established beyond reasonable doubt. This, he claims, is because retributive punishment is only justifiable if the offender had free will with respect to his offence. He argues that the fact that the free will debate is widely regarded as being 'philosophically valuable' indicates that it can reasonably be doubted that anyone ever has free will. He suggests that consequentialist arguments for punishing people may meet his proposed standard. However, this claim is dubious, for there is also a philosophically valuable debate about the soundness of consequentialism.[172] By parity of reasoning, this indicates that consequentialist justifications of punishment also fail to meet the beyond reasonable doubt standard.

The problem of moral uncertainty surrounding the soundness of punishment theories has received insufficient attention. Existing approaches to moral uncertainty encounter serious difficulties. In the next two chapters I will argue that Vilhauer is right to say that the arguments for punishing people must be held to a high standard of credibility. However, his precise definition of the beyond reasonable doubt standard is inappropriate to the context of theories of punishment. In Chapter 6 I will argue that a person should only be punished if the main theories of punishment agree

[172] Elsewhere, Vilhauer defends a Rawlsian contractarian theory of punishment. However, there is surely also a philosophically valuable debate about the soundness of this theory. B Vilhauer, 'Free Will Skepticism and Personhood as a Desert Base' (2009) 39 (3) Canadian Journal of Philosophy 489.

that punishing that person is appropriate – I call this 'the convergence requirement'. If an argument for punishing someone satisfies the convergence requirement then it has reached a sufficiently high standard of credibility – the nearest arguments for punishment can get to being established beyond reasonable doubt.

Chapter Five: The Rationale for the Beyond Reasonable Doubt Standard

Introduction

In Part One, I drew attention to the serious doubts about the adequacy of retributivism that are raised by the free will debate and I have also pointed to difficulties with retributivism's main rival – consequentialism. In this chapter, I will relate these doubts about punishment to the 'beyond reasonable doubt' (BRD) standard of proof in criminal trials. I argue that it would be arbitrary to maintain that an accused's criminal responsibility must be proved beyond a reasonable doubt before it is fair to punish her, but to fail to take seriously the issue of doubt surrounding the soundness of the conception of responsibility and punishment being applied. I argue that we should therefore hold the entire moral argument for punishing a person to a high standard of credibility. (Although, as I will explain in Chapter 6, it does not follow that all elements of the argument must be held to *precisely the BRD standard*).

I will consider two ways in which this proposal could be challenged: Firstly, the beyond reasonable doubt standard itself might be rejected. If we can endorse a lower standard of proof in criminal trials, then we might also be satisfied with a similarly low credence in the moral justification for punishing people. Secondly, perhaps there are good reasons why the beyond reasonable doubt standard should only apply in its current context and should not be extended in the way I propose. In order to address these challenges it is necessary to look at the underlying justification(s) for the beyond reasonable doubt standard.

The BRD standard has extremely widespread support among legal theorists[173] and is a central principle of most adversarial legal systems.[174] Therefore, the first strategy of abandoning this principle altogether would be a tough bullet to bite. In addition, the BRD standard is supported by a number of key non-consequentialist principles including: the doing/allowing distinction; the foreseen/intended distinction; and arguments relating to the state's moral authority to punish. These principles, in turn would be costly for non-consequentialists to abandon, since they each have a plausible moral basis and they help to explain widespread intuitions in a range of different cases. The non-consequentialist principles that can justify the BRD standard are also *general ethical principles* (that can be endorsed by a variety of punishment theorists). The rationale for the BRD standard does not stem directly from any specific *theory of punishment,* e.g. retributivism. I will also argue that the most plausible justifications for the BRD standard provide no basis for restricting the BRD solely to its current context in criminal trials. The entire moral argument for inflicting serious harm on offenders should be held to a high standard of credibility.

Consequentialists on the whole are also reluctant to reject the BRD standard (with some notable exceptions).[175] However, this chapter will primarily provide reasons why non-consequentialists should support the BRD standard. In the next chapter, I

[173] Patrick Tomlin notes that even theorists who strongly disagree about how courts should interpret the BRD standard still agree on the importance of the principle and agree that its underlying normative justification is based on the grave wrongness of mistaken convictions: 'Extending the Golden Thread? Criminalisation and the Presumption of Innocence' The Journal of Political Philosophy (Forthcoming).

[174] See e.g. A v HM Advocate 2003 S.L.T. 497; Woolmington v DPP [1935] AC 462; Re Winship 397 U.S. 358 (1970), at 364. In these jurisdictions the BRD standard is considered to be required by the presumption of innocence. Hock Lai Ho points out that even in inquisitorial systems, the presumption of innocence still requires that the state must *prove* guilt: 'The Presumption of Innocence as Human Right' in Roberts P (ed) *Criminal Evidence and Human Rights: Reimagining Common Law Procedural Traditions* (Hart Publishing, Oxford 2012), 259, p262.

[175] For a consequentialist defence of the BRD standard see: Rizzolli M and Saraceno M, 'Better That Ten Guilty Persons Escape: Punishment Costs Explain The Standard Of Evidence' (2011) Public Choice DOI: 10.1007/s11127-011-9867-y (online first). For a consequentialist critique of the BRD standard see: Laudan L, 'The Rules of Trial, Political Morality, and the Costs of Error: Or, is Proof Beyond a Reasonable Doubt Doing More Harm than Good' in L Green and B Leiter (eds) *Oxford Studies in Philosophy of Law* (OUP, Oxford 2011).

will address those who are not committed to the non-consequentialist principles discussed in the present chapter and will argue that they too should demand that the state's moral argument for punishing people is held to a high standard of credibility.

Arguments Against a Retributive Basis for the BRD Standard

I will begin by critiquing a possible justification for the BRD standard that derives specifically from retributivism. My thesis aims to show that there are serious doubts about the soundness of pure retributive and pure consequentialist conceptions of responsibility and punishment; that (by analogy with the BRD standard) we should hold arguments for punishing people to a high standard of credibility; and that, therefore, such arguments should not rely on either pure consequentialism or pure retributivism alone. If the 'real' reason for the BRD standard were retributive, then this could pose problems for my argument. A critic might object that since the BRD standard would only be defensible if retributivism were correct, and there is little doubt about the appropriateness of the BRD standard, then is little doubt about the appropriateness of retributivism. Or, conversely, the critic might argue that if the beyond reasonable doubt principle is sound, then retributivism must be sound, but retributivism is faulty and therefore the beyond reasonable doubt standard is unsound. My thesis aims to question the soundness of retributivism, without undermining the BRD standard. So it is important to show that the BRD standard does not depend on retributivism.

Jeffrey Reiman tries to derive the BRD standard directly from retributivism itself.[176] Retributivism contains two requirements – 1) punish the guilty (positive

[176] Reiman J and Van Den Haag E, 'On the Common Saying that it is Better that Ten Guilty Persons Escape than that One Innocent Suffer: *Pro* and *Con'* 7(2) (1990) Social Philosophy and Policy 226. (Henceforth: Reiman and Van den Haag, 'On the Common Saying'.)

retributivism); and 2) do not punish the innocent (negative retributivism). Confining punishment to only those people whose guilt has been established to the BRD standard seems to privilege the second duty over the first. Reiman argues that retributivism has the conceptual tools to justify this. He tries to find a way of comparing the relative stringency of the two retributive obligations. He does this by comparing the seriousness of failing to fulfil these obligations with the seriousness of criminal offences. He claims that punishing an innocent person for a crime (i.e. a failure to fulfil duty 2) is about as seriously wrong as the crime itself - for example, punishing an innocent person for murder is about as wrong as murder. In contrast, he claims, failing to punish an actual murderer (i.e. a failure in duty 1) is not as bad as murder.

In order for Reiman's account to succeed he must first explain (in retributive terms) why he has chosen the seriousness of crimes as the 'comparator' for measuring the seriousness of failures to achieve retributive justice. Secondly, he must find a rational basis for his claim that punishing the innocent and committing crimes are equally bad, but that it is less bad to let guilty people go free - to do this he must find a way of assessing the degree of badness which each of these things involves.

He justifies his choice of comparator with reference to the *lex talionis* – the idea (historically associated with retributivism) that making the punishment 'fit' the crime requires that 'criminals ought to be punished (as nearly as is feasible) with harm equivalent to that which they intentionally caused their victims'.[177] The *lex talionis* invites us to compare the seriousness of punishments to the seriousness of crimes (in common with other retributive proportionality requirements).

[177] Reiman and Van den Haag, 'On the Common Saying', p230.

The *lex talionis* also seems to assess the gravity of the offence in terms of the harm done to the victim. Similarly, Reiman proposes that the badness of failing to fulfill the two retributive duties can be measured in terms of the 'palpable harm' that this inflicts on the people to whom the duties are owed; i.e. the extent to which this failure 'subtracts…from his ability to pursue his own purposes'.[178] According to Reiman, the duty not to punish the innocent is owed to the innocent. He is less sure to whom the duty to punish the guilty is owed, but he considers three possibilities: guilty criminals themselves (who may have a 'right' to be punished); victims (and their friends and family) and the law-abiding population.[179] Reiman states that punishing an innocent person for crime X harms the punished person (i.e. interferes with their purposes) to roughly the same degree as committing crime X against an innocent victim harms that victim. However, failure to punish a guilty person for crime X *per se* does not harm the criminal, the victim or the law-abiding population to the same degree as crime X harmed the victim. Reiman's calculation might work out differently if we took into account factors such as failure to prevent the mistakenly acquitted person from re-offending. However, Reiman justifies his narrow focus on the 'direct negative impact' of these failures of duty to the persons to whom they are owed, because any consideration of 'further losses' would 'cross the boundary' that distinguishes consequentialism from retributivism.[180] If Reiman's reasoning were correct, then this would provide a retributive basis for the claim that punishing the innocent is worse than acquitting the guilty and thus a retributive justification for the BRD standard.

However, there are problems with Reiman's account. The *lex talionis* measures the *amount of punishment* that criminals deserve by comparison with harm to individual

[178] Reiman and Van den Haag, 'On the Common Saying', p234.

[179] Reiman states that the duty to punish is owed to 'victims' without explicitly mentioning whether this includes friends and family of direct victims. If it does not, it is puzzling what Reiman would say about murder victims. See V Halvorsen, 'Is it Better that Ten Guilty Persons Go Free than that One Innocent Person be Convicted?' 23 (2004) Criminal Justice Ethics 3. (Henceforth: Halvorsen, 'Ten Guilty Persons').

[180] Reiman and Van den Haag, 'On the Common Saying', p232.

crime victims. Even if it were acceptable to focus on 'harm' in that context, it is dubious to measure the stringency of the retributive *duty to punish* or the *wrongness of failing to punish* in terms of 'harm' to individuals. The harm to individuals that results from the failure to punish a particular wrongdoer can vary, depending on factors including how much the victim or others want to see the wrongdoer harmed (which sometimes depends on how vengeful they are, or how popular the victim was); whether the victim is still alive and whether they have any living friends or relatives. But all of these factors seem largely irrelevant to the question whether there is a duty to punish the murderer of an unpopular orphan, or how stringent that duty is or how bad a breach of that duty would be.

Furthermore, measuring the stringency of the duty to punish in terms of the 'harm' caused by failing to punish seems to involve a significant departure from the spirit of retributivism. The retributive duty to punish the guilty (and its stringency) seems more plausibly to stem from the demands of abstract justice, or perhaps, as Ernest van den Haag puts it, from an obligation to the 'perpetual moral community' (even at the *expense of* harm to the contemporary community).[181] Reiman's focus solely on narrow class of harms seems arbitrary and does not succeed in transforming his account into a retributive one.

Even if Reiman could justify his reliance on the notion of 'harm' (in his restrictive sense) this would not necessarily help his argument. Reiman asks: 'for a given crime and its appropriate punishment, does failing to punish the guilty criminal impose a loss on his victim that is as bad as the loss that the crime imposed on the victim?' He answers that the crime itself would always cause the victim to 'suffer more' than the mere failure to punish the victimizer. He then asks whether punishing an innocent

[181] Reiman and Van den Haag, 'On the Common Saying', p242.

person would cause that person to suffer as much as the original victim of crime and answers that it would. Therefore, he concludes that punishing innocent people is worse than failing to punish guilty people. However, his statements about the relative suffering caused by failure to punish, being a crime victim and being unjustly punished are not always correct. Consider the following example:

Imagine that a criminal, motivated by prejudice, assaults a victim. The assault was minor. The main harm the victim suffered was psychological distress. In this case, the failure of the system to bring that bigot to justice may be as distressing for the victim as the original offence. The victim might rather be assaulted again and have both bigots punished, than let the bigot get away with this crime.

Furthermore, one can imagine that some unjustly punished people suffer less than crime victims. Some unjustly punished people may be innocent of a specific crime, but may also have a string of just convictions. For some such people, the suffering caused by a single false conviction may not be as great as the suffering that crime victims experience. (How much unjustly convicted people suffer can plausibly be affected by how accustomed they are to the prison environment). Yet the BRD standard rightly protects people who may be innocent of the particular crime with which they are charged, even though they are not wholly 'innocent' in a broad sense.

Perhaps people would *typically* suffer more from being subject to crime or to unjust punishment than from seeing guilty people go free. However, Reiman claims that the relative stringency of the duty to punish the guilty and the duty to acquit the innocent depends on the relative amounts of harm that result from failures in these duties. If this were correct, then presumably in the 'exceptional' cases where failure to punish the guilty would cause more harm than inflicting undeserved punishment, the duty to punish the guilty would become more stringent than the duty not to punish the innocent. This would lead us to lower the standard of proof in such cases if that would increase the chance of the guilty person being punished, even at the expense

of risking punishing an innocent person. Some consequentialists might be content to vary the standard of proof in this way, but it does not seem to be something that a retributivist could endorse and it does not seem just.[182]

My own defence of the BRD standard is not based on the essential nature of the retributive duties. Instead, I will invoke general non-consequentialist ethical principles such as the requirement that inflicting active, intentional harm must be justified to a very high standard. It is certainly *open* to retributivists to appeal to these principles, but they are *independent* from retributivism, constraining the positive retributive duty to punish. There is nothing in the nature of the retributive duties *per se*, which tells us which retributive duty is more stringent. Reference must be made to these broader principles. My defence of the BRD standard does not imply that the standard of proof should vary from case to case. Punishment in every case involves the active and intentional infliction of harm. True, the severity of the punishment varies. But all punishment is above the threshold which attracts the protection of these non-consequentialist safeguards**.** Reiman cannot appeal to this threshold idea, because his theory depends on measuring the precise *level* of harm that failing to acquit the innocent causes versus failure to convict the guilty. On his theory, the first failure is *only* worse than the second, because the *level* of harm that the first failure involves is (allegedly) greater. It follows that, in cases where this is not true, punishing the guilty should be prioritised over acquitting the innocent. In contrast, on my account, we need a high standard of proof because when the state punishes innocent people it *always* inflicts harm actively and intentionally and this needs to be strongly justified. Whereas when it fails to punish guilty people because the evidence has not met the required standard, the harm that flows from that is merely allowed and is unintentional.

[182]Cf: Lilquist E, 'Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability' 36 (2003) University of California Davis Law Review 85.

In the next two sections I will develop my justification for the BRD standard, based on the doing/allowing and intention/side-effect distinctions. As, I have said, retributivists (together with other non-consequentialists) can consistently appeal to these distinctions in order to justify the BRD standard. The fact that most retributivists recognise the importance of the BRD standard could motivate them to endorse this type of justification for it. But since there is nothing about these distinctions that ties them to the context of the criminal trial, they also provide a rationale for applying a high standard of credibility to theories of punishment, including retributivism.

The Doing/Allowing Distinction and the BRD Standard

The purpose of this section is to argue that theorists who endorse the doctrine of doing and allowing have reason to support the BRD standard and that such theorists also have reason to apply the BRD standard (or a similarly high standard of justification) to the entire moral argument for inflicting serious harm on offenders. I will begin with a brief description of the doctrine of doing and allowing and with an indication of why many theorists consider the doctrine to be important. I will then rebut some of the main arguments that have been presented by those who doubt that the doctrine can provide a basis for the BRD standard. My discussion will be focused on how the doctrine of doing and allowing relates to the BRD standard. I will not attempt to engage with the debate on whether it is possible to distinguish between doing and allowing at all, or with the wider debate about whether this distinction (if it can be drawn) is ethically significant. This section, is therefore, primarily addressed to theorists who find the doctrine of doing and allowing plausible. However, this chapter also criticises some of the main alternative strategies for

defending the BRD standard that do not depend on the doctrine of doing and allowing.[183] These criticisms may provide some motivation for those who endorse the BRD standard also to endorse the strategy for defending it that draws on the doctrine of doing and allowing. The position of those who reject the doctrine of doing and allowing will be further discussed in chapter 6, where I will present some different reasons (including meta-theoretical reasons) for holding the moral argument for inflicting serious harm on offenders to a high standard of justification.

The doctrine of doing and allowing (or DDA) states that it is harder to justify doing than allowing harm. Fiona Woollard has recently provided a sophisticated account of what this doctrine is and why it is important, which draws together themes from various other proponents of the doctrine.[184] Here is Woollard's brief summary of what she takes to be the doctrine's significance: 'The DDA should be understood as a principle that protects us from harmful imposition. When an agent does harm, he imposes on the victim. When an agent is required to prevent harm he is imposed upon by the potential victim. Protection against both types of imposition, as provided by the DDA is required to recognise our authority over what belongs to us.'[185] In other words, firstly, if the DDA were false and deciding actively to harm another person did not require particularly strong justification, then we would all become excessively vulnerable to being harmed by other agents in the course of pursuing their goals. Having authority over oneself and one's resources, however, implies that we have a powerful right not to be interfered with by other agents, which would take

[183] See my discussion of Reiman's retributive strategy in the previous section, as well as my discussion of Tadros's deterrence strategy, Lee's social contract strategy, and a condemnation-based strategy discussed in subsequent sections.
[184] See e.g., Woollard, 'If This Is My Body…: A Defence of the Doctrine of Doing and Allowing' (2013) 94(3) Pacific Philosophical Quarterly 315; Woollard F, 'The Doctrine of Doing and Allowing I: Analysis of the Doing/Allowing Distinction' 7(7) (2012) Philosophy Compass 448 [henceforth: Woollard, 'DDA I']; Woollard F, 'The Doctrine of Doing and Allowing II: The Moral Relevance of the Doing/Allowing Distinction'7 (7) (2012) Philosophy Compass 459 [henceforth: Woollard, 'DDA II'].
[185] Woollard, 'DDA II', p466.

considerable justification to overcome. Secondly, without the DDA, morality becomes excessively demanding: if our obligations to prevent harm were as powerful as our obligation not to do harm, we would have to give up our resources and make ourselves liable to injury or death in a very wide range of circumstances. But if, as the DDA states, our obligations to prevent harm are weaker than our obligations not to do harm, this sets limits to the claims others can make on us and thus helps to safeguard our authority over our bodies and resources.

Supporters of the DDA should hold justifications for harming offenders to a high standard of credibility for the following reason: When the state inflicts serious harm on offenders in response to their crimes, it does so *actively,* whereas the harms that result from failing to impose hardship on offenders are merely allowed to occur.

Reiman dismisses the act/omission distinction (and presumably also the doing/allowing distinction) as a basis for determining whether the state should prioritise sparing the innocent over punishing the guilty.[186] He points out that the state has a *pre-existing obligation* to punish the guilty. Where someone fails to discharge a pre-existing obligation, they cannot 'get an automatic moral discount' because that failure is an omission. For example, doctors who intentionally refuse to

[186] Reiman and Van den Haag, 'On the Common Saying', p229. Jeff McMahan points out that the act/omission distinction is not identical to the doing/allowing distinction. He gives the following example: Imagine a thief steals my wallet because he needs the money to pay for an operation to prevent him becoming disfigured. I run after him and take the wallet back; the thief cannot afford the operation and so becomes disfigured. He has become disfigured partly as a result of my act. But (assuming I know his motives) I have merely allowed him to become disfigured, by withholding my resources from him: 'A Challenge to Common Sense Morality' (1998) 108 (2) Ethics 394, p411. Similarly, the acquittal of a factually guilty person can result from various acts by state officials (including the choice of the high standard of proof), but the harm to victims and others that results from this is merely allowed by the state. Cf Cass Sunstein's and Adrian Vermeule's argument that the state cannot appeal to the act/omission distinction to diminish its responsibility for the harm criminals cause to victims, because some of those crimes could have been prevented if the state had made different policy choices and policy choices are 'acts': 'Is Capital Punishment Morally Required? Acts, Omissions and Life-Life Tradeoffs' (2005) 58 Stanford Law Review 703. Youngjae Lee also observes that the Sunstein-Vermeule argument ignores the doing/allowing distinction: 'Deontology, Political Morality, and the State' 8 (2011) Ohio State Journal of Criminal Law 385, pp388-390.

treat their patients or parents who intentionally fail to feed their children are, according to Reiman, 'morally indistinguishable' from doctors who actively kill their patients or parents who actively take food away from their children.[187] Similarly, he concludes that:

> ' [If the state's] obligation to punish the guilty is as strong as their obligation not to punish the innocent, then failure at the first obligation is as bad as failure at the second. This leaves us with the task of determining the relative strength of these obligations, which is where we were before taking up the acts/omissions distinction.'[188]

Thought-experiments designed to undermine the doing/allowing distinction, typically involve agents who are motivated by clearly inadequate or immoral considerations.[189] The agents in Reiman's examples apparently want or intend their victims to come to harm and that is what explains their decision actively to cause or allow this to happen. Even where agents have such bad motives, some theorists have argued that the doing/allowing distinction is still morally relevant. Halvorsen argues that a lifeguard who fails to rescue a swimmer (though still very blameworthy) is less culpable than a lifeguard who actively holds a swimmer's head under the water until he drowns.[190] This seems plausible even where both lifeguards are motivated by hostility towards the swimmer. However, the significance of the doing/allowing distinction becomes even clearer when the agent is motivated by something of significant moral value. Consider the following cases:

Allowing Harm – Diverting Resources

[187] Reiman and Van den Haag, 'On the Common Saying', p229.
[188] Reiman and Van den Haag, 'On the Common Saying', p229.
[189] E.g. James Rachels' famous 'Wicked Uncle' case: 'Active and Passive Euthanasia' (1975) 292 New England Journal of Medicine 78.
[190] Halvorsen, 'Ten Guilty Persons', p11.

Imagine a doctor who diverts resources away from some patients, thereby allowing some harm to come to them, in order to treat even more needy patients. We normally think that this is permissible. [191]

Doing Harm – Medicine Machine

Now imagine that there is a machine in a room in a hospital that can produce medicine for some very sick patients. However, the doctor knows that there is a patient in that room who cannot immediately be moved and, if the machine is switched on, it will emit toxic fumes (as a side-effect of the production process) which would harm that patient. It seems that the doctor should not turn on the machine until the patient can be moved, even though this delay is likely to harm the patients who need their medicine.[192]

In these examples, the doctor has a duty of care to all the patients concerned, but how he carries out his duty is constrained by the powerful considerations against active harming. Similarly, the state has a duty to respond to crime in a just manner – to coercively interfere with offenders who ought to be subject to such interference and not interfere with individuals where this is unjustified, e.g. where the individual is innocent - but how it carries out its duty is likewise constrained by the powerful considerations against active harming. These considerations are so powerful that they can only be rebutted by an argument that is at least as compelling. Hence the prosecution's case for convicting the accused will only succeed if it meets the demanding BRD standard. Likewise, more generally, seriously harming offenders is only permissible if the moral justification for doing so meets a high standard of credibility.

---

[191] V Tadros, *The Ends of Harm: The Moral Foundations of the Criminal Law* (OUP, Oxford 2011), p120.

[192] This example is a variation on a thought-experiment discussed in Philippa Foot, 'The Problem of Abortion and the Doctrine of Double Effect' in P Foot, *Virtues and Vices and Other Essays in Moral Philosophy* (OUP, Oxford 2002), 19.

Youngjae Lee attempts to formulate a better counter-example, in which the agent's motives are good and where she is justified in choosing to do harm to someone instead of allowing harm to befall others.[193] He imagines several variations on a scenario from the novel Sophie's Choice.[194] These variations are progressively altered to make them more analogous to the state's decision about where to set the standard of proof.

Sophie's Choice

In the scenario, there is a war and Sophie is taken to a prison camp. In the first variation, a sadistic prison guard tells Sophie that he will release five children (whom she has never met) if she kills Bruno (a man whom she has never met). Otherwise he will release Bruno and the five children will stay in the dangerous prison camp. Lee then replaces the unknown children with Sophie's own children, on the basis that Sophie's parental duty is analogous to the state's obligation to protect victims of crime. He also adds that at the beginning of the war Sophie's husband was murdered and 'it is more likely than not' that Bruno is the murderer, although this could not be established beyond reasonable doubt.

Lee argues that it may be harder for Sophie to justify killing Bruno in the first variation, where the children she is trying to save are strangers to her. However, when we imagine that Sophie is their mother, Lee claims that her positive duty to protect them may make it 'mandatory for her to violate the negative duty owed to Bruno [not to kill him]', and this conclusion, he claims, is strengthened when we factor in that Bruno is probably a murderer.[195]

[193] 'Deontology, Political Morality, and the State' 8 (2011) Ohio State Journal of Criminal Law 385. (Henceforth: Lee, 'Deontology'.)
[194] William Styron, *Sophie's Choice* (The Modern Library, New York 1998).
[195] Lee, 'Deontology', p392.

Sophie must choose between the lives of her own children, and the life of Bruno - a stranger and probable murderer. Lee claims this is analogous to the state's choice between protecting 'its citizens' and giving probable 'criminals' the benefit of the BRD standard.[196] This way of framing the problem is incorrect, because the state has also got a duty to protect accused people. To make the Sophie case somewhat more analogous, imagine that Bruno is one of her children. The claim that she is required to kill her child in order that her other children will be released seems less persuasive. The idea that Sophie is morally required to base this decision on the 51% chance that her child is a murderer is also problematic. It is far from clear that Sophie would be doing something wrong if she refused to believe that her child had done such a terrible thing and therefore dismissed this consideration. As CS Lewis writes, 'to love involves trusting the beloved beyond the evidence, even against much evidence...Such confidence...is in fact almost universally praised as a moral beauty....'. [197] As I will argue in more detail below, trust is an important part of other kinds of relationships too, in particular, between the state and citizens. It might be argued that given the terrible choice that Sophie is forced to make, the normal attitude of trust should be abandoned. Even if this were true, the loss of trust would be a tragic feature of this kind of emergency situation, and hardly a principle that a society under ordinary conditions should enshrine in its institutions.

In a footnote, Lee envisages the objection that the state's duty of protection is not limited to innocent victims of crime, but extends to criminals and to accused people. However, he denies that this undermines his analogy. Even if Bruno were Sophie's child, he argues, the 'the fact that she has a duty to protect Bruno from harm does not give her an additional reason not to kill Bruno'.[198] This seems very counter intuitive.

[196] Lee, 'Deontology', p392-393.
[197] CS Lewis, 'On Obstinacy in Belief' (1955) 63(4) The Sewanee Review 525, p535.
[198] Lee, 'Deontology', p393, fn28.

While he insists that Sophie's positive obligation to protect her children is more stringent than her duty to prevent other people's children from being harmed, he denies that Sophie's protective role as a mother strengthens her negative obligation not to kill her children. This negative obligation is, according to Lee, no more stringent than her general negative duty not to kill strangers. Therefore, if we assume that Sophie has done her best to protect all her children including Bruno (e.g. by attempting to help them escape) and has 'run out of such options', she has discharged her parental duty to Bruno and it is no longer relevant to her decision to kill him.[199] Lee's argument rests on a peculiar asymmetry between positive and negative duties: the duty to protect, on his account, only strengthens the positive duty to aid the protected person, and in no way strengthens the negative duty not to harm the protected person. He compares the protector to a debtor who pays back his creditor and then steals from him. The debtor's positive duty to pay the creditor has been discharged and does not give him an additional reason not to steal from his creditor. The debtor simply has the general negative duty we all have not to steal. A flaw in this analogy is the fact that protectors (such as parents, or the state) have a duty to protect their children/citizens, but the debtor does not have a positive duty to *protect* the creditor's assets. A closer analogy would be a security guard who steals the thing he is meant to be protecting from thieves. His position of responsibility does seem to give him an additional reason not to steal. Similarly, as I will argue below, the state's duty to protect its citizens does give it an additional reason to refrain from convicting and punishing them unless there is extremely good evidence that this is justified.

Even if Sophie were clearly required actively to kill one of her children in order to save the others, this cannot undermine the general validity of the doing/allowing distinction. Nor can it undermine the use of this distinction to defend the BRD

[199] Lee, 'Deontology', p393, fn28.

standard in the context of punishment. Sophie makes her choice in an emergency situation where civil society has broken down. In such emergency situations important values are often sacrificed. These values may include trust, freedom and the doing/allowing distinction among others. But we should not model our criminal justice system according to the moral norms governing captives struggling for survival in a concentration camp. Social life would be hellish if we did. My earlier hospital analogy (which supports the doing/allowing distinction) taps into intuitions that are of greater relevance to state punishment. Hospitals, like the penal system need to adopt policies that are appropriate for everyday life in relatively peaceful societies. In such societies we cannot be willing actively to inflict serious harm on others, unless this is justified to a very high standard.

Larry Laudan attempts to find examples from our ordinary social practices that undermine the doing/allowing distinction.[200] He argues that, in fact, such examples can be found within the criminal justice system. He cites several cases where he claims we tolerate the risk of unjustifiably doing harm to individuals, rather than allow a greater harm to occur. For instance, confessions are admissible in evidence. But we know that accused people sometimes make false confessions, so admitting confession evidence at all means that we are willing to risk convicting innocent people. Even the BRD standard itself, does not make convicting innocent people impossible. High as it is, the BRD standard demands less than absolute certainty. To completely avoid the risk of the state doing harm to innocent people, punishment would need to be abolished altogether.

However, Laudan's examples do not provide evidence against the doing/allowing distinction as generally understood. In fact, he is arguing against a straw man. Few

[200] He actually uses the terms 'omission/commission': Laudan L, 'The Rules of Trial, Political Morality, and the Costs of Error: Or, is Proof Beyond a Reasonable Doubt Doing More Harm than Good' in L Green and B Leiter (eds) Oxford Studies in Philosophy of Law (OUP, Oxford 2011).

proponents of the doing/allowing distinction hold that actively causing harm is *never* justified, or is only permissible if we are *completely* certain that we are doing the right thing (assuming we can be completely certain of anything). It is clear that our society does not, and has never taken this approach to harmful acts. As I have explained above, the doing/allowing distinction just states that it is *harder* to justify doing harm than allowing harm.

Whether a harmful act can be justified, in a particular context, depends on a variety of considerations. For example, it can be permissible to cause harm in the course of protecting a particularly powerful right, e.g. the right to self-defence. But even in self-defence cases, the doing/allowing distinction is still relevant. Because doing harm to others is so serious, strict constraints are placed on the use of force in self-defence. Punishment is analogous to self-defence in some respects (it can protect innocent people from wrongdoers) but there are key differences. For instance, criminal courts have more time for deliberation than individuals in self-defence situations. Secondly, acts of convicting and punishing criminals have symbolic force (i.e. they communicate messages about society's core values), but an individual's decision to repel an attacker does not have this same symbolic force. Thirdly, (as I will discuss in more detail below) the state has a special relation to the offender, which makes miscarriages of justice particularly egregious. These differences mean that the constraints on harming people for societal self-defence must be stricter than the constraints on individual self-defence – the BRD standard is appropriate for decisions about punishment; but the 'reasonable belief' standard is appropriate for decisions about individual self-defence.

As I explained at the start of this section, there is not scope within this thesis to provide a full defence of the doctrine of doing and allowing. Instead, I have focussed on the relevance of the DDA to the justification of the BRD standard. I

have rebutted some of the main arguments that have been presented by those who doubt that the doctrine can provide a basis for the BRD standard. Many of the arguments that I have criticised so far share a similar strategy: they deny the relevance of the distinction, by producing examples where the distinction purportedly fails to govern our judgements about the agent's behaviour. Without considering every supposed counter-example against the DDA I will make some general remarks about common flaws with this strategy: Firstly, the DDA states that it is *harder* to justify doing harm, than allowing harm, because doing harm is *prima facie* particularly wrong. (Hence punishment, which undeniably involves doing harm, is not absolutely prohibited by this principle, but must be strongly justified). So examples that merely show that doing harm is sometimes permitted/required do not automatically undermine the distinction. Furthermore, given that the principle is concerned with *justifying* harm, purported counter-examples (such as Reiman's examples, discussed earlier in this section) involving behaviour which is very clearly *un*justified, and which the agents themselves do not try to justify, are not obviously relevant. As Warren Quinn has pointed out, the DDA does not imply that badly-motivated agents who do harm will always seem more culpable than badly-motivated agents who allow harm.[201] Furthermore, it is a mistake to think that the difference between doing and allowing harm is only morally significant if *every* pair of cases that differ in this respect, also differ as to permissibility. Sometimes a factor that is normally morally significant can have its force negated or outweighed by other aspects of the particular case.[202] In addition, it is often difficult, or impossible to be sure if the doing/allowing distinction makes a moral difference in a particular case, because we cannot be confident that our intuitions are entirely accurate, especially if the case involves certain distracting features (e.g. the extreme wickedness of an

[201] Quinn W, 'Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing' (1989) 98 (3) Philosophy and Public Affairs 287.
[202] Kagan S, 'The Additive Fallacy' 99 (1988) Ethics 5.

actor's motivations). Thus, it is hard to tell if the DDA makes any difference to James Rachels's case of the wicked uncles. In this famous example, one uncle allows his young nephew to drown in a bathtub because he desires the money that the nephew stands to inherit, and the other uncle actively drowns his nephew with the same motivation.[203] Our intuitions may enable us to judge that both uncles behave wrongfully, but our intuitions may not be sensitive enough to detect subtle moral differences between the two cases. Our extreme abhorrence at the uncles' wicked motives may 'swamp' responses we might otherwise have to the examples.[204] It would therefore seem premature to conclude on the basis of such examples that our intuitions are misguided in those cases (discussed above) where the doing/allowing distinction does seem relevant.

To conclude this section: those who endorse the doctrine of doing and allowing have reason to endorse the BRD standard. When the state inflicts serious harm on offenders in response to their crimes, it does so *actively,* whereas the harms that result from failing to impose hardship on offenders are merely allowed to occur. The BRD standard, as it is currently applied in the courtroom, helps to ensure that *part* of the state's purported justification for harming offenders (that they have committed a crime) is sufficiently credible. But since the doctrine of doing and allowing is a general ethical principle, implying that all instances of doing harm require particularly strong justification, there is no reason to restrict the application of this doctrine to the process of establishing factual guilt in a criminal trial. The *entire moral argument* for seriously harming an offender should be held to a high standard of credibility.

[203] Rachels, 'Active and Passive Euthanasia' (1975) 292 New England Journal of Medicine 78.
[204] Woollard F, 'The Doctrine of Doing and Allowing II: The Moral Relevance of the Doing/Allowing Distinction'7 (7) (2012) Philosophy Compass 459.

The Doctrine of Double Effect (or the Intention/Side-effect Distinction)

The state's infliction of serious harm on offenders is not only active; it is also intentional. In this section I will argue that theorists who endorse the doctrine of double effect – also known as the intention/side-effect distinction - have reason to support the BRD standard and that such theorists also have reason to apply the BRD standard (or a similarly high standard of justification) to the entire moral argument for inflicting serious harm on offenders. Like the previous section, my focus will be on how this non-consequentialist principle relates to the BRD standard and I will not attempt to provide a full defence of the doctrine of double effect. I will, however, rebut some of the main arguments that have been presented by those who doubt that the doctrine can provide a basis for the BRD standard. Chapter 6 will provide some different reasons for holding the moral argument for inflicting serious harm on offenders to a high standard of justification, which may appeal to those who reject the intention/side-effect distinction.

The doctrine of double effect (DDE) is a widely held non-consequentialist principle. The formulation that I will focus on states that it is often harder to justify harming people intentionally than to justify harming them as a side-effect; and that there are cases where it is permissible to harm people as a side-effect, where it would have been impermissible to harm them intentionally. Like the doctrine of doing and allowing, the importance of the doctrine of double effect may lie in its recognition of the individual's independence from other agents and of each individual's authority over herself. If we were permitted, without particularly strong justification, to intentionally harm someone (e.g. because doing so would further our goals, or remove an obstacle to our goals), then we would all be vulnerable to being co-opted

into other agents' plans. As Warren Quinn puts it, 'people have a strong *prima facie* right…not to be pressed, in apparent violation of their prior rights, into the service of other people's purposes. Sometimes these additional rights may be justifiably infringed… but in all cases they add their own burden to the opposing moral argument. The Doctrine of Double Effect thus gives each person some veto power over a certain kind of attempt to make the world a better place at his expense.'[205]

When criminals are punished, they are harmed intentionally. This is true regardless of which theory of punishment is taken to justify punishing them. On retributive theories (elements of) the hardship involved in punishment is taken to be (partly) the end that is aimed at. On deterrence theories, the hardship of punishment is a means to the end of preventing crime through deterrence. On incapacitation theories, the harm of interfering with the offender's liberty is a necessary evil, which is justified because the threat offenders pose to society needs to be eliminated. Incapacitation theorists do not want the offender to suffer, but they do intend that the offender is deprived of his liberty and this deprivation is a harm to the offender. In contrast, if a court mistakenly fails to convict an offender and he then goes on to harm another victim, the authorities do not intend that victim to be harmed. Similarly, when offenders are sent to prison, this often harms their families, but the authorities do not harm offenders' families intentionally.

---

[205] Quinn W, 'Actions, Intentions, and Consequences: The Doctrine of Double Effect' (1989) 98 (4) Philosophy and Public Affairs 334, pp350-351. This rationale for the DDE is similar to Tadros's rationale for the *prima facie* prohibition on harming others as a means (discussed in Chapter 3 above). The DDE and the means principle are closely related, but the DDE is wider – it applies to all forms of intentional harming including, for example, harming people who pose a direct threat to oneself or others. Tadros claims that harming direct threats counts as 'eliminative harming' and is outwith the scope of the means principle – a principle that applies to 'manipulative harm'. Even if Tadros is right that it is permissible to inflict 'eliminative harm' in a wider range of circumstances than 'manipulative harm', the DDE still implies that eliminative harm requires strong justification, since it is a type of intentional harm. Furthermore, in order for eliminative harm to be justified, we must be very confident that the harm we are intending to inflict genuinely is 'eliminative', i.e. we must have good grounds for believing that the person is a direct threat and that our action will remove/reduce the threat.

Those who endorse the doctrine of double effect therefore have a reason to hold purported justifications for *intentionally* inflicting serious harm on offenders to a high standard of credibility. If the state intentionally harms someone and it turns out that it lacked adequate justification for doing so, for instance because the person was actually innocent, this is a particularly serious kind of injustice. The BRD standard is meant to act as safeguard against this kind of injustice.

Some theorists might question my claim that punishment of the innocent (under all systems of punishment) counts as 'intentional' harm. For instance, Victor Tadros assumes that, under a retributive system of punishment, when the innocent are punished, they are harmed unintentionally.[206] He therefore argues that retributivists cannot use the intention side/effect distinction to justify the BRD standard (and he claims that, more generally, retributivists have 'problems' justifying procedural protections for accused people). This is because, he argues, retributivists believe that giving offenders their just deserts is good, and 'normally, when we aim at some significant good, we are quite tolerant of bad side effects that we would bring about in achieving that good'.

This objection is flawed. When the state mistakenly punishes an innocent person, it does harm that person intentionally, and not merely as a side-effect. True, the state does not mean to punish him *qua* 'innocent person'. But it does intentionally inflict hardship on the particular individual who has been convicted. Consider the analogy of mistaken self-defence. A person who harms someone whom she mistakenly believes to be attacking her cannot claim that this mistake meant that she harmed the supposed attacker unintentionally.

[206] Tadros, *The Ends of Harm: The Moral Foundations of the Criminal Law* (OUP 2011) pp328-329. See also Reiman and Van den Haag, 'On the Common Saying', p245 for a similar argument.

Tadros is correct to argue that the BRD standard cannot be derived solely from the retributive principles that the state should punish the guilty and should not punish the innocent.[207] However, *pace* Tadros, retributivists can consistently defend the BRD standard on the basis of general ethical principles including the DDA and the DDE. Any punishment theorist who accepts these general principles can use these doctrines to justify the BRD standard. This type of justification for the BRD standard is not tied to one particular theory of punishment.

Another potential objection to my argument cites the state's *positive obligation* to protect those who may be harmed by criminals who escape punishment. If the state has an obligation to protect potential victims from harm, it might be claimed, it is morally irrelevant that those victims are not *intentionally harmed by the state*. Therefore, the critic claims, the intention/side-effect distinction provides no moral justification for the BRD standard – a standard which in effect favours allowing harm to befall potential victims (as a side-effect of mistakenly acquitting criminals) over intentionally harming possible offenders when there is reasonable doubt as to the justification for inflicting such harm.[208]

This objection is misconceived. The Intention/side-effect distinction is still relevant despite the presence of a positive obligation to protect or care for the individuals who may be harmed. Consider the following cases:

Diverting Resources

This example (which I also cited in the previous section) involves doctors who divert resources from one group of patients to another larger or needier group of patients.

[207] Tadros, fn 206, pp328-329.
[208] Cf. Y Lee , 'Deontology, Political Morality, and the State' 8 (2011) Ohio State Journal of Criminal Law 385.

Human Guinea Pigs[209]

The doctors deliberately leave a certain group of patients untreated in order to study the progress of the disease, so they can treat a larger group of patients more effectively.

Intuitively, it seems that the doctor's conduct in Diverting Resources is morally permissible, because the harm that results to the patients is only an unintended side-effect of the withdrawal. In contrast, the conduct of the doctors in *Human Guinea Pigs* seems morally unacceptable, because they intend harm to befall the patients. In both of these examples the doctors have a positive obligation to care for all the patients. Yet the intention/ side-effect distinction still seems morally relevant. Similarly, the state has a duty to respond to crime in a just manner – to coercively interfere with offenders who ought to be subject to such interference and not interfere with individuals where this is unjustified, e.g. where the individual is innocent - but how it carries out its duty is likewise constrained by the powerful considerations against intentional harming. These considerations are so powerful that they can only be rebutted by an argument that is at least as compelling. Hence the prosecution's case for convicting the accused will only succeed if it meets the demanding BRD standard. Likewise, more generally, seriously harming offenders is only permissible if the moral justification for doing so meets a high standard of credibility. The DDE is not irrelevant in situations where there is a positive obligation to protect or care for individuals who may be harmed.

Some writers might attempt to challenge my argument by claiming that the intention/side-effect distinction is only relevant to blameworthiness not to permissibility.[210] In other words, according to this claim, when the state intentionally

[209] W Quinn, 'Actions, Intentions, and Consequences: The Doctrine of Double Effect' (1989) 98 (4) Philosophy and Public Affairs 334.

[210] For an interesting discussion of this issue see D Husak, 'The Costs to Criminal Theory of Supposing that Intentions are Irrelevant to Permissibility' 3 (2009) Criminal Law and Philosophy 51.

harms offenders despite lacking adequate moral justification, it may show itself to be especially morally deficient. But the fact that the harm is intentional does not, on this view, make the state's action especially impermissible, or any more impermissible than allowing harm to occur as a side-effect of failing to punish. Even if this were true, however, the intention/side effect distinction could still help to justify the high standard of proof in criminal trials. If the state consistently showed itself to be morally deficient, by intentionally harming offenders, despite lacking adequate justification, the state could lose the moral authority to punish at all.

It might be objected that courts do not (and should not) *intend to harm* offenders when they convict and sentence them. For instance, on a communication theory of punishment, the intention is to call the offender to account for his wrongdoing, to attempt to persuade him of the wrongfulness of his criminal conduct and to restore him to the moral community. While it might be conceded that the things we do to offenders (publically condemning them as wrongdoers, depriving them of their liberty and/or some of their resources etc.) would normally count as harmful, it might be objected that the court does not impose such burdens on offenders *qua* harms. In response, it is submitted that the relationship between what the state intends to do to the person who is convicted and punished and the harm which that involves is a *constitutive relationship* not a causal one. Where one state of affairs is causally downstream from another state of affairs, it can be meaningful to speak of intending the former, while foreseeing the latter as an unintended side-effect. However, where one state of affairs constitutes another state of affairs, the intention side-effect distinction cannot apply. Consider an analogy from the literature. A group of explorers are trapped in a cave – one of their number, who is particularly fat, is stuck in the entrance and cannot be moved.[211] The only way to escape is to use dynamite to

[211] See e.g. W FitzPatrick, 'The Intend Foresee Distinction and the Problem of 'Closeness' 128 (2006) Philosophical Studies 585.

blow him to bits. Can the explorers claim that they did not intend to kill him; they only intended to blow him to bits? It seems not, because the relationship between those two states of affairs ("killing him" and "blowing him to bits") is too close. "Blowing him to bits" constitutes "killing him".[212] Similarly, when the state intentionally imposes severe burdens on an offender, such as labelling him a murderer and sending him to prison for life, this constitutes harming him.

Some theorists might claim you do not really 'harm' a person if what you do to him is justified. However this idea can sometimes seem counterintuitive. If somebody breaks an attacker's leg in self-defence, or breaks a bystander's leg in order to rescue five people from certain death, it would seem odd for him to say, 'I did not harm anyone, because what I did was justified'. It would be more natural to say, 'yes, I harmed the attacker/bystander, *but* what I did was justified'. Perhaps, if the justification is a paternalistic one, then it might seem more natural to deny that the conduct was harmful. But even then, it does not seem clearly wrong to speak of 'harm'. Imagine that a person is trapped beneath a collapsed building and a rescuer cuts off the person's arm, which is pinned by some debris, because this is the only way to rescue him. The rescuer might say, 'I didn't do him any harm, because, overall, my actions benefited him'. On the other hand, the rescuer might say, 'yes I harmed him, but I did it for his own good'. The latter statement still seems to be a legitimate use of the word 'harm'. Or contrast the following two lives: Person A leads a law-abiding, comfortable, healthy life, pursuing his own goals. Person B commits a crime and is sentenced to 20 years in prison (under fairly grim conditions) in order to help him see the error of his ways and restore him to the moral community. It would seem counterintuitive to say that B suffered no more harm in his life than A. In any case, this semantic issue can be sidestepped, when it comes to

[212] Ibid.

justifying the beyond reasonable doubt standard. We could say that imposing measures such as fines or imprisonment involves the intentional infliction of what is *prima facie* harmful. It therefore requires strong justification. It is only after that justification has been provided and has met the relevant standard of credibility that we can be satisfied that depriving an offender of liberty or property does not 'truly' harm him. Alternatively, one might say that such measures involve the intentional infliction of 'hardship'. Although I will continue to refer to the principle that the active, intentional infliction serious harm requires strong justification, those who are still sceptical about whether justified harm really counts as harm, may substitute the terms '*prima facie* harm' or 'hardship' where I use the term 'harm'.[213]

To conclude this section: I have argued that those who endorse the doctrine of double effect have reason to endorse the BRD standard. When the state inflicts serious harm on offenders in response to their crimes, it does so *intentionally,* whereas the harms that result from failing to impose hardship on offenders are unintended. Since the doctrine of double effect is a general ethical principle, implying that all instances of doing harm require particularly strong justification, there is no reason to restrict the application of this doctrine to the process of establishing factual guilt in a criminal trial. The *entire moral argument* for seriously harming an offender should be held to a high standard of credibility.

[213]In opting for the term 'harm' I am following the terminology of key writers in this area including B Vilhauer 'Free Will and Reasonable Doubt' (2009) 46 (2) American Philosophical Quarterly 131and V Tadros, *The Ends of Harm: The Moral Foundations of the Criminal Law* (OUP, Oxford 2011). I am grateful to Antony Duff for raising, in conversation, the objection that the state should not intend to harm offenders. It should be noted that on Duff's account, the courts certainly *intend* to subject offenders to *hardship* (as a means of enabling them to reform). He writes that '[h]ard treatment' is the 'appropriate method of pursuing' the 'aim of persuading people to repent the wrongs they have done'. See RA Duff, *Punishment, Communication and Community* (OUP, Oxford University Press 2001), p30.

Special Obligations

The state has a special obligation to protect citizens from unjustified harm. The state can breach this duty if it allows criminals to continue offending and makes insufficient efforts to bring them to account. This breach involves *failing to carry out* the duty to protect. However, there is an even more serious way in which this duty can be breached – if the supposed protector actually becomes the threat from which people need protection. This is an '*inversion*' of the original duty.

John Gardner makes this distinction in order to show why killings by police officers are among the worst types of killing.[214] He states that certain 'public officials....are in a special moral position because they are officials.'[215] Sometimes police kill people in the belief that this is necessary in order to carry out their duty to protect people from harm. Gardner cites the example of Jean-Charles de Menezes, whom the police killed, having mistaken him for a terrorist. Even if their belief was reasonable, this does not neutralise the moral awfulness of what happened (although it might render the police officers blameless). They ended up doing the opposite of their duty – they were bound to protect Mr de Menezes from harm (including the harm of being killed) and they themselves killed him.

When supposed protectors attack the people who had trusted and depended on them, they violate their victims' legitimate expectations. Gardner makes this point especially vivid, by citing reactions from a victim of the Utoya massacre to the sight of the killer dressed as a policeman: 'Just think of it, he dressed himself in a police uniform, the symbol of safety and support.'[216] Similarly, when courts convict and sentence people without an adequate justification for doing so, the state has betrayed

[214] 'Worst' does not necessarily mean 'most blameworthy'. Gardner J, 'Criminals in Uniform' in R.A. Duff, Lindsay Farmer, S.E. Marshall, Massimo Renzo, and Victor Tadros (eds), The Constitution of Criminal Law (OUP, Oxford 2013) (forthcoming).
[215] Ibid, p8.
[216] Ibid, p10.

the people it was meant to protect. The husband of Sally Clark (a mother who was wrongly convicted of murdering her children) commented on the devastating effect of this kind of betrayal. He writes: 'We were all people who had complete faith in the justice system and still find it hard to believe it could have let us down. .... both of us [were] numb with shock and disbelief as we heard the appeal refused.'[217]

Gardner notes that time constraints limit how carefully the police can assess the justifications for their actions. But, as I have noted above, at the trial stage there is much more time to deliberate than in police operations. Therefore, the state's argument for inflicting harm on offenders, who have already been apprehended, should be held to an even higher standard of justification than decisions made by police officers.

Gardner also observes that that the police are often the victim's 'last line of protection'. A similar statement could also be made concerning parents and spouses. This exacerbates the wrongfulness of unjustified harm inflicted by the police, parents or spouses on those who depend on them. This feature also characterises the state's relation to offenders and possible offenders. The accused and the convicted criminal are at the mercy of the courts. Where else can they turn for protection?

Gardner also notes that failure to recognise the particular moral awfulness of protectors doing the opposite of their duty means that one misses what is especially tragic about Sophie's Choice (discussed above). This is another reason why Lee's reliance on this type of case fails to undermine the rationale for the BRD standard.

To summarise my argument so far, I have claimed that those non-consequentialists who endorse the DDA, the DDE and the notion of special obligations should support

[217] http://innocent.org.uk/cases/sallyclark/

the BRD standard because punishment involves the *active*, *intentional* infliction of serious harm *by the state*. According to the DDA the active infliction of harm requires strong justification; according to the DDE the intentional infliction of harm requires strong justification; and the argument from special obligations suggests that the state needs strong justification when it proposes to harm someone whom it has a duty to protect. It seems that nothing about this rationale for the BRD standard restricts its application to the process of proving an accused's factual guilt in a trial. Instead this rationale implies that the entire moral justification for harming offenders should be held to a high standard of credibility.

I have also argued that the BRD standard is not tied to one particular justification of punishment. I have already argued against Reiman's retributive rationale for the BRD standard. I will now critique three other purported rationales: one based on a communication theory of punishment, one derived from a social contract theory of punishment and one derived from a deterrence theory of punishment. It is important to expose the difficulties with these purported rationales for the following reasons. Firstly, if the rationale for the BRD standard did derive *solely* from one theory of punishment then theorists who rejected that theory of punishment would have no reason to accept the BRD standard.[218] However, my thesis claims that all theorists have reasons to hold arguments for punishment to a high standard of justification. (This chapter gives reasons that should appeal to non-consequentialists; the next chapter also addresses those who may not be committed to non-consequentialism.) Secondly, the difficulties with the alternative strategies for defending the BRD that I will now discuss may provide some motivation for theorists to endorse the rationale for the BRD standard that I have defended in this chapter.

[218] I claim that the BRD standard cannot be derived *solely* from any particular theory of punishment. Nor does any particular theory of punishment provide the 'main' reason why the BRD is important. However, I do not deny that some theories of punishment may provide *additional support* for the BRD standard.

A Communicative Theory of Punishment and The BRD Standard

Some theorists have claimed that the rationale for the BRD derives from a communication theory of punishment.[219] On this view, convicting and punishing offenders involves condemning them as wrongdoers. Condemning someone without being really sure that such condemnation is justified shows disrespect for the person condemned. Therefore, according to this approach, we need the BRD standard to avoid the wrongfulness of unjustifiably condemning people, *not* to avoid the wrongfulness of unjustifiably imposing hard treatment on them. This view faces certain difficulties.

Firstly, it is possible to imagine a system of punishment (or quasi-punishment) where the state does not intend to condemn wrongdoers. For instance, imagine a consequentialist system that viewed morally condemning wrongdoers as counterproductive, so inflicted hard treatment on them, without subjecting them to moral condemnation. If there were two such consequentialist systems and one of them upheld the BRD standard and the other did not, we would have reason to prefer the former system. The condemnation approach cannot explain why one system is preferable to the other. The communication/condemnation theorist might criticise *both* systems for getting the *positive justification* for punishment wrong. However, this criticism is separate from the criticism that the system, in failing to uphold the BRD, fails adequately to *constrain* punishment.

[219] RA Duff seems to support this approach, see e.g. 'Presuming Innocence' in J Roberts and L Zedner (eds.) *Principles and Values in Criminal Law and Criminal Justice: Essays in Honour of Andrew Ashworth (OUP, Oxford forthcoming),* Electronic copy available at: http://ssrn.com/abstract=2103337 - see especially pp3-4. See also, V Tadros, 'The Ideal of the Presumption of Innocence', paper presented at Fraying the Golden Thread: The Presumption of Innocence in Contemporary Criminal Law (Aberdeen, 2012). Tadros's main reason for endorsing the BRD standard seems to be his 'manipulative harm' argument discussed below. He has not yet provided an account of how precisely that argument relates to the 'condemnation argument' for the BRD standard.

Secondly, the idea that people need protection from unjustified condemnation, but not from unjustified hard treatment seems arbitrary, fails to reflect the concerns of many accused people and fails to capture our intuitions about the wrongness of miscarriages of justice. Typically, an accused person who was told that they could be convicted on the balance of probabilities, or on the basis of a mere suspicion, would not only be concerned about the risk of being unjustifiably condemned, but would also be concerned about the hard treatment aspect of punishment e.g. being falsely imprisoned. One of the terrible aspects of miscarriages of justice is the unjustified infliction of hard treatment. It would be odd to say, on discovering that an innocent person had spent 20 years in prison or (in a country which imposed the death penalty) had been executed that the *really* objectionable thing about the case was the unjustified condemnation, not the hard treatment aspect.

Having said this, the 'condemnation approach' does have an element of truth in it. The risk of being unjustifiably condemned is one of the things accused people would have reason to be concerned about if the standard of proof were lowered and it is one element of what is wrong about punishing the innocent. It is plausible to claim that condemning a person without good evidence fails to respect that person. However, that idea derives from the more basic principle that it is wrong to inflict serious harm on someone without strong justification. Publically condemning someone as a criminal wrongdoer is a way of seriously harming that individual (or at least involves imposing serious negative consequences or hardship on her).[220]

[220] See above for a discussion of the meaning of 'harm'.

Lee's Social Contract Theory and the BRD Standard

Although Lee denies that the doing/allowing and intending/foreseeing distinctions could justify the beyond reasonable doubt standard, he proposes another rationale for the standard, which he claims is more promising.[221] His account has two key elements – a slippery-slope-type argument and a social-contract-type argument. Firstly, he points out that the requirement that the state must prove guilt beyond reasonable doubt constrains the state's power to punish. If the state were permitted to depart from this standard in order to increase deterrence and to reduce the number of people who are victimised by falsely acquitted people, the state might end up punishing people even where these social benefits will not be achieved by such punishment or where these benefits are outweighed by other consequentialist considerations. He writes: 'The government enjoys an enormous amount of power to interfere with peoples' lives with force and to stigmatize individuals with its stamp of blameworthiness....Unless we treat the constraints against convicting without sufficiently convincing proof...as close to inviolable, such limitations on criminalization and punishment will give too often and will not be able to provide meaningful limitations of the government's power to criminalize and punish'.[222]

Lee's account seems to cite the wrong kind of reason for guarding against false convictions.[223] Punishing someone whom we know may well be innocent (who perhaps has a 49% chance of being innocent) seems wrong because that individual has been treated unjustly, not merely because it could lead to further bad

[221] He argues that his rationale is *better* than the alternative justifications, but leaves it open whether the beyond reasonable doubt standard is ultimately defensible. His main claim is that this question can only be settled with reference to his proposed framework.
[222] Lee 'Deontology, Political Morality, and the State' 8 (2011) Ohio State Journal of Criminal Law 385, p399.
[223] This element of Lee's account seems similar to Bentham's utilitarian defence of a high standard of proof, which is liable to the same objection. J. Bentham, *A Treatise on Judicial Evidence* (Paget, London 1825). Even if Bentham's argument provides a good reason for the BRD standard, it cannot be the sole or the main reason for that standard.

consequences. It seems odd to characterise this kind of injustice as the beginning of a slippery slope to bad consequences – rather, a society with such practices is already at the bottom of the mountain.[224]

The other feature of his account relates to the source of the state's duty and power to punish. The state is authorised and obligated to protect citizens through punishment because those citizens have consented to render the power to punish to the state (and to refrain from taking the law into their own hands). However, they only give the state this power if it abides by strict conditions, including upholding the beyond reasonable doubt standard. According to Lee, the state's protective duty towards it citizens differs from Sophie's duty towards her children in the following respect – Sophie's duty stems from her role as a mother, whereas the state's obligation is conferred on it by the people. When Sophie is deciding whether to sacrifice one child to save the others she must take into account her children's competing rights and interests. But when the state is faced with a similar choice regarding its citizens, it must also take into account the restrictions imposed on it by the people, who are the source of its authority. The state has its power to protect '*only on condition* that it respect[s] such limitations'.[225] The people 'demand that the state be able to justify the acts it is about to take by correctly identifying wrongdoers. The proof beyond a reasonable doubt requirement is generated from this demand.'[226]

This raises the question of whether the people's demands have a rational basis. It seems likely that the people do not support the beyond reasonable doubt standard for purely utilitarian reasons. The shock people feel at miscarriages of justice seems to

[224] Relying heavily on the slippery slope argument in this context seems almost as bizarre as someone who says, 'It is wrong for me to murder Tom, because that would be the start of dangerous slippery slope – I might end up killing Dick and Harry as well'. Even if the slippery slope consideration is one reason for endorsing the BRD standard, it does not seem to be the main reason.
[225] Lee 'Deontology, Political Morality, and the State' 8 (2011) Ohio State Journal of Criminal Law 385, p400.
[226] ibid, p399.

be a reaction to the wrong done to the individual, not primarily due to 'their subsequent realisation that a false conviction is going to reduce the overall utility of the system'.[227] If the deontological arguments in favour of the beyond reasonable doubt principle are as inadequate as Lee has argued, and the people generally support that principle for deontological and not merely utilitarian reasons, then the state's supposed duty to respect their irrational demands seems a flimsy basis on which to defend the beyond reasonable doubt standard.

If Lee's account were accepted it might be thought that we could keep the beyond reasonable doubt principle and not be forced to apply it to issues like free will or retributivism. Lee's account seems to imply that if the people do not (yet) demand that the BRD standard be applied to these issues there is no reason to do so. This kind of account seems too complacent. It does not have the resources to say how the system should be improved. The people might be content with an unjust system, e.g. one that unfairly discriminated against minority groups.

Tadros's Deterrence Theory and The BRD Standard

Tadros does not succeed in refuting the defence of the BRD standard that relies on the DDE or the DDA. His alternative method of defending the BRD standard, I will argue, has considerable difficulties of its own. His preferred strategy is a non-consequentialist one. Although, he accepts the importance, in general, of the intention/side-effect distinction and of the DDA, he employs neither of these doctrines to defend the BRD standard. He argues that it is the *means*/side-effect distinction, rather than the intention/side-effect distinction that can justify the BRD standard. He claims that we need the BRD standard, because we need protection

[227] Halvorsen, 'Ten Guilty Persons', p8.

against being harmed as a means (or 'manipulative harm').[228] He claims that manipulative harm is normally prohibited, with the following exception: punishment that is inflicted for the purpose of general deterrence is a justified form of manipulative harm. He claims we need to avoid mistaken convictions because inflicting unjustified manipulative harm is extremely wrong.

Tadros's strategy fails to get to the heart of why the BRD standard is important. We would need the protection afforded by the BRD standard even if punishment were not inflicted for instrumental reasons (e.g. to deter crime) and thus did not count as manipulative harm. For instance, consider a system under which punishment were inflicted for purely retributive reasons. According to Tadros, since retributivists regard the suffering involved in punishment as being the *end* they seek to achieve (not a means of achieving some further end), it does not count as manipulative harm. Yet it would still be wrong for a purely retributive system to abandon the BRD standard. If there were two such purely retributive systems and one of them upheld the BRD standard and the other did not, we would have reason to prefer the former system. Tadros's approach cannot explain why one of these retributive systems is preferable to the other. Tadros might criticise *both* retributive systems for getting the *positive justification* for punishment wrong. However, this criticism is separate from the criticism that the system, in failing to uphold the BRD, fails adequately to *constrain* punishment. Similarly, we would need the BRD under a purely incapacitative system of dealing with offenders, i.e. one that interfered with their liberty purely to reduce the harm they might pose to others. However, Tadros classes incapacitation as a form of 'eliminative harm' (i.e. harming someone to eliminate a threat they directly pose), not as 'manipulative harm'. Thus his defence of the BRD standard (based on the need to protect people from manipulative harm) could not

[228] V Tadros, *The Ends of Harm: The Moral Foundations of the Criminal Law* (OUP, Oxford 2011). Chapter 3 of this thesis also discussed 'manipulative harm'.

explain why an incapacitative system of dealing with offenders that upheld the BRD standard would be preferable to one that did not.

The most that can be said for the deterrence-based rationale is that were we to accept deterrence as a justification for punishment this would provide *additional* reasons for endorsing the BRD standard. However, it is submitted that even if we were persuaded by Tadros's argument that it is all right to punish culpable wrongdoers to deter others, we should not rely on his deterrence theory of punishment alone because we should have a reasonable doubt about whether the people whom we subject to deterrence are culpable in the right sense.[229]

Conclusion

In this chapter I have argued that non-consequentialists should support the BRD standard because this standard is implied by the principle that the active, intentional infliction of serious harm by the state requires strong justification. This principle draws on the DDA, the DDE and the notion of special obligations. I have also rejected various alternative rationales for the BRD standard.

[229] See my discussion of this issue in Chapter 3.

Chapter Six: The Convergence Requirement

Introduction

In the previous chapter I examined the moral basis for the beyond reasonable doubt standard in criminal trials. I argued that the justification for this standard of proof does not stem directly from any particular theory of punishment, and is not tied to the specific context of the criminal trial. Instead, I argued that the most plausible non-consequentialist rationale for the beyond reasonable doubt standard relies on four considerations (which a variety of different punishment theorists could invoke): 1) the doctrine of doing and allowing; 2) the intention/side-effect distinction; 3) the seriousness of the hardship involved in convicting and punishing someone; and 4) the fact that unjustifiably punishing someone is an inversion of the state's duty to protect that person. These considerations imply that we should only punish a person if we have a high degree of certainty that doing so is justifiable. This does not merely mean that certain facts (e.g. that the accused was the person who committed the crime) should be proved beyond reasonable doubt. Rather, we should also require that the *entire moral argument* for punishing a person be established to a high standard of credibility. (I call this the "cautious approach" to punishment.) Given the significant room for doubt about the soundness of each theory of punishment, it seems unlikely that any single theory of punishment can satisfy this standard.

I propose that we should try to reduce the risk of inflicting unjustified punishment by only punishing someone if the main theories of punishment agree that punishing that person is appropriate. (I call this the "convergence requirement".) This chapter will begin by giving some reasons why people from different theoretical perspectives have reason to endorse my cautious approach to punishment and my convergence requirement, rather than simply relying on their own favoured theory of punishment.

I will then further clarify why exactly the convergence minimises the risk of unjustified punishment. Next, this chapter will address two potential objections to my view. One challenge stems from the existence of uncertainty about which theory of moral uncertainty is sound. The other potential objection concerns my reliance on the idea of agreement among 'experts' on punishment. The final section will briefly outline some issues that are relevant to the implementation of the convergence requirement in practice.

Why Should Theorists from Different Philosophical Perspectives Accept The Convergence Requirement?

Theorists from different philosophical perspectives have somewhat different reasons for endorsing the convergence requirement. Those non-consequentialists who are also retributivists have three reasons for doing so: a consideration that is internal to non-consequentialism; a reason that stems specifically from their adherence to retributivism; and finally a consideration that derives from their status as ethical decision-makers under conditions of uncertainty. Consequentialists on the other hand, should accept my approach because of this final consideration and because of another consideration, which is internal to consequentialism.

A Non-Consequentialist Basis for the Convergence Requirement

Firstly, non-consequentialists have a reason *qua* non-consequentialists for endorsing my approach. If they support the beyond reasonable doubt standard (as they almost certainly will) then, as I have argued in the previous chapter, the non-consequentialist considerations in favour of the beyond reasonable doubt standard also count in favour of holding substantive theories of punishment to a similarly high

standard and count against relying on a single, hotly-contested theory of punishment alone.

A Retributive Basis for the Convergence Requirement

Secondly, retributivists have some motivation *qua* retributivists for endorsing my approach, for pragmatic reasons. This is because, as I will explain, there is some reason to think that the convergence requirement could constrain emotions and biases that interfere with the administration of justice (conceived in retributive terms). It is a fact about human nature that a desire for retribution very easily slides into a desire for vengeance.[230] Indeed, it is often psychologically very difficult for individuals to *tell* whether their desires are genuinely retributive or merely vengeful (assuming that this distinction is valid in principle). A desire to find a scapegoat on whom to vent vengeful feelings could result in wholly undeserved punishment. Crimes with particularly emotionally-distressing features could also attract punishments that exceed a person's true desert. The literature on moral psychology and cognitive biases also sheds some light on the ways in which people's judgements about retribution can be distorted. For instance, there is some evidence that people may confuse a feeling of (non-moral) disgust for a judgement that an action is morally wrong and deserving of retributive punishment.[231] Another bias called the

[230]According to some recent psychological research, a tendency to hold attitudes in favour of personal vengeance is positively correlated with endorsing retributive punishment, see, e.g., I. McKee and N. Feather, 'Revenge, Retribution and Values: Social attitudes and Punitive Sentencing' (2003) 21 Social Justice Research 138.

[231] For instance, people who are in a dirty room make harsher condemnatory judgements than people who are passing moral judgments in a clean room. People who were hypnotically induced to feel a brief pang of disgust at an innocuous word, such as 'often', judged transgressions to be morally worse when the description of the transgression included the word 'often' than when a synonym was used. Some subjects even judged morally innocuous behaviour to be morally wrong when the behaviour was described using the disgust-inducing word. People who are particularly sensitive to non-moral disgust (e.g. noxious smells) also make harsher judgements about when punishment is deserved and about its severity. See, e.g. T. Wheatley and J. Haidt, 'Hypnotic Disgust Makes Moral Judgments More Severe' (2005) 16 Psychological Science 780; for an overview of this literature see: Y. Inbar and D. Pizarro, 'Grime and Punishment: How Disgust Influences Moral, Social, and Legal Judgments' (2009) 21 The Jury Expert 11.

'fundamental attribution error' causes people to underplay the causal importance of circumstances in bringing about an event and to exaggerate the role of human agency.[232] Empirical studies on judges and mock juries reveal that cognitive biases and emotional factors can distort their decision-making about guilt and sentencing and there is also reason to believe that such biases can affect legislators when they decide to criminalise behaviour (and render those who engage in it liable to punishment).[233] Even if some day a retributive theory were developed capable of convincing people beyond reasonable doubt that retributivism is correct in principle, there would still always be uncertainty about whether retributive punishment is appropriate in a given case or whether it only appears to be appropriate due to biases and distorting emotional factors and hence there would still be a danger of misapplying retributivism in practice. However, my convergence approach could make this less likely. If my approach were adopted, a supposedly retributive basis for punishing someone (which might be mistaken in an individual case) would not be enough to justify punishing the person - a consequentialist basis would also be required. This would be beneficial from the perspective of negative retributivism, as certain individuals who appear (e.g. due to cognitive biases) to satisfy the retributive requirement (but who really do not) would fail to satisfy the consequentialist requirement and these would be spared retributively undeserved punishment. Of course, from the perspective of positive retributivism the convergence requirement would have the disadvantage of allowing some guilty people to escape their 'just deserts'. However, since most retributivists place more importance on the negative element, overall they would have reason to endorse the convergence requirement.

[232] D. Dripps, 'Fundamental Retribution Error: Criminal Justice and the Social Psychology of Blame' (2003) 56 Vanderbilt Law Review 1383.
[233]See e.g., E. Peer and E. Gamliel, 'Heuristics and Biases in Judicial Decisions' (2013) 49 Court Review 114; J Salerno and B Bottoms, 'Unintended consequences of toying with jurors' emotions: The Impact of Disturbing Emotional Evidence on Jurors' verdicts' (2010) 22.

Furthermore, some research suggests that our cognitive biases and emotional factors in favour of punishment are more powerful and numerous than any similar factors that might lead us to refrain from punishing people who actually deserve it. Punitive biases therefore seem to be *more in need of restraint* than leniency biases. Evidence about psychological biases and moral judgements indicates that, 'human beings are predisposed to give affirmative answers to questions about personal responsibility'.[234] Moreover, studies have shown that people with a 'leniency' bias are more likely to set their bias aside when evidence about guilt is reliable, whereas a bias in favour of guilt-attribution is less affected by the strength of the evidence.[235] Anger is probably the emotion most likely to be aroused by contemplating serious crimes. Strong negative emotions tend to cause people both to become more punitive and to trigger the use of cognitive shortcuts or stereotypes to reach a decision (increasing the likelihood of unjust punishment). In contrast, studies of mock jurors have revealed that emotional considerations which one might expect to result in leniency (e.g. vividly describing the accused's history of being abused as a child) sometimes produce a 'backfire effect' – causing jurors to judge the accused to be guilty and deserving of severe punishment.[236] This may be because the anger at the accused's abuser triggered reliance on stereotypes and the accused belonged to a negatively stereotyped group.[237] It might also be explained by the fact that anger can increase punitiveness in general, regardless of who was the original cause of the anger.[238] Another study found that 'E-processors' - individuals who were particularly susceptible to emotional considerations - were also prone to allow legally irrelevant

[234] Dripps, fn232, p1437.
[235] Kaplan and L. Miller, 'Reducing the Effects of Juror Bias' (1978) 36 (12) Journal of Personality and Social Psychology 1443, p1450.
[236] See e.g., M. Stevenson, B. Bottoms and S. Diamond, 'Juror's Discussions of a Defendant's History of Child Abuse and Alcohol Abuse in Capital Sentencing Deliberations' (2010) 16 (1) Psychology, Public Policy, and Law 1.
[237] Salerno and Bottoms, fn233; D. DeSteno et al., 'Prejudice from Thin Air: The Effect of Emotion on Automatic Intergroup Attitudes' (2004) 15 Psychological Science 319.
[238]Salerno and Bottoms, fn233.

information to bias their judgements about an accused's guilt.[239] Although occasionally these irrelevant factors led E-processors to be unduly lenient, the overall effect was to bias them in favour of conviction and harsh sentences. My approach would provide a counterbalance against such punitive biases and therefore should be welcomed by retributivists.

A retributivist might wonder whether the convergence requirement would only *indirectly* reduce instances of retributively unjust punishment, as a side effect of reducing the *overall* number of persons punished. However, there is also reason to think that the convergence requirement would *specifically* minimise retributively unjust punishment. The most *central cases* of retributive wrongdoing (about whose wrongfulness retributivists are most certain) are very often actions that consequentialists would also condemn and consider worthy of punishment. This is because part of what makes an action blameworthy from a retributive perspective is often that it infringes some of the victim's fundamental interests (e.g. life, bodily integrity, or autonomy) – interests that consequentialists also recognise as important and wish to protect. In contrast, there are certain actions which might strike some retributivists as intuitively wrong, but whose wrongfulness is most in doubt, or which, over time, retributivists have realised are actually not immoral at all, e.g. certain non-harmful consensual sexual practices. It is often harder to find a plausible consequentialist justification for condemning these actions. Relying on retributive intuitions about the moral status of such actions carries a particularly high risk of moral error (including retributive injustice) because such intuitions often stem from misleading emotional factors such as disgust. The convergence requirement would avoid punishing people based on such intuitions alone, but would demand that a

[239]J. Gunnell and S. Ceci, 'When Emotionality Trumps Reason: A Study of Individual Processing Style and Juror Bias' (2010) 28 Behavioral Sciences and the Law 850.

person should only be punished if doing so is also necessary to achieve a forward-looking purpose (such as preventing him from causing serious harm in the future). Thus, this consequentialist element could actually minimise the chance of retributive injustice. Of course, the convergence requirement is not an infallible method of eliminating the effect of biases – people can always come up with spurious *post hoc* consequentialist rationalisations for their original biased judgements - but it seems to be one important way in which such biases could sometimes be corrected. Studies on moral decision-making provide empirical support for this claim, showing that when people revise their initial judgements that certain (non-morally) disgusting actions are morally wrong, this is often the result of deliberating about whether such behaviour is harmful.[240]

---

[240] M. Feinberg et al., 'Liberating Reason from the Passions: Overriding Intuitionist Moral Judgments Through Emotional Reappraisal' (2012) 23(7) Psychological Science 788.

I am not arguing that retributivism is necessarily unsound to the extent that it involves reliance on emotions. The capacity to experience certain emotions may be useful or even essential to making correct moral judgements. However, reliance on emotional reactions can sometimes cause retributivism to be misapplied, when a morally irrelevant emotional reaction is mistaken for a morally relevant one (and the studies cited above, fns 231 and 237, suggest that such mistakes are quite easy to make). Nor am I suggesting that consequentialist reasoning does not involve emotion. Compassion, for instance, may be one reason why consequentialists adopt their theory in the first place. However, the process of *applying* consequentialism (i.e. working out what will maximise good consequences) may be less influenced by emotions such as anger and disgust (and hence less vulnerable to the biases connected with these emotions) than the process of *applying* retributivism, which typically relies heavily on intuition or gut-feelings. Indeed retributivists often explicitly encourage reliance on such feelings when deciding who deserves punishment and how severe that punishment should be. Michael Moore calls intuition and certain emotions (e.g. outrage) 'our best heuristic guide to moral truth' and P.F. Strawson famously endorsed the role of 'resentment' in retributive blame and punishment (see Moore, *Placing Blame* (OUP, Oxford 1997)). and Strawson, 'Freedom and Resentment' (1962) 48 Proceedings of the British Academy 187.). In contrast, consequentialists such as Peter Singer often advocate caution about relying on intuitive/emotional responses, as these are often incompatible with consequentialist moral principles (see P Singer, 'Philosophers are back on the job,' New York Times Sunday Magazine 7 July 1974 pp19-20.). Neuroscientific research provides some support for these claims about the differences between retributive and consequentialist reasoning processes. When people make retributive judgements, areas of the brain associated with emotion are activated. In contrast, areas of the brain that are activated when people engage in consequentialist reasoning are also associated with cognitive processes that modify the influence of emotional factors. (See e.g., J. Greene, 'The Cognitive Neuroscience of Moral Judgment' in M. Gazzaniga (ed.), *The Cognitive Neurosciences* (4th ed. MIT) pp. 987-1002; K. Ochsner and J. Gross, 'Cognitive Emotion Regulation: Insights From Social Cognitive and Affective Neuroscience' (2008) 17 (2) Current Directions in Psychological Science 153.) The fact that, typically, retributivists tend to trust their intuitions about punishment (but are prone to confusing indisputably irrelevant emotional reactions with genuine moral disapproval) means that pure retributivism would be likely to result in over-punishment in practice, even by retributivists' own lights. Insisting that both consequentialist and retributivist criteria must be satisfied before punishment was justified could help correct such pro-punishment biases.

A Consequentialist Basis for the Convergence Requirement

Similarly, consequentialists, *qua* consequentialists, have a reason to adopt the convergence requirement. If officials had the power to punish someone whenever they judged that this would produce good consequences, regardless of the individual's desert, this would probably cause bad consequences in practice. Given the widespread support for negative retributivism, the public would lose faith in the criminal justice system if they were aware that the state *knowingly* punished the factually innocent.[241] If the state attempted to keep the factual innocence of such individuals secret, in order to maintain citizens' faith in the system, then such secrecy would almost certainly lead officials to abuse their power for political or personal reasons.[242] Thus, punishing someone only when the main theories of punishment converge could well produce the best overall consequences.

A Meta-Theoretical Argument

Thirdly, theorists have grounds for endorsing my approach *qua* ethical decision-makers under conditions of uncertainty about which substantive moral principles are correct. If retributivists and consequentialists step back from their commitment to their own favoured theories and survey the structure of the debate about punishment they should realise that there are considerable grounds for uncertainty about which theory of punishment is correct. It is important to emphasise the sheer number and

[241] My claim that there is widespread support for negative retributivism (and for the principle that it is worse to punish the innocent than acquit the guilty which underlies the BRD standard) is consistent with my claim that people are subject to powerful pro-punishment biases. Support for negative retributivism and the BRD standard are generally conscious commitments, whereas pro-punishment biases are unconscious factors that often cause people to depart unknowingly from such commitments in practice. When people inflict unjust punishment because of these biases, they are not consciously aware that they are acting unjustly.

[242] C.f. J. Bentham, *A Treatise on Judicial Evidence* (Paget, London 1825), 197. R.M. Hare also argued that although consequentialism was correct in principle, due to human fallibility, attempting to apply pure consequentialism to all of one's everyday actions could lead to bad consequences: Hare, *Moral Thinking* (OUP, London 1971). Furthermore, I should qualify my remarks in fn240, by acknowledging that consequentialist reasoning is not immune from biases, e.g. self-interest. It is possible that adopting the convergence requirement could help to correct consequentialists' pro-punishment biases as such biases may come to light when reflecting on whether the desert criterion was satisfied.

complexity of the philosophical arguments on the different sides of the debate. In Part One, I outlined some of the difficult hurdles that retributivists and consequentialists would have to overcome in order to make their positions credible. With every hurdle new possibilities for error arise and the cumulative effect of this is to create considerable room for doubt.

As I argued in the previous chapter, theorists who support the BRD standard, because they think that it is particularly bad to punish a person without adequate justification, should also take seriously the doubts about which theory of punishment is correct. They have reason to adopt my convergence requirement because it minimises the chance of unjust punishment.[243]

This argument invoked one set of substantive ethical principles (the non-consequentialist basis for the BRD standard) to deal with moral uncertainty about another set of substantive ethical principles – theories of punishment. However, some theorists will doubt the first set of non-consequentialist ethical principles and the BRD standard itself. Therefore, I will now present a *meta-theoretical* argument for taking a cautious approach to punishment.

The first step in this argument is to recognise that there are some ethical considerations/principles about whose soundness there is a relatively high degree of certainty and that there are other moral principles whose soundness is much less certain. Secondly, it seems sensible to give a privileged status in our decision-making to those ethical considerations about whose soundness there is most certainty, i.e. we should safeguard these them from being overridden too easily. Thirdly, it is submitted that we have grounds for having a relatively high degree of credence in the importance of an ethical consideration if respected experts who have thought seriously about the matter agree that it counts as an important consideration. It is

[243] As I will explain further below.

useful to introduce Benjamin Vilhauer's term 'basic reason' to describe moral considerations about whose validity we can be most certain.[244] According to Vilhauer, a reason is basic if all mainstream ethical theories agree that we have this reason. To 'safeguard' a basic reason (about whose soundness we are very certain) involves only allowing it to be overridden in limited circumstances. Specifically, we should only allow other considerations to override a basic reason if we have a similarly high degree of credence that it is justified to do so. Again, it is submitted that we can have grounds for such credence if experts agree that overriding the basic reason in these circumstances is justified.

A plausible candidate for a basic reason is the following: People have a strong reason not to deliberately inflict serious harm on others. Call this reason 'Harm Avoidance'. Different theories provide different explanations of why we have this reason. For instance, as Vilhauer puts it: 'Kantian deontologists explain [Harm Avoidance] as an imperative to which we must conform unless in harming someone (e.g. by punishing him) we do not use him as a mere means. Virtue ethicists might explain it as a principle followed by those who possess the virtue of justice except in cases where justice permits or requires harm. Act-utilitarians explain [Harm Avoidance] as derived from our reason to maximise overall happiness….' However, despite the difference in the explanations that these theorists would give for *why* Harm Avoidance counts as a reason, it is difficult to deny that all mainstream ethical theories agree *that* we have a strong reason to avoid deliberately inflicting serious harm on others. It would be very difficult (or impossible) to imagine revising any of

[244] Vilhauer introduced the concept of a basic reason in order to argue that free will scepticism (unlike scepticism about induction and about other minds) does not undermine our reason against intentionally harming others. He further argues that free will scepticism actually strengthens this reason. B Vilhauer, 'Taking Free Will Skepticism Seriously' (2012) 62 *The Philosophical Quarterly* 833, p849. However, unlike the approach adopted in this thesis Vilhauer does not argue that Harm Avoidance should only be overridden (in the context of punishment) when there is a high level of agreement among theorists that doing so is justified. Instead, in other work, he defends a Rawlsian theory of punishment: Vilhauer B, 'Free Will Skepticism and Personhood as a Desert Base' (2009) 39 (3) Canadian Journal of Philosophy 489.

these ethical theories in a way that would remove this reason and yet preserve the theory as a whole. Ethical theories differ regarding what reasons, if any, could override Harm Avoidance – call these 'overriding reasons'. For instance, in the context of punishment, retributivists believe that Harm Avoidance can be overridden in the case of a blameworthy offender who deserves punishment, and consequentialists about punishment believe that Harm Avoidance can be overridden in situations where punishing someone would promote the general welfare. Because there is a lack of agreement among experts about the retributivist's overriding reason (and hence uncertainty about its status) my meta-theory of moral uncertainty implies that we should not override Harm Avoidance, based on the retributive consideration alone. Equally, it implies that we should not override Harm avoidance based on the consequentialist consideration alone.

A consequentialist would object to the idea that it is intrinsically worse to inflict active, intentional harm than to allow harm to occur unintentionally. It is therefore important to make clear that my meta-theoretical approach to moral uncertainty is not based on the idea that there is convergence on the principle that the active, intentional infliction of harm is *worse* than allowing harm to occur unintentionally. Rather, I argue i) that there is convergence on the idea that there is *a strong reason against* the active, intentional infliction of harm and ii) that there is a lack of convergence on the principle which consequentialists claim can override this reason (their 'overriding principle'), i.e. that inflicting active intentional harm is better than unintentionally allowing a greater harm to occur. Propositions i) and ii), when taken together with the idea that we should safeguard reasons about which there is convergence and only allow them to be overridden by principles about which there is a similar level of convergence, imply, in practice, that avoidance of active, intentional harm is given a privileged status. However, this argument for its privileged status is a meta-theoretical argument based on considerations about

uncertainty, not an argument that assumes the truth of a non-consequentialist theory or of a consequentialist theory.

So, given moral uncertainty about theories of punishment, in what circumstances is it justifiable to override Harm Avoidance and inflict harm on an offender? Vilhauer proposes that theorists who give Harm Avoidance a privileged status in their decision-making should opt for a theory of punishment which (compared with other theories) would recommend subjecting a relatively small number of people to punishment and which would, overall, recommend relatively lenient types of punishment. Vilhauer believes that a Rawlsian theory of punishment is the most promising candidate.[245] However, minimising the instances where Harm Avoidance is overridden is not the only thing that those who wish to safeguard Harm Avoidance should be concerned about. It is also important that, in those cases where Harm Avoidance is overridden, we can have a high degree of confidence that this is justified. In chapter 3 I mentioned one reason for doubting that Vilhauer's Rawlsian approach can provide such justification (the doubt pertained to the ability of Vilhauer's theory to explain what was wrong with punishing the innocent). More generally, it seems that we cannot have a sufficiently high degree of credence in any highly contentious theory of punishment to warrant relying on that theory of punishment alone as a basis for overriding Harm Avoidance. To determine what considerations could justify overriding Harm Avoidance, we should again attempt to minimise uncertainty by looking to areas of agreement among ethical theories.

There are two types of agreement that might be invoked here. I will call these contrasting kinds of agreement 'convergence' and 'consensus'.[246] I will use the term

[245] I briefly discussed aspects of Vilhauer's account in Chapter 3.

[246] My use of these terms is similar to their usage in the literature on 'public justification'. For a general overview of this literature see: Vallier, K and D'Agostino, F, 'Public Justification' (2013) in The Stanford Encyclopaedia of Philosophy, available at: http://plato.stanford.edu/entries/justification-public/ [Accessed August 2013]. Roughly speaking, theorists who advocate public justification believe that the state should only use coercive force on the basis of reasons that all citizens can accept

'convergence' to refer to agreement on decisions about what to do – in this context, the decision about which individuals may be harmed. There is considerable agreement that the state may impose hardship on certain types of offender. Both retributivists (of all varieties) and non-retributivists (including consequentialist and non-consequentialist theorists) would agree that the state is entitled to constrain the liberty of dangerous violent and sexual offenders. However, retributive and non-retributive theories provide different reasons for this decision. A pure positive retributivist, *qua* positive retributivist, would only endorse the decision to impose hardship on dangerous violent offenders *because those offenders are violent criminals* (i.e. because they have culpably committed violent crimes), not because they are dangerous.[247] Many non-retributivists, on the other hand, do believe that the decision to interfere with such offenders is justified, at least partly, because the offenders are dangerous and need to be incapacitated.[248] To say that ethical theories converge on a decision to impose hardship on an offender implies that these theories each provide a different, positive justification in favour of imposing hardship on that offender (and that these justifications are logically independent of each other).

In contrast, I will use the term 'consensus' to refer to agreement on the rationale for a decision. In addition to convergence on decisions, there may also be some degree of

---

as being valid *for them* (i.e. the state's reasons count as reasons within the citizens' own value-systems). Public justification theorists might favour my convergence requirement. However, my defence of the convergence requirement is based on *epistemic* considerations – i.e. given uncertainty about the justification for coercive force, the convergence requirement maximizes the chance that such force is justified. In contrast, public justification theorists are typically motivated by considerations of what is required in order to show *respect* to citizens within a liberal democracy, the idea being that such respect requires the state to justify its use of force to all citizens in terms they can see as valid.

[247] 'Positive' retributivism refers to the idea that the offender's moral culpability provides a positive reason *in favour* of punishing him. 'Negative' retributivism is the idea that we should *refrain* from punishing someone if he is not morally culpable.
Culpability plays a different role in different types of retributive justification for punishment. For instance, according to Michael Moore's retributive theory, the offender's moral culpability makes it intrinsically good that he suffers for his crime. In contrast, according to RA Duff's variety of retributivism, the offender's culpable commission of a crime means that it is appropriate to punish him in order to bring him to recognize and repent the wrong he perpetrated.

[248] They may also justify such interference on the basis that it promotes goals such as deterrence, reform or rehabilitation.

consensus among theorists about the reasons to override Harm Avoidance. I have argued that, due to moral uncertainty (e.g. doubts about whether we have free will), retributive reasons alone are not credible enough to override Harm Avoidance. Similarly, we should not override Harm Avoidance whenever unconstrained consequentialism recommends this, because there is also moral uncertainty about consequentialism (e.g. the counterintutiveness of some of the implications of pure consequentialism and concerns about using people as a means). A theorist who accepts these arguments about moral uncertainty might look for a reason to override Harm Avoidance that would still be available if pure consequentialism and pure retributivism were rejected. If a theorist, who had previously endorsed retributivism, comes to believe that there is not enough certainty about the soundness of retributivism to allow it to override Harm Avoidance, what should this theorist do? Surely, it would not be rational for such a theorist to say that, if retributivism can no longer provide a sufficient basis for interfering with the liberty of rapists and murderers, such offenders may not be interfered with at all. True, a retributivist, *qua* retributivist, would not take dangerousness to be part of the justification for interference with an offender's liberty. But that same theorist, *qua* ethical theorist under conditions of moral uncertainty, would almost certainly do so. As Murtagh, puts it:

"Everyone seems to agree [that]…. even if they are not morally responsible for their crimes, it will still be justifiable to incarcerate habitually violent and dangerous offenders in order to protect the rest of society. That point is difficult to argue with, given the absurdity of allowing serial murderers and rapists to roam free."[249]

Derk Pereboom defends his incapacitation approach to dealing with offenders in this way. He reasons that there is at least a reasonable doubt about the soundness of

[249] K. Murtagh, 'Free Will Denial and Punishment' (2013) 39 (2) Social Theory and Practice 223.

retributivism and that therefore retributivism should not be relied on to justify punishment. He then claims that punishing people for reasons of general deterrence uses them merely as a means and should be ruled out on that basis.[250] He then concludes that the need to protect people from dangerous offenders is something that any rational person will agree can warrant state incapacitation of such individuals. He draws an analogy with quarantine: if the state can justifiably protect itself against carriers of dangerous diseases by subjecting them to quarantine, then it has the right to protect itself from dangerous offenders by incapacitating them. However, there is an important consideration that Pereboom fails to recognise and which has important implications for the kind of evidence of 'dangerousness' that can justify incapacitation.

The consideration that Pereboom overlooks is that we do not know beyond reasonable doubt that retributivism is false. As Stephen Kearns has recently explained, when we survey the state of the free will debate, there is a strong argument for 'free will agnosticism' rather than for *certainty that we lack* free will in the sense required for retributive responsibility.[251] If we wish to safeguard Harm Avoidance, therefore, it is rational to treat appeals to free will and retributive responsibility differently depending on whether they feature in arguments for overriding Harm Avoidance or for strengthening Harm Avoidance. Free will may be invoked as a reason for overriding harm avoidance when it features in a 'positive' retributivist argument *for inflicting hardship* on someone because he deserves it, or *for increasing the penalty* to proportionately reflect his desert. In contrast, free will can be used in order to strengthen Harm Avoidance (or in order to undermine a

[250] Although Pereboom uses the language of uncertainty and reasonable doubt when arguing against a retributive system he does not invoke uncertainty when arguing against a system of general deterrence. Considerations relating to uncertainty would considerably strengthen the latter argument as well as the former, since even those theorists who are sympathetic to a general deterrence approach must recognize that there is at least a reasonable doubt about whether such an approach is ethically defensible.) See, e.g., Pereboom, Living Without Free Will (OUP, Oxford 2001), chapter 6.
[251] S. Kearns, 'Free Will Agnosticism' (2013) Nous (Online First). DOI: 10.1111/nous.12032

potential overrider of Harm Avoidance) when it features in a 'negative' retributivist argument *against harming* someone who does not deserve it, or *against* inflicting a form of hardship that exceeds the individual's desert.

The possibility that we might have free will and that negative retributivism might be sound casts doubt on the justifiability of incapacitating people on the basis of predictions of dangerousness that are unrelated to their past conduct (e.g. predictions based on the results of a genetic test or a brain scan revealing biological traits that are associated with a disposition to act violently.) The negative retributivist might argue that detaining a person (at least a sane, rational agent) on the basis of such data would show disrespect for that person's free will and would inflict an *undeserved* hardship on him.[252] It would be reasonable for a theorist acting under conditions of moral uncertainty to take the above considerations into account in the following way. Firstly, the theorist should have a high degree of credence that there is a strong reason to avoid deliberately inflicting harm others (due to the agreement among mainstream theorists that we have this reason); secondly the theorist should also have a high degree of credence in the common sense idea that violent criminals cannot simply be allowed to roam free; and, thirdly, the theorist can give due weight to doubts about pre-emptive incapacitation by insisting that assessments of dangerousness are based at least partly on evidence of past seriously harmful behaviour. Thus, arguably, theorists under conditions of moral uncertainty would reach consensus about the proposition that Harm Avoidance can justifiably be overridden to the extent that this is necessary in order to incapacitate individuals who, through their own conduct, have shown themselves to be dangerous.

[252] There are also other considerations, which are independent of free will and retributivism, which cast doubt on the justifiability of incapacitating those who have never yet offended, such as trust-based and personhood-based considerations. See chapter 3, above.

To summarise this section on convergence and consensus: both of these ideas involve trying to minimise the chance of inflicting unjust harm by only imposing hardship on an offender if there is agreement among different theorists that doing so is appropriate. As I have said, convergence involves agreement on the decision to impose harm on someone and consensus involves some degree of agreement on the reasons for this decision. It is submitted that the convergence approach and the consensus approach are both rational responses to moral uncertainty. They both reduce the chance of moral error, but in different ways. The convergence approach minimises the risk of inflicting unjust harm by providing more than one plausible rationale for a decision to impose hardship on an offender (e.g. positive retributivism and societal protection). This provides a 'theoretical safety net' – i.e. even if positive retributivism fails to deliver an adequate justification for imposing hardship, the decision may still be justified on the basis of societal protection and *vice versa*. The consensus approach minimises the risk of inflicting unjust harm by only harming offenders if there is a reason for doing so in which all theorists can have a high degree of credence.

Both approaches have the same results in practice, i.e. they make the same recommendations about when the state may interfere with an offender's liberty. The consensus approach would recommend state interference when an offender has already engaged in conduct which constitutes a seriously wrongful infringement of the fundamental interests of others and when the offender is, at least partly on that basis, judged to be dangerous. The convergence approach would also make this recommendation, because, if the state were only to impose hardship on offenders when mainstream theories converged on that outcome, then, in effect, state intervention would be constrained by those theories which are most likely to object to harm imposition (call these 'harm-restrictive' theories). These harm-restrictive theories are negative retributivism (which would require proof that the offender

engaged in seriously wrongful conduct) and incapacitation theory (which would require proof of the offender's dangerousness).

For ease of exposition, I will continue to refer to my proposal for dealing with moral uncertainty about punishment as the 'convergence requirement', i.e. the requirement that offenders are only punished when mainstream theories agree that this is appropriate. However, it should be noted that this requirement could also, in effect, ensure that offenders are only subject to state coercion when there is a single reason for doing so that ethical theorists under conditions of moral uncertainty would agree on (i.e. when there is 'consensus').

In the next section, I will explain further how precisely the convergence requirement minimises the chance of punishing someone unjustifiably.

How Does The Convergence Requirement Minimise the Risk of Punishing Someone Unjustifiably?

The idea that, under conditions of moral uncertainty, it is rational to seek convergence among mainstream theories relies on certain assumptions about probability. It assumes that the probability of a disjunction as a whole being true is a function of the probability of the truth each of its disjuncts taken together with the number of those disjuncts. So if retributivism recommends punishing a particular person (and retributivism has a certain probability of being true) and consequentialism also recommends punishing that person (and consequentialism has a certain probability of being true); then the probability that this person ought to be punished is higher than if we relied on one of the disjuncts alone.

This point can be illustrated with the following example:

Betting Example

Imagine you have two options:

Option 1: You win £100 if A turns out to be true.

Option 2: You win £100 if either A or B turn out to be true.

Clearly, it would not be rational to bet on Option 1 in preference to option 2. This is because of a basic rule of probability theory called the 'disjunction rule', which states that:

"….the probability of A-or-B can be smaller than neither the probability of A nor the probability of B, since it contains both."[253]

[253] M. Bar-Hillel and E. Neter, 'How Alike Is It Versus How Likely Is It: A Disjunction Fallacy in Probability Judgments' (1993) 65 Journal of Personality and Social Psychology 1119, p1119. The disjunction rule is a special case of the extension rule, which states that, 'if A is a subset of B, then the probability of A cannot exceed that of B'. Bar-Hillel and Neter, p1119.

Furthermore, one can make the stronger claim that the probability of the truth of pure-retributivism-or-pure-consequentialism is *greater* than the probability of the truth of pure retributivism, or the probability of pure consequentialism individually. This follows from the fact that we have grounds for according a probability greater than zero to each of the mainstream theories of punishment and from the 'restrictive disjunction rule', which states that P(A-or-B) = P(A) + P(B).[254] The restrictive disjunction rule applies when the options are mutually exclusive, which pure retributivism and pure consequentialism are, by definition. (The options can still be thought of as mutually exclusive, even if we include a 'mixed theory' as one option, since, by definition, if a mixed theory is true, then neither *pure* retributivism, nor *pure* consequentialism can be true.)

We have grounds for according an epistemic probability significantly greater than zero to each of the mainstream justifications of punishment, because they have struck people who have thought seriously about the matter and who are well-informed about the various arguments (i.e. 'experts' on punishment) as being the most plausible theories. Relying on one of these theories alone is more likely to result in unjust punishment than if we were to adopt the convergence requirement, because it is more probable that any one of the most plausible theories will be wrong than that *every one of them* will be wrong.

This point can also be put in terms of 'reasonable doubt'. The convergence requirement decreases the chance of punishing someone unjustifiably through *reducing grounds for reasonable doubt* about the justifiability of punishing that person. All theorists, regardless of their own theoretical commitments, should recognise that consequentialism and retributivism are both 'reasonable' positions to take, in the sense that the tenets of these theories are intellectually respectable, not

[254] i.e. the probability of A-or-B is equal to the probability of A plus the probability of B.

'fanciful' or 'frivolous'. If, one of these theories recommends punishing someone, but another theory opposes this, then this latter theory gives rise to a reasonable doubt about whether punishing this person is justifiable. However, if the mainstream, reasonable theories (i.e. the best theories we have) agree that punishing someone is justified, then the overall case for punishing that person can be said to reach the beyond reasonable doubt standard.

Uncertainty About The Convergence Requirement

It might be objected that my convergence requirement does not satisfy itself. I have argued that we should only punish someone if the moral argument for punishing that person is established to a high degree of credibility and that this standard has been reached if the main theories of punishment agree that the person should be punished. However, this is just one possible approach to moral uncertainty about punishment theories and, as I have acknowledged, there is a debate about which theory of moral uncertainty is correct.

Firstly, in response, it is important to emphasise that my convergence requirement is not meant to be a requirement that all moral principles must satisfy. Rather it is a requirement that arguments for punishment should satisfy. But the convergence requirement is itself not an argument for punishment. It is a constraint on punishment. It is therefore not obvious that it needs to satisfy itself.

Secondly, if second order uncertainty (i.e. uncertainty about which theory of moral uncertainty is correct) is a problem for my approach, it is also a problem for every theory of punishment, since all punishment theorists must necessarily adopt some theory of moral uncertainty. All punishment theorists must recognise the grounds for

being uncertain about which theory of punishment is correct (unless they are epistemically arrogant). Given this recognition, *any* approach that the punishment theorist then advocates is a response to moral uncertainty. Even if she advocates the 'simple approach' of acting on her preferred theory of punishment, despite the uncertainty about its soundness, this is itself a response to moral uncertainty. By default, the most common approach to moral uncertainty is the simple approach. However, the simple approach often involves internal inconsistency. As I have noted, many punishment theorists have an *even higher* degree of credence in the beyond reasonable doubt standard and in the proposition that punishment requires strong justification, than the degree of credence they have in their own substantive theory of punishment, e.g. retributivism. If they advocate acting on retributivism, despite the doubts about its soundness, they are ignoring principles in which they have an even higher degree of credence. So they are deviating from their own approach to moral uncertainty: that one should act on those principles in which one has the highest degree of credence.

Thirdly, my approach seems to be more internally consistent that the simple approach (which is currently the dominant approach). My approach recommends seeking convergence between theories of punishment and there is convergence on at least part of my theory of moral uncertainty itself. Most punishment theorists agree that punishment requires considerable justification, even though they disagree about which theory of punishment is the best. They also agree about the beyond reasonable doubt standard. My convergence requirement, unlike the simple approach, incorporates these agreed elements.

Who Counts As An Expert On Punishment?

I argue that we ought to minimise the risk of punishing someone unjustifiably and that, given the moral uncertainty about which theory of punishment is correct, it is

unacceptably risky to punish someone based on a single theory of punishment alone. I claim that we have firmer grounds for believing that it is justifiable to punish someone if experts on the subject agree that punishing that person is justifiable. In order to defend this claim, I need to say more about the idea of an 'expert'.

Alvin Goldman sets out a number of plausible criteria for identifying who should count as an expert within a certain domain of knowledge (the 'E-domain'). His first criterion relies on the idea that the concept of expertise is partly *comparative:*

1. An expert must have 'more beliefs (or higher degrees of belief) in true propositions and/or fewer beliefs in false propositions within that domain than…the vast majority of people do.'[255]

However, expertise (in an objective sense) is not entirely comparative. If the vast majority of people's thinking about a certain matter is entirely riddled with false beliefs and the putative 'expert' is superior to them in having a slightly smaller number of false beliefs, she still does not count as a genuine expert in the matter, in an objective sense. So Goldman's second requirement for a person to qualify as an expert is:

2. The possession of 'a substantial body of true beliefs' in the E-domain.[256]

In addition to having this body of accurate information, an expert must also have:

3. The disposition and set of skills necessary to use this information to form true beliefs in answer to new questions about the E-domain.[257]

[255] A. Goldman, 'Experts: which ones should you trust' (2001) 63 (1) Philosophy and Phenomenological Research 85-110, p91.
[256] Ibid.
[257] Ibid.

It might be objected at this point that the second criterion poses a problem for my approach. Given the (seeming) incompatibility between various theories of punishment (say, a pure consequentialist theory versus a pure retributive theory), presumably punishment theorists from opposing camps cannot all be right. It might therefore be thought that some of these theorists must lack a 'substantial body of true beliefs' about the justification of punishment and cannot qualify as experts. Given the uncertainty about which theory is right, there would seem to be a corresponding uncertainty about who would count as an expert, and hence it might be claimed that relying on the notion of expertise, as part of my strategy for minimising this uncertainty, is problematic.

In response to this objection, it is helpful to clarify the second criterion by invoking Goldman's distinction between 'primary' and 'secondary' questions in a domain of expertise.[258] Primary questions are the main, substantive questions that researchers on a topic are ultimately trying to find answers to (e.g. what are the prerequisites for criminal responsibility and fitness for punishment; what aims should a system of punishment serve etc.?). Secondary questions concern the arguments or evidence that are relevant to answering the primary questions (e.g. the various arguments for and against the compatibility of determinism and retributive responsibility; people's intuitions about punishing the innocent etc.). Secondary questions also concern the views of the leading researchers in the field about such arguments and evidence (e.g. what do prominent theorists of punishment say on the issue of compatibilism/incompatibilism; or the significance of intuitions about punishing the innocent).

To count as an 'expert' for the purposes of my argument, a punishment theorist must have a substantial body of true beliefs about the secondary questions about

[258] ibid.

punishment, i.e. she must be familiar with the main arguments and evidence for the different positions and the views of prominent researchers about such arguments and evidence. The views of two opposing camps of theorists may sharply diverge regarding some of the conclusions that these arguments and this evidence support (i.e. they diverge about some of the primary questions) and yet both sets of theorists remain experts. In addition, even though the ultimate primary question (what is the correct justification of punishment?) is unsolved and hence at least one camp of opposing experts must lack true beliefs about this question, all of the theorists have true beliefs about *some* of the primary questions. These include conclusions about what characteristics a system of punishment should not have, e.g. that people should not be punished purely because of morally irrelevant characteristics such as poverty or height. In addition, as I have emphasised in this thesis, there is considerable agreement between theorists about certain conclusions about justified punishment – i.e. which specific people should be punished. Therefore, the mere fact that experts hold opposite views about theories of punishment does not mean that their views about the justifiability of punishing specific individuals will conflict, and therefore they could all be right about *these* conclusions.

Implementing The Convergence Requirement

The main aim of this thesis is to make the moral case for taking seriously the uncertainty about the soundness of retributive responsibility and about theories of punishment more generally. I have attempted to show that, in principle, the convergence requirement is one promising response to this uncertainty. There is not scope within this thesis to discuss the numerous practical issues that would need to be addressed to demonstrate precisely how the requirement should be implemented. Instead, in this section, I will lay the foundations for future work by briefly outlining some considerations which are relevant to deciding at which stages of the criminal

justice process the convergence requirement should be taken into account and how it should be applied at each stage. Part 3 will then examine in more depth the practical implications of my approach for one specific example of a possible response to criminal behaviour – the use of direct brain interventions in rehabilitation programmes.

Sentencing

The convergence requirement is relevant at the sentencing stage, since at this stage serious hardship may be imposed on offenders. Some of the most severe forms of punishment, such as the death penalty, should be ruled out entirely, because a number of mainstream theories fundamentally reject these forms of punishment. Some mainstream theories also oppose confining offenders under the harsh conditions that currently prevail in American and British prisons.[259] However, agreement could be reached on the permissibility of subjecting certain offenders to confinement under less severe conditions.

1) Confinement

As, indicated above, incapacitation theories (of the kind advanced by Derk Pereboom) and negative retributivism are the theories that are most likely to object to a proposal to deprive an offender of liberty. Negative retributivists would require that, for confinement of a (sane) offender to be permissible, the offender must have committed a sufficiently serious offence and must have met the retributivist's criteria for moral responsibility. Negative retributivists would oppose consequentialist arguments for confining sane individuals who have not met these responsibility criteria, who have committed no offence, or whose offence is minor. Negative

[259] For a discussion of the effects of harsh prison conditions on offenders see: D Bierie, 'Is Tougher Better? The Impact of Physical Prison Conditions on Inmate Violence' (2012) 56 International Journal of Offender Therapy and Comparative Criminology 338. Regarding conditions in English prisons see: C Kruttschnitt and A Dirkzwager, 'Are There Still Contrasts in Tolerance? Imprisonment in the Netherlands and England 20 Years Later' (2011) 13(3) Punishment and Society 283-306.

retributivists would also oppose confinements that were lengthier or imposed more physical or psychological hardship than the offender deserved. Incapacitation theorists would oppose consequentialist or retributive proposals to confine non-dangerous individuals. Incapacitation theorists would also reject proposals to confine individuals for longer than is necessary to prevent them from reoffending, or under conditions that are designed to be harsh. Therefore, with certain qualifications, the convergence requirement implies that confinement is only permissible in the following circumstances: 1) the offender meets the retributivist's responsibility criteria (something that should be determined at the conviction stage, discussed below); 2) the offender is dangerous; 3) the confinement is not a disproportionately severe response to the wrong the offender committed; and 4) the confinement is not longer than needed to prevent the offender from being a danger to society and the conditions of confinement, in themselves, are not designed to inflict severe physical or psychological suffering.

One qualification that should be added at this point is that the convergence requirement must be balanced against the need to preserve society. One might worry that a system that confined offenders under relatively benign conditions would lack a deterrent effect and that this might seriously threaten society by leading to a drastic increase in offending and widespread chaos. If implementing the convergence requirement strictly would seriously threaten the existence or functioning of civilised society then the requirement should be diluted. We should seek the greatest level of agreement on decisions to inflict severe hardship on offenders that is compatible with the need to preserve society. If insisting on unanimity between mainstream theories is incompatible with the preservation of society, then my approach would recommend seeking a level of agreement short of unanimity. Placing importance on the need to preserve civilised society is consistent with my general approach to moral

uncertainty. I have proposed that when deliberating under conditions of moral uncertainty we should recognise that there are some ethical considerations about whose importance there is a relatively high level of certainty and that we should privilege these considerations in our decision-making. We can be very certain that there is a strong reason against seriously harming others and that we should only override that reason if the justification for doing so has reached a very high standard of credibility. However, we can be equally certain that it is important that serious crime does not escalate to a level where society is at risk of disintegration. Therefore, we should not hold justifications for imposing hardship on offenders to such a high standard that this scenario would occur.[260]

Having said this, it is doubtful that the reduction in the use and severity of confinement that would result from the convergence requirement would seriously threaten society. As Pereboom points out, people generally place such a high value on their liberty that the fear of being deprived of it would probably be an adequate deterrent, even if the conditions of confinement were relatively benign. Furthermore, evidence from numerous studies indicates that a potential offender's belief in the likelihood of being caught is much more important in deterring crime than the severity of the sanctions that would be imposed if he were convicted.[261] If these studies are correct, they provide a reason for diverting the resources that are currently spent on keeping offenders in prison for very long periods into policies that increase the likelihood that those who fit the criteria for confinement are apprehended. It should also be noted that Scandinavian criminal justice systems have more benign prison regimes and lower rates of imprisonment than many other Western countries.

[260] For an example of the 'social disintegration' objection to hard incompatibilism see: Smilansky, S, 'Hard Determinism and Punishment: A Practical Reductio' (2011) 30 Law and Philosophy 353.
[261] For an overview of such evidence see: D Nagin, 'Deterrence in the Twenty-First Century' (2013) 41(1) Crime and Justice 199. For a hard incompatibilist reply, see: N Levy, 'Skepticism and Sanction: The Benefits of Rejecting Moral Responsibility' (2012) 31 Law and Philosophy 477.

Yet these policies have not resulted in high levels of crime in Scandinavian countries.[262]

Another qualification concerns how to assess the 'severity' of the confinement. *Prima facie*, a longer period of confinement seems more severe than a shorter period. However, it is important to take account of the conditions under which an offender is confined, including the presence of rehabilitative/therapeutic interventions, and the effect that such interventions may have on the length of confinement necessary to protect society. A particular rehabilitative intervention might allow offenders to be released earlier into society, but may adversely affect offenders in other ways. Part Three of this thesis will address this issue, focussing on the example of neurological interventions.

One criticism that has been made of Pereboom's hard incompatibilist model of responding to criminal behaviour is that it recommends incapacitating offenders until they are judged to be no longer dangerous.[263] This seems to imply the use of indeterminate sentences. Such sentences may be thought unjust (because they are potentially disproportionate to the offence) and it may be thought unfair or inhumane to leave offenders in a state of complete uncertainty as to when they may be released. Unlike Pereboom's model, the convergence requirement would suggest that a determinate maximum limit should be set on an offender's period of confinement and that the offender should be informed of this when the sentence is passed. This is because the convergence requirement takes account of the possibility that negative retributivism might be sound and therefore requires that the sentence of confinement

[262] See e.g. T Lappi-Seppala, and M Tonry, 'Crime, Criminal Justice and Criminology in the Nordic Countries' (2011) 40 (1) Crime and Justice 1.

[263] M Corrado, 'Why Do We Resist Hard Incompatibilism? Thoughts on Freedom and Punishment' in T Nadelhoffer (ed.) *The Future of Punishment, ed. Thomas Nadelhoffer* (Oxford OUP 2013)

is not a disproportionately severe response to the wrong the offender committed. However, the offender should be released early, if he is no longer dangerous.

2) Rehabilitation and Community Service

Participation in rehabilitation programmes or community service may be ordered instead of confinement. The convergence requirement implies that a more lenient sentence (e.g. a non-custodial one) should be preferred to a harsher one (e.g. confinement) where a more lenient sentence would sufficiently reduce the chance of the offender reoffending. Such rehabilitation programmes or community service orders may involve an element of moral communication. For instance, they may be designed to bring the offender to recognise the wrong he has done; to reflect on the psychological dispositions that led him to commit that wrong; to attempt to modify those dispositions; and, where appropriate, to make symbolic reparation to his victims. As will be explained below, when discussing the conviction stage, such moral communication can be acceptable to hard incompatibilitsts, since their theory does not object to holding others morally responsible in the 'moral appraisability' sense.

Furthermore, the cautious approach to punishment defended in this thesis would recommend measures that involve an element of moral communication over ones that do not. This is because the cautious approach to punishment is designed to protect the interests of the individual who may be subject to punishment (or similar coercive measures). Among the key interests people have is that they are not made to suffer and that they are not deprived of their liberty. Hence arguments for infringing such interests must be justified to a high standard of credibility. However, offenders plausibly have other interests that are worthy of protection. These include their interest in being respected as a rational agent. Rehabilitation programmes can show respect for offenders' rational agency by *explaining* to offenders why their liberty

needs to be interfered with (e.g. by pointing out the impact of criminal conduct on victims) and which present offenders with good *reasons* why they should not reoffend. (I will discuss the idea of respecting rational agency further in Part 3.)

3) Fines

Since mainstream theories of dealing with criminal behaviour tend to endorse the use of fines in some circumstances, it seems that this measure could, in principle, satisfy the convergence requirement. If a fine can prevent an offender from reoffending, then theories that restrict the use of confinement to incapacitation of the dangerous would recommend fines over confinement. Pereboom defends one such incapacitation theory. He goes even further than this and suggests that fines may even be imposed for reasons of general deterrence. It might be wondered whether this suggestion is compatible with his rationale for using coercive measures against offenders (which, as I mentioned earlier, relies on an analogy with quarantining carriers of infectious diseases). He objects to imprisoning people to deter others, because this involves 'using' them, but claims that incapacitating those who pose a threat to others is an exception to the prohibition on using people, if it is done strictly to eliminate that threat.[264] However, he defends imposing fines for reasons of general deterrence on the basis that this is fairly minor infringement of the prohibition on using others, given that the right to property is less weighty than the right to freedom of movement. This seems plausible provided that limits are set on the levels of fines that take into account the offender's income, so that the fine does not inflict harm

[264] D Pereboom, *Living without Free Will* (CUP, Cambridge 2001), chapter 6. On Tadros's theory, in contrast, harming direct threats does not violate the prohibition on using people as a means at all. Such harm involves 'eliminating' the threat they pose. Tadros's argument is plausible, since 'use' is most closely associated with 'manipulation' or 'exploitation', whereas eliminating the harm posed by a direct threat does not seem to be accurately described by those terms. V Tadros, The Ends of Harm: The Moral Foundations of the Criminal Law (OUP, Oxford 2011). Similarly, Quinn argues that it is not plausible to classify eliminating a threat as an example of 'using' someone. According to Quinn, using someone is typically 'opportunistic'. The 'user' sees the other person as presenting an opportunity positively to further the user's goals – the user benefits from the presence of the person who is used. In contrast, when someone eliminates a direct threat, they are removing an obstacle to their goals. W Quinn, 'Actions, Intentions, and Consequences: The Doctrine of Double Effect' (1989) 98 (4) Philosophy and Public Affairs 334.

severe enough to jeopardise the offender's chance of leading a 'reasonably good life'.[265] This should not be too controversial, since it would be unreasonable to hold that the prohibition on using people is absolute. For example, if an attacker wielding a weapon were pursuing someone, it would be wrong for the intended victim to grab a bystander and use him as a human shield - resulting in the bystander being stabbed to death; but it would not be wrong to push a bystander over, merely causing him a minor injury, if this were the only means by which the intended victim could save his own life. To prohibit, on the basis of the means principle, pushing the bystander over would seem fanatical.

Conviction

I have said that one implication of the convergence requirement is that the state should only punish those individuals who appear to have satisfied the retributivist's responsibility criteria. It seems sensible for this to be determined at the conviction stage. I now need to clarify what would satisfy the retributivist's responsibility criteria. I argued in Part 1 that, strictly speaking, there is good reason to doubt whether the criteria for retributive responsibility can ever be satisfied. I argued that retributive responsibility requires libertarian free will and there is little evidence that we have this kind of free will. However, as I noted earlier in the present chapter, the possibility that retributivism is sound and that the kind of responsibility it requires exists provides a reason against harming people who clearly could not be retributively responsible for an offence – which most obviously includes those who have committed no offence at all. Most libertarians believe that possessing free will in the compatibilist sense is also necessary for moral responsibility. In addition, most libertarians seem to assume that those who have compatibilist free will also have libertarian free will. When I refer to 'the retributivist's criteria for responsibility' I am referring to the compatibilist responsibility criteria, which are

[265] Pereboom, ibid, p177.

clearly capable of being satisfied in many cases. The convergence requirement thus protects those whom all retributivists would regard as non-responsible. It is submitted that this way of framing the convergence requirement gives appropriate weight to the concerns of negative retributivism.

The next question is whether the forward-looking aspect of the convergence requirement should be taken into account at the conviction stage. So far, I have argued that the serious hardship involved in *punishment* should only be inflicted if the main theories of punishment agree that this is appropriate. Usually, a conviction is a precursor to the infliction of punishment. However, a conviction, in itself, involves intentionally imposing a hardship on someone (since identifying someone as a wrongdoer is a hardship); even if that person is then spared punishment altogether. It is therefore important to consider under what circumstances forward-looking theories of punishment will agree with retributive ones that a conviction is appropriate.

The conviction by itself is an act of communication. It is a public declaration that the offender has committed a criminal offence. This involves declaring that the offender committed a serious moral wrong, since the offence-definition must specify a serious moral wrong in order to meet the retributivist's demand that *only* wrongdoers are convicted and punished. Making such a declaration can also serve forward-looking purposes. It allows society to express and reinforce their commitment to the values that have been breached, it helps the community to develop its understanding of those values and it can help criminals to reform, by bringing them to recognise the wrong that they have done.

Under what circumstances might a forward-looking theorist object to the conviction (considered by itself, leaving aside the issue of punishment) of serious wrongdoers? It might be thought that hard incompatibilists would object to convicting offenders to

the extent that this involves holding them morally responsible. However, in fact, hard incompatibilists would not oppose holding offenders responsible, provided that responsibility is understood in the sense of 'moral appraisability'. Although there is fierce controversy (and considerable grounds for uncertainty) about the justification of *punishment,* there is very widespread agreement about the appropriateness of *morally appraising* the conduct of offenders. Pereboom, who is at the most sceptical extreme of the spectrum of views on free will, claims that the arguments against free will (while undermining retributivism) give us no reason to doubt our right to morally criticise wrongdoers' conduct. Free will sceptics accept the legitimacy of engaging the wrongdoer in dialogue about the moral quality of his actions and about his reasons for performing them; evaluating what these actions reveal about his attitudes or character traits; and demanding that he engages in self-reflection and apologises for his behaviour. Pereboom writes:

> "The moral responsibility invoked here has been called the *moral answerability* or *the fittingness of providing a moral explanation* sense, and it is the variety of moral responsibility that is most thoroughly ingrained in our practice and least controversial… It may well characterise human interactions across cultures…The main thread of the historical free will debate does not pose determinism as a challenge to moral responsibility as answerability, and free will sceptics accept that we are morally responsible in this sense."[266]

Hard incompatibilists would only object to the practice of condemning offenders purely in order to make them experience psychological suffering. However, if

[266] Pereboom, 'Free Will Skepticism and Criminal Punishment', p51. Pereboom recognises that his conception of moral answerability cannot justify the 'hard treatment' aspect of punishment, without relying on retributive ideas that he rejects. In line with his incapacitation theory, Pereboom proposes that non-dangerous 'criminals' should be allowed to go free, but he does not say whether, in his view, they should be 'acquitted'. Indeed he does not mention how we are to determine whether an individual is a 'criminal' – through a trial or through some other procedure. It is submitted that in most cases the forward-looking benefits that can flow from morally appraising offenders at the conviction stage, justify convicting offenders who have committed serious moral wrongs, even if they are not dangerous. This conclusion seems to be consistent with Pereboom's overall position. (See the text regarding convictions that would produce very bad consequences.)

convicting offenders serves the forward-looking purposes just described, as well as applying only to those whom retributivists regard as responsible, then this would satisfy the convergence requirement.

Prosecution

In some cases where a trial and conviction would clearly fail to meet the convergence requirement (e.g. because it would produce very bad consequences) prosecutorial discretion could be exercised so that the case was not tried.

Criminalisation

Legislators should take into account the convergence requirement when deciding whether to criminalise behaviour (since enacting a criminal law renders individuals liable to punishment if they breach it) and when deciding what criminal penalties to provide for in legislation. Legislators should consider both i.) whether the legislation identifies individuals who would be deserving of punishment in the retributive sense (if retributivism were sound) and ii.) whether interfering with the liberty of those individuals is necessary in order to remove the threat which they pose to society.

Even if the legislature has attempted to take the convergence requirement into account at the criminalisation stage, and has succeeded in drafting an offence that, in general, only applies to individuals whose punishment would satisfy the convergence requirement, there will almost certainly be exceptional cases that the legislature has not foreseen. For instance, the accused may have personal characteristics or circumstances which mean that, in this particular case, the consequentialist rationale for punishment is not satisfied. The judge should be able to take this into account at the sentencing stage, and even impose no punishment on the individual, if one mainstream theory of punishment would recommend that course.

Conclusion

In this chapter I have defended the convergence requirement – the idea that we should try to reduce the risk of inflicting unjustified punishment by only punishing someone if the main theories of punishment agree that punishing that person is appropriate. This chapter began by explaining why people from different theoretical perspectives have reason to endorse my convergence requirement, rather than simply relying on their own favoured theory of punishment. I argued that theorists from different philosophical perspectives have somewhat different reasons for endorsing the convergence requirement. Non-consequentialists have the reasons outlined in the previous chapter for holding arguments for punishment to a high standard of credibility. Retributivists and consequentialists each have some reasons for adopting the convergence requirement, because, due to human fallibility, acting on a single, unconstrained theory of punishment may fail to serve the aims of either theory. However, the main argument this chapter advanced in favour of the convergence requirement was a meta-theoretical argument. After clarifying why exactly the convergence requirement minimises the risk of unjustified punishment, this chapter addressed two potential objections to my view. One challenge stemmed from the existence of uncertainty about which theory of moral uncertainty is sound. The other potential objection concerned my reliance on the idea of agreement among 'experts' on punishment. The final section briefly outlined some issues that are relevant to the implementation of the convergence requirement in practice.

The convergence requirement is designed to protect individuals from unjustified state coercion. It assumes that punishment (and related coercive responses to criminal behaviour) infringes these individuals' interests – in particular their interest in not being seriously harmed - and therefore requires strong justification. This chapter primarily considered the question of which individuals should be subject to coercive

measures at all. The answer it proposed was: only those individuals whom mainstream theories of punishment agree should be subject to such measures. Part Three will explore the following issue: what should be done when mainstream theories agree that an individual should be subject to some sort of coercive measures, but disagree as to which specific measures are appropriate? This question is complicated by the fact that offenders have a number of different interests that should be protected, but these interests can sometimes pull in different directions. In general, offenders have an interest in not being deprived of liberty. Therefore, where one mainstream theory recommends a sentence that involves less interference with liberty, that sentence should usually be preferred. However, I have also indicated that offenders have an interest in being treated as rational agents. Certain rehabilitative interventions might allow offenders to be released earlier into society, but may fail to respect the offender as a rational agent. This problem arises most acutely in connection with the possibility of using direct brain interventions to modify criminal behaviour. Therefore, Part Three will be devoted to examining this example in detail.

It is also important to explore this topic, since it is of particular relevance to the arguments for free will scepticism which I advanced in Part One, when raising doubts about the justifiability of relying solely on a retributive theory punishment. It is tempting to try to explain why certain direct brain interventions are troubling by claiming that they threaten free will. If free will scepticism implied an acceptance of these troubling interventions then this could undermine free will scepticism and could possibly strengthen the case for a retributive system that stressed the importance of free will. However, in Part Three I will argue that, in fact, the objectionable nature of certain direct brain interventions has very little to do with free will and that there are in fact non-retributive reasons for opposing such interventions.

Part Three: Free Will and 'Manipulative' Responses to Criminal Behaviour

Part Three: Overview

Part Two focussed mainly on the following two questions: '*Which* individuals may the state subject to coercive measures within the criminal justice system?' and *'on what basis* are such measures justified?' This section focuses on the question of which *methods* of responding to criminal behaviour are morally acceptable. Theorists often express the concern that without the notions of 'free will' and 'retributive responsibility' we would be unable to explain what is wrong with measures that are intuitively 'manipulative' (e.g. attempting to control the offender's behaviour using biomedical means, such as direct brain interventions). This part of my thesis aims to show how a non-retributive, hard incompatibilist approach to criminal behaviour can be both just and humane.

In Chapter Seven, I argue that the objectionable nature of certain direct brain interventions cannot persuasively be explained in terms of the interventions' effect on free will (regardless of whether 'free will' is understood in a libertarian or a compatibilist sense). In Chapter Eight, I argue that certain forms of biomedical interventions are objectionable and should not be used within the criminal justice system. Unlike many traditional approaches to the issue, my objection to such interventions does not rely on the idea that they necessarily violate the offender's free will (conceived of as a capacity that we can identify just by examining the individual's psychology and actions carefully enough). Instead, the objection is based on the problematic nature of the *relationship* between the intervener and the subject of the intervention. I argue that the state's way of dealing with offenders must be constrained by the principle that the offender must not be objectified, and that his personhood must be respected. Objectifying a group of people typically involves creating a division between 'them' and 'us', which excludes the objectified group and which portrays them as radically deficient in some fundamental respect.

The distinction between objectification (in this sense) and respecting personhood is not an 'all-or-nothing' matter. The state's coercive response to offenders inevitably excludes them *to some extent* and inevitably highlights differences between offenders and other citizens. But the state should not *entirely* exclude offenders and should emphasise commonalities between the offender and other citizens as well as differences. I argue that entirely medicalising the problem of criminal behaviour and attempting to re-shape offenders' values/goals via biomedical means would fall too much towards the objectification end of the spectrum. Rather, attempts to reform offenders should involve engaging the offender in rational dialogue, for example through victim-offender mediation programmes.

I specifically oppose the idea of using biomedical interventions to deal with offenders who are basically rational in an attempt to ensure that the offender adopts one particular set of values rather than another. However, I do not oppose the use of *all* biomedical interventions within the criminal justice system. For instance, such techniques might legitimately be used in order to treat offenders who have mental illnesses. A more controversial question is whether biomedical interventions (including neurological interventions) might be used to enhance the rational capacities of basically normal offenders so that they are better able to *decide for themselves* which goals they should pursue, or which values they should endorse.

In Chapter Nine, I argue that, *in principle,* it would be morally permissible for the state to employ certain types of biomedical interventions (such as 'cognitive enhancements') in a limited way within the criminal justice system, provided that effective enhancements can be developed in the future that have minimal side-effects. This chapter sets out the considerations that need to be balanced when deciding which types of techniques should be permissible and under what circumstances they may be used. It argues that offenders have three main interests

that should be taken into account here. Firstly, the offender's rationality should be respected. This, I argue, requires that the offender gives informed consent to any enhancement and that enhancements are only permissible if they are reversible at the request of the offender. Furthermore, techniques are only permissible if they do not restrict (but rather aim to positively enhance) the offender's capacities for critical reflection. Secondly, the offender should be treated as a member of the moral community. This entails that reforming offenders should occur primarily through relationships with others. Biomedical interventions should never be used to circumvent the need for dialogue with the offender, but rather to enable effective dialogue to occur. Finally, the offender has an interest in not being made to suffer unnecessarily. This may be a consideration in favour of employing biomedical interventions. Attempts to reform criminals that do not involve such interventions may be much more burdensome to the criminal (e.g. requiring a longer time of incarceration) than if these interventions were used. However, this consideration, can also count against the use of biomedical interventions that might expose the individual to (possibly unknown) side-effects.

Chapter Nine then goes on to discuss how we can distinguish techniques that enhance rational capacities from interventions that fundamentally change the person's character, and the extent to which this distinction matters. Examples of particular interventions which I consider include methods of reducing the strength of an offender's violent urges and increasing control over these urges, treatments for sex offenders and techniques for enhancing offenders' empathy. Finally, I consider under what conditions a valid, informed consent to such interventions may be obtained within the coercive context of the criminal justice system.

Chapter Seven: Direct Brain Interventions and Free Will

Introduction

A seemingly obvious objection to attempting to alter criminals' thought-processes through direct brain interventions is the idea that such interventions would deprive the offender of free will. Free will theorists, however, have found the issue of brain interventions surprisingly problematic. This chapter begins by briefly explaining why the libertarian conception of free will does not provide a secure basis for objecting to direct brain interventions. The remainder of this chapter focuses on compatibilist accounts of freedom since compatibilism is probably the dominant approach to free will among philosophers.[267] According to compatibilists, a person can still have free will even though *all* of her thoughts, values and choices are entirely determined by forces beyond her control (e.g. the facts of the remote past and the laws of nature); forces that *completely guarantee* that she would have precisely those thoughts and values and make precisely those choices.[268] Compatibilists persuasively argue that, even if determinism is true, many agents still possess a number of characteristics and abilities that are necessary for free will (some of which are discussed below). One can imagine certain forms of brain manipulation that clearly deprive agents of these characteristics and abilities. However, not all brain interventions have this effect. Yet even in these latter cases, using these techniques to modify behaviour can still seem problematic. This chapter aims briefly to summarise some of the main trends in compatibilist thinking and to explain why it is not always possible to use

[267] Bourget D and Chalmers D (eds.) *The Philpapers Survey 2009*, available at http://philpapers.org/surveys/results.pl Accessed 31st May 2011.

[268] Compatibilists maintain that even if a person lacks control over the factors that determine her action, she may still control the action itself.

compatibilist accounts of free will as the basis for distinguishing 'normal' agents from those who have been subject to troubling types of intervention.

Libertarian Freedom and Direct Brain Interventions

Most libertarians would oppose any type of direct intervention that guaranteed that the agent would act in one particular way. For most libertarians, freedom consists partly in the ability to choose between different alternatives for action, without the outcome of one's decision being guaranteed in advance by prior events.

In a very recent article, John Harris has developed an argument against using direct neurological interventions to morally enhance offenders, which seems to rely on libertarianism. He begins by referring to the image of a forking path, stressing that if an agent is genuinely free then it is possible to choose to go down any one of the available paths. After quoting a passage from *Paradise Lost*, he, agrees with Milton that for an agent to be capable of virtue it must be possible for him to do wrong or to 'fall'. He writes:

> 'Without the freedom to fall, good cannot be a choice; and freedom disappears and along with it virtue. There is no virtue in doing what you must....
> ....[Liberty could be] threatened by any measures that make the freedom to do immoral things impossible....sufficiency to stand is worthless, literally morally bankrupt, without freedom to fall....'[269]

Determinism entails that, given the facts of the past and the laws of nature, nobody could have acted differently from the way in which they actually did act. If a person actually refrained from doing an immoral thing, then given these facts and laws, it was physically impossible for them to have done the immoral thing. Harris's arguments seem to imply that causal determinism is incompatible with true virtue,

[269] J Harris, 'Moral Enhancement and Freedom' (2011) 25 (2) Bioethics 102, pp.105-111.

because virtue requires that it was genuinely possible for the agent to have been vicious. There are, of course, compatibilist interpretations of alternatives possibilities (which will be discussed below). However, Harris's arguments against direct interventions only seem to make sense on a libertarian interpretation. He states that it is a conceptual truth that God himself could not have *guaranteed* that human beings would behave virtuously and still have left us free.[270] This is implied by libertarianism, which states that nothing, not even God, can ensure in advance that a free agent will decide to do one thing rather than another – the future is open right up until the agent makes her choice. However, according to compatibilism, individuals who are predetermined to behave virtuously are still free (all that matters on this compatibilist view is that in some *hypothetical* scenario the agent would have behaved differently). It seems perfectly conceptually possible to imagine a world which God had created in such a way that *every* individual was guaranteed to develop into a virtuous agent and yet retained freedom in the compatibilist sense.[271]

There are various reasons why it is problematic to rely on libertarianism as one's only basis for opposing intuitively objectionable types of direct intervention. Firstly, Libertarian free will requires that certain empirical facts obtain. It requires that human deliberations are (at least sometimes) undetermined. It also requires that they are undetermined in a way that does not merely introduce randomness into our deliberations. Most libertarians themselves concede that we lack epistemic justification for these beliefs.[272] Libertarianism therefore makes the question of whether anyone ever has free will a hostage to empirical fortune. For this reason,

[270] Ibid, p105.
[271] For a discussion of determinism and Christian theology, see: Pereboom D, 'Free Will, Evil, and Divine Providence', in Chignell A and Dole A (eds) God and the Ethics of Belief: New Essays in Philosophy of Religion (CUP, Cambridge 2005) 77.

[272] R Double 'The Moral Hardness of Libertarians' (2002) 5 (2) Philo 226.

many philosophers and legal theorists do not wish to rely on the libertarian notion of free will.

A second difficulty with libertarianism is that it is far from obvious that the 'freedom to fall' is as crucially important as libertarians make it out to be. If someone does good things (e.g. helping others, telling the truth, speaking out against injustice etc.) because they genuinely recognise that there are good moral reasons for doing the right thing then it does not seem wholly inappropriate to call them 'virtuous', without enquiring into whether they were capable of doing morally obnoxious deeds. It may be that certain people have such a vivid awareness of the good (due perhaps to having received an inspiring moral education) that leading an immoral life is not a genuine psychological option for them. It does not seem obvious that such people necessarily lack a freedom that is really worth having.

Compatibilism Part One: The Rational Flexibility Approach

Compatibilist accounts of freedom often emphasise that intentions, beliefs, desires, motives, decisions etc can still have an important role in explaining our actions even if determinism is true. As I explained in the introduction to this thesis, determinism just implies that, if our actions are to be explained by reference to such psychological phenomena as mentioned above, then these phenomena were themselves produced by prior events that were causally sufficient for the occurrence of those psychological phenomena and that those prior events were themselves produced in the same manner by even earlier events etc. in an unbroken chain of cause and effect that can be traced back to before the person was even born. Determinism does not imply that people will not modify their behaviour in response to good reasons for doing so. It merely implies that whether a person recognises and responds to one particular reason for action rather than another at any given time is determined by

prior events in the manner described above. Flexibility – the ability to adapt one's behaviour in an appropriate way to changes in circumstances – is generally agreed to be a hallmark of rationality.[273] If a person would perform a particular action (e.g. eating lunch) no matter what reasons there were against this (e.g. she knew the meal was poisoned, the house was on fire etc.) one would question her rationality. Determinism implies that given the *actual circumstances* it is inevitable that the agent will behave in one particular way. However, determinism does not imply that the agent must be irrationally inflexible, because there may still be a range of considerations that *would* induce the person to behave differently *if* those considerations were present.[274] This capacity to respond to relevant reasons lies at the heart of several influential compatibilist accounts of freedom.[275]

Certain types of direct intervention may undermine rational flexibility. For instance, a person might have been subjected to conditioning that was so intense that it instilled in her a literally irresistible desire – one which she would not resist under any circumstances. Or the intervention may impair/destroy her understanding so that she cannot grasp any sane reasons for action, or understand how reasons relate to each other (e.g. why one reason is more weighty than another).

However, it is possible to imagine certain intuitively troubling types of intervention that seem to leave rational flexibility intact. For example imagine that the intervener instils a particular desire in the offender. The intervener ensures that the desire is not irresistible. There is a range of incentives that would induce the offender to resist the

[273] See D Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting* (OUP, Oxford 1984).
[274] Strictly speaking, Fischer and Ravizza emphasise that, in order for the agent to be free, it must be true that *the 'mechanism'* underlying the agent's action would react to at least one reason for behaving differently, rather than that the *agent herself* would react to that reason (although, in most cases, if the mechanism would react then the agent herself would also react). The text in the next section contains further discussion of mechanisms. See J Fischer and M Ravizza, Responsibility and control: A theory of moral responsibility. (CUP, Cambridge 1998) [henceforth: 'Fischer and Ravizza 1998'].
[275] See, e.g. Fischer and Ravizza 1998; Vihvelin K, 'Free Will Demystified: A Dispositional Account' (2004) 32 Philosophical Topics 427.

desire. However, imagine that the intervener herself has selected *which* incentives (w, x, y and z) will induce the offender to resist the desire.[276] The intervener also ensures that the offender believes that the presence of w, x, y and z would be sufficient reasons against acting on the desire and that the offender has no insane beliefs or insane methods of 'reasoning'. The intervener has made sure that no other incentives (apart from w, x y and z) will induce the agent to resist the desire. The offender will definitely act on the desire when she is in situations where none of the incentives selected by the intervener are present.

Now it seems plausible that the offender still has rational flexibility, since she is not prey to irresistible impulses or obviously distorted, 'crazy' thinking. However, the intervener has interfered with her in a manner that many would find disturbing. There are three possible ways in which one might try to explain what is wrong with this kind of intervention: Firstly, one might insist that a sufficiently refined account of rational flexibility would show that such interventions actually do deprive the agent of rational flexibility. However, even some leading compatibilists have come to acknowledge that this approach is unlikely to be able to deal with all kinds of problematic interventions.[277] Secondly, one might argue that rational flexibility is not sufficient (although it is perhaps necessary) for free will and that such interventions are wrong because they interfere with some *other aspect* of free will, apart from rational flexibility. I will shortly consider and reject various other types of free-will-based objection to direct interventions. The third option is to argue that

[276] In this example there are only four possible considerations that would induce the agent to act differently. But the intervention would still be troubling even if the intervener had selected a larger number of considerations. It should be noted, however, that it would be unreasonable to demand that the agent must be responsive to a very wide range of considerations for acting differently. Many normal agents adhere to certain courses of action in a very rigid way, and would only depart from their course under fairly extreme conditions, and yet are considered free (if sometimes fanatical, or sometimes principled). They may even be blamed or praised for their rigidity. Fischer and Ravizza 1998, p70 stipulate that it is only necessary for there to be *one* consideration that would cause the agent's mechanism to react differently.

[277] See e.g. Fischer and Ravizza 1998 pp230-231: '...ahistorical....accounts cannot adequately treat such cases... What seems relevant is not only the fact that the mechanism issuing in the action is suitably reasons-responsive; what also matters is *how* that mechanism has been put in place."

these interventions are wrong for some reason that is not connected with free will. This is the approach that I will ultimately defend in chapter 2.

Compatibilism Part Two: Freedom as Authenticity

i) Authenticity and Psychological Coherence

The 'freedom as authenticity' approach defines free will in terms of whether the agent's actions express her 'real self'. Compatibilists differ over which psychological states are to be identified with the agent's 'real self'. Probably the most influential real self view was developed by Harry Frankfurt.[278] Frankfurt defined free will in terms of whether the agent's first order desires 'cohered' with the agent's second order desires. First order desires are desires to perform actions, e.g. the offender may want to lash out violently. Second order desires have first order desires as their objects, e.g. the offender might want to have his violent impulses; he endorses them. According to Frankfurt, in order to be free, an action must flow from a desire that the agent wants to have and which she wants to be executed in action. She must 'wholeheartedly identify' with the desire that results in her action. Other compatibilists, such as Gary Watson, focus instead on coherence between the agent's desires and values, rather than between different orders of desire.[279]

Direct interventions could undermine an offender's psychological coherence. An intervention might cause the offender to have strong desires or aversions which jar with his values or second-order desires. For example, the intervention might cause the offender to experience powerful feelings of disgust at the idea of re-offending. The offender may not endorse or identify with these feelings of disgust. This kind of intervention creates an internal conflict between fundamental constituents of the

[278] E.g. H Frankfurt, 'Alternate Possibilities and Moral Responsibility' (1969) 66 Journal of Philosophy 829. [Henceforth: 'Frankfurt 1969']
[279] G Watson, 'Free Agency' in G Watson (ed), *Free Will* (OUP, Oxford 1982).

person's agency – between his values and his desires/feelings. Alienation from his desires and feelings can threaten the person's identity, as it seems that an important part of his mental life is not truly his own.

It is important to remember, however, that direct interventions need not create psychological conflict within the offender. The offender may welcome the change in his motivations. In fact, a direct intervention might enhance an offender's psychological coherence, by bringing his feelings and desires more into line with his values. For instance, prior to intervention, the offender may have felt deeply ashamed of his violent impulses and may feel that interventions which reduce the strength of those impulses help him to become the sort of person he wants to be.

Furthermore, if we define 'free will' in terms of psychological coherence, then it seems that the following approach would preserve the offender's free will: employ direct interventions in order to modify *both* the offender's first-order desires *and* his second-order desires and values, in a way that ensures psychological harmony. Some philosophers, such as Harry Frankfurt, accept this conclusion.[280] Yet it would strike many people as counterintuitive to suggest that interfering to a *greater* extent in an individual's mental life and modifying aspects of the person that are *particularly central* to the individual's agency (i.e. their values) allows the individual *more* free will than interventions that only affect first-order desires/aversions. Some compatibilists have tried to avoid this counterintuitive conclusion by including a historical dimension in their theories.

ii) Historical Authenticity

According to historical compatibilists, whether a person's mental states are authentically hers at a given time depends on how she came to have those mental states. Her current mental states are only authentic, on this view, if they are

[280] Frankfurt 1969.

connected in an appropriate way to the agent's earlier mental states. Thus, even if a direct intervention left the agent with a set of desires, beliefs and values etc. that were coherent and not in conflict, historical compatibilists might still find the intervention objectionable if the individual's post-intervention mental states were not appropriately connected to her prior mental states. What counts as an appropriate connection? At least three different types of connection have been suggested.

1) Similarity with previous mental states

Historical compatibilists often focus on cases where a significant alteration to the brain brings about a very sudden, dramatic change in the agent's motivations. Many different scenarios have been discussed, including: a very good woman who, after being manipulated by an evil neuroscientist, acquires the values of a serial killer[281] and a saintly nurse who, after receiving a blow to the head becomes cruel and reckless towards her patients.[282] They cite these examples as central cases where the individual's free will has been eliminated. There are also documented real-life examples of sudden personality changes, e.g. acquired paedophilia[283] and acquired sociopathy.[284]

Now, historical compatibilists acknowledge that sometimes ordinary people, whom we normally regard as possessing 'free will', undergo fundamental changes in their character, values, and desires. However, when such fundamental changes occur, they typically emerge gradually over time. Even if a person's motivational set-up when the agent is twenty years old differs considerably from her motivational-set-up at fifty years old, this often is the result of a very gradual transformation where each incremental stage in the person's development resembles the previous stage in

[281] A Mele, Free Will and Luck (Oxford University Press, New York 2006).
[282] V Tadros, Criminal Responsibility (OUP, Oxford 2005).
[283] See J Burns and R Swerdlow, 'Right Orbitofrontal Tumor with Pedophilia Symptom and Constructional Apraxia Sign' (2003) 60 Archives of Neurology 437-440.
[284] A Damasio, *Descartes's Error: Emotion, Reason and the Human Brai*n (Putnam, New York 1994).

important respects, but where the final stage in the series is very different from the initial stage.

There are two problems with this version of historical compatibilism. Firstly, there are cases of individuals who undergo very fundamental changes in their values over quite a short period of time, and are still considered to be free. For instance, the individual may have a 'road to Damascus experience' – an inspired insight into important moral truths, which lead her to reject her previous values. This suggests that incremental change is not, in fact, a necessary condition for free will. Therefore, the fact that a direct intervention brings about a sudden change in the offender's values does not *in itself* render the offender unfree. Secondly it is possible to imagine a type of direct intervention that successfully alters the offender's values but which takes effect gradually over time. This version of historical compatibilism lacks the resources to explain why such an intervention is intuitively objectionable.

2) A connection in terms of deliberation

On this view, if an agent's values alter, the agent's new value is only authentic if the acquisition of this value was preceded by deliberation in the light of the person's prior value system.[285] However, road to Damascus cases provide a challenge for this view as well. Imagine that an agent, Denise, was a thoroughly selfish person with a corrupt value-system. One day a natural disaster strikes her town. She is unharmed but encounters numerous victims of the disaster. Denise experiences an unfamiliar experience of compassion accompanied by a sudden insight into the reasons for helping others. She acts on her new moral insight and performs some good deeds. However, she did not deliberate about her new insight in the 'light' of her old corrupt value-system. The new moral insight just displaced the old corrupt values. Is

[285] Haji I and Cuypers S, 'Magical Agents, Global Induction and the Internalism/Externalism Debate' (2007) 85 Australasian Journal of Philosophy 343.

Denise's insight therefore inauthentic and are her subsequent actions unfree? It does not look that way.

Imagine Denise's community decides to present her with a medal for her good deeds. At the awards ceremony, a psychologist stands up and says, 'As part of my research into why people perform heroic acts, I have looked very carefully into Denise's case. I discovered that when Denise acquired her new, emotionally-charged awareness of the need to alleviate human suffering, she did not evaluate this insight in the light of her earlier corrupt value-system. In fact, her corrupt evaluative scheme was completely idle! Hence her new good moral values are inauthentic and the actions that flowed from them were not an exercise of free will. Denise therefore does not deserve a medal.' This reaction would seem bizarre. The 'deliberation connection' does not seem to be a necessary condition for free will and the supposed absence of this connection *per* se cannot provide a convincing basis for objecting to direct neurological interventions.

It might be objected that the sudden change in Denise's attitudes was still a response to a reason, whereas changes that occur due to direct interventions by-pass the agent's rational faculties. In the above example the rush of empathy was caused by something that could provide an appropriate basis for it – i.e. the sight of human suffering. However, a person's pro-attitudes can change even though she is not presented with a *new reason* for changing. Consider a person who on one occasion comes to feel the force of a consideration of which he had long been aware, but which had never moved him before. He may suddenly feel that now he 'gets it'. A person might have this experience because of a non-rational factor, such as an alteration in hormones or neurotransmitters.[286] These factors might, for instance,

[286] N Arpaly, Unprincipled Virtue (OUP, Oxford 2003).

make the consideration more emotionally salient.[287] But if all such experiences were considered to result in lack of freedom and authenticity then many ordinary people would have to be described as unfree and inauthentic – which seems implausible. Therefore, a change in pro-attitudes cannot be considered inauthentic and freedom-undermining just because it was brought about partly by non-rational factors.[288] A direct intervention might conceivably operate in a similar way to the case of the ordinary person who suddenly 'gets it'. If the intervener knew that a person was on the verge of changing his mind, if only the relevant consideration was a bit more emotionally salient, the intervener might alter the person's brain accordingly. Intervening in another's mind in this manner remains intuitively troubling, yet for the reasons already given, authenticity approaches have difficulty accounting for this.

3) Mental states connected in virtue of sharing the 'same kind of mechanism'

In response to some of the problems with the rational flexibility approach discussed above, Fischer and Ravizza added a 'historical' dimension to their account of free

[287] Or indeed non-rational factors could influence behaviour by making certain factors less salient. It is conceivable that a reduction in testosterone might somewhat reduce an individual's attraction to anti-social behaviour (relative to their other desires) and this might on a particular occasion 'tip the scales' for an individual, causing them to decide to engage in a more law abiding activity instead. For instance, there is some evidence that testosterone levels can influence pro-attitudes connected with anti-social behaviour. It has been suggested that a natural decline in testosterone in men as they age may sometimes partly account for a reduced inclination to reoffend, or a reduction in certain kinds of sexual reoffending in particular. See, respectively: Barbaree H, Blanchard R, and Langton C, 'The Development of Sexual Aggression Through The Life Span. The Effect Of Age On Sexual Arousal And Recidivism Among Sex Offenders', (2003) 989 Annals of the New York Academy of Sciences 59-71; Quinsey V, Evolutionary Theory and Criminal Behaviour (2002) 7 (1) Legal and Criminological Psychology 1. If it turns out that fluctuating testosterone levels do indeed influence pro-attitudes, this surely does not *by itself* show that the individuals concerned lack free will. It is therefore implausible to suggest that influencing a person's behaviour partly through a non-rational means necessarily deprives them of free will.

[288] In the example I envisage, the change in pro-attitudes is only *partly* brought about by non-rational factors, because it is also partly explained by factors such as the presence of the consideration that the person now 'gets'. The non-rational factor (e.g. alteration in hormones) *by itself* is not sufficient to bring about the new pro-attitude in the absence of the consideration. However, the person was aware of (and unmoved by) the consideration before the alteration in hormones occurred, the agent would not have been moved by the consideration if not for the alteration in hormones, and the awareness of the consideration itself did not bring about the alteration in hormones.

will.[289] According to their theory, in order for an agent's actions to be genuinely her own, the agent must have previously 'taken responsibility' for the mechanisms from which her actions arise, by *viewing herself* as being responsible for actions that flow from these mechanisms. By 'mechanisms', they mean the features of her agency that play a causal role in her actions (including, but not limited to, mental states such as intentions, desires and beliefs). On Fischer and Ravizza's view, when an agent, at a particular time, comes to take responsibility for a certain type of mechanism, she *thereby* takes responsibility for her future behaviour that results from the *same* kind of mechanism. They claim that motivations resulting from direct neurological interventions (almost invariably) involve a *different kind of mechanism* from ordinary motivations. Therefore, they maintain, when an individual takes responsibility for her ordinary mechanisms she does *not* thereby typically take responsibility for motivations or actions that arise from neurological interventions.

Yet Fischer and Ravizza's account still attributes free will to individuals who have received intuitively troubling kinds of direct interventions. For instance, they appear to argue that merely telling a person that she has been manipulated can restore her free will. They write:

> '[Imagine that a] scientist induces (via his direct manipulation of Judith's brain).... a desire [to punch Jane] that is not literally irresistible but is nevertheless extremely strong [1998, p232].... Now it is possible that Judith becomes aware of the stimulation of her brain by the scientist. She thus can understand the provenance of her previously inexplicable desire to punch. She now has two choices. Presumably, she will modify her desire so that she returns to her normal state (antecedent to the implantation of the desire). Alternatively, it is conceivable that she will decide to keep the new desire, upon reflection. (Perhaps she will have decided that she likes it.) *Either way, awareness and reflection returns Judith to the*

[289] Fischer and Ravizza 1998.

> *mechanism of ordinary practical reasoning, and she can subsequently be held morally responsible*. Awareness and reflection of a certain sort returns Judith to a situation in which she is acting from *her own* mechanism.'[290]

The above passage implies that a person is still responsible, even if she does not endorse the effects of the treatment. Apparently, as soon as she is aware of the manipulation she is free again, free to reject or accept the desire, and she is responsible, regardless of which choice she makes. Fischer and Ravizza assert that 'awareness' of the treatment will (somehow) enable the person to 'return to her normal state', provided there are no literally irresistible impulses.

Fischer and Ravizza's ownership requirement (as they have defined it) cannot be relied upon to object to the use of intuitively troubling kinds of direct intervention. There are legitimate worries about the state using brain interventions to instil 'extremely strong' desires in offenders (particularly desires of the kind mentioned by Fischer and Ravizza - desires so strong that they will only be resisted if the offender knows that large numbers of innocent people will die if she acts on the desire). Such interventions would not automatically violate Fischer and Ravizza's ownership requirement, provided the agent is made aware of the intervention from the start. [291]

Another, more fundamental problem arises when Fischer and Ravizza try to explain what makes a mechanism belong to one 'kind' rather than another. They do not simply maintain that actions which flow from psychological states like 'desires', 'beliefs' and 'intentions' arise from one type of mechanism and actions that have nothing to do with such psychological states (such as epileptic seizures) belong in a different category. If they settled for this simple account then it would not help them

[290] Fischer and Ravizza 1998, p235, emphasis added.
[291] Fischer and Ravizza 1998, pp235-236 do acknowledge that a person is non-responsible if she has received further manipulation designed to make her endorse the desires that have been implanted in her. However, I think that implanting extremely strong desires in offenders can be objectionable *per se* whether or not the offender has been programmed to endorse them.

to differentiate reliably between cases of 'ordinary' mechanisms and mechanisms produced by intuitively objectionable types of direct interventions. For it is possible to imagine mental states such as desires and beliefs being induced by direct interventions. Fischer and Ravizza rely heavily on intuition to differentiate between different kinds of mechanism.[292] They maintain that, *intuitively*, motivations resulting from direct stimulation of the brain belong (in most cases) to a different kind of mechanism from motivations that are determined in the 'ordinary' way by one's genes and environment. This approach is open to challenge. For it seems that the notion of 'different mechanisms' is no longer doing the work it was supposed to do. This notion was meant to help *explain* why we intuitively feel that certain types of direct intervention are problematic. But instead it seems that our intuitions that certain types of direct interventions are problematic dictate whether one mechanism counts as belonging to a 'different kind' of mechanism from another. In order for the notion of 'different mechanisms' to have explanatory power, Fischer and Ravizza need to have a principled basis for individuating mechanisms, which is derived from *'independent reflection on the nature of these mechanisms'*.[293] Otherwise, it seems that they are merely stipulating that certain mechanisms are different from others in an *ad hoc* way in order to generate the conclusions they want about direct interventions. Unfortunately, it is far from obvious that truly independent criteria for individuating mechanisms (e.g. derived from psychology, or neurology) will produce the results that Fischer and Ravizza desire.

So far I have criticised various attempts to distinguish agents whom we normally regard as free from agents who have been subject to intuitively troubling brain manipulation. I will now critique a final compatibilist strategy that has gained recent

[292] Fischer and Ravizza 1998, p40.
[293] D Pereboom, 'Reasons-Responsiveness, Alternative Possibililities, and Manipulation Arguments Against Compatibilism: Reflections on John Martin Fischer's My Way' (2006) 47 Philosophical Books 198, p200. See also, M McKenna, 'Book Review: Responsibility and Control: A Theory of Moral Responsibility, by John Martin Fischer and Mark Ravizza' (2001) 98 Journal of Philosophy 93.

attention, before defending an alternative type of objection to direct interventions that is firmer than free-will based objections.

Compatibilism Part 3: the Nature/Person Distinction

One simple response to the problem of direct interventions is to argue that it is just a basic moral fact that 'normal' influences on a person's psychology (e.g. stemming from one's genes and standard environmental influences) do not undermine free will, whereas free will is undermined by direct brain interventions by other human beings who are trying to shape the person's psychology to suit their own ends. On this view, the *source* of the influence on the individual's psychology is the crucial factor, even if the actual impact that these influences have are identical. Free will theorists have generally tried to avoid adopting this approach, because of its seeming arbitrariness. However, versions of this response have recently been defended against this charge. I will argue that although these attempts ultimately fail, nevertheless, they point towards something that is important.

a) The Responsibility-Shifting Defence

One rationale for the nature/person distinction[294] is based on the idea that when one person intervenes in the mind of another using brain interventions the former takes over responsibility for the results. However, if a person's mind is shaped by 'natural' causes then there is no particular individual who can relieve her of responsibility. Therefore, on this view, responsibility for her actions must rest entirely on her own shoulders.[295] This position is defended by Jan Christoph Bublitz and Reinhard Merkel:

[294] The term is taken from Bublitz and Merkel (2009)

[295] Daniel Dennett calls a similar line of argument "The Principle of Default Responsibility" which states that "*If no other agent is responsible for your condition and the acts that flow from it, you are.* The buck stops there, if you are competent." (Dennett 2011, p11, emphasis in original).

> 'Some claim that pro-attitudes transformed by direct brain interventions such as neuroenhancements derive from mechanisms that are not the agent's own; hence, the resulting actions are nonautonomous. This is plausible only insofar as agents are manipulated by other agents who then bear primary responsibility, thus exempting the manipulated agent.'[296]

It is important to note that, according to Bublitz and Merkel, what the manipulator actually does to the manipulated person is not *intrinsically* freedom-undermining. They launch a powerful attack on the notion of 'authenticity', arguing that a person is not rendered unfree even if she undergoes a radical transformation due to a process that bypasses her rationality. They state that

> 'Having self-arranged for all of these bypassing transformations is too demanding a condition [for free will]. If we take that criterion seriously, then the majority of our pro-attitudes would have to be declared inauthentic and all the resulting actions nonautonomous. There is no self-creation *ex nihilo*. From one's sex and other bodily constitutions through to moods, core character traits, behavioural dispositions, social environments and natural endowments, there exist myriad influences on the formation of pro-attitudes that bypass rational control, depend on natural contingencies and are not self-arranged.'[297]

Bublitz and Merkel even go so far as to say that a person who takes prescribed medication that (as an entirely unforeseen side effect) drastically alters her character and values remains free and responsible provided that she satisfies Harry Frankfurt's requirement – that she identifies with her new desires. Given these views, Bublitz and Merkel's approach to manipulation seems rather odd. It seems to involve the claim that manipulators function as 'blame magnets'.[298] If a person is 'lucky' enough to have been influenced by a blame magnet then she can violate norms with impunity - the responsibility for all these acts sticks to the blame magnet. However, if a person

[296] J Bublitz, and R Merkel, 'Autonomy and Authenticity of Enhanced Personality Traits' (2009) 23(6) Bioethics 360-74, p373.
[297] Ibid, p371.
[298] My term.

is unlucky enough to have been influenced by forces not manipulated by a blame magnet (forces that may have been equally powerful and equally outwith her control), then she must take all the blame. It is also puzzling why the fact that the manipulator bears some degree of responsibility should thereby *relieve* the manipulated person of responsibility. Why cannot they *both* be held fully responsible?

A further problem arises if there is any question over whether the manipulator herself possessed the capacities that are required for responsibility. Perhaps the manipulator was mentally ill. In this case it does not seem as if she can function as a blame magnet, because she is an inappropriate candidate for blame. We therefore get the paradoxical conclusion that the question of whether X (the manipulated person) is responsible for her actions depends on whether Y (the manipulator) was sane. One would have to assess the capacities of Y in order to establish the responsibility status of X![299]

b) Do existing practices support the nature/person distinction?

According to Bublitz and Merkel "from a normative perspective, there is a widely agreed difference" between nature and persons.[300] Bublitz and Merkel are of course correct in this assertion. In fact, there are many widely agreed differences between nature and persons. For instance, you cannot reason with a storm about whether it is a good idea to damage people's property and you cannot deter a volcano from erupting by threatening to punish it unless it behaves itself. But Bublitz and Merkel's

[299] It might be objected that the victim of manipulation is non-responsible regardless of whether the manipulator was sane or insane. However, the objector must then explain why an insane manipulator is relevantly different from 'ordinary' deterministic forces. The objector cannot use Bublitz and Merkel's argument that the manipulator's own culpability deflects blame away from the manipulated agent.

[300] Supra, p372.

case will only be persuasive if they can identify a difference between nature and persons that is *relevant* to their argument. Here are the examples they rely on:

> Case 1
> "Doctor D asks patient P for consent to remove his cancerous kidney – otherwise, it is certain that P will shortly die, say within the next month. P consents and the kidney is removed."[301] Here P's consent is valid.
> Case 2
> "Being held at gunpoint, P is 'asked' to consent to the removal of his kidney in order to transplant it to the coercer's son; otherwise P will die with a bullet through his head. P consents and the kidney is removed."[302] Here P's consent is invalid.

These examples would be relevant to Bublitz and Merkel's argument if the examples could bear the following interpretation: In case 1, the patient is still exercising his free will when he agrees to receive medical treatment, despite the fact that his consent arises from an intense pressure, whereas the patient in case 2 is not exercising his free will, despite the fact that the pressure to which he is subjected is of no greater intensity. On this interpretation, what makes the difference between the patient being free in case 1 and lacking freedom in case 2 is the source of the pressure that they are under – nature in the first case, a person in the second. This appears to be the conclusion that Bublitz and Merkel want us to draw from their examples. In support of their argument, they invoke Joel Feinberg's view that pressures from nature are usually just "background conditions" and do not deprive the person of free will or voluntariness, whereas a human threat amounts to an "intervening force, rendering his decision involuntary".[303]

However, this is not the correct way to analyse these examples. It is a mistake to infer that P's consent in case 2 is necessarily invalid *because it was involuntary*. Lack of voluntariness is one ground on which consent may be invalidated, but it is

[301] Ibid.
[302] Ibid.
[303] J Feinberg, *The Moral Limits of the Criminal Law: Harm to Others* (OUP 1989).

not the only ground. To see what other grounds there may be it is necessary to consider what role a valid consent serves in cases such as the kidney transplant examples. In such cases, a valid consent provides the doctor with a defence to assault. If the patient's consent to treatment is valid, the doctor has not wronged the patient by treating him. Why does consent serve this function? By acting with the patient's consent, the doctor shows respect for the patient's preferences. The doctor treats the patient as a being worthy of moral consideration and does not treat him 'merely as a means'.[304]

Sometimes a person *completely lacks* the capacities necessary for free will/voluntariness: for instance he might be insane, so that any apparent 'consent' he gives is not valid. Typically the best way to respect such a person's status as a being with moral worth is to act in the person's 'best interests'. On other occasions, voluntariness is *severely restricted*, although not completely lacking. In such cases, giving weight to the degree of voluntariness that still exists can show respect for the person as an end in himself. Consent can then be used as a defence. However, sometimes the person who wishes to rely on the consent is not genuinely showing respect for the degree of voluntariness that exists, but, rather, has associated herself with the restriction on voluntariness.

In case 1 the patient's freedom is severely restricted by the threat of death from the kidney disease, but his consent is not entirely involuntary (assuming that he is not so overwhelmed by the fear of death that he is incapable of rational thought.) The doctor, however, has not associated herself with the restriction on the patient's voluntariness – the doctor is not responsible for it, nor is she exploiting it. She is showing respect for the patient as a being worthy of moral consideration, both by paying attention to his preferences and by trying to save his life. In case 2, the patient

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

has *the same degree of voluntariness* as in case 1 (again assuming that he is not so overwhelmed by the fear of death that he is incapable of rational thought.) However, the gunman *stands in a different relation to the patient's voluntariness* from the doctor. The gunman has associated himself with the *restriction* on voluntariness – he was responsible for it. The gunman cannot rely on the degree of voluntariness that the patient still has as a defence, because the gunman is not showing respect for the patient as an end in himself.

On my interpretation of the above cases, the source of the restriction on X's voluntariness is relevant to whether Y can *rely* on the degree of voluntariness that X still has as a defence. If Y is the source of the restriction, then Y is barred from defending herself by saying "look, X's decision was restricted but not entirely involuntary, and I respected what little freedom X had left", because Y has already shown immense *disrespect* for X by being the cause of the illegitimate restriction on X's freedom. The approach defended in this chapter explains the difference between case 1 and case 2 in a way that makes sense and which appeals to an already widely accepted principle – the idea that people should be respected as ends in themselves.[305]

In contrast, the approach advocated by Bublitz and Merkel (and Feinberg) relies on a counter-intuitive stipulation - the idea that the presence or absence of a person's free will depends on whether the person is being influenced by a human pressure or a

[305] I do not claim that this is the only basis for deciding whether or not someone can rely on another person's apparent consent. The rules governing consent vary depending on the context, e.g. whether consent is being used as a criminal defence, or in the context of the validity of different types of contract. Public policy considerations may often be influential, e.g. the idea that it is important not to undermine stable contractual relationships by too frequent challenges concerning the freedom of the contracting parties. It may be argued that certain constraints on freedom are so widespread that it would be impossible for any contract to be relied on if these constraints were allowed to render the contract invalid. Yet these pragmatic considerations cannot shed light on whether people subjected to such constraints really are free or unfree. It is impossible within the scope of this thesis to give a full discussion of all the different considerations that may have a bearing on the validity of consent in every context. However, the proponents of the nature/person distinction have not produced an example which clearly supports their position.

natural one. By what mysterious alchemy does a human pressure render a decision unfree which would otherwise be free if only the pressure came from a natural source? Normally we consider free will, or voluntariness, to depend on capacities and opportunities. But natural pressures can affect a person's capacities and opportunities just as much as human pressures and natural pressures can be just as much beyond the individual's control as pressures from other human beings. There are many contexts in which it is obvious that the nature/person distinction makes no difference to a person's freedom. If a person is paralysed, her freedom is just as constrained whether the paralysis arose from an attack by another person or from a disease. If a person commits a crime and argues that she should be excused because she was threatened with death or serious injury, the success of her defence does not depend on whether the threat arose from nature (necessity) or from a human being (coercion).

For the reasons stated above, examples involving the issue of valid consent do not appear to provide support for the claimed significance of the nature/person distinction. This distinction seems a shaky basis for objecting to directly intervening in the minds of others. An intervention is no more likely to turn someone into a puppet merely because it is caused by a human being rather than nature. However, the consent examples do point to another basis for objecting to direct interventions. They suggest that the reason why a person may not rely on another's consent sometimes depends on the nature of the relationship between the people concerned rather than on the fact that the 'consenting' person entirely lacked free will. The basis of my objections to certain types of direct interventions (discussed in the next chapter) is also concerned with the *relationship* between the intervener and the subject of the intervention.

Chapter Eight: Objectionable Types of Brain Intervention

Introduction

In this chapter, I argue that certain forms of biomedical intervention are objectionable and should not be used within the criminal justice system. I specifically oppose the idea of using biomedical interventions to deal with offenders who are basically rational in an attempt to ensure that the offender adopts one particular set of values rather than another. Unlike many traditional approaches to the issue, my objection to such interventions does not rely on the idea that they necessarily violate the offender's free will. Instead, the objection is based on the problematic nature of the *relationship* between the intervener and the subject of the intervention. I argue that the state's way of dealing with offenders must be constrained by the principle that the offender must not be objectified, and that his personhood must be respected.

A Thought Experiment

Consider the following thought experiment: One day an angel appears on earth. The angel possesses a magic flute. Anyone who hears the flute will suddenly have a powerful insight into fundamental moral truths. This vivid recognition of the reasons for behaving morally will motivate the agent to act in accordance with these reasons. Flute in hand, the angel marches off to the nearest prison. The authorities get to hear about this before the angel reaches the prison. What should they do? It seems that the free-will-based objections to direct interventions would apply equally to the magic flute scenario –if the recognition of moral reasons and the subsequent commitment to act accordingly, *guarantees* that the offender will act virtuously (in the actual world) then this violates incompatibilist freedom; if the offender's new values are

disconnected from her prior values (in any of the senses of 'disconnection' mentioned above) then this violates a version of 'freedom as authenticity'; given that a causal factor behind the change of values (listening to the flute music) does not provide the agent with any new reason for changing her behaviour, then this arguably goes against a rationality-based conception of free will. If these approaches to free will are correct, then it seems that the authorities have great cause for concern - the free will of a large number of offenders is in jeopardy. Yet it seems counter-intuitive to suggest that the authorities would have a pressing obligation to rush to prevent the offenders from being affected by the music's reformative powers, or that it would be such a terrible thing if the authorities failed to take action in time to prevent the prisoners from being reformed.

The 'magic flute' thought experiment is intended to cast doubt on the claim that changing an offender's values using direct interventions, rather than moral dialogue, necessarily violates the offender's free will in an objectionable way. This thought experiment features a means of altering values that does not involve moral dialogue and yet does not seem to violate the offenders' free will, or even if it does so, it does not seem seriously morally objectionable. However, this thought experiment does not show that it is all right *for us* to use interventions other than moral dialogue. It is submitted that ordinary human beings do not have the *moral status* to directly re-shape a person's values or goals using means other than rational persuasion. The objection to direct interventions presented in this thesis does not rely on the idea that these interventions violate the offender's 'free will', conceived of as a capacity that we can identify just by examining the individual's psychology and actions carefully enough. Rather, it is submitted that an objection to such interventions can be based on the problematic nature of the *relationship* between the intervener and the subject of the intervention. It is possible to identify the objectionable features of this relationship by highlighting the ways in which it departs from a model of an

appropriate type of relationship between the state and offenders. I will not attempt to fully describe and defend such a model within the scope of this thesis. Rather, I will present certain principles concerning how the state ought to relate to offenders, which have some intuitive plausibility. If my account is accepted, it provides a basis for objecting to certain kinds of direct neurological intervention, which does not rely on the notion that these interventions violate the offender's free will.

Objectification, Personhood and Dialogue

In order to see why some direct interventions are objectionable it is useful to return to the idea that a person who has been subjected to direct brain interventions is transformed into a mere 'puppet', an 'automaton' or a 'robot'. I have argued that to the extent that this charge is meant to convey the idea that the person now has as little free will as a puppet then this is inaccurate. However, it is sometimes true that subjecting a person to direct brain interventions would amount to *treating her as if* she were a puppet, an automaton or a robot – as something less than human. In other words it would 'objectify' her.

The term 'objectification' can be used in different ways. The conception of objectification that this thesis adopts is influenced by discussions of the ways in which disfavoured groups within society have historically been objectified.[306] This kind of objectification typically involves creating a division between 'them' and 'us' which excludes the objectified group. It also typically involves portraying the disfavoured group as radically deficient in some fundamental respect. This idea of objectification can be usefully contrasted with the idea of respecting personhood. Personhood can be respected by preserving connections between the group in question and other members of society and by highlighting commonalities between members of the group and other citizens. There is a danger that society's (often)

[306] S Reicher , 'Saving Bulgaria's Jews: An Analysis of Social Identity and the Mobilisation of Social Solidarity' (2006) 36 European Journal of Social Psychology 49.

justified horror at and condemnation of criminal acts will lead to objectification of offenders. It is therefore particularly important to have clear restrictions on the ways in which the state may treat offenders, in order that society does not lose sight of their personhood.

Respect for an offender's personhood can be shown through engaging rationally with the offender as he is, and by challenging his mistaken views with arguments, without using direct neurological interventions to fundamentally re-shape his values. There are several ways in which rational dialogue affirms commonalities between offenders and other moral agents.

Dialogue and Equality

Engaging in dialogue with the offender includes him within the moral community by allowing the offender to voice his criticisms of the community's norms, which can potentially contribute to a shift in those norms. Dialogue leaves open the possibility that *either* party may change the other. As Lawrence Stern writes:

"[Dialogue] involves the recognition of a certain equality between oneself and the other. There is, in general, no point in reasoning unless the other person is capable of seeing reason, getting the point. If he can do that, he can also correct *me* if I am mistaken."[307]

In contrast, attempting to re-shape the offender's values using direct neurological interventions is a one-way street. It seeks only to change the offender, to ensure that he will think and act in a particular way.

The most appropriate way for members of a moral community to attempt to change one another's values is through dialogue. Engaging offenders in dialogue, rather than re-shaping their values through direct interventions, assumes that there is a

[307] L Stern , 'Freedom, blame, and moral community' (1974) 71 The Journal of Philosophy 72, p75.

commonality between the offender and other moral agents. It implicitly acknowledges that the authorities (and majority opinion) are fallible, as is the offender. It allows that the offender (as he is, without neurological modification) may have useful insights, as other agents do. It also allows for the fact that the pursuit of moral understanding is a shared process. People need to interact with other people and to consider different points of view before they can form reliable judgements about how they should act.

The above considerations do not apply to the case of the angel in the thought experiment. The angel, as the embodiment of rationality and virtue, never stands in need of 'correction'. In contrast, the authorities do not have the moral status to portray themselves as the embodiment of rationality and virtue. Re-shaping offenders' values through direct neurological interventions replaces the acknowledgement that the authorities (like the offender) are human and fallible with the inappropriate assumption that the authorities are absolutely certain about what the 'right' values are. Furthermore, the angel is not a fellow member of the offender's human community, so the lack of dialogue between the angel and the offender does not convey the message that the offender is excluded from the community. However, if other human beings were to re-shape offenders' values through direct neurological interventions, rather than engaging them in dialogue, this would be an act of excluding offenders from the moral community.

Focussing on the principles that should govern the *relationship* between the state and the offender, helps to explain why the use of direct interventions by the state is more intuitively troubling than the intervention employed by the angel in the thought experiment. The relationship-based approach also produces other intuitively-appealing results. Unlike some of the free-will-based approaches discussed above, the relationship-based approach implies that more extensive modifications of the

offender's motivations are worse than less extensive interventions. For instance, interventions that just enhance the offender's control over his violent impulses, or reduce the strength of those impulses do not alter the offender's values, and so leave open the possibility that he will criticise the authorities on the basis of those values.[308] Modifying the offender's values precludes this possibility. Modifying the offender's values also sends out the strong message that the authorities view themselves as having hugely privileged access to knowledge of what the 'right' values are.

It might be objected that I have taken an unrealistic view of the potential for offenders to make a valuable contribution through moral dialogue to society's understanding of moral norms. Surely the authorities can be very confident that some offenders are completely in the wrong and that some of society's norms are very well-founded. In response, it is important to remember that, historically, a number of values which society has now come to reject once seemed self-evidently sound and that individuals who were very widely condemned by the rest of society have ultimately been vindicated.

Furthermore, instituting a policy of trying to instil acceptable moral values in offenders through direct neurological modification would create a disturbing relationship between the state and offenders, even if the policy were restricted to offenders who were genuinely in the wrong, and even if it succeeded in instilling values that were genuinely well-founded. Such a policy would mark a huge shift towards characterising these offenders as 'the other' and thus towards objectifying them. It would express the attitude that they are a group of people to whom we need not listen, (or at least that we need not listen to them until we have modified their

[308] Although dialogue has an advantage over even this technique, in that dialogue, unlike direct interventions, positively reaffirms the offender's status as a moral agent and includes him within the moral community.

brains such that they are likely to tell us what we want to hear). If all attempts to change offender's values involve entering into a relationship with the offender, rather than relying on direct neurological interventions, then society is less likely to lose sight of the personhood of the offender. In addition, even if society's condemnation of a particular offender is justified and the offender is completely in the wrong, dialogue with the offender can still make a useful contribution to other agents' moral understanding. For the attempt, through rational dialogue, to reform a wrongdoer who is very unwilling to be persuaded can cause the would-be reformer to try to make his arguments as compelling as possible, which can lead to a clearer understanding of the justification for society's norms.

It might also be objected that this chapter adopts an excessively rosy view of the available alternatives to direct interventions. No society responds to criminal behaviour by relying on dialogue alone. A prison sentence, for instance, 'is more than an appeal to sweet reason and morality'.[309] Furthermore, it might be argued, punishing criminals necessarily involves highlighting the differences (rather than commonalities) between offenders and law-abiding citizens, by condemning the offender as a wrongdoer. Punishment also excludes offenders from the community. It can do this in terms of the moral stigma that attaches to a criminal conviction and sentence. It can also physically exclude the offender from the community, e.g. by putting him in prison.

It should be acknowledged that society's response to criminal behaviour does involve coercion, exclusion and the highlighting of differences between offenders and law-abiding citizens. It is perhaps impossible to conceive of a practicable approach to the problem of crime which does not involve these elements to some degree. But it is submitted that society's response to criminal behaviour can and

[309] Ibid, p82.

should *also* involve dialogue (and not just coercion); that it should emphasise the commonalities between offender and other moral agents (and not just the differences); and that it should preserve some connections between the offender and the rest of the community (and not exclude the offender entirely).

It is important to emphasise that this thesis is not a defence of our current system of punishment and rehabilitation. Some of our current approaches to dealing with criminal behaviour are objectionable and fail to treat the offender as a member of the moral community. In order for our practices to be justifiable they would have to include much more sustained attempts to engage with offenders, to present them with moral reasons for changing their behaviour and to re-integrate them into the community.[310]

Nevertheless, measures that interfere with offenders' liberty, such as restrictions placed on their freedom of movement, can be *compatible* with continuing to view offenders as members of the moral community, provided that, among other things, these measures still permit the offender to challenge the authorities on the basis of his pre-existing value-system. In contrast, as argued above, the technique of re-shaping the neurological basis for offenders' values would take a significant step towards characterising the offender as 'the other'. It would vastly increase the (already considerable) powers for controlling offenders' behaviour which the authorities have at their disposal. This would set the authorities on a completely different plane from offenders, greatly increasing the inequality of power between them.

[310] For some criticism of the current system and for one account of ways in which it should be reformed which emphasises the importance of moral dialogue with offenders see R Duff, Punishment, Communication and Community (OUP, Oxford University Press 2001).

Dialogue and Offenders' Better Natures

There is a further way in which attempting to change offenders' values through dialogue rather than through direct brain interventions, emphasises the commonalities between the offender and the rest of the community. Dialogue aimed at persuading offenders to reform typically involves appealing to the offender's 'better nature'. This presupposes that, in common with most law-abiding citizens, the offender has certain positive qualities and that, although he committed a serious wrong, he is not completely corrupt.[311] In contrast, altering values via direct brain Interventions imply that offenders are different from law-abiding individuals in a very fundamental way. It implies that offenders are so inferior to the rest of the community in terms of their moral characters that these offenders will not respond appropriately to the most compelling moral reasons for changing their behaviour (unless the offenders receive radical neurological modifications). Most moral agents assume that, even though they may have certain vices and may sometimes behave wrongly, they would respond to really compelling reasons for improving their behaviour, provided that they were given sufficient time to reflect on the matter, that the reasons were put to them persuasively enough and the issue at stake was really important. They further assume that responding in this way is possible for them because they are not thoroughly bad; that they respond to these compelling moral reasons because they already have certain good qualities, which are brought out by sufficiently persuasive arguments. Viewing oneself in this way is particularly valuable, because it provides an important basis for self-respect. The preparedness to re-shape offenders' values through direct neurological interventions suggests that offenders lack the qualities that provide this basis for self-respect.

It should be noted that possessing these positive moral qualities is not the same thing as 'having free will'. It is conceivable that an individual might improve his

[311] A similar point is made in Duff, Trials and Punishments (CUP, Cambridge 1986).p. 266.

behaviour of his own free will, even if hitherto he had been thoroughly corrupt. It is not essential to the common sense notion of free will that a person's moral improvement was partly caused by the fact that the individual already had certain good moral qualities. But, as a matter of fact, most instances of moral improvement probably do build on pre-existing good qualities and it is part of a positive self-conception to view one's moral development in this way. Extensively re-shaping an offender's values through direct neurological interventions strongly suggests that the authorities consider the offender's existing character to be so comprehensively morally inadequate that positive moral change is unlikely to emerge from it. This carries the message that offenders are fundamentally not like 'us'. This message is much more extreme than the alternative message (conveyed by moral dialogue) that the offender behaved wrongly on a particular occasion, or that he demonstrated a particular vice.

A critic of my view might raise the following objection. My argument stresses that membership of the moral community is valuable. It also accepts that, in some cases it seems fairly likely that the offender's capacities for practical reasoning are such that they will never lead the offender to be reformed and to be genuinely restored to the moral community. Yet, if this is the case, then it would surely benefit such an offender if it were possible to use direct neurological interventions to re-shape his psychology such that he is much more likely to fully grasp and take to heart the moral reasons for reforming. Would not increasing the probability that the offender will *actually* be restored to the moral community in this way be better for the offender than maintaining the *fiction* that it is possible that he will reform, when in fact it seems that he never will?

In response, while it may be true that the individual offender might benefit from this kind of intervention, my objection to re-shaping offenders' values through direct

interventions is not based primarily on the idea that this violates the individual's rights or interests in every case. Rather, this chapter argues that the use of certain types of direct intervention would create a troubling relationship between different groups within society. A policy of employing direct interventions to re-shape offenders' values would be based on the assumption that these offenders' existing capacities for moral agency are so fundamentally inferior to the capacities of the rest of the moral community that these offenders will not respond appropriately to the most compelling reasons for changing their behaviour. Basing social practices and institutions on the assumption that a particular group of individuals are radically incomplete as moral agents goes against the ideal that the moral community should be as inclusive as possible and that it should emphasise its members' common humanity. Incorporating into our social structures the message that a particular group is so different from the rest of us that they require radical neurological modification to enable them to be part of the moral community is *prima facie* objectionable even if such a system would end up (in a sense) benefitting certain offenders.

For the reasons stated above, it is also submitted that altering an offender's values using direct neurological interventions would be unacceptable even if the offender requested such treatment. The offender's consent could not legitimise this practice because the practice affects society's stance towards offenders as a group. The very act of offering this type of intervention to offenders would send out the message that all offenders who are offered the intervention stand in need of it, whether or not they ultimately agree to it. This practice has the potential to be socially divisive and its effects are not limited to those offenders who give their consent. Therefore the offender's consent is not sufficient to make it morally acceptable.

Conclusion

These considerations suggest that the state should not use neurotechnologies to try to ensure that the offender adopts the state's favoured values. Direct interventions should not be employed in an attempt to create citizens who are models of 'responsibility' in the virtue sense.[312] Efforts to reform the offender should be through rational dialogue. However, I am not suggesting that we should oppose the use of all direct interventions within the criminal justice system. Neurotechnologies could potentially play a role in enabling certain offenders to engage in moral dialogue and could help the offender to become reintegrated back into the community. However, clear limitations must be imposed on the use of such technologies. First, they should only be used in order to *increase* the offender's *capacity responsibility* by restoring/enhancing his ability to engage in moral dialogue and practical reasoning, and should never aim to restrict his powers of critical reflection, or to directly re-shape his values. Secondly, brain interventions should never *replace* attempts to engage the offender in human relationships. Thirdly, direct interventions are only permissible with the offender's consent.

This proposal raises the following questions: 1) to what extent is it possible to distinguish between interventions that enhance an offender's capacities and those that re-shape his values? 2) Is it possible to obtain valid consent to direct interventions within the coercive context of the criminal justice system?[313] These issues will be addressed in Chapter Nine.

---

[312] (For more on the importance of this distinction, see N Vincent, 'Capacitarianism, responsibility and restored mental capacities' in B van den Berg and L Klaming (eds), Technologies on the Stand: Legal and Ethical Questions in Neuroscience and Robotics (Wolf Legal Publishers, Nijmegen 2011) pp41-65.

[313] My position on this second issue is informed by L Bomann-Larsen, 'Voluntary Rehabilitation? On Neurotechnological Behavioural Treatment, Valid Consent and (In)appropriate Offers' (2011) Neuroethics (Online First) doi:10.1007/s1215201191059.

Chapter Nine: A Role for Direct Brain Interventions?

Examples of potentially useful enhancements

Increasing Empathy

As noted above there is some evidence to suggest that the ability to empathize is key to understanding moral norms.[314] For example, individuals with markedly reduced levels of empathy have exhibited difficulties in distinguishing conventional rules (such as rules of etiquette) from moral rules and in ranking wrongs in order of seriousness. Philosophers differ as to whether empathy is essential for moral understanding. However, even those who believe that it is not essential, often maintain that it is indirectly helpful in moral development. If techniques were produced which increased empathy in individuals who appear to be deficient in it, then this might play a useful role in reforming offenders.

Decreasing Violent Urges

Certain offenders may experience repetitive violent fantasies and powerful surges of anger which they find difficult to control. As discussed in greater detail below, these factors can impair offenders' ability to think clearly about how they should act and may distort their moral judgments. Research is beginning to uncover certain neurological factors that seem to have an impact on individuals' dispositions to anger and violence. There is some evidence that selective serotonin reuptake inhibitors

[314] J Blair et al, The Psychopath: Emotion and the Brain (Blackwell, Oxford 2005). Studies concerning empathy and moral understanding have often been carried out on individuals with psychopathic personality disorder. There is some disagreement concerning whether psychopathy is a mental illness and hence whether interventions to increase psychopaths' emotional deficits would count as a treatment or an enhancement. For opposing views on the mental illness question see: L Mealey, 'The Sociobiology of Sociopathy: An Integrated Evolutionary Model', in S Baron-Cohen (ed), The Maladapted Mind: Classical Readings in Evolutionary Psychopathology (1997 Psychology Press, East Sussex) 133. R Kendell, 'The Distinction Between Personality Disorder and Mental Illness' (2002) 180 British Journal of Psychiatry 110.

(SSRI's) may reduce aggression[315]. It may become possible in the medium term future to develop techniques which can reduce the strength of offenders' volatile impulses or which increase their control over these impulses. John Harris argues that such developments may not be morally desirable (Harris 2011). He cites an example of an individual who attacked a terrorist who was about to detonate a bomb, thereby rescuing a plane full of people. According to Harris, if the rescuer had been given SSRI's to reduce his aggression he might not have managed to save the plane. Harris does highlight a genuine concern – moral understanding and morally-motivated behavior are complex phenomena. Even the well-intentioned use of biomedical interventions risks causing undesirable consequences. However, when assessing whether offenders should receive such interventions, it is important to take into account the *likelihood* of the relevant scenarios occurring. In the case of many violent offenders, the risk that the intervention will prevent them from heroically rescuing a crowd of innocent people may seem relatively small compared with the risk that without the intervention they will reoffend.

Anti-libidinal medication

Drugs have already been developed to help reduce deviant sexual urges and thoughts. This can create an opportunity for offenders to concentrate on the reasons why they should change their behavior and the steps they need to take, without being distracted by their impulses. These medications are already being used to some extent within the criminal justice system.[316]

---

[315] P Ferari et al, 'Escalated Aggressive Behavior: Dopamine, Serotonin and GABA', (2005) 526 European Journal of Pharmacology 51; T Douglas, 'Moral Enhancement' (2008) 25 (3) Journal of Applied Philosophy 228; M Crockett et al, 'Serotonin Selectively Influences Moral Judgment and Behavior Through Effects on Harm Aversion' (2010) 107 (40) Psychological and Cognitive Sciences 17433.

[316] Regarding the use of these medications in Scotland see: http://www.forensicnetwork.scot.nhs.uk/Medication%20for%20Sex%20Offenders/medication%20for%20sex%20offenders%20protocol.pdf Regarding the use of these medications in England see: http://www.insidetime.org/resources/Publications/Use-of-Med-to-treat-SexOff_PSJ176.pdf

Decreasing Racist Sentiments

Certain individuals experience a strong negative emotional reaction to members of different races. Such emotional reactions may stem from early childhood experiences, e.g. parents who taught them to fear members of a different race. Such deeply-rooted emotional reactions may help to fuel racially motivated crimes and may interfere with the racist's ability to see why racism is wrong. Some research has been undertaken into the neural basis for racial stereotyping.[317] Potentially this might lead to interventions which could attenuate such emotional responses. Harris has criticized this proposal on the basis that racism is likely to involve a complex network of beliefs and not merely emotional reactions.[318] Although this is almost certainly true, it does not demonstrate that ingrained emotional reactions do not contribute to the tendency to hold stubbornly onto ill-founded beliefs in the face of the evidence. Attenuating such emotional responses might help the offender to assess the issues dispassionately and realize that his racist views are ill-founded.

Delaying gratification

Difficulties with delaying gratification may lie behind some individuals' tendency to break the law. Cognitive enhancements could potentially help to rehabilitate criminals through enabling them to work out and implement strategies to delay gratification.[319]

Increasing the ability to focus on relevant issues

Recent studies suggest that individuals who score highly on measures for psychopathy may suffer from a kind of attention-deficit disorder which may help to

---

[317] A Hart et al, 'Differential Response in the Human Amygdala to Racial Outgroup Vs. Ingroup Face Stimuli', (2000) 11 'Neuroreport: For Rapid Communication of Neuroscience Research' 2355.
[318] J Harris, 'Moral Enhancement and Freedom' (2011) 25 (2) Bioethics 102.
[319] This issue is discussed in J Kennett 'Do Psychopaths Really Threaten Moral Rationalism?' (2006) 9(1) philosophical Explorations 69.; E Phelps et al, 'Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation' (2000) 12 Journal of Cognitive Neuroscience 729; W Cunningham et al, 'Separable Neural Components In The Processing Of Black And White Faces' (2004) 15 Psychological Science 806.

explain their characteristic anti-social behavior.[320] It seems that when presented with incentives for performing an action these individuals lose sight of the reasons against performing the action. Cognitive enhancements might enable these individuals to focus on all the relevant considerations (and in particular the reasons against breaking the law).

The need to take into account the interests of the offender

My approach places considerable weight on protecting the interests of the offender. Some theorists may object that it gives the offender's interests too much weight. It might be thought that the criminal has (to a large extent) forfeited his right to our moral concern. I do not accept this forfeiture view, partly because of the considerations that I mentioned earlier about free will and determinism. However, my arguments in the rest of this chapter do not *depend* on any particular position about free will. Whatever their views on free will, many people will find it intuitive that certain basic rights are inalienable, held in virtue of being human.[321] A society which regards certain members as worthless, not only wrongs those individuals but also degrades itself by treating them as worthless. The remainder of this chapter considers three ways in which society must respect offenders' moral worth – by treating them as members of the moral community, by recognizing their status as rational agents and by refusing to subject them to needless suffering.

Membership of the Moral Community

Society's response to criminal behaviour should recognize that offenders are members of the moral community, albeit members who have breached the community's norms. This principle is supported by the intuition that the state should

[320] J Newman et al, 'Attention Moderates the Fearlessness of Psychopathic Offenders' (2010) 67 Biological Psychiatry 66.
[321] For a defense of this idea from the point of view of a free will sceptic see: B Vilhauer , 'Free Will and Reasonable Doubt' (2009) 46 (2) American Philosophical Quarterly 131.

not ‘objectify’ law-breakers – that offenders should be treated as persons. Objectifying a group of people can involve emphasizing that ‘they’ are fundamentally unlike ‘us’. It can involve focusing on the idea that a deep division exists between the objectified group and the rest of society. One way of respecting offenders’ membership of the moral community is to preserve connections between the offender and other moral agents. This suggests that reforming offenders should occur primarily through relationships with others. Cognitive enhancements should never be used to circumvent the need for dialogue with the offender, but rather to enable effective dialogue to occur.

In order to treat the offender as a member of the moral community, limits must be set on the types of biomedical intervention that are permissible. Interventions that attempt to radically re-shape the offender’s basic character, goals or values are morally impermissible. The ability to employ such interventions would vastly increase the authorities’ (already considerable) powers for controlling offenders’ behavior. It would give the authorities a significant level of control over the individual’s *inner life.* This obviously has potential for abuse, e.g. it could be used to suppress legitimate dissenters. Even if this technique were only used to prevent offenders from engaging in uncontroversially wrongful activities it would still be morally objectionable because of the troubling *relationship* that would be created between the state and offenders. It would set the authorities on a completely different plane from offenders, greatly widening the *inequality of power* between them. It would imply that the offender is so radically morally deficient, and so unlike the rest of ‘us’, that the state needs to take over control of fundamental aspects of the offender’s personality. Such a policy would mark a huge shift towards characterizing offenders as ‘the other’ and thus towards objectifying them.

In contrast, interventions that merely reduce the strength of the offender's impulses, or increase his capacity for self control, or increase his capacity to empathize with others seem less likely to interfere with the core of the offender's personality. They do not instil in him particular values, or deprive him of the ability to decide for himself what values and beliefs he should adopt. Rather, a person who has become less impulsive and more self controlled seems to be in a better position to think about what his values really are and to translate those beliefs into action. The capacity for empathy could also put him in a better position to appreciate the reasons that are relevant to such decisions. He can criticize the authorities on the basis of his values and beliefs. The more limited interventions that this chapter advocates do not give the state the power to guarantee that the offender will behave in one particular way. The aim of these interventions is just to put the offender in a better position to *understand* the relevant reasons for changing his behaviour and to act effectively on these decisions.

One potential objection to my view states that using biomedical interventions to help change offenders' conduct inappropriately 'medicalizes' the problem of crime. According to this objection, medicalizing criminal behaviour implicitly *separates* criminals from the rest of the community – sending out the message that the problem is with 'them' and not with the rest of 'us' and that offenders are the only ones who need to change. Criminal behaviour is partly caused by social factors. It might be thought that giving cognitive enhancement to offenders obscures this fact, by sending out the message that the causes of crime lie solely within the individual (perhaps stemming from a biological defect) rather being the product of the offender's circumstances. This, it may be argued absolves the rest of the community from responsibility for helping to create or failing to alleviate these unfortunate circumstances.

In response, while it is important to acknowledge and seek to remedy the social causes of crime, this should not lead us to ignore the factors that can make particular individuals likely to reoffend. Pretending that these factors do not exist would distort the truth and would disadvantage the offender and wider society by putting obstacles in the way of effective rehabilitation.

Respecting the offender's personhood and rationality

It is essential that any attempt to change offenders' behavior respects the offender's status as a rational human being. This chapter only endorses those biomedical interventions that do not restrict (but rather aim to positively enhance) the offender's capacities for critical reflection.

Furthermore, cognitive enhancements should only be used if the offender gives his free and informed consent. By according weight to the offender's preferences, the state treats the offender as a person who still has moral worth and whose wishes are not completely discounted. The state shows respect for the offender's rationality, by allowing him to weigh the advantages and disadvantages for him of enhancement versus, for instance, spending a longer time in prison, and trusting that he is able to make an appropriate decision. I will argue, below, that the state should inflict no more distress on the offender than is needed to achieve its legitimate aims. Where such aims can be achieved by different methods, it is appropriate to give the offender some choice between those methods. Provided he is given adequate information, the offender is best placed to determine which method is likely to cause him more distress.

It might be objected that even if an offender agreed to accept cognitive enhancements in preference to other methods of reform/rehabilitation, this would not amount to *genuine, free* consent, given the coercive situation in which the offender

finds himself. In response, it is submitted that allowing the offender some say in the matter still shows respect for his preferences, even though the offender's options are limited. For the reasons stated in the previous section, limiting the offender's options can be justified by the need to protect society and by the value of reforming the offender and restoring him to the community. The 'consent requirement' strikes a balance between these interests and the offender's interests in not being forced to receive biomedical interventions. Provided officials do not exert *additional pressure* on offenders to receive biomedical interventions (rather than longer detention or other modes of rehabilitation), the fact that the offenders' options are limited does not seem to render their consent involuntary. After all, patients are often faced with hard choices where none of the available options are attractive and yet this does not make voluntary consent impossible. Some offenders have reported desperately wanting biomedical interventions, e.g. to help control destructive thoughts and urges, and have fought hard to have access to such medication.[322] If an offender voluntarily requests a biomedical intervention it seems more disrespectful to the offender's autonomy to refuse this request than to grant it.[323]

Some theorists still find such interventions troubling, regardless of whether the offender consents. They claim that the use of biomedical interventions treats the offender as a being without rights and not as a rational agent. Eric Matthews raises this type of objection to biomedical interventions designed to suppress offenders' violent or deviant sexual impulses. He writes:

> 'The harm done to [the offender] would be that of treating him like a thing, not a human being.... He would have been reduced to the level of a robot... People's bad behaviour can be legitimately changed only by persuasion to see that what they

[322] See, e.g. this section below and J Fischer, My Way: Essays on Moral Responsibility (OUP, Oxford 2006), p1-4.
[323] Bomann-Larsen L, 'Voluntary Rehabilitation? On Neurotechnological Behavioural Treatment, Valid Consent and (In)appropriate Offers' (2011) Neuroethics (Online First) doi:10.1007/s1215201191059.

> have been doing or proposing to do is unacceptable....Even if [the offender] chose to undergo this treatment, that would not necessarily make it morally tolerable. To choose to be dehumanized is choosing to be in a state where one can make no more choices, where one's existence is determined not by one's own will but by the requirements of others, and that does not seem like a morally legitimate choice to make. A sex offender or someone prone to outbursts of anger can legitimately seek help in learning to control his own impulses, but not treatment designed to remove those impulses altogether: the former is compatible with his continuing humanity, the latter is not.'[324]

Matthews seems to begin by taking an absolutist stance against using any form of biomedical means of trying to reform criminals. He states that persuasion is the '*only'* morally acceptable technique. However, the arguments he then goes on to produce do not support such an absolutist position. According to Matthews, biomedical interventions dehumanize offenders by putting them in a state where they can 'make no more choices'. This is simply not true of all types of biomedical interventions and particularly not those that can be classed as forms of 'cognitive enhancement'.

For one thing, suppressing an offender's urges to commit violent or sexual crimes would not thereby deprive him of the ability to make choices concerning all of his other activities. If the offender were released into the community after receiving this intervention, he would certainly have greater scope for making choices about how to lead his life than if he remained in prison. Furthermore, biomedical interventions which aim to suppress offenders' destructive impulses can actually *increase* the accused's ability to make informed and meaningful choices, by reducing the impediments to effective practical reasoning. Frequent, intense surges of violent anger can cloud an individual's judgment making it extremely difficult for him to assess issues such as whether or not his anger is justified and to appreciate the

[324] E Matthews, *Body-Subjects And Disordered Minds. Treating The Whole Person in Psychiatry* (OUP, Oxford 2007), pp181-182.

reasons why he should control himself. It can also make it difficult for him to enter into the kinds of relationships that are crucial to bringing home to him the wrongfulness of criminal conduct and to assisting him to lead a law-abiding productive lifestyle in future. He may alienate those who are trying to help him, or he may not take on board their advice because he is overwhelmed by the feeling that they are a threat to him that must be warded off. Reducing the offender's volatility through biomedical interventions can help him to focus more clearly on what he needs to do to improve his conduct.

Matthews's critique of using biomedical interventions in the criminal justice system relies on a dubious conception of what it means to be 'human' as opposed to being a 'thing' or a 'robot'. According to Matthews, an offender who received medication to suppress his urges to commit violent or sexual offences would be 'dehumanized'. But is a person really 'reduced to the level of a robot' just because she lacks a powerful urge to commit horrific crimes of violence or sexual exploitation? Many individuals, by nature, find the idea of performing these acts utterly repulsive and disturbing. Their failure ever to experience temptations to commit such crimes does not render these individuals robot-like or less than human.

There are alternative, more plausible conceptions of what capacities are important to leading a full, human life. Distinctively human capacities include the power to accord appropriate weight to relevant reasons for action and to conform one's actions to one's considered judgments about what is the best thing to do. Biomedical interventions, including those that reduce offender's destructive urges, could potentially enhance these capacities. It shows respect for the offender's existing rational capacities to allow him to decide for himself whether to avail himself of these enhancements. Giving him this choice treats him as an agent and not as a being without rights.

The following comments by a perpetrator of violent sexual offences illustrate these points. He vividly explains how powerful urges can interfere with an individual's thoughts and reasoning powers and how medication can help to restore the offender to a freer, more rational, and more human state.

> 'Basically, I am plagued by repetitive thoughts, urges, and fantasies.... I cannot get those thoughts out of my mind.... The best way for the average person to try to understand this is to remember a time when a song played over and over again in your head. Even if you liked the melody, its constant repetition was quite annoying, and the harder you tried to drive it out of your head, the harder it seemed to stick. Now replace that sweet melody with noxious thoughts of degradation, rape and murder and you will begin – and only just begin – to understand what was running rampant through my mind uncontrollably....I was tired of being tormented by my own...mind. So unbelievably tired.....Having those thoughts and urges is like living with an obnoxious roommate. You cannot get away from him because he is always there. What Depo-Provera[325] did was to move that roommate down the hall to his own apartment. The problem was still there, but it was a whole lot easier to deal with because it wasn't always in the foreground. He didn't control me anymore – I was in control of him. It was an unbelievable sense of freedom. It made me feel as if I were a human being again, instead of some sort of horrible monster.'[326]

A critic of biomedical interventions might concede that a person would not be in a robot-like state just because some of his negative urges were suppressed by medication. Nevertheless, the critic might object to giving this medication to the offender, because doing so would involve a *failure to recognize* the offender's humanity. On this view, the only way to treat an offender like a human being is by

---

[325] Depo-Provera is a branded drug that was originally developed as a progesterone-only female contraceptive. The active ingredient is medroxyprogesterone acetate (MPA). It is given as an injection every three months. When used in males, it can reduce compulsive sexual fantasies and sex drive. (B B Maltzky and Field G, 'The Biological Treatment Of Dangerous Sexual Offenders, A Review And Preliminary Report Of The Oregon Pilot Depo-Provera Program.' (2008) 8 Aggression and Violent Behavior 391.
B Maltzky, A Tolan and B McFarland B, 'The Oregon depo-provera program: A five-year follow-up' (2006) 18 Sex Abuse 303.
who also list possible side-effects.)

[326] Michael Ross, quoted in Fischer 2006, pp. 2-3.

seeking to reform him using the power of rational argument alone, rather than trying to give him a chemical fix as if he were a broken 'machine' (Freedman 2000, p136). Traditional techniques of reform and rehabilitation, such as victim-offender mediation present him with reasons to change his behavior, e.g. by trying to convince him that his behavior cannot be justified and by showing him the suffering of those affected by his crime. In contrast, biomedical interventions, it is claimed, alter the offender's thought-processes and/or behavior by directly affecting how his brain works without giving the offender any additional *reason* to think or act differently.

This criticism would have some force if biomedical enhancements were used as a *substitute* for reasoning with the offender. The idea of treating someone like a 'machine' sounds so sinister mainly because it suggests that the individual is being *excluded* from rational dialogue and from relationships with others. However, this chapter advocates using cognitive enhancements to enable or facilitate rational dialogue to take place. Without the aid of cognitive enhancements the offender may never fully access certain reasons for action and may be cut off from relationships which could help him to develop as a rational human being. Furthermore, it is implausible to suggest that any method of altering behavior that does not involve rational argument is necessarily morally intolerable. Imagine that there was good evidence that putting offenders on a regular exercise regime would help to reform them, by increasing their serotonin levels, enabling them to feel more empathetic and less defensive. Imagine that this opens up a window of opportunity for offenders to enter into relationships that they had been emotionally resistant to entering, e.g. victim-offender mediation schemes. Having begun such relationships, the offenders could become better able to appreciate why their criminal actions were wrong and why they should change their conduct. The exercise regime *per se* would not give offenders a reason for changing their behavior. But this does not seem to make it morally intolerable.

Showing compassion to an individual by helping to alleviate his distress can surely be an important way of treating him as a fellow human being. As I will discuss in the next section, cognitive enhancements have the potential to spare individuals needless suffering.

Suffering

In the future, cognitive enhancements may have the potential to reduce the offender's suffering significantly. As noted above, the offender himself may find the factors that impede his practical reasoning, such as repetitive thoughts and powerful, irrational urges intensely distressing. Cognitive enhancements might help to relieve this distress. Such interventions may also make the process of reform and rehabilitation itself less burdensome to the criminal. For instance, attempts to reform criminals that do not involve enhancements may require a longer time of incarceration than if enhancements were used. However, the offender's interest in not being made to suffer can count against using enhancements that might expose the individual to serious side-effects.[327]

But why should the state prefer methods of dealing with criminal behavior that involve less suffering for the law-breaker? According to retributivists, it is intrinsically *good* that the offender is made to suffer in proportion to his moral guilt. Retributivists would oppose giving enhancements to offenders if doing so would diminish the distress involved in punishment to a level that is lower than the amount of suffering that the offenders' 'deserve'. In response, it should be recalled that retributivism faces considerable difficulties for the reasons indicated in Chapters One

[327] Loss of bone density is one possible long-term side-effect of anti-androgens that are currently given to sex offenders: H Greely, 'Direct Brain Interventions to 'Treat' Disfavored Human Behaviors: Ethical and Social Issues' (2012) 91(2) Clinical Pharmacology & Therapy 163. As Greely notes, the possibility of this side-effect is known due to studies on the use of this medication as a form of birth control for women -'Sex offenders receive the drug at much higher doses. What is effect on their bones? No one knows, as it was never tested on men, and because the recipients are sex offenders, almost no one cares.' P 163.

and Two. Secondly, many of those who believe in retribution also believe that this is only *one* of the functions of the criminal justice system, alongside reform and rehabilitation. Once the offender has served the part of his sentence that is designed to inflict the suffering he deserves, there is no reason, on this view, why the part of the sentence which is aimed at reform and rehabilitation should purposely aim to impose still more suffering on the criminal.

Pure consequentialist theories place certain limitations on the amount of suffering which should be imposed on offenders. They state that the offender should only suffer to the extent that this is necessary in order to prevent crime, or to promote the general welfare in some other way.[328] This is sometimes referred to as 'economical prevention'. However, the protection that this principle affords to the offenders' interests does not seem to go far enough. The principle of economical prevention is compatible with inflicting levels of suffering on offenders which are intuitively far too severe. It is compatible with inflicting an extremely harsh penalty on the offender if this will prevent each of very many other people from suffering some very slight hardship. What matters for the consequentialist is the *total* level of distress to be prevented. For example, if subjecting a group of offenders to an extremely distressing form of treatment would prevent a much larger number of people from each suffering a tiny inconvenience, then the total amount of distress to be prevented could, according to the utilitarian calculation, be enough to justify forcing the offenders to undergo the painful treatment.

This chapter advocates providing greater protection for the offender's interests than that implied by traditional consequentialism. It is unacceptable to impose a sentence on the offender which would cause him much greater distress than the distress which any particular individual would suffer if the sentence were not imposed. The

[328] J Bentham, *Principles of Penal Law* (ebooks@adelaide, Adelaide, South Australia 2011), available at: http://ebooks.adelaide.edu.au/b/bentham/jeremy/principles_of_penal_law/. Accessed 26 June 2012.

proposed level of protection is justified by considerations of equality. It is *prima facie* wrong to create a situation where individuals suffer grossly unequal levels of distress (I will refer to this as 'the equality principle'). [329] This principle could have a similar effect to the retributive principle that the severity of the penalty must be proportionate to the gravity of the crime. For instance, both principles imply that more burdensome interventions may be imposed on dangerous violent offenders than on shoplifters. However, the 'equality principle' has a distinctly different basis from the retributive doctrine. Unlike retributivism, the principle that I advocate is not based on the moral responsibility of the individuals concerned. For instance, the equality principle applies even if the law-breaker is mentally ill. It is permissible to detain dangerous psychotic individuals in an institution for relatively long periods of time, despite the fact that they are not morally responsible, if this is necessary in order to prevent them from being seriously violent to those around them. However, it would not be justifiable to take such an extreme measure to prevent a mentally ill person from committing relatively minor disturbances. This can be explained by the equality principle. It is better that each of a larger number of individuals should suffer a slight hardship than that one individual should bear an extremely heavy burden in order to prevent others from suffering this slight hardship. This principle should act as a constraint when deciding between more or less burdensome methods of reform and rehabilitation.

Some theorists may be concerned that attempting to limit the amount of suffering which the offender undergoes may prevent him from genuinely reforming. Experiencing remorse, it may be argued, which is necessarily painful, is an essential element of the process of true reform. In reply, it should be acknowledged that, in order to achieve the legitimate aim of bringing the offender to recognize that his actions were wrongful and that he needs to change, it will indeed be necessary for the

[329] This principle is defended in Honderich 1984, p. 78.

offender to experience some distress. Cognitive enhancements could potentially play an important role in helping some offenders to appreciate the wrongfulness of their conduct and to experience remorse. It should also be noted that certain offenders experience greater distress than others, not because they have greater cause to feel remorseful for their crimes, but because the process of reform and rehabilitation is more prolonged and difficult for them due to factors which impede their powers of practical reasoning. Cognitive enhancements could reduce these impediments, helping to ensure that offenders go through no more distress than is necessary for genuine reform.

It might be objected that if the process of reform and rehabilitation is more difficult for some offenders (e.g. because they are prone to outbursts of anger) this is due to their own moral shortcomings and so it is fair that they suffer more distress than more even-tempered offenders. In response, this claim seems to rest on the assumption that individuals are responsible for creating their own flawed characters. For the reasons indicated in Chapter Two, this is a dubious claim. Furthermore, it is important to bear in mind that it can be an extremely difficult and slow process to try to undo character traits that have been laid down early in life – a process with many relapses along the way and which in some cases, despite the individual's considerable efforts, is never wholly successful. While the offender is learning to control his anger without medication, those around him may be at risk from or actually suffering the consequences of his outbursts. An offender who seeks to receive cognitive enhancements to facilitate the process of rehabilitation, rather than expose others to this increased risk shows a willingness to take responsibility for his conduct, which ought to be encouraged.

A related concern is the idea that the experience of struggling with conflicting desires, and resisting temptation is intrinsically valuable, and that this might be lost if

biomedical interventions were used in order to reduce the strength of offenders' urges. JM Olsen writes:

> 'So, then, what is inherently valuable in moral effort? The answer is that moral effort is required if we are to have morality at all. Morality, I maintain, requires agency, and if no moral action ever requires any effort, then we would be, in Kantian terms, mere slaves of inclination. Put another way, there is something inherently valuable about agency, but agency is empty without resistance—that is, temptation.'[330]

It is important to bear in mind, however, that it is highly unlikely for it to become technically possible to *eliminate* all of an individual's temptations to do wrong. The question is whether it is permissible to use biomedical interventions to reduce somewhat the force or number of these temptations, or to increase the ability of the offender to deal with them. As noted above, many people due to their upbringing and/or natural predispositions experience little or no temptation to commit serious crimes. This does not seem to indicate that their agency is somehow deficient in comparison with someone who feels strong temptations to break the law. Furthermore, Olsen puts forward a doubtful interpretation of the Kantian idea of being a slave to inclination. Olsen suggests that a person's good deeds are morally 'empty' unless she feels tempted to perform bad actions. Kant, in contrast, required that, to have moral worth, a good deed must not be motivated merely by an inclination to do it, but by the recognition that it is the right thing to do. If this recognition is sufficient to motivate the agent to do the good deed, then she does not seem to be enslaved to her inclinations. Feeling tempted to do bad actions, however, does not seem to be strictly necessary to enable the agent to recognize or be motivated by the reasons for doing good actions.

[330] JM Olsen, 'Depression, 'SSRIs, and the Supposed Obligation to Suffer Mentally' (2006) 16 (3) Kennedy Institute of Ethics Journal 283, p289.

To summarise the argument so far: It is possible to identify several kinds of cognitive enhancement which may in the future, if sufficiently refined, play a useful role in reforming and rehabilitating offenders. However, certain restrictions must be placed on the means which state may employ to achieve its rehabilitative goals. Society's approach to dealing with criminal behavior must treat the offender as a member of the moral community and a rational agent and must respect the offender's interests in not being made to suffer unnecessarily.

Distinguishing Values from Capacities

So far I have argued that the state should not attempt to control an offender's values using direct interventions, but that it may be permissible to enhance his capacity to grasp the relevant considerations, so that he is better able to decide for himself which values to adopt. The next question is how this distinction between enhancing capacities and re-shaping values is to be drawn in practice.

a) Re-shaping Values – Central cases

It is possible to imagine certain types of intervention that clearly aim to re-shape the offender's values. For instance, the state might try to influence offenders by sending out subliminal messages promoting the state's favoured values, while the offender is watching TV in his cell. Or it might, in the future, become possible to develop a device which transmits such messages that might be installed in the offender's brain. Another clearly unacceptable technique would be to try to modify the offender's brain to make him very suggestible, impair his powers of critical reflection and then bombard him with propaganda. (Even more extreme interventions have been discussed in the free will literature –e.g. assuming that psychological properties such as those that are involved in valuing are identical with or are nomically paired with brain states, it is conceivable that the authorities could operate on the offender's brain in order to render his values qualitatively identical to the values of a 'model

citizen'. However, thankfully, it seems unlikely that knowledge of the brain will advance enough to make that technically possible in the foreseeable future.)

b) Enhancing Capacities – Central Cases

Certain types of intervention seem to be relatively straightforward instances of enhancing capacities. One example is the idea of increasing the offender's power of attention. As, I mentioned earlier, recent studies suggest that individuals who score highly on measures for psychopathy may suffer from a kind of attention-deficit disorder that may help to explain their characteristic anti-social behaviour. [331] Neurological enhancements might enable these individuals to focus on all the relevant considerations (and in particular the reasons against breaking the law). Neurological enhancements may also potentially help to rehabilitate criminals through enhancing their ability to delay gratification. [332]

Certain offenders seem to lack the normal bodily responses to stimuli. There is some evidence to suggest that the emotional quality of an experience (e.g. whether it was rewarding or aversive) is normally 'remembered' by the body and when the person is contemplating facing the stimulus again they experience a bodily reaction in anticipation of the stimulus, like a kind of warning system. People whose warning system is lacking or defective may be more likely repeatedly to engage in self-defeating behaviour, and may also be more likely to reoffend.[333] Direct interventions aimed at helping such people seem to fall into the enhancing capacities category.

[331] For an interesting discussion of this issue see J Newman et al, 'Attention Moderates the Fearlessness of Psychopathic Offenders' (2010) 67 Biological Psychiatry 66.
[332] This issue is discussed in J Kennett (2006).
[333] Blair et al, The Psychopath: Emotion and the Brain (Blackwell, Oxford 2005).

c) Borderline Cases

There is not always a razor sharp line between using neurotechnologies to directly re-shape offenders' values and the use of these techniques to enhance offenders' capacities for responsible agency. For example, an intervention might reduce the strength of an offender's violent and/or deviant sexual impulses. It might be argued that this is a method of enhancing offenders' rational capacities, because intense, repetitive urges or fantasies can cloud an individual's judgement, making practical reasoning difficult. Reducing the strength and frequency of these urges could put the offender in a better position to focus on the reasons that are relevant to his decision about how he should act. Alternatively, it might be argued that interfering with offenders' urges is a method of directly re-shaping their values, because an offender who values violence or deviant sexual conduct might do so partly as a result of experiencing these impulses and urges.

Another borderline case is the capacity for empathy. There are both conceptual and empirical reasons for thinking that this capacity is necessary genuinely to appreciate what is wrong about harming others. For instance, individuals with markedly reduced levels of empathy have exhibited difficulties in distinguishing conventional rules (such as rules of etiquette) from moral rules and in ranking wrongs in order of seriousness.[334] However, it also seems likely that one's degree of empathy plays a role in *moral motivation* (as well as understanding) and in which values one ends up adopting.

d) Dealing with Borderline Cases

The issue of borderline cases can be decided partly on the basis of the principles that I have already outlined in the previous two chapters. One relevant consideration is the amount of control which the intervention would allow the state to exert over the

[334] J Blair et al, The Psychopath: Emotion and the Brain (Blackwell, Oxford 2005), p57-59.

agent's decisions about what he should do. The greater the state's level of control, the greater the inequality between the offender and the rest of the community. Interventions which merely reduce the strength of an offender's violent impulses do not give the state the power to ensure that the offender endorses the state's favoured values. The offender may still reject society's demands. Similarly, it seems unlikely that interventions that increase a person's empathy to within normal levels would thereby determine which values the individual will adopt. People with normal levels of empathy often behave callously and have less than caring values. One possible objection to enhancing the capacity for empathy is that there are some situations when the person cannot help but *exercise* this capacity. However, there are many capacities of which this is true, such as the capacity to read or to understand a language. People rarely raise objections to literacy courses in prisons or to teaching non-native speakers English. Furthermore, exercising such skills can also plausibly affect people's values, perhaps allowing them to become more integrated in the community. There are also various ways in which people can repress their capacity for empathy. But such interventions are less troubling than interventions that allow the state to shape the offender's behaviour and inner life to a greater extent.

Interventions that would alter an attribute which is *central* to who the person is, as an agent, are particularly troubling. A particularly fundamental alteration sends out a strong message that the offender is radically defective, and unlike the rest of 'us'. Again, it is submitted that a momentary impulse or urge is less central to the offender's agency than, say, a firm commitment to a particular principle or course of action. Interventions that directly target 'second order desires' are also particularly problematic. It is plausible that second order desires are at least partly constitutive of values, since they concern what kind of person the agent wants himself to be. Bublitz and Merkel note that some pharmaceuticals seem to promote a positive view of oneself and an experience of authenticity – a feeling of 'really being oneself'. If an

intervention instilled new first-order desires, and was accompanied by authenticity-enhancing medication, then this could cause the individual to *identify* with his new first order desires. This could amount to an objectionable interference with the offender's second order desires.

Even interventions which only target first order desires may give the state an unacceptable level of control over the agent's values and character if the agent's first order desires are *extensively* altered. For instance, imagine an offender who has a corrupt value system according to which acts of terrible cruelty are morally permissible. This offender also, by nature, has an aggressive temperament and has always been extremely insensitive to others' distress. Imagine that the state managed, through direct interventions, greatly to reduce his aggressive feelings, so that he became more placid than most non-criminals, and that direct interventions were also used in order to greatly increase his empathy so that he could not bear the slightest sign that another was suffering. Such a course of treatment seems to go beyond an attempt to put the offender in a better position to understand what is wrong about harming others. It seems likely that this use of direct interventions could have a significant impact on his higher order desires and values, even though this is not *inevitable* (an even-tempered person who does not like to witness violence first-hand could still endorse cruelty and violence). Furthermore, the fact that the offender's aggressiveness and empathy have both shifted from one end of the spectrum to the other suggests that he has undergone a fundamental personality change and, as argued above, implementing extreme changes to the offender's character sends out a stronger message that the offender's pre-existing character is fundamentally defective and that the offender is incapable of change through normal social interaction.

Nicole Vincent is sceptical about whether we are currently able to distinguish reliably between capacities and fundamental character traits/values. She is particularly concerned that altering offenders' values/character traits might undermine the offender's authenticity or even transform the offender into a different 'self'. She concedes that we cannot be *certain* that direct interventions would have this effect. However, she insists that we do not need certainty about this in order to be justified in ruling out the use of direct interventions to modify 'borderline traits'. She writes:

> '...we currently have no way to distinguish character flaws from capacity deficits, and thus ... to be on the safe side we should abstain from 'treating' people with direct brain interventions until we have gathered more empirical data on this topic and analysed the conceptual basis of the distinction between capacity and character.'[335]

Despite the critique of the notion of authenticity presented in Chapter 1, Vincent is right to insist that all such arguments still leave room for doubt. (Furthermore, some idea of authenticity may still be of value even if it is not essential for free will.) Granted that this area involves uncertainty, it is less clear that abstaining from all direct interventions amounts to staying on the 'safe side'. A number of different interests need to be balanced. Given that offenders are liable to state interference of some sort, it may be difficult to determine whether treatment or traditional punishment is the safer option. What is safer for the offender may not be safer for the public. Even if we give more weight to the offender's interests than to the interests of the state, it is far from clear that abstaining from direct interventions would be safer than, say, prison. As Lawrence Stern notes: "It is true that prison does not aim to subvert rational or moral capacity... But it can break a man. A man can emerge from

[335] N Vincent, 'Capacitarianism, responsibility and restored mental capacities' in B van den Berg and LKlaming (eds), Technologies on the Stand: Legal and Ethical Questions in Neuroscience and Robotics (Wolf Legal Publishers, Nijmegen 2011) 41-65, p52.

prison no more able to commit a crime than to walk into a fire."[336] Even if we are concerned primarily with promoting the authenticity of the offender, there are compelling reasons in favour of treatments such as empathy enhancement, or treatment to reduce impulsivity or violent urges.

On one plausible interpretation of authenticity, an authentic individual has an appropriate degree of self-knowledge. If someone is completely deluded about such things as her own virtues, vices, abilities and limitations, that seems to undermine her authenticity. For instance, the character Cordelia in Rebecca West's novel, *The Fountain Overflows* is deluded that she is a talented violinist.[337] She has tremendous technical skill and perfect pitch but she is deeply unmusical – her skill is merely mechanical, she lacks musical sensitivity. Her life is centred around her supposed musical talent and the sycophantic people who pretend to admire it. Imagine that one day she 'wakes up' from her delusion, realises that she will never be a great violinist but then discovers she has a genuine talent for something else and builds on that. Her new life would seem more authentic than her old lifestyle.

This emphasis on the self-discovery element of authenticity fits with the experience of some patients with ADHD who have reported feeling that taking Ritalin to reduce their impulsivity helped them to feel authentic. Bolt and Scherner provide the following examples:

> One respondent said ..."It's not that you're not yourself anymore. I believe I have always been myself, but because the medication makes you more tranquil you start to look differently at yourself. You take more time for yourself. And you discover things that you did not expect of yourself." In fact, she discovered that she was a good painter and enjoyed painting a lot. Another respondent also said that he felt more 'himself' on medication. He was more able to control his impulses and his life moved more smoothly. He also felt

[336] L Stern, 'Freedom, blame, and moral community' (1974) 71 The Journal of Philosophy 72, pp84-85.
[337] R West, *The Fountain Overflows* (Virago Press, London 1984).

> calmer on medication and this gave him more 'time for himself': "I haven't read a book in years because I couldn't concentrate. But now I'm reading again. I used to read a lot when I was younger".[338]

Enhancing an offender's empathy is likely to increase her self-knowledge. People who are deficient in empathy frequently have limited self-understanding since they are unable to see themselves as others see them. Some such people also seem to have a limited conception of who they are due to a related inability to put themselves in the shoes of their future selves.

Conclusion

I have argued that compatibilists face considerable difficulties in objecting to direct interventions on the basis that they violate free will. However, there is an alternative objection to certain types of intervention, which does not rely on an account of free will. Attempting to enhance virtue responsibility by using neurological interventions to modify offenders' values would risk creating the wrong kind of relationship between the state and offenders. However, I have not argued that we should oppose *all* neurological interventions within the criminal justice system, and I have outlined some relevant considerations for assessing techniques that may emerge in the future.

[338] I Bolt and M Scherner, 'Psychopharmaceutical Enhancers: Enhancing Identity?' (2009) 2 Neuroethics 103-111, p106, emphasis in original.

Part Four: Free Will and the Criminal Law

Part Four: Overview

Part Four will discuss some of the practical implications of my approach for the criminal law. The arguments presented in this thesis suggest that any aspects of the criminal law whose only justification depends on the concepts of free will and retribution should be revised. Chapter 10 examines the criminal law's current position on the questions of free will and retributive responsibility. It argues that the dominant view among criminal law theorists - that the criminal law is thoroughly compatibilist - rests on dubious arguments. There is at least as much reason for thinking that principles of criminal law embody libertarian, incompatibilist assumptions as there is for thinking that they make only compatibilist assumptions. Chapter 11 examines what revisions to criminal law doctrines would enable these doctrines to be justifiable even if retributivism (and the notion of free will on which it depends) were regarded as unsound. It suggests several changes to our *understanding* of the rationale for the provocation defence, self-defence and mental disorder defences; and for the overall *structure* of criminal defences. My intention in Part Four is not to give a comprehensive treatment of these topics, but rather to lay the groundwork for future work.

Chapter Ten: The Criminal Law's Current Position on Free Will

Introduction

The aim of this chapter is to evaluate two rival accounts of the criminal law. According to one account, the current law presupposes a conception of responsibility that is entirely compatible with determinism**.** The other account represents the law as endorsing incompatibilism. There is no clear legal rule that explicitly supports either side in this debate. However, this chapter aims to provide some reasons for thinking that those rules which concern retributive responsibility implicitly rest on principles that are incompatible with determinism. It is important to consider this issue for two reasons. Firstly, the argument that retributivism is compatible with determinism would be strengthened if it could be shown that determinism is completely irrelevant to the legal doctrines according to which individuals are currently excused from or subjected to retributive punishment. So this claim should be challenged in order to make a convincing case that, in the light of the free will debate, the soundness of retributivism is genuinely uncertain. Secondly, this chapter aims to lay the groundwork for the final chapter which will consider what reforms would be needed in order to make the criminal law compatible with determinism. In order to do this, it is relevant to examine the extent to which existing legal doctrines are compatible with determinism.

The greatest impact that the free will debate has had on criminal law theory is in relation to the 'causal theory of excuses'. ('Excuse' in this context is used in a broad sense to refer to situations where an actor is considered not to be blameworthy, even though her behaviour was prohibited and was not justified. The term is used in the literature on causal theory to cover certain defences that some writers do not regard as 'true' excuses, including defences which involve denying that the accused had

mens rea or that her behaviour constituted an 'action', or that the accused possessed the general capacities necessary to qualify as a moral agent.[339]) According to causal theory, the criminal law presumes that agents generally have libertarian free will.[340] However, causal theory continues, the law recognises that, on rare occasions, factors outwith the agent's control are either causally *sufficient* to produce the agent's behaviour or exert such a heavy causal *influence* on her conduct that she is not blameworthy for that conduct. (Examples often cited by the causal theorist include reflex 'actions' and coercion.) On such occasions the accused does not deserve retributive punishment and (to the extent that the law upholds retributivism) the law does not hold the agent criminally responsible.

The opponents of causal theory – compatibilists - vary in their positive accounts of criminal responsibility and excuse, but they all agree that causal determinism has nothing to do with liability to retributive punishment.[341] 'Choice' theorists, for instance, argue that people are responsible for their choices (even if those choices

---

[339] See, e.g., M Moore, 'Causation and the Excuses' (1985) 73(4) California Law Review 1091.

[340] The following writers endorse this view. On English law: A Ashworth, 'Justifying the Grounds of Mitigation' (1994) 13 Criminal Justice Ethics 5, p 8, stating that 'there are a few defences in which elements of determinism play a significant role (involuntariness, duress, perhaps insanity)..'. On North American law: A Kaye, 'Resurrecting the Causal Theory of Excuses' (2005) 83 Nebraska Law Review 1116 ; N Morris, *Madness and the Criminal Law* (University of Chicago Press, Chicago 1982). On Scots law: G Gordon, *The Criminal Law of Scotland* (W Green, Edinburgh 1984), Volume 1, pp 118-119, stating that, 'Voluntary human actions are...regarded as themselves uncaused. This is a necessary inference from the doctrine of freewill; and without some form of that doctrine, however restricted, there can be no moral responsibility in the sense of praise or blame.' He cites coerced and reflex 'acts' as instances where behaviour is regarded as a mere effect of prior causes and where the actor is not held legally responsible. However, in an earlier passage he argues that practices of praise, blame, reward and punishment can still be justified even if determinism is true, since such practices can still be an effective means of improving behaviour (pp51-53). Perhaps the best way of reconciling the two passages is to interpret Gordon as arguing that without free will there can be no moral responsibility in the sense of praise and blame *without pragmatic justification.* On Australian law: D Hodgson, 'Criminal Responsibility, Free Will and Neuroscience' in N Murphy et al (eds), *Downward Causation and the Neurobiology of Free Will* (Springer-Verlag, Berlin 2009).

[341] Compatibilist accounts of criminal law include: J. Horder, 'Determinism, Liberalism and Criminal Law' (1996) 49, Current Legal Problems, 159; P Litton, 'The Abuse Excuse in Capital Sentencing Trials: Is it Relevant to Responsibility, Punishment or Neither?'(2005) 42 American Criminal Law Review 1027*;* M Moore, 'Causation and the Excuses' (1985) 73(4) California Law Review 1091; S Morse, 'Culpability and Control' (1994) 142 University of Pennsylvania Law Review 1587; S Pilsbury, 'The Meaning of Deserved Punishment: An Essay on Choice, Character, and Responsibility' (1991) 67 Indiana Law Journal 719 ; G Vuoso, 'Background, Responsibility, and Excuse' (1986) 96 Yale Law Journal 1661.

were the inevitable product of factors beyond their control).[342] Behaviour is excused, according to this theory, if it is not the result of an agent's choice; or if the agent was not sane, or mature enough to be blamed for her 'choices'; or if the agent made the choice for acceptable reasons, e.g. to avoid a 'substantial evil'. 'Character' theorists claim that people are only responsible for actions that reflect their (predetermined) characters. On one version of this theory, an action does not reflect an agent's character if it springs from a desire that the agent does not accept in the light of her value system.[343] According to some character theorists, an agent may be excused if her action does reflect on her character, but does not show her character to have unacceptable flaws.[344] 'Attitude' theorists claim that people are punished for actions that reflect certain attitudes of hostility/disrespect (regardless of whether the person was predetermined by factors outwith her control to have those attitudes). On this theory, a person may, for example, be excused if her conduct does not in fact express an unacceptable attitude (because, for instance, it was involuntary) or if she is not the kind of agent from whom the criminal law demands an attitude of respect (e.g. an individual incapable of practical reasoning).[345]

According to all of these compatibilist theories the fact that a person's action was determined is *entirely irrelevant* to whether the person should be excused or punished for that action. Clearly, in order to establish this strong claim, compatibilists must do more than simply produce a description of the excuses that does not mention determinism. For it is possible that in a particular situation where an actor is excused there are several factors that are relevant to the question of the actor's responsibility. The compatibilist may have named one or more of these

[342] E.g., M Moore, *Placing Blame* (OUP, Oxford 1997).
[343] E.g., V Tadros, *Criminal Responsibility* (OUP, Oxford 2005).
[344] E.g., J Horder, *Excusing Crime* (OUP, Oxford 2004).
[345] E.g., P Westen, 'An Attitudinal Theory of Excuse' (2006) 25 Law and Philosophy 289. See also P Strawson, 'Freedom and Resentment' (1962) 48 *Proceedings of the British Academy*, 187.

factors, but may still have left a relevant factor out of his account, in leaving out determinism. The causal theorist can accept that the features of agency that compatibilists consider important (such as voluntariness and rationality) really are essential to free action and that their absence should result in an excuse. But the causal theorist would also insist that certain capacities which are incompatible with determinism (such as the ability to do otherwise and the ability to be the 'originator' of one's choices) are also required for genuine freedom and responsibility. Factors such as insanity, epileptic seizures and coercion could deprive a person of *both* compatibilist and incompatibilist freedom and the absence of both kinds of freedom can explain why such a person is excused. After all, one can think of other situations where several exculpatory factors exist, each of which being sufficient on its own to excuse, e.g. an accused might have been both seriously mentally ill and coerced and entitled to an excuse based on either of these factors.

What would be needed in order to settle the issue conclusively is a case where the accused satisfied the compatibilist's prerequisites for criminal responsibility, but where the law clearly acknowledges that the accused lacked the conditions for incompatibilist responsibility. If the law holds an accused criminally responsible (on a retributive basis) in these circumstances, then this would provide a clear counterexample against incompatibilism. If the person is relieved from criminal responsibility then this supports the view that the law is in fact incompatibilist. The first section of this chapter will argue that purported counterexamples against incompatibilism fail. Nor, however, can the incompatibilist point to any real life case where an accused is fully responsible in the compatibilist sense, but is excused because she lacked incompatibilist freedom.

A legal positivist might conclude that the lack of a clear rule supporting either compatibilism or incompatibilism means that it is impossible to say that the law

favours one side over the other. This chapter, however, asks the further question of whether there is any legal *principle* that can settle the issue. According to Ronald Dworkin, a legal principle is an explanation of a body of rules that fulfils the following requirements: 1) it 'fits' best with the legal system's institutional history and 2) it shows the rules which it explains in their best possible moral light.[346] The second part of this chapter will compare compatibilist and incompatibilist explanations for current excuses. It will argue that explanations which include the premise that determinism and responsibility are compatible do not satisfy either requirement for being a legal principle. Rather, it is much more plausible that our legal system includes incompatibilist, libertarian legal principles.

In assessing whether a principle meets the two criteria for being a legal principle, the second part of this chapter will pay particular attention to whether the principle is likely to be in accordance with the moral intuitions of those who have shaped, applied, accepted and obeyed the laws which the principle seeks to explain. For, to say that a principle fits with a system's 'institutional history' does not simply mean that it is consistent with the system's rules. It is also essential to consider whether the principle reflects judges', legislators' and citizens' understanding of the rationale behind the rules. It is submitted that a moral principle which happens to fit with the rules, but which is absent from judicial and political rhetoric and which is completely alien to the thinking of ordinary citizens cannot count as *belonging to their legal system.* Furthermore, insisting that such a principle is actually a *legal* principle is unlikely to show the law in a particularly good moral light. If there is no real indication of this principle in the rationales that judges and legislators have publically given for the rules then this raises serious questions of procedural justice – for surely citizens are entitled to know on what principles the laws they are expected to obey are based. As Dworkin comments, 'The political history of the community is

[346] R Dworkin, *Law's Empire* (Harvard University Press, Cambridge, Massachusetts 1986).

pro tanto a better history....if it shows judges making plain to their public, through their opinions, the path that later judges ....will follow'.[347] In addition, an explanation of the rules will tend to reveal the law in a better moral light if it 'shows judges making decisions that give voice as well as effect to convictions about morality that are widespread throughout the community', rather than enforcing a moral view which goes 'against the wishes of the people'.[348] This chapter will argue that causal theory (the incompatibilist, libertarian account of criminal responsibility) seems to be more in tune with widespread convictions about morality than the opposing theory, which states that determinism and retributive responsibility are entirely compatible. This is not to say, however, that libertarianism is acceptable. On the contrary, as explained in the previous chapter, libertarianism faces considerable empirical and conceptual difficulties. Nevertheless, it is submitted that, of all the positions in the free will debate, the libertarian conception of responsibility seems to be the most plausible candidate for being a legal principle, since it probably fits best with ordinary people's moral intuitions.

Counterexamples

Counterexamples against Incompatibilism

Compatibilist legal scholars frequently claim to have produced counterexamples against the causal theory of excuses. In order to see why these counterexamples fail it is necessary to be clear about what causal theory actually is. It is a libertarian incompatibilist theory. It is not a hard incompatibilist position. Causal theory holds that it is sometimes fair to subject a person to retributive punishment for criminal behaviour. However, it maintains that a person does not deserve retributive punishment for conduct that was causally determined by factors wholly outwith the

[347] Ibid, p248.
[348] Ibid, p249.

agent's control. According to causal theory, the law excuses such conduct (or at least, if the law fails to do so then it fails to uphold retributivism).

Causal theory is not essentially committed to an all or nothing view according to which a particular piece of behaviour must be either entirely determined or entirely free. Causal theorists can also maintain that responsibility is incompatible with or at least diminished by 'near determinism'. In other words, an agent deserves little or no retributive blame/punishment for behaviour if the agent were subject to causal pressures that were so powerful that they rendered her behaviour *extremely* probable. Whether a person is excused (or partially excused) will, according to causal theory, depend partly on how powerful the pressures were. Compatibilists sometimes caricature all libertarian theories as dismissing 'near determinism' as irrelevant to responsibility. For instance, John Martin Fischer represents libertarians as excusing a person only if factors outwith the agent's control 100% guaranteed that her behaviour would occur and holding her fully responsible if these factors made her behaviour even 99.9999% more likely to occur. Fischer then asks, 'how could *this* sort of difference (the difference between 100 percent and even 99.9999 percent) make such a difference (a difference between being robustly responsible and merely responsible in some attenuated sense or not responsible at all)?'.[349] However, few (if any) libertarians would in fact hold a person fully responsible if causes outwith the agent's control made it 99.9999% certain that she would break the law – i.e. only one in a million people subjected to these pressures would have done differently. Libertarians do not think that people should be punished for failing to be that one in a million – a saint or a hero. Furthermore causal theory does not claim that *every* excuse involves determinism, or near determinism. So pointing to an excuse that has nothing to do with determinism is not enough to refute causal theory.

[349] J Fischer, *My Way: Essays on Moral Responsibility* (Oxford, OUP 2006), p6.

Many purported counterexamples against causal theory's libertarian conception of responsibility fail because they do not involve determinism or anything approaching determinism. For instance, Steven Morse argues that accomplice liability is inconsistent with libertarianism. According to Morse, a libertarian would insist that 'the accomplice did not in any way cause the perpetrator to commit the crime, because only the perpetrator caused himself or herself to commit the crime... We punish accomplices derivatively, however, because we do believe that an accomplice's behavior does or potentially does causally contribute, thus undermining a fully libertarian basis for criminal liability.'[350] Morse's comments misrepresent libertarianism. Libertarians do not deny that a person can be held responsible even if she was, in a sense 'caused' to commit her crime - if 'cause' simply means that another person encouraged her to do the crime, or provided her with the means of doing it. Nor does the law view the accomplice's input as a deterministic (or near deterministic) cause of the perpetrator's behaviour. The accomplice's input is not, by itself, seen as being anything approaching causally *sufficient* to ensure that the perpetrator's behaviour occurred. On the contrary, it is more plausible that the law assumes that the agent's exercise of free will was an important necessary condition for her criminal behaviour. Therefore, accomplice liability is perfectly consistent with libertarianism as it does not in any way challenge the view that determinism (or near determinism) is incompatible with criminal responsibility.

Morse repeats the same mistaken strategy when listing other cases which he claims refutes causal theory. He writes 'Consider the following examples, which demonstrate the implausibility of the simple causal theory.... Assume that a writer is working at her desk by a window as the sunset approaches. When the natural light becomes insufficient to continue working she turns on the desk lamp. According to any coherent account of causation, her turning on the light was caused primarily by

[350] M Morse, Reason, Results and Criminal Responsibility (2004) Illinois Law Review 363, p436.

her perception of the increasing darkness. Now take an example of an internal, psychological cause for the behaviour. Suppose the same writer works straight through the usual dinner hour. Later that evening she notices she is very hungry and eats something. Her eating is clearly caused. The writer is caused to turn on the lights and caused to eat, but there seems no reason to excuse her from responsibility for her acts in either case.'[351]

These attempts to refute causal theory are unsuccessful as they merely describe situations where someone is held responsible for an action despite the presence of a single 'but for' cause of the action which was outwith the agent's control. True, she might not have eaten when she did but for her hunger pangs, or turned on the lamp when she did but for the growing darkness. However, it is hardly obvious that the darkness guaranteed that the agent would put on the desk lamp rather than, say, going to bed, or that the hunger pangs rendered it inevitable that the agent would eat then, rather than deciding to diet. No libertarian or causal theorist would take the absurd position that a person should be excused just because she did not have complete control over every single 'but for' cause of her action. What compatibilists like Morse need is an example where it is obvious that a person's action was produced by a set of causally *sufficient conditions* (or something approaching very close to it) that was wholly outwith the agent's control and where it is equally obvious that the agent is responsible in the retributive sense. They have not yet found such an example.

Consider also Michael Moore's attempts to refute causal theory, which exhibit the same flaws as Morse's: 'Merely because behavior is caused does not mean that the

[351] S Morse, 'Psychology, Determinism and Legal Responsibility' in G Melton (ed) *The Law as a Behavioural Instrument* (University of Nebraska Press, Nebraska), p48.

law will ....excuse it. Suppose, for example, I know that Z has a limited repertoire of jokes and that if reminded of one of them in a social setting, he will tell it. Suppose further that I trigger one of his known jokes with a paraphrase of its first line. The responsibility for telling the bad joke is still Z's (even if it is also mine)....Causation is equally irrelevant to other proposed ways of negating voluntary action. Suppose high correlations are found between crime and certain environmental factors, or between crime and an extra Y chromosome in some men, or between crime and premenstrual tension in some women. Suppose that further, it is established that a defendant would not have committed a certain crime but for one of these "criminogenic" factors. We can then say that the factor caused the crime. We still have said nothing relevant...'[352]

Moore even uses the term 'but for' and yet does not seem willing to acknowledge that the debate about determinism is not about mere 'but for' causes. He recognises that the sense of the word 'cause' he relies on might be criticised by the causal theorist as being irrelevant, but responds that such a critic is 'gerrymandering his concept of causation in an *ad hoc* manner so as to include only his examples and to exclude counterexamples like those given earlier'.[353] This response is very puzzling. The difference between a necessary, 'but for' cause of an event as opposed to conditions that were sufficient for the occurrence of that event is recognised in a wide variety of contexts. It is hardly a novel, far-fetched distinction plucked out of the air by some desperate causal theorist in a last attempt to save his theory. And yet it is surprising that such eminent writers repeatedly come up with examples like these.[354] It is as if they think that citing a large enough number of cases which only

[352] M Moore, 'Causation and the Excuses' (1985) 73(4) California Law Review 1091, p1115
[353] Ibid, p1134-1135.
[354] Other writers who also seem to rely on such examples include: S Pilsbury, 'The Meaning of Deserved Punishment: An Essay on Choice, Character, and Responsibility' (1992) 67 Indiana Law Journal 719, p729.

involve necessary conditions will eventually defeat an argument that is concerned with sufficient conditions. The philosopher, Daniel Dennett has suggested one possible explanation for why some writers adopt such unpromising strategies when arguing about free will and responsibility:

'Some of them may...be seduced by the following quite reasonable consideration: when we consider whether free will is an illusion or reality, we are looking into an abyss. What *seems* to confront us is a plunge into nihilism and despair. Our whole reason for living is jeopardized. What to do? If it is really as important as all that, perhaps what it would be rational to do is *blow more smoke*. Whatever you do, don't try to get clear about this! Don't let the cat out of the bag.'[355]

Many legal compatibilists are convinced that a scientific view of the world reveals that human actions are really unlikely to be 'free' in the indeterministic libertarian sense. If the law rests on libertarian intuitions, however, then our practices may need to be radically revised. This chapter aims to make it plausible that the law does rest on such intuitions (however scientifically dubious they may be). However, this thesis attempts to show that a revision of our practices in the light of the free will problem should not be such a terrifying prospect.

Counterexamples against Compatibilism

Counterexamples against compatibilism tend to resemble the 'Professor Plum' case, discussed in Chapter 2.[356] These cases seem genuinely to involve agents who possess compatibilist freedom but who lack incompatibilist freedom. Many people find it intuitive that such agents are non-responsible. In order to ensure that these hypothetical agents truly possess compatibilist freedom, the examples involve

[355]D Dennett, 'Some Observations on the Psychology of Thinking about Free Will', in J Baer et al, *Are We Free? Psychology and Free Will* (OUP, Oxford 2008)
[356] D Pereboom, *Living without Free Will* (CUP, Cambridge 2001)

unrealistic science-fiction scenarios (e.g. mad scientists creating human beings). By the very nature of these cases, therefore, they are not the type of thing that comes up before the courts. How then could they shed light on whether the law is compatibilist or incompatibilist? One might speculate about whether, if Professor Plum were put on trial for his crimes, he would be entitled to an excuse under existing law. The answer seems to be no. But it is far from clear that this reveals the law to be compatibilist. It is important to ask *why* Professor Plum would be denied an excuse.

If the causal theory of excuses is correct then some of our current excuses represent instances where the law recognises that human behaviour is subjected to deterministic or near-deterministic pressures. There is clear empirical evidence that certain categories of behaviour (such as reflex responses) constitute such pressures. There is obviously no evidence that anyone is actually created in exactly the way that Professor Plum was created. So it is not surprising that the law has not developed a category of excuses to cover him. Similarly, science has not yet fully revealed the mechanisms by which ordinary peoples' genetics and environment shape (or determine) their characters and choices. In the absence of such evidence, the law may simply presuppose that, as an *empirical* fact about the world, people who do not fall into the traditional excusatory categories were not predetermined to commit their crimes by factors outwith their control. This does not show that the law presupposes that it is a *moral* truth that determinism is compatible with criminal responsibility and liability to retributive punishment.

Although Professor Plum examples do not provide the kind of direct help to the causal theorist that real life cases would provide, they are of indirect assistance. To the extent that ordinary people find it intuitive to excuse Professor Plum, this suggests that ordinary people are incompatibilists. As argued above, if most ordinary people within the legal system are incompatibilists, then it is more plausible that the

law endorses the principle that criminal responsibility and determinism are incompatible, rather than the competing compatibilist principle.

Whether actors within the legal system are compatibilists or incompatibilists is an empirical question. A number of studies have been carried out to try to shed light on the 'folk's' (i.e. non-philosophers') beliefs about free will. Unfortunately these studies have suffered from a number of serious flaws.[357] These flaws include a lack of clarity as to the term 'moral responsibility'. The way the questions are framed do not make it clear whether 'moral responsibility' is being used in the 'moral appraisability' sense or in the retributive sense.

Conclusion

This chapter has argued that the dominant view among criminal law theorists - that the criminal law is thoroughly compatibilist - rests on dubious arguments. There is no clear legal rule that explicitly states whether the criminal law is compatibilist or incompatibilist. Furthermore, there is no more reason for supposing that the law rests on compatibilist *principles,* than that it relies on libertarian principles. In fact the latter supposition may be more plausible than the former. Criminal law theorists seem unable to come up with genuine counterexamples against incompatibilism – i.e. examples where it is intuitive to hold people legally responsible, despite their criminal behaviour being pre-determined by factors outwith their control. Instead, purported counterexamples against incompatibilism miss their target, because they involve 'but for' causes rather than *deterministic* causes. This tendency among compatibilist criminal law theorists to equivocate over the word 'cause' instead of producing examples that that clearly involve determinism, suggests the difficulty of finding examples that elicit the intuition among ordinary readers that determinism

[357] For an excellent analysis of these shortcomings see: T Sommers, 'Experimental Philosophy and Free Will' (2010) 5 Philosophy Compass 199–212. doi: 10.1111/j.1747-9991.2009.00273.x).

and legal responsibility are compatible. In contrast, thought-experiments like the one involving Professor Plum, genuinely do challenge compatibilist assumptions and the apparent intuitiveness of these thought-experiments suggests that ordinary people within the legal system may be libertarian incompatibilists and that the law may contain libertarian legal principles. This is an area that needs further empirical investigation.

Chapter Eleven: The Criminal Law without Retributivism

Introduction

Non-retributive approaches do not necessarily imply that we should abolish criminal trials and stop holding law-breakers criminally responsible for their actions. There are forward-looking justifications for these practices. As noted above, moral dialogue is an important means of enabling sane offenders to reform themselves. The trial process can serve as a vivid form of moral communication, which can help the offender to appreciate more fully the impact of her conduct on others and to resolve to change her behaviour.[358] It also shows respect for the offender's rationality and membership of the moral community to allow her to give an account of her conduct in court, before other members of the community.[359]

However, a non-retributive, hard incompatibilist approach would recommend alterations to various legal doctrines. This chapter will suggest several changes that such an approach would imply for a) our understanding of the provocation defence, self-defence and mental disorder defences; and b) for the overall structure of criminal defences.

Implications for the Rationale behind Certain Defences

Provocation

According to the currently accepted rationale for the provocation defence, the person who kills in response to provocation is partially *excused*, but in order for the defence to succeed the anger that motivated the killing must have been *justified*. If hard incompatibilism is correct then is anger in response to someone's provocative

[358] See e.g. R Duff, *Punishment, Communication and Community* (OUP, Oxford 2001).
[359] See Duff (2001).

behaviour ever truly justified? 'Anger' in the sense of 'retributive outrage' cannot be justified. However, other similar emotional states can be. If, for instance, a person is physically attacked, it is appropriate for the person to feel very upset (as a person might not feel if she were hurt by a non-moral being e.g. an animal). In addition, it would be appropriate to feel a sense of repulsion at the moral 'defect' that led to the attacker's action, even if that defect was not something for which the attacker was retributively responsible. It is also appropriate to feel a strong sense of disapproval that a moral wrong has been committed. (I have argued that the concepts of moral 'right' and 'wrong' are compatible with determinism, even though the concept of 'moral responsibility' construed in the retributive sense is not.) The provocation defence should require that the above types of emotion are justified, but not that retributive anger is justified.

This approach might have the additional advantage of reducing the risk that the provocation defence is seen as 'blaming' the victim. Retributive anger is only appropriate if the victim is blameworthy. So saying that the provoked person's anger is justified implies that the victim deserved blame. However, the negative emotions that are justifiable on the hard incompatibilist view do not imply that the victim was retributively blameworthy.

A hard incompatibilist provocation defence would not entail that offenders should be excused merely because their offence was due to the effect of intense emotions, because it would still be necessary that the circumstances warranted a strong emotional reaction of the kind described above.

Self-defence

Self-defence theorists are divided over whether it is ever permissible for an attacked person to harm/kill an 'innocent aggressor'. An aggressor may be considered innocent if, for instance, he was psychotic, and therefore not morally responsible at

the time of the attack. Many theorists maintain that the self-defence justification should cover both innocent and responsible offenders alike. However, a few argue that harming/ killing an innocent aggressor is never justified, although it may sometimes be excusable. A third, 'compromise' view holds that harming/killing an innocent aggressor is only justified under much more restrictive circumstances than when the aggressor is morally responsible.

Self-defence theorists who consider the 'innocence' or 'culpability' of the aggressor to be morally relevant, often appeal to the following considerations. They may argue that innocent aggressors had the *bad luck,* to be subject to circumstances *outwith their control* which *caused* them to attack. They may argue that we should therefore have particular *compassion* for innocent aggressors, which should be reflected in the legal definition of self-defence. In contrast, the culpable aggressor *chose* to initiate the attack. It may be argued that it is therefore justifiable to use harmful or lethal force against the culpable aggressor, because his moral guilt entails that his interests are now less valuable than the interests of the attacked person, or because the culpable aggressor 'forfeited' his right to life/bodily integrity as a result of his immoral actions.

This type of account has a distinctively retributive flavour.[360] Its judgement about the culpable aggressor is strongly reminiscent of PF Strawson's description of the attitude of retributive indignation – an attitude which entails a 'partial withdrawal of goodwill' towards the wrongdoer and which entails a 'modification...of the general demand that another should, if possible, be spared suffering'.[361] The hard incompatibilist could not endorse this type of reasoning, but would insist that *all*

[360]It is, however, not *exactly* the same as the retributive justification for punishment. Even theorists who place a lot of weight on the aggressor's 'guilt' still accept that the use of self-defensive force against an aggressor is justified in order to *prevent* harm to the attacked person, not in order to *punish* the aggressor for his attack.
[361] P Strawson, 'Freedom and Resentment' (1962) 48 *Proceedings of the British Academy*, 187.

aggressors are ultimately victims of bad luck, even if the aggressor's behaviour flowed from his 'choice' to do wrong. For that choice was either completely determined by factors wholly outwith the aggressor's control, or else it was due to a random occurrence for which the aggressor was not responsible.

Would a hard incompatibilist theory of self-defence treat all aggressors equally, regardless of whether their attack was the product of psychosis or a rational choice? Not necessarily. A hard incompatibilist might distinguish between these different types of aggressor *if* there were non-retributive (e.g. consequentialist) reasons for doing so. However, this distinction must not be based on the idea that one type of aggressor is merely unlucky and so worthy of special legal protection, whereas another type is undeserving of compassion and legal protection, purely because of the nature of his act of aggression.

If all aggressors are unlucky, does this mean that killing/harming an aggressor in self-defence is only ever excusable and never justified? Hard incompatibilism does not necessarily entail this conclusion either. Many self-defence theorists argue that self-defence can be a justification regardless of whether the aggressor's behaviour was the result of circumstances outwith his control. This position accords with widespread intuitions. 'Common sense' seems to tell us that (all other things being equal) people are entitled to defend themselves against an attack which poses an immediate threat of significant physical harm or death, regardless of whether the attacker made himself into the kind of person who would attack people, or was just unlucky to have turned out that way. It seems absurd to suggest that, when faced with an aggressor who is in some sense 'innocent'[362], attacked people have a 'duty of

[362] (because, e.g., he lacked either libertarian or compatibilist free will)

martyrdom' and should meekly submit to the attack, rather than harm the aggressor.[363]

However, the hard incompatibilist's arguments should make us reconsider what precisely is meant by saying that self-defence is a 'justification' defence. On a 'modest' interpretation, saying that it is justified to use force to defend oneself, means that it is morally and legally *permissible* to use force; that, all things considered, the use of force is acceptable. Theorists who adopt this interpretation often emphasise that harming/killing someone is always regrettable, but that in extreme circumstances it can be the least bad option. In contrast, Boaz Sangero argues that self-defence should be portrayed as a 'real and strong' justification. He writes '...private defence should not be viewed as evil, and not even as the lesser evil, but as the "best possible good." It concerns a desirable action...'.[364] Having characterised self-defence in this way, Sangero then argues that it should be reserved for the killing/harming of 'guilty' aggressors, and that it shows a 'lack of compassion' to allow self-defence to cover the killing/harming of 'innocent' aggressors, such as a children, or psychotic individuals. Now, it does indeed seem somewhat heartless to characterise harming/killing child-aggressors or psychotic-aggressors as positively 'desirable', or to appear to welcome their injury/death by calling it the 'best possible good'. Nevertheless, it can still be appropriate to say that harming/killing such aggressors is 'justified' in the more modest sense described above. The modest account of justification in terms of an all-things-considered judgement of permissibility better captures the moral complexity of this kind of situation than Sangero's description of a justified action as something which is unequivocally 'desirable'. Given the considerations advanced by the hard

[363] See W Kaufman, 'Self-Defense, Innocent Aggressors and the Duty of Martyrdom' (2010) 91 *Pacific Philosophical Quarterly* 78.
[364] B Sangero, 'In Defense of *Self-Defence in Criminal Law*; and on *Killing in Self-Defence* – A Reply to Fiona Leverick' (2008) 44 (6) Criminal Law Bulletin 3, p17.

incompatibilist, it seems inappropriate to label the killing/injury of *any* aggressor as 'justified' in Sangero's sense of the word. All aggressors are victims of bad luck and are entitled to compassion. Maintaining that harming/killing another in self-defence is only justified as 'the lesser of two evils' acknowledges that every life is valuable and that causing death or injury is always a terrible thing (even when justified).

Mental disorder

The forward-looking approach to dealing with criminal behaviour which is defended in this thesis provides a method for determining the scope of mental disorder defences. The state should aim (among other things) to communicate to offenders that that they have committed a serious wrong, to make clear the reasons why the behaviour was wrong and to enter into dialogue with the offender with the aim of reforming him. This suggests that offenders who are incapable of engaging in genuine dialogue or undergoing reform, should not be held responsible. Other theories of punishment often leave the question of the scope of mental disorder defences largely down to intuition. E.g. some people have the intuition that an offender 'deserves' punishment just as long as he 'knew' that his behaviour was wrong in the sense of being aware what the legal or moral rules are, without any depth of understanding. Others have the intuition that a person is only retributively blameworthy if he knew the difference between right and wrong in a deeper sense. A forward-looking communication theory can help to settle this dispute, by interpreting 'knowledge' in terms of the ability to undergo moral dialogue. This is not an advantage that is *unique* to a hard-incompatibilist communication theory. Duff's retributive theory, for instance, also has this advantage. Nevertheless it is still a merit of the approach advanced in this thesis.

Implications for the structure of defences

Currently, defences are organised according to the cause of the accused's behaviour. Coercion involves behaviour was caused by a another's threats, necessity involves behaviour that was caused by a non-human threat, self-defence involves behaviour that was caused by fear of an attack, insanity involves behaviour caused by a mental illness, automatism involves behaviour caused by an 'external factor' (such as a spiked drink); provocation involves behaviour caused by anger at another's provocative behaviour. Wider categories of justification, excuse, lack of capacity etc. can be imposed on top of this scheme (by academic lawyers).

The current organisation of defences makes sense (up to a point) if one adopts a libertarian perspective. According to libertarianism, it is unfair to punish a person whose behaviour was caused by factors outwith the person's control. So categorising the defences according to the cause of the accused's behaviour does draw our attention to something that the libertarian claims to be morally relevant.

In contrast, if compatibilists or hard incompatibilists had designed the law then the key factor to be emphasised would not be the causes of the accused's behaviour (since compatibilists believe that causal determinism should provide no basis for a defence). Rather, compatibilists and hard incompatibilists would wish to focus our attention on whether the accused lacked morally relevant capacities and on which kinds of capacity were impaired (e.g. control, or understanding); or alternatively whether the accused was justified.

I therefore propose that criminal defences should be restructured in a way that emphasises factors that would be morally relevant irrespective of the truth of determinism. If the best rationale for criminal defences concerns issues like 'capacity' and 'justification', rather than causal pressures, then the structure of these defences should reflect this. This is necessary in order for these defences to be

rational and transparent. The law should not appear to rest on a dubious, libertarian conception of free will.

There are various different ways in which defences could be categorised. An in-depth treatment of this issue is outwith the scope of this thesis. I merely aim to put forward one option that is compatible with the truth of determinism and which is more rational than the current approach of categorising defences according to causes of behaviour.

Existing defences could be replaced with the following categories: 'Justification' (subdivided into different grounds for justification); 'unjustified behaviour by a non-culpable agent' (subdivided into lack of appreciation and lack of control, regardless of what caused this); a partial defence of 'unjustified behaviour by a less than fully culpable agent' (including impaired appreciation or diminished control); and 'The state is barred from punishing' (e.g. entrapment, or time bar).

'Defences' based on the absence of *actus reus* or *mens rea* would remain unaltered, except for automatism which would fall into the category of 'unjustified behaviour by a non-culpable agent'. By 'non-culpable' I mean that the action does not indicate that the agent had a serious moral defect. This would allow the court to order that the individual should be treated (if necessary), rather than simply allowed to go free. The current distinction between someone who breaks a law while severely mentally disordered and someone who breaks a law while in an automatistic state (although philosophically interesting) does not seem to be relevant to how the state should respond to such behaviour. In both cases, the behaviour fails to be a genuine response to the agent's reasons. This is what matters when it comes to deciding between a response that involves an element of moral communication on the one hand; and a non-condemnatory disposal (e.g. treatment or simply releasing the person) on the other.

Advantages of this Approach

My proposed approach focuses the court's attention on issues that are relevant to how we should respond to the offender. If the offender knowingly did wrong and was in control, then attempts to reform the offender can be useful and appropriate. Convicting an offender under these circumstances would send out the message that this sort of conduct should not be done. In contrast, if the accused lacked understanding or control, then attempts at reform are unnecessary and pointless. Acquitting someone based on my proposed 'non-culpability' defence would send out the appropriate message - that while this behaviour is unjustified, the actor is not in need of reform. In contrast, basing excuses on whether the person's behaviour was caused would fail to distinguish between different law-breakers since causal determinism may well be true of everyone.

The capacity-based approach also avoids the complex problems involved in distinguishing automatism from insanity. It avoids irrelevant disputes about whether something counts as a 'mental disorder' or whether behaviour was caused by an 'external' or an 'internal' source. It just looks at what is morally relevant – capacities.

Distinguishing clearly between a 'non-culpability' defence and a 'justification' defence would communicate to citizens something that they really need to know – whether or not the type of conduct that the accused engaged in was justified (i.e. permissible). This is necessary in order for the law properly to fulfil its action-guiding function.

The traditional approach focused a lot of attention on whether behaviour was due to a mental disorder, or whether it was due to ordinary human nature. In contrast, a

capacity-based approach does not categorise defences according to whether the offender's behaviour was caused by 'mental disorder'. Behaviour that is not reasons-responsive falls into the 'unjustified conduct, non-culpable agent' category regardless of whether it was caused by a 'normal' reaction, such as fear in response to a threat, or by a mental disorder. This could also help to reduce the stigma associated with mental disorder defences.

My approach does not draw the excuse/exemption distinction advocated by some theorists. This distinction purports to draw a line between non-agents who, due to their general lack of capacity, are outside the moral community; and persons who merely lack understanding or control on a particular occasion. While this might be a valid distinction, it is not one that the law should announce. Why flag up the fact that a person is not a member of the moral community? This seems cruel, especially with regard to people who will *never* be a full member of the moral community. Even regarding children, it is unhelpful because it reinforces the prejudice that children cannot have moral insights. Furthermore, publically declaring that children are not full members of the moral community does not help to shape children's behaviour in a positive way.

My approach also helps to avoid justificatory drift, e.g. in the context of provocation. Instead of merely announcing that the accused received a lesser sentence and was convicted of a lesser offence because he was provoked (which people might interpret as a justification), the court would explicitly state that the accused's action was unjustified but that he was not fully culpable, because his capacities were temporarily impaired. Furthermore, the court would not be forced into the position of saying that battered women have an 'abnormality of mind' (which is currently a defining feature of diminished responsibility). Rather, my proposed partial defence is

compatible with the idea that a basically normal person's capacities can be understandably impaired under extreme and abnormal circumstances. The criminal actions of a battered woman would also be considered unjustified, but she would not be fully culpable.

Categorising defences according to the causes of behaviour makes the law excessively inflexible. A person who lacks the morally relevant capacities can be denied a defence just because the factors that caused the person to lack these capacities do not fall into any of the existing legally recognised categories. For instance, it is arguable that in certain (rare cases) an individual from a culture whose norms are extremely different from the norms of this society might not have been able to appreciate what is wrong with committing a particular crime, at the time he committed it. Or perhaps someone who never received any kind of moral education at all may have been unable to appreciate what was wrong with committing the crime she committed. Such individuals would not be able to rely on existing defences because this kind of incapacity must be caused by mental disorder. A capacity-based approach would not deny these individuals a defence merely because their lack of capacity was not caused by the 'right' type of factor.

Conclusion

In this chapter I argued that a non-retributive, hard incompatibilist approach would recommend alterations to various legal doctrines. This chapter suggested several changes that such an approach would imply for a) our understanding of the provocation defence, self-defence and mental disorder defences; and b) for the overall structure of criminal defences.

Conclusion

In Part One of this thesis I argued that there is at least a reasonable doubt about the soundness of retributivism (which is probably the dominant theory of punishment among contemporary criminal justice theorists). This doubt arises from retributivism's reliance on a hotly contested conception of free will. Chapter One provided reasons for doubting that retributivism could justifiably rely on the assumption of libertarian free will. Chapter Two provided reasons for doubting the adequacy of compatibilist retributivism. Chapter Three responded to an argument for retributivism, based on the implications of that theory for our practices: the argument that we need retributivism because it is the only theory that implies that accused people and offenders should be protected by considerations of justice. I argued that in fact there are good reasons for thinking that our practice of upholding these principles of justice could be defended on non-retributive grounds.

In Part Two I related the doubts about the soundness of retributivism that I raised in Part One to the literature on moral uncertainty (i.e. uncertainty about which moral theory should guide our conduct). I pointed out that, in the light of this literature, adopting a purely consequentialist theory of punishment would not be a rational response to the doubts about the soundness of retributivism since there is also uncertainty about the soundness of this consequentialism. Chapter Four presented an overview of some of the main theories of moral uncertainty and highlighted some of their key shortcomings (at least in relation to dealing with uncertainty about theories of punishment). Chapters Five and Six defended my own approach to moral uncertainty about theories of punishment. On my approach, the entire moral argument for punishing a person should be held to a high standard of credibility. I called this the 'cautious approach to punishment'. In Chapter Five I argued that one reason for adopting a cautious approach to punishment stems from the underlying

rationale for the beyond reasonable doubt standard in criminal trials – a standard that has widespread support. In Chapter Six, I argued that in order to minimise the risk of punishing someone unjustifiably, we should only punish that person if the main punishment theories agree that doing so is justifiable. I called this 'the convergence requirement'. If the convergence requirement is satisfied, then the state's argument for punishing a person has met the required standard of credibility. I acknowledged in Chapter Six that, notwithstanding the arguments against retributivism presented in Part One, when we survey the state of the free will debate, there is a strong argument for 'free will agnosticism' rather than for *certainty that we lack* free will in the sense required for retributive responsibility.[365] Given that the cautious approach recommends giving people who may be liable to punishment 'the benefit of the doubt', such people should only be punished if retributive theories (as well as forward-looking theories) would recommend this; i.e. in effect this means that punishment should be constrained by negative retributivism. The possibility that negative retributivism might be sound therefore provides a reason against such intuitively unjust practices such as punishment of the innocent – a reason in addition to those non-retributive reasons outlined in Chapter Three. Chapter Six also outlined reasons why theorists from different philosophical perspectives should endorse the convergence requirement and defended this requirement against certain potential objections.

The convergence requirement assumes that punishment infringes the interests of the individuals who are punished – in particular their interest in not being seriously harmed - and therefore requires strong justification. Part Two focussed primarily on which individuals should be punished *at all*. Part Three addressed the question of which method of responding to an individual's criminal behaviour should be

[365] S. Kearns, 'Free Will Agnosticism' (2013) Nous (Online First). DOI: 10.1111/nous.12032

preferred, once it is determined that some response is required, in situations where several possible alternatives are available. This question is complicated by the fact that offenders have a number of different interests that should be protected, but these interests can sometimes pull in different directions. In general, offenders have an interest in not being deprived of liberty. Therefore, where one mainstream theory recommends a sentence that involves less interference with liberty, that sentence should usually be preferred. However, I have also indicated that offenders have an interest in being treated as rational agents. Certain rehabilitative interventions might allow offenders to be released earlier into society, but may fail to respect the offender as a rational agent. This problem arises most acutely in connection with the possibility of using direct brain interventions to modify criminal behaviour. Therefore, Part Three was devoted to examining this example in detail. It is also important to explore this topic, since it is of particular relevance to the arguments for free will scepticism which I advanced in Part One, when raising doubts about the justifiability of relying solely on a retributive theory punishment. It is tempting to try to explain why certain direct brain interventions are troubling by claiming that such interventions threaten free will. If free will scepticism implied an acceptance of these troubling interventions then this could undermine free will scepticism and could possibly strengthen the case for a retributive system that stressed the importance of free will. However, in Part Three I argued that, in fact, the objectionable nature of certain direct brain interventions has very little to do with free will and that there are in fact non-retributive reasons for opposing the most intuitively-objectionable interventions. Chapter Seven identified certain forms of biomedical interventions that are objectionable and should not be used within the criminal justice system. Chapter Eight argued that, *in principle,* it would be morally permissible for the state to employ certain types of biomedical interventions (such as 'cognitive enhancements') in a limited way within the criminal justice system, provided that effective

enhancements can be developed in the future that have minimal side-effects. Chapter Nine considered how we can distinguish interventions that enhance rational capacities from interventions that fundamentally change the person's character, and the extent to which this distinction matters.

In Part Four, I discussed some of the practical implications of my approach for the criminal law. Chapter 10 examined the criminal law's current position on the questions of free will and retributive responsibility. It argued that the dominant view among criminal law theorists - that the criminal law is thoroughly compatibilist - rests on dubious arguments. There is at least as much reason for thinking that principles of criminal law embody libertarian, incompatibilist assumptions as there is for thinking that they make only compatibilist assumptions. Chapter 11 examined what revisions to criminal law doctrines would enable these doctrines to be justifiable even if retributivism (and the notion of free will on which it depends) were regarded as unsound. It suggested several changes to our *understanding* of the rationale for the provocation defence, self-defence and mental disorder defences; and for the overall *structure* of criminal defences. My intention in Part Four was to provide a general overview of directions for future work that would need to be done in order to reconcile criminal law doctrines with hard incompatibilism and not to give a comprehensive account of the criminal law doctrines discussed.

Bibliography

A

- Adebowale V, 'Diversion Not Detention' (2010) 17 (2) Public Policy Research 71.
- Allen R and Laudan L, 'Deadly Dilemmas' 41 (2009) Texas Tech Law Review 65.
- Alper, J 'Genes, Free Will and Criminal Responsibility' (1998) 46 (12) Social Science and Medicine 1599.
- Arpaly N, *Unprincipled Virtue* (OUP, Oxford 2003).

B

- Barbaree H, Blanchard R, and Langton C, 'The Development of Sexual Aggression Through The Life Span. The Effect Of Age On Sexual Arousal And Recidivism Among Sex Offenders', (2003) 989 Annals of the New York Academy of Sciences 59-71.
- Bentham J, *Principles of Penal Law* (ebooks@adelaide, Adelaide, South Australia 2011), available at: http://ebooks.adelaide.edu.au/b/bentham/jeremy/principles_of_penal_law/. Accessed 26 June 2012.
- Berofsky B (ed), *Free Will and Determinism* (Harper & Row, New York 1966).
- Bierie, D, 'Is Tougher Better? The Impact of Physical Prison Conditions on Inmate Violence' (2012) 56 International Journal of Offender Therapy and Comparative Criminology 338.
- Blair et al, *The Psychopath: Emotion and the Brain* (Blackwell, Oxford 2005).
- Bolt I and Scherner M, 'Psychopharmaceutical Enhancers: Enhancing Identity?' (2009) 2 Neuroethics 103-111.
- Bomann-Larsen L, 'Voluntary Rehabilitation? On Neurotechnological Behavioural Treatment, Valid Consent and (In)appropriate Offers' (2011) Neuroethics (Online First) doi:10.1007/s1215201191059.
- Bourget D and Chalmers D (eds.) *The Philpapers Survey 2009*, available at http://philpapers.org/surveys/results.pl Accessed 31st May 2011.
- Bublitz, J C and Merkel R, 'Autonomy and Authenticity of Enhanced Personality Traits' (2009) 23(6) Bioethics 360-74.
- Burns, J M and Swerdlow R H, 'Right Orbitofrontal Tumor with Pedophilia Symptom and Constructional Apraxia Sign' (2003) 60 Archives of Neurology 437-440.

C

- Campbell C, *In Defence of Free Will* (Allen & Unwin, London 1967).
- Carlsmith K, 'The Roles of Retribution and Utility in Determining Punishment' (2006) 42 Journal of Experimental Social Psychology 437.

- Clarke R, 'Libertarian Views: Critical Survey of Noncausal and Event-Causal Accounts of Free Agency', in R Kane (ed.), *The Oxford Handbook on Free Will* (OUP, Oxford 2002).
- Clarke R, 'Dispositions, Abilities to Act and Free Will: The New Dispositionalism' (2008) 118 Mind 323-351.
- Copp D, 'Defending the Principle of Alternate Possibilities: Blameworthiness and Moral Responsibility' (1997) 31 Nous 441.
- Corrado M, 'Why Do We Resist Hard Incompatibilism? Thoughts on Freedom and Punishment' in T Nadelhoffer (ed.) *The Future of Punishment, ed. Thomas Nadelhoffer* (Oxford OUP 2013)
- Corrado M, 'Notes on the Structure of a Theory of Excuses' (1992) 82(3) The Journal of Criminal Law and Criminology 465.
- Corrado M, 'Addiction and Causation' (2000) 37 San Diego Law Review 913.
- Corrado M, 'Responsibility and Control' (2005) 34 Hofstra Law Review 59.
- Cotton M, 'A Foolish Consistency: Keeping Determinism Out of the Criminal Law' (2005) 15 Public Interest Law Journal 1.
- Crocker L, 'Ethics and the Law's Burden of Proof' 18 (1) (2008) Philosophical Issues 272.
- Crockett M et al, 'Serotonin Selectively Influences Moral Judgment and Behavior Through Effects on Harm Aversion' (2010) 107 (40) Psychological and Cognitive Sciences 17433.
- Crouch W, *Moral Uncertainty and Intertheoretic Comparisons of Value* (unpublished dissertation).
- Cunningham W et al, 'Separable Neural Components In The Processing Of Black And White Faces' (2004) 15 Psychological Science 806.

D

- Damasio A, *Descartes's Error: Emotion, Reason and the Human Brain* (Putnam, New York 1994).
- Dennett D, *Elbow Room: The Varieties of Free Will Worth Wanting* (OUP, Oxford 1984).
- Dennett D, 'My Brain Made Me Do It. (When Neuroscientists Think They Can Do Philosophy)' (2011) European University Institute: Max Weber Lecture Series, no. 2011/1, 1.
- Double R, 'The Moral Hardness of Libertarians' (2002) 5 (2) Philo 226.
- Douglas T, 'Moral Enhancement' (2008) 25 (3) Journal of Applied Philosophy 228.
- Dressler J, 'Reflections on Excusing Wrongdoers: Moral Theory, New Excuses and The Model Penal Code' (1987) 19 Rutgers Law Journal 671
- Duff RA, *Trials and Punishments* (CUP, Cambridge 1986).
- Duff RA, *Punishment, Communication and Community* (OUP, Oxford University Press 2001).

F

- Fara M, Masked Abilities and Compatibilism (2008) 117 Mind, 843.

- Farah M, 'Emerging Ethical Issues in Neuroscience' (2004) 5(11) Nature Neuroscience 1123.
- Ferari P et al, 'Escalated Aggressive Behavior: Dopamine, Serotonin and GABA', (2005) 526 European Journal of Pharmacology 51.
- Fischer J, 'The Cards that are Dealt You' (2006) 10 The Journal of Ethics 107.
- Fischer J, *My Way: Essays on Moral Responsibility* (OUP, Oxford 2006).
- Fischer J, *The Metaphysics of Free Will* (Blackwell, Oxford 1994).
- Fischer J and Ravizza M, *Responsibility and control: A theory of moral responsibility.* (CUP, Cambridge 1998).
- FitzPatrick W, 'The Doctrine of Double Effect: Intention and Permissibility' 7 (3) (2012) Philosophy Compass 183.
- FitzPatrick W, 'The Intend Foresee Distinction and the Problem of 'Closeness' 128 (2006) Philosophical Studies 585.
- Foot P, 'The Problem of Abortion and the Doctrine of Double Effect' in P Foot, *Virtues and Vices and Other Essays in Moral Philosophy* (OUP, Oxford 2002) 19.
- Frankfurt, H, 'Alternate Possibilities and Moral Responsibility' (1969) 66 Journal of Philosophy 829.
- Freedman C, 'Aspirin for the Mind? Some Ethical Worries about Psychopharmacology', in: Parens E (ed) *Enhancing Human Traits: Ethical and Social Implications* (Georgetown University Press, Washington DC 2000) 135.

G

- Gardner J, 'Criminals in Uniform' in R.A. Duff, Lindsay Farmer, S.E. Marshall, Massimo Renzo, and Victor Tadros (eds), *The Constitution of Criminal Law* (OUP, Oxford 2013) (forthcoming).
- Ginet C 'In Defence of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing' (1996) 10 Philosophical Perspectives 404.
- Greely H, 'Direct Brain Interventions to 'Treat' Disfavored Human Behaviors: Ethical and Social Issues' (2012) 91(2) Clinical Pharmacology & Therapy 163.
- Guerrero A, 'Don't Know, Don't Kill: Moral Ignorance, Culpability and Caution' (2007) 136 (1) Philosophical Studies 59.

H

- Haji I and Cuypers S, 'Magical Agents, Global Induction and the Internalism/Externalism Debate' (2007) 85 Australasian Journal of Philosophy 343.
- Halvorsen V, 'Is it Better that Ten Guilty Persons Go Free than that One Innocent Person be Convicted?' 23 (2004) Criminal Justice Ethics 3.
- Harris, J 'Moral Enhancement and Freedom' (2011) 25 (2) Bioethics 102.
- Harrison, G 'Hooray! We're Not Morally Responsible!' (2009) 8 Think 87.
- Harrison G, 'A Challenge for Soft Line Replies to Manipulation Cases' (2010) 38 Philosophia 555-568.

- Hart A et al, 'Differential Response in the Human Amygdala to Racial Outgroup Vs. Ingroup Face Stimuli', (2000) 11 'Neuroreport: For Rapid Communication of Neuroscience Research' 2355.
- Hart H, *Punishment and Responsibility* (Clarendon Press, Oxford 1968).
- Ho H, 'The Presumption of Innocence as Human Right' in Roberts P (ed) *Criminal Evidence and Human Rights: Reimagining Common Law Procedural Traditions* (Hart Publishing, Oxford 2012), 259.
- Hodges A, *Alan Turing: The Enigma,* The Centenary Edition (Random House, London 2012).
- Honderich T, *A Theory of Determinism: The Mind, Neuroscience and Life-Hopes* (Clarendon Press, Oxford 1988).
- Honderich T, *How Free Are You?,* 2nd edn. (OUP, Oxford 2002).
- Honderich T, *On Determinism and Freedom* (Edinburgh University Press, Edinburgh 2005).
- Honderich T, *Punishment: The Supposed Justifications* (Penguin Books, Middlesex 1984).
- Howard-Snyder F, 'Doing vs. Allowing Harm', in *Stanford Encyclopaedia of Philosophy* (2011), available at http://plato.stanford.edu/entries/doing-allowing/
- Hudson J, 'Subjectivization in Ethics'(1989) 26 American Philosophical Quarterly 221.
- Husak D, 'The Costs to Criminal Theory of Supposing that Intentions are Irrelevant to Permissibility' 3 (2009) Criminal Law and Philosophy 51.

J

- Jones, M 'Overcoming the Myth of Free Will in the Criminal Law: The True Impact of the Genetic Revolution' (2003) 52 Duke Law Journal 1031.

K

- Kadish S, 'Excusing Crime' (1987) California Law Review 257.
- Kagan S, 'The Additive Fallacy' 99 (1988) Ethics 5.
- Kagan S, *The Limits of Morality* (OUP, Oxford 1989).
- Kamm F, *Morality, Mortality Volume II: Rights, Duties, and Status* (OUP, Oxford 2001).
- Kane R (ed), *The Oxford Handbook on Free Will* (OUP, Oxford 2002).
- Kane R, *The Significance of Free Will* (OUP, New York 1996).
- Kant I, Paton H (tr), *The Moral Law: Groundwork of The Metaphysic of Morals* (Routledge, London 1948).
- Kaye A, 'Resurrecting the Causal Theory of Excuses' (2005) 83 Nebraska Law Review 1116.
- Kendell R, 'The Distinction Between Personality Disorder and Mental Illness' (2002) 180 British Journal of Psychiatry 110.
- Kennett J, 'Do Psychopaths Really Threaten Moral Rationalism?' (2006) 9(1) philosophical Explorations 69.
- Kirkmeier J, 'A Tear in the Eye of the Law: Mitigating Factors and the Progression Toward a Disease Theory of Criminal Justice' (2004) 83 Oregan Law Review 632.
- Kruttschnitt C and Dirkzwager A, 'Are There Still Contrasts in Tolerance? Imprisonment in the Netherlands and England 20 Years Later' (2011) 13(3)

Punishment and Society 283-306.

L

- Lappi-Seppala, T and Tonry, M, 'Crime, Criminal Justice and Criminology in the Nordic Countries' (2011) 40 (1) Crime and Justice 1.
- Laudan L, 'The Rules of Trial, Political Morality, and the Costs of Error: Or, is Proof Beyond a Reasonable Doubt Doing More Harm than Good' in L Green and B Leiter (eds) *Oxford Studies in Philosophy of Law* (OUP, Oxford 2011).
- Lee Y, 'Deontology, Political Morality, and the State' 8 (2011) Ohio State Journal of Criminal Law 385.
- Levy, N, 'Skepticism and Sanction: The Benefits of Rejecting Moral Responsibility' (2012) 31 Law and Philosophy 477.
- Levy N, 'Why Frankfurt Examples Don't Beg the Question: A Reply to Woodward' (2004) 35 (2) Journal of Social Philosophy 211.
- Levy N, *Neuroethics: Challenges for the 21st Century* (CUP, Cambridge 2007).
- CS Lewis, 'The Humanitarian Theory of Punishment' (1953) 6 Res Judicatae 224.
- CS Lewis, 'On Obstinacy in Belief' (1955) 63(4) The Sewanee Review 525.
- Lilquist E, 'Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability' 36 (2003) University of California Davis Law Review 85.
- Litton P, 'The Abuse Excuse in Capital Sentencing Trials: Is it Relevant to Responsibility, Punishment or Neither?'(2005) 42 American Criminal Law Review 1027.
- Lockhart T, *Moral Uncertainty and its Consequences* (OUP, Oxford 2000).
- Loewy A, 'Taking Reasonable Doubt Seriously' 85 (1) (2010) Chicago-Kent Law Review 63.

M

- Mackie J, 'Morality and the Retributive Emotions' (1982) Criminal Justice Ethics 3.
- Maibom H, 'The Mad, The Bad And The Psychopath' (2008) 1 Neuroethics 167.
- Maletzky B and Field G, 'The Biological Treatment Of Dangerous Sexual Offenders, A Review And Preliminary Report Of The Oregon Pilot Depo-Provera Program.' (2008) 8 Aggression and Violent Behavior 391.
- Maltzky B, Tolan A and McFarland B, 'The Oregon depo-provera program: A five-year follow-up' (2006) 18 Sex Abuse 303.
- Matthews E, *Body-Subjects And Disordered Minds. Treating The Whole Person in Psychiatry* (OUP, Oxford 2007).
- McKenna M, 'A Hard-Line Reply to Pereboom's Four-Case Manipulation Argument' (2008) 77 Philosophy and Phenomenological Research 142.
- McKenna M, 'Book Review: Responsibility and Control: A Theory of Moral Responsibility, by John Martin Fischer and Mark Ravizza' (2001) 98 Journal of Philosophy 93.
- McKenna M, 'Alternative Possibilities and the Failure of the Counterexample Strategy' (1997) 28 Journal of Social Philosophy 71.

- McKenna M, 'Compatibilism', in *Stanford Encyclopaedia of Philosophy* (2009), available at http://plato.stanford.edu/entries/compatilism/.
- McKenna M, 'Compatibilism: The State of the Art', in *Stanford Encyclopaedia of Philosophy* (2009), available at http://plato.stanford.edu/entries/compatilismsupplement/.
- McMahan J, 'A Challenge to Common Sense Morality' (1998) 108 (2) Ethics 394.
- Mealey L, 'The Sociobiology of Sociopathy: An Integrated Evolutionary Model', in S Baron-Cohen (ed), *The Maladapted Mind: Classical Readings in Evolutionary Psychopathology* (1997 Psychology Press, East Sussex) 133.
- Mele A, *Free Will and Luck* (Oxford University Press, New York 2006).
- Ministry of Justice, Compendium of re-offending statistics and analysis (2010), available at http://www.justice.gov.uk/publications/docs/compendium-of-reoffending-statistics-and-analysis.pdf Accessed 2/08/11.
- Mitchell E, *Self-Made Madess* (Ashgate Publishers, Aldershot 2003).
- Moore G, *Ethics* (Williams and Norgate, London 1912).
- Moore M, *Placing Blame* (OUP, Oxford 1997).
- Moore M, 'Causation and the Excuses' (1985) 73(4) California Law Review 1091.
- Moore M, 'The Determinist Theory of Excuses' (1985) 95(4) Ethics 909.
- Moore M, 'Causation and the Excuses' (1985) 73 California Law Review 1091.
- Morse S, 'Culpability and Control' (1994) 142 University of Pennsylvania Law Review 1587.

N

- Nagin, D, 'Deterrence in the Twenty-First Century' (2013) 41(1) Crime and Justice 199.
- Newman J et al, 'Attention Moderates the Fearlessness of Psychopathic Offenders' (2010) 67 Biological Psychiatry 66.
- Norrie A, 'Freewill, Determinism and Criminal Justice' (1983) 3(1) Legal Studies 60.

O

- Oddie G, 'Moral Uncertainty and Human Embryo Experimentation' in K Fulford et al (eds), *Medicine and Moral Reasoning* (CUP, Cambridge 1994).
- Olsen M, Depression, 'SSRIs, and the Supposed Obligation to Suffer Mentally' (2006) 16 (3) Kennedy Institute of Ethics Journal 283.

P

- Penney S, 'Impulse Control and Criminal Responsibility: Lessons from Neuroscience' (2012) 35 International Journal Law and Psychiatry 99.

- Pereboom D, 'Reasons-Responsiveness, Alternative Possibililities, and Manipulation Arguments Against Compatibilism: Reflections on John Martin Fischer's *My Way*' (2006) 47 Philosophical Books 198.
- Pereboom D, *Living without Free Will* (CUP, Cambridge 2001).
- Pereboom D, 'Free Will, Evil, and Divine Providence', in Chignell A and Dole A (eds) *God and the Ethics of Belief: New Essays in Philosophy of Religion* (CUP, Cambridge 2005) 77.
- Phelps E et al, 'Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation' (2000) 12 Journal of Cognitive Neuroscience 729.

Q

- Quinn W, 'Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing' (1989) 98 (3) Philosophy and Public Affairs 287.
- Quinn W, 'Actions, Intentions, and Consequences: The Doctrine of Double Effect' (1989) 98 (4) Philosophy and Public Affairs 334.
- Quinsey V, Evolutionary Theory and Criminal Behaviour (2002) 7 (1) Legal and Criminological Psychology 1.

R

- Rachels J, 'Active and Passive Euthanasia' (1975) 292 New England Journal of Medicine 78.
- Reicher S, 'Saving Bulgaria's Jews: An Analysis of Social Identity and the Mobilisation of Social Solidarity' (2006) 36 European Journal of Social Psychology 49.
- Reid T, *Essays on the Active Powers of the Human Mind* (MIT Press, Cambridge, Massachusetts 1969, [1788]).
- Reiman J and Van Den Haag E, 'On the Common Saying that it is Better that Ten Guilty Persons Escape than that One Innocent Suffer: *Pro* and *Con'* 7(2) (1990) Social Philosophy and Policy 226.
- Rizzolli M and Saraceno M, 'Better That Ten Guilty Persons Escape: Punishment Costs Explain The Standard Of Evidence' (2011) Public Choice DOI: 10.1007/s11127-011-9867-y (online first).
- Ross J, 'Rejecting Ethical Deflationism' (2006) 116 Ethics 742.

S

- Sarkissian H et al, 'Is Belief in Free Will a Cultural Universal' (2010) 25(3) *Mind and Language* 346.
- Scanlon T, *Moral Dimensions: Permissibility, Meaning, Blame* (Harvard University Press, Cambridge, Massachusetts 2008).
- Sepielli A, 'What to Do When You Don't Know What to Do' in R Shafer-Landau (ed.), *Oxford Studies in Metaethics, Volume Four* (OUP, Oxford 2009) 5.
- Sepielli A, 'Review of Ted Lockhart's Moral Uncertainty and its Consequences' (2006) 116 Ethics 601.
- Shaw E, 'Psychopaths and Criminal Responsibility' (2009) 13(3) Edinburgh Law Review 497.
- Shaw E, 'Free Will, Punishment and Neurotechnologies' in van den Berg B and Klaming L (eds), *Technologies on the Stand: Legal and Ethical*

Questions in Neuroscience and Robotics (Wolf Legal Publishers, Nijmegen 2011) 177-194.
- Shaw E, 'Direct Brain interventions and Responsibility Enhancement' (2012) Criminal Law and Philosophy, DOI: 10.1007/s11572-012-9152-2 (online first).
- Shaw E, 'Cognitive Enhancement and Criminal Behaviour' in E Hildt and A Franke (eds), *Cognitive Enhancement: An Interdisciplinary Perspective* (Springer, Dordrecht 2013).
- Singer P, 'Philosophers are back on the job,' New York Times Sunday Magazine 7 July 1974 pp19-20.
- Smart J, 'Free Will, Praise and Blame' (1961) Mind 291.
- Smart J, *Philosophy and Scientific Realism* (London, Routledge 1963).
- Smilansky, S, 'Hard Determinism and Punishment: A Practical Reductio' (2011) 30 Law and Philosophy 353.
- Smilansky S, 'Compatibilism: The Argument from Shallowness' (2003) 115 Philosophical Studies 257.
- Smith M, 'Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion', in Stroud S and Tappolet C (eds), *Weakness of Will and Practical Irrationality* (OUP, New York 2003) 17.
- Sommers T, 'The Objective Attitude' (2007) 57 (228) The Philosophical Quarterly 321.
- Stern L, 'Freedom, blame, and moral community' (1974) 71 The Journal of Philosophy 72.
- Strawson G, *Freedom and Belief* (Clarendon Press, Oxford 1986).
- Strawson P, 'Freedom and Resentment' (1962) 48 Proceedings of the British Academy 187.
- Styron W, *Sophie's Choice* (The Modern Library, New York 1998).
- Sunstein C and Vermeule A, 'Is Capital Punishment Morally Required? Acts, Omissions and Life-Life Tradeoffs' (2005) 58 Stanford Law Review 703.

T

- Tadros V, 'The Ideal of the Presumption of Innocence', paper presented at Fraying the Golden Thread Conference, Aberdeen University (2012).
- Tadros V, *The Ends of Harm: The Moral Foundations of the Criminal Law* (OUP, Oxford 2011).
- Tadros V, *Criminal Responsibility* (OUP, Oxford 2005).
- Thomson J, 'Physician-Assisted Suicide: Two Moral Arguments' 109 (1999) Ethics 497.
- Tomlin P and Barry C, 'Uncertainty Permissibility and Compromise' (unpublished paper).
- Tomlin P, 'Could the Presumption of Innocence Protect the Guilty', paper presented at Fraying the Golden Thread Conference, Aberdeen University (2012)
- Tomlin P, 'Extending the Golden Thread? Criminalisation and the Presumption of Innocence' The Journal of Political Philosophy (Forthcoming).

V

- Vallier, K and D'Agostino, F, 'Public Justification' (2013) in *The Stanford Encyclopaedia of Philosophy*, available at: http://plato.stanford.edu/entries/justification-public/ [Accessed August 2013].
- Vihvelin K, 'Free Will Demystified: A Dispositional Account' (2004) 32 Philosophical Topics 427.
- Vilhauer B, 'Taking Free Will Skepticism Seriously' (2012) 62 The Philosophical Quarterly 833, p849.
- Vilhauer B, 'Free Will and Reasonable Doubt' (2009) 46 (2) American Philosophical Quarterly 131.
- Vilhauer B, 'Free Will Skepticism and Personhood as a Desert Base' (2009) 39 (3) Canadian Journal of Philosophy 489.
- Vincent N, 'Responsibility: Distinguishing Virtue from Capacity' (2009) 3(1) Polish Journal of Philosophy 111.
- Vincent N, 'Capacitarianism, responsibility and restored mental capacities' in van den Berg B and Klaming L (eds), *Technologies on the Stand: Legal and Ethical Questions in Neuroscience and Robotics* (Wolf Legal Publishers, Nijmegen 2011) 41-65.
- Vuoso G, 'Background, Responsibility, and Excuse' (1986) 96 Yale Law Journal 1661.

W

- Watson G (ed), *Free Will* (OUP, Oxford 1982).
- West R *The Fountain Overflows* (Virago Press, London 1984).
- Westen, P 'Getting the Fly out of the Bottle: The False Problem of Free Will and Determinism' (2004) 8 Buffalo Criminal Law Review 599.
- Westen P, 'An Attitudinal Theory of Excuse' (2006) 25 Law and Philosophy 289.
- Widerker D, 'Frankfurt's Attack on Alternative Possibilities' (2002)Z 14 Philosophical Perpectives 181.
- Woollard F, 'The Doctrine of Doing and Allowing I: Analysis of the Doing/Allowing Distinction' 7(7) (2012) Philosophy Compass 448.
- Woollard F, 'The Doctrine of Doing and Allowing II: The Moral Relevance of the Doing/Allowing Distinction'7 (7) (2012) Philosophy Compass 459.

Z

- Zimmerman, B *Living With Uncertainty* (CUP, Cambridge 2009).

Cases

A v HM Advocate 2003 S.L.T. 497
Erdemovic v Prosecutor, IT-96-22, October 7, 1997
James Gibson (1844) 2 Broun 332
McNaghten's Case (1843) 10 Cl & F 200
People v Leopold and Loeb, Cook County Crim. Ct. III [1924]
R v Hamilton, [2004] 186 C.C.C. (3d) 129

Thomson v HM Advocate 1983 JC 69.
Woolmington v DPP [1935] AC 462
Re Winship 397 U.S. 358 (1970), at 364.