Punishment and Personal Responsibility

GÖRAN DUUS-OTTERSTRÖM

DEPARTMENT OF POLITICAL SCIENCE
GÖTEBOURG UNIVERSITY
2007
Distribution

Göran Duus-Otterström
Department of Political Science
Göteborg University
P.O. Box 711
SE 405 30 Göteborg
Sweden
e-mail: goran.duus-otterstrom@pol.gu.se

Punishment and Personal Responsibility
Göran Duus-Otterström
ISBN: 978-91-89246-34-8
ISSN: 0346-5942
© 2007 Göran Duus-Otterström
Printed by Grafikerna Livrén AB, Kungälv 2007

This dissertation is included as number 108 in the series Gothenburg Studies in Politics, edited by Bo Rothstein, Department of Political Science, Göteborg University.
ACKNOWLEDGEMENTS.................................................................................................................1

PART I. THE PROBLEM

1. THE PRACTICE OF PUNISHMENT......................................................................................7
   1.1 INTRODUCTION........................................................................................................7
   1.2 PERSPECTIVES ON PUNISHMENT........................................................................11
   1.3 CLASH OF INTUITIONS........................................................................................15
   1.4 WHERE WE ARE TODAY: CONFLICTING TENDENCIES......................................27
   1.5 MY ARGUMENT..................................................................................................34
   OUTLINE OF THE BOOK..............................................................................................42

2. RULES, WRONGS AND PUNISHMENT..........................................................................45
   2.1 RULES AND RULE BREAKING............................................................................45
   2.2 PUNISHMENT, WHAT?........................................................................................48
   2.3 CRIME AND WRONGDOING...............................................................................51
   2.4 IS CRIME SOCIALLY CONSTRUCTED?...............................................................53
   2.5 PENAL REGIME, WHAT?...................................................................................58

3. THREE THEORIES OF PUNISHMENT........................................................................61
   3.1 DETERRENTISM.................................................................................................63
   3.2 REHABILITATIONALISM....................................................................................72
   3.3 RETRIBUTIVISM.................................................................................................80
   3.4 SOME MISCONCEPTIONS ABOUT RETRIBUTIVISM..........................................86
   3.5 AN ENDURING CHOICE: COMPARING THE THEORIES....................................101
   3.6 GOING FORWARD: QUESTIONS OF RETRIBUTIVISM.....................................106

PART II. A DEFENCE OF RETRIBUTIVISM

4. THE STRANGE ‘RIGHT TO BE PUNISHED’..............................................................119
   4.1 WOULD WANT AND SHOULD WANT.............................................................121

5. THE INSTITUTIONAL REASON FOR RETRIBUTION........................................127
   5.1 THE INSTITUTIONAL REASON TO PREFER PUNISHMENT..............................127
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2 RACHELS’S TEST</td>
<td>129</td>
</tr>
<tr>
<td>5.3 RACHELS’S TEST – AN APPRAISAL</td>
<td>137</td>
</tr>
<tr>
<td>5.4 RULE-UTILITARIANISM: RAWLS’S ATTEMPT TO DISSOLVE THE CONFLICT</td>
<td>141</td>
</tr>
<tr>
<td>5.5 CONCLUSIONS</td>
<td>145</td>
</tr>
<tr>
<td>6. THE SYMBOLIC REASON FOR RETRIBUTION</td>
<td>149</td>
</tr>
<tr>
<td>6.1 THREE MODELS OF RULE BREAKING</td>
<td>150</td>
</tr>
<tr>
<td>6.2 THE DISORDER MODEL</td>
<td>152</td>
</tr>
<tr>
<td>6.3 THE AUTONOMY MODEL</td>
<td>158</td>
</tr>
<tr>
<td>6.4 THE RATIONALITY MODEL</td>
<td>167</td>
</tr>
<tr>
<td>6.5 COMPARING THE MODELS: THE SYMBOLIC SUPERIORITY OF RETRIBUTIVISM</td>
<td>174</td>
</tr>
<tr>
<td>6.6 RESPECT FOR PERSONS AND THE SIGNIFICANCE OF CHOICE</td>
<td>176</td>
</tr>
<tr>
<td>6.7 THREATS AND AUTONOMY</td>
<td>180</td>
</tr>
<tr>
<td>6.8 THE IMPORTANCE OF BEING HELD RESPONSIBLE</td>
<td>183</td>
</tr>
<tr>
<td>6.9 THE SYMBOLIC REASON FOR RETRIBUTION</td>
<td>191</td>
</tr>
<tr>
<td>6.10 TAKING STOCK: THE OBJECTION FROM INSUFFICIENCY</td>
<td>194</td>
</tr>
<tr>
<td>6.11 GOING FORWARD: THE OBJECTION FROM NON-EXISTENCE</td>
<td>196</td>
</tr>
<tr>
<td>PART III. AN OBJECTION TO RETRIBUTIVISM</td>
<td></td>
</tr>
<tr>
<td>7. A THEORY OF EXCUSES</td>
<td>203</td>
</tr>
<tr>
<td>7.1 THE RECEIVED VIEW: THE DIFFERENCE BETWEEN EXPLANATIONS AND EXCUSES</td>
<td>204</td>
</tr>
<tr>
<td>7.2 EXCUSES, JUSTIFICATIONS, AND MERCY</td>
<td>207</td>
</tr>
<tr>
<td>7.3 EXPLANATIONS AS EXCUSES</td>
<td>210</td>
</tr>
<tr>
<td>7.4 THE OBJECTION FROM THE NON-DETERMINATE USES OF EXPLANATIONS</td>
<td>216</td>
</tr>
<tr>
<td>7.5 VARIETIES OF EXPLANATIONS: WHY DIFFERENT EXPLANATIONS ARE EQUALLY EXCUSING</td>
<td>224</td>
</tr>
<tr>
<td>7.6 DOES IT MATTER IN PRACTICE? P.F. STRAWSON ON THE PSYCHOLOGICAL IMPOSSIBILITY OF THE OBJECTIVE ATTITUDE</td>
<td>232</td>
</tr>
<tr>
<td>7.7 NORMAL AND ABNORMAL ACTION</td>
<td>240</td>
</tr>
<tr>
<td>7.8 CONCLUSIONS</td>
<td>243</td>
</tr>
<tr>
<td>8. THE THESIS OF HARD DETERMINISM</td>
<td>245</td>
</tr>
<tr>
<td>8.1 DETERMINISM</td>
<td>246</td>
</tr>
<tr>
<td>8.2 HARD DETERMINISM</td>
<td>254</td>
</tr>
<tr>
<td>8.3 OBJECTION 1: COMPATIBILISM</td>
<td>258</td>
</tr>
<tr>
<td>8.4 OBJECTION 2: PROBABILISM</td>
<td>264</td>
</tr>
<tr>
<td>8.5 OBJECTION 3: THE DIFFERENCE BETWEEN REASONS AND CAUSES</td>
<td>275</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

Viewed from the outside, the process of writing this dissertation must have appeared uneventful, indeed even boring: I was accepted as a PhD student; I went by tram to the department of political science a lot; I sat around in my office, visited the occasional seminar and drank a lot of coffee; I kept on playing for my beloved Hybris Basketball Club; I bought a scrappy old car; and I did some travelling (though not as much as I would have liked). Most of all, I read a lot and wrote a lot.

From a first-person perspective, however, the years I have spent working on this book have been anything but uneventful. Though writing a dissertation can be mind-numbingly boring at times – especially when you after a period of frenzied writing have to go back and get all the poorly specified footnotes right – the placid and tedious impression it can give the outside observer is entirely wrong. Scratch the surface of a PhD student and you’ll find a person in turmoil: beneath are doubts, grandiose hopes and aspirations, bright and not so bright ideas, and all kinds of practical and theoretical problems. Some people are under the frankly ridiculous impression that the reason why PhD students are often reluctant to discuss their projects with non-academics is that they feel superior to them – that we do not want to sully our bright research by communicating it to laypersons. This is by no means the case. Rather, the reason why we become shifty-eyed and evasive when you ask us about our work is that, most of the time, we firmly believe that what we are producing is an instance of the subject matter discussed by Harry Frankfurt in that small and cherished book of his. A PhD student generally walks under a cloud of imminent doom. Having completed this book, I do hope that cloud will now disappear.

This dissertation is about the importance of the notion of personal responsibility, and the attractive and empowering nature of being judged as praise- or blameworthy. It is curious to see how often academics are confused about these concepts. The typical academic book begins by listing the people or organisations that have helped the project, either intellectually or financially, and then proceeds, in an exercise of supreme humility, to state that most of the credit belongs to those acknowledged, while the blame remains entirely the author’s. This strikes me as a ques-
tionable asymmetry, which seems to place the author in a lose-lose situation. I am enormously grateful for all the advice I have received over the years. But if a book is a collective enterprise between the author and those who try to influence the author (as surely it is), then, in order to be consistent with the ideas endorsed in this book, it must be said that if the book before you has any merit, some of the praise belongs to me; and if it has none, some of the blame belongs to you!

The Department of political science at Göteborg University has been kind enough to have me over the past few years, and I would like to extend a collective thanks to everyone working there for making my time as a PhD student pleasant. My greatest debt is of course owed my supervisors. My main supervisor, Gunnar Falkemark, has with seemingly unlimited patience read and commented my drafts. Although twice my age, his irrepressible energy and high spirit often left me feeling like a whining and tired old man in comparison. I have benefited greatly from Gunnar’s incisive and detailed comments, as well as from his unwavering support. My second supervisor, Peter Esaiasson, has been equally supportive and has provided excellent advice throughout. His good judgement and no-nonsense approach to political inquiry, combined with the fact that he is not a political theorist, has been most welcome, for sometimes it takes an “outsider” to ask the best questions. My warmest thanks to both of you. The only gripe I have has to do with your handwriting, which can be almost sublimely unreadable and at times made me feel more like an Egyptologist than a political theorist when getting commented drafts back from you!

Apart from my supervisors, numerous people have commented on my work over the years. At the Department of political science, a special thanks goes out to the class of 2002 (“the egos on parade”): Stefan Dahlberg, Johan Karlsson, Anna Persson, Maria Pettersson, and Matthias Zannakis. These are some of the funniest, brightest and kindest people you’ll ever meet, and I have enjoyed enduring post-graduate school with them. I would also like to thank Bengt-Ove Boström, Marie Demker, Thomas Gelotte, Henrik Friberg-Fernros, Andreas Johansson, Andrej Kokkonen, Fredrika Lagergren, and Ulrika Möller for helpful advice on various drafts over the years. Bengt-Ove and Marie read the whole manuscript and offered detailed comments, the accommodation of which has improved the book greatly. Last but not least I would like to thank Sverker C. Jagers, a good friend and an inspiring collaborator.
Outside the Department of political science, I am grateful for the comments and advice offered, at various stages, by J. Angelo Corlett, Lina Eriksson, Victor Galaz, Niklas Juth, Ed Page, and Andrew Williams. Ed has been a good friend throughout and deserves special thanks for his encouragement and support.

While this project has primarily been funded by the Swedish taxpayers – a fact which still can leave me with a sense of wonder - I have gratefully received extra financial support from Adlerbertska Stipendiefond, Jubileumsfonden at Göteborg University, The Swedish Royal Academy of Sciences, Hierta-Retzius Stipendiefond, Stiftelsen Paul och Marie Berghaus donationsfond, and Stiftelsen Oscar Ekmans Stipendiefond.

The final and most heart-felt thanks goes out to my family and friends. With this book having been completed, you will now hopefully get to see more of me. I for one am really looking forward to seeing more of you.

Göteborg, 2007
Part I. The Problem
1. The Practice of Punishment

“All punishment is evil.”
(Bentham 1988: 170)

“The injury which the criminal experiences is inherently just […] and is his right.”
(Hegel 2005: 36-7)

1.1 INTRODUCTION

Among peacetime exercises of state power, the most vivid and principally problematic is arguably the practice of punishment. Consider the following figures from Sweden, where I live: In 2006, approximately 63,400 were issued a fine, either by the courts or prosecutors. 14,598 were sentenced to imprisonment. 326 were deemed unaccountable for their crimes and were sentenced to coercive psychiatric care. 9,918 received a suspended sentence, many of whom were obligated to do community work. On top of this, around 311,700 people were summarily fined directly by the police.1 These numbers represent, on reflection, a staggering amount of interference with people’s freedom and property; a staggering amount of fundamental interventions in the lives of individuals. They add up to the practice of (legal) punishment – the state’s practice of sanctioning those who break common rules of behaviour.

The practice of punishment may appear a given; an aspect of social life so natural that it hardly warrants closer attention. And certainly the

---

1 Figures taken from Criminal Statistics 2006, Report 2007:20 by the National Council of Crime Prevention. Two qualifications: these are brut numbers, which means that the number of individuals being convicted does not account for that the same person may be convicted more than once a year.
right to punish, from which the legitimacy of the prevalent interference in the lives of individuals must derive, is central to most accounts of the state. Weber famously included the successful monopolization of legitimate force into the very definition of a state, and even ultra-minimalist theorists about the state retain the right to punish as the core function of the overarching political authority in society. But the tangibility and seriousness of the practice of punishment warrant that we closely examine not only why the state should be taken to have a right to punish, but also in what ways, and for what purposes, its practice of punishment should be designed. For those 14,598 people sent to Swedish prisons in 2006, the need for such scrutiny is perhaps especially keenly felt.

What makes punishment a central topic for political and moral analysis is of course that punishment, as classically conceived, makes people worse off. Punishment thus seems to challenge the widely held belief that it is wrong to induce harm – punishment is simply about treating people in ways that we generally believe is wrong. Some might protest that the criminal, in virtue of having broken the law of society, has waived his or her normally held rights, and that there is consequently no need to assume that we have the same kind of obligations to criminals as we do to law-abiding people. But this is not an argument for punishment, but only another way of saying that punishment is permissible.

That punishment harms people suggests that we should take a strong normative presumption against it. As Rawls once noted, however, it is a striking fact that most believe that the practice of punishment is somehow justified: “Most people have held that, freed from certain

---

2 Weber offers this definition in Weber 1978: 54. The standard example of an ultra-minimalist about the state who still retains punishment as a central state function is Nozick 1974.

3 To be sure, there are also those who prefer to see punishment as good for the person being punished. According to such paternalist theories of punishment, it is incorrect to say that punishment harms people, for while punishment may inflict pain or deprivation, the end result is nevertheless to the benefit of the criminal. For paternalist theories of punishment, the normative question instead becomes, as in other examples of paternalism, why the state should to be taken to have right to interfere with the life of a rule breaker for his or her own benefit.

4 Hobbes, for instance, held that the criminal in breaking the law has declared war on society, and thus is no longer part of it. There is therefore no need to justify harming him or her (see Hobbes 1998; Brettschneider 2007).
abuses, it is an acceptable institution”, says Rawls. “Only a few have re-
jected punishment entirely, which is rather surprising when one consid-
ers all that can be said against it” (Rawls 2001: 21). I shall be no excep-
tion. I too shall assume that there is some principle or set of principles,
and some way of acting on them, which justifies the practice of punish-
ment. Another way of saying this is that this book is not an abolitionist
one. But while I will assume that the practice of punishment is somehow
justifiable, I shall take seriously that the nature of that justification is
very unclear and in need of scrutiny. Seeking to alleviate this state of
affairs, this book intends to offer an answer to the following question (Q):

(Q) Which principle or theory (or principles or theories) should
serve as the basis for a state’s penal regime?

(Q) is here understood as an explicitly normative question. It asks which
principle or theory (or principles or theories in combination) the practice
of punishment ought to be fashioned after: in accordance with what prin-
ciple or theory (I shall use these terms interchangeably) it should be de-
signed. Inherent in the question is the notion that the practice of punish-
ment can be fashioned in different ways, and that our task is to deter-
mine the best way of fashioning it; a way which ensures that the practice
of harming or otherwise interfering in the lives of rule breakers is morally
justified. This is the classical question of the philosophy of punishment.

What does it mean to justify something? Colin Bird convincingly
argues that to justify something is to give reasons to value it. To be sure,
this differs from merely expressing one’s personal likes and dislikes, and
from defending ethical claims on parochial grounds. We justify some-
thing by demonstrating “its value in a rational and impartial way” (Bird
2006: 13). If we succeed, we have presented general reasons which count

5 Abolitionists are people who believe that the practice of punishment cannot be
justified even in principle. For examples of contemporary abolitionists, see Duff
2004.

6 “Rational and impartial” for Bird means, sensibly enough, that we cannot jus-
tify something on narrow non-principled grounds, such as by the fact that we
very much like them: “I might […] very much prefer to have a coterie of slaves at
my beck and call, but we would not think that personal likes and preferences of
this kind (mine or anyone’s) decide the questions whether slavery is justified.
in favour of the action or practice for which the justification was sought, for a justified action or practice is one which deserves our approval or support. Of course, one must keep separate the quality of a justification and whether it wins our backing. We often support things for which there is no justification, and good arguments often fall on deaf ears. The quest for justification consists in presenting reasons which we ought to be persuaded by (ibid.).

The classical quest of the philosophy of punishment, and the quest which animates this book, is precisely this: to search for what it is about the practice of punishment, when it is conceived of and administered in a certain way, which makes it morally right and deserving of our support. To answer (Q), then, means first and foremost to search for the principle (or principles) which can, if it serves as the basis for it, justify the practice of punishment.\footnote{By saying that a principle “serves as the basis for” a practice of punishment (or the penal regime), I mean that it determines the overall functioning and nature of that practice. A principle of need “serves as the basis” for the health care system, then, if it determines what the health care system aims to do (i.e. help those in need), and the way it operates (i.e. gives all access to health care according to their medical needs, not their desert, ability to pay, etc.).}

What can we learn from an answer to (Q)? We can obviously learn something about what a just or justified penal regime would be like: we can learn something about according to which principle the crucial and contentious practice of punishment ought to be fashioned. The meaning of the term “penal regime” will become clearer in chapter 2, but let me here say that this choice of wording reveals a solid focus on punishment as a general and durable state practice. We can all agree that the overall quality of that practice can be better or worse, and that it is preferable if it as good as it can be. The quest of answering (Q) boils down to searching for the best penal regime.

Once (or if) we have an account of upon which principle or theory the penal regime should be based, we can say a number of additional things. By implication we can say something about what might be wrong with punishment as it is actually practiced by various states. We can also draw some conclusions about the specific institutional architecture that
follows from the nature of the answer to (Q). While this book will mainly analyze different justificatory principles of punishment - principles which purport to be the appropriate basis of the penal practice - on a more practical note I shall also have some things to say about the concrete institutional make-up of a just penal regime. The bulk of the attention, however, will be paid more abstract and principled issues.

Let’s begin the task of answering (Q) by taking a brief look at the history of theorising about punishment.

1.2 PERSPECTIVES ON PUNISHMENT

For a long time, it seems, the practice of punishment was largely unchallenged. The widely influential passages of the Bible, which sketched the outlines of a penal code that was to influence penal practice for centuries, did not question the right to punish. Since authority coincided with God’s will, its actions were by definition good and right. In Romans 13:1-13:4, for instance, we read:

“Let every person be subject to the governing authorities. For there is no authority except by God’s appointment, and the authorities that exist have been instituted by God. So the person who resists such authority resists the ordinance of God, and those who resist will incur judgment (for rulers cause no fear for good conduct but for bad). Do you desire not to fear authority? Do good and you will receive its commendation, for it is God’s servant for your good. But if you do wrong, be in fear, for it does not bear the sword in vain.”

Though there was some valuable prior work on punishment - most notably perhaps Plato’s treatment of the practice in the Laws (2000) - critique of the notion of punishment as natural and incontestable did not pick up pace until the writings of reformists such as Cesare Beccaria and Jeremy Bentham. Once considered as an instance of divine justice, these authors wondered, on wholly secular grounds, about the justification of punish-

---

8 I will do this mainly in the concluding chapter 10.
9 I return to (Q) in greater detail in chapter 2.
ment: where the right to punish came from (if it existed) and what purpose (if any) it served.

The conclusion of these early reformists was that punishment indeed can be justified, because it can promote compliance with common rules and thereby uphold social peace and order. But they also concluded that punishment, taken in and of itself, is an evil. Cesare Beccaria, whose _On Crime and Punishments_ (1986 [1764]) was widely held to be the Enlightenment treatise on punishment, held that all punishment not deriving from absolute necessity was tyrannical. A contractarian, Beccaria understood authority as what each individual would consent to in forming a social contract. Individuals would consent to public authorities having a right to punish, Beccaria argued, since this would promote compliance with common rules, which in turn promotes happiness for all. This goal qualifies as an “absolute necessity”. But unless punishment serves this goal, and unless punishment is impersonally, reliably and moderately administered, punishment cannot be justified. It would then be tyrannical, and the contracting parties would not consent to it (ibid. 7-17).

The idea of punishment as a _prima facie_ evil was most succinctly put by Jeremy Bentham in _The Principles of Morals and Legislations_ (1988 [1781]). “All punishment”, Bentham stated, “is mischief: all punishment in itself is evil. Upon the principle of utility, if it ought at all to be admitted, it ought only to be admitted as far as it promises to exclude some greater evil.” (ibid. 170) The principle of utility referred to by Bentham is the idea that the rightness or wrongness of every action or practice is determined by its consequences for general utility. If it leads to more utility than disutility than feasible alternatives, the action or practice is right. But Bentham saw no justification in excluding the people being punished from the calculus of utility. Their suffering would also have to be taken into account. This meant that without compensating gains in utility, punishment would only leave a negative mark in the balance of utility. In imposing pain or deprivation on the person being punished, it

---

10 An action or institution is right if and only if there are no (feasible) alternative which would generate better consequences. It thus does not suffice to show that punishment generates more utility than disutility – unless punishment generates better consequences than, say, a system where every rule breaker is pardoned, punishment would not be right (and hence not justified). See e.g. Tännö 1998.
would merely serve to produce disutility. Therefore, said Bentham, punishment is in itself evil, and can only be justified if it produces some outweighing amount of good.

Beccaria and Bentham are influential representatives of what I in chapter 3 call deterrentists. They both held that punishment taken by itself is tyrannical or evil, but argued that punishment could be justified in virtue of its beneficial overall consequences, in particular its capacity to deter future rule breaking.

Another influential position in the theory of punishment, which consolidated itself in the 19th century, also came to the conclusion that punishment in itself is evil, but in a different way. According to this position, penal policy should eradicate the causes of crime within offenders, not scare them into compliance or subject them to publicly sanctioned revenge. On this view, offenders engage in crime because they suffer from some disorder or the like, whether it be social, economic, psychological, genetic (or whatever) in origin. What justifies penal interference in the lives of rule breakers is that it may transform them into law-abiding and well-functioning citizens. A sensible penal policy addresses these disorders and rehabilitates or reforms the offender, to the benefit of offenders and the public alike.

We may refer to this position as the rehabilitative or rehabilitationalist one. Rehabilitationalists have perennially argued that traditional conceptions of punishment, which involve the infliction of pain on rule breakers, are primitive and irrational. Penal policy should be designed to effectively reduce crime, in particular repeated crime (or recidivism). Proponents of rehabilitation have tended to argue that this requires that we drop old ideas about punishment, especially ideas connected to guilt and revenge. A classical statement to this effect is Karl Menninger’s The Crime of Punishment. Menninger here writes:

“We, the agents of society, must move to end the game of tit-for-tat and blow-for-blow in which the offender has foolishly and futilely engaged himself and us. We are not driven, as he is, to wild and impulsive actions. With knowledge comes power, and with power there is no need for the frightened vengeance of the old penology. In its place there should go a quiet, dignified, therapeutic program for the rehabilitation of the disorganized one, if, possible, the protection of society during the treatment period, and his guided return to useful citizenship, as soon as this can be effected” (Menninger 1959: 63-4).
A psychiatrist, Menninger espoused a particularly ambitious therapeutic approach to crime, according to which penal policy should be entirely directed towards therapeutic treatment rather than punishment. Central to this important approach is the idea that the very rationale of punishment is outdated, unjust and ineffective. For the idea of harming offenders into becoming better people must be taken to be absurd, some argue. As Bertrand Russell, his usual witty self, once said:

“No man treats a motorcar as foolishly as he treats another human being. When the car will not go, he does not attribute its annoying behaviour to sin; he does not say: ‘You are a wicked motorcar, and I shall not give you any more petrol until you go.’ He attempts to find out what is wrong and to set it right” (Russell, cited in Rachels 1997: 475).

Deterrence and rehabilitation as justifications of punishment have perennially been advanced against the backdrop of the theory known as retributivism. Retributivism is possibly the oldest and certainly the most controversial among theories of punishment. It is associated with the writings of Kant (1998 [1785]) and Hegel (2005 [1821]), but has continued to garner support from such contemporary philosophers as Robert Nozick, Jeffrie Murphy, and J. Angelo Corlett. The precise meaning and definition of retributivism is notoriously contested. It will suffice, for now, to define retributivism as the theory which states that punishment is justified if, when, and to the extent that, it is morally deserved. Retributivism holds that in order for a punishment to be justified, it must give the rule breaker what he or she deserves in virtue of his or her actions. Unlike preventive theories of punishment, retributivism adopts a backward-looking approach to punishment. What makes the state’s punishing criminals right is neither that doing so sets an example for others, nor that incarcerating him or her facilitates rehabilitation. It is justified simply in that a wrongdoer gets what he or she deserves, no more and no less.

Why is it important to give rule breakers what they deserve? Retributivists are far from always clear on this key point. Suffice to say, for now, that it is somehow important in itself and not because of its future consequences. Giving rule breakers what they deserve is somehow intrinsically good or just.
1.3 CLASH OF INTUITIONS

The theories of deterrence, rehabilitation, and retribution are by no means the only theories of punishment out there. But they are the main ones, and in my experience represent quite well the different intuitions people have about what punishment is about. If we reflect on why the state should be taken to have a right to interfere in the lives of rule breakers, its deterrent, rehabilitative or justice-making effects seem like plausible answers.

Deterrence, rehabilitation and retribution are also the theories I shall consider in this book. They represent the weightiest and in my mind most compelling principles that can serve to justify the practice of punishment, and are as such fitting candidates for answering (Q). I will present the theories and their respective (competing) answer in fuller detail in chapter 3. Here it is time to say that while much of the book in some sense can be read as an evaluative comparison of the rehabilitative, deterrence-based, and retributive answer to (Q), my intention is to defend one particular answer over its rivals. I will defend the thesis that retributivism should serve as the basis of the penal regime: that we should prefer the retributive answer to (Q) to the competing answers. That is, I will defend the claim that a retributive penal regime, which aims to give rule breakers what they deserve, is preferable to – more justified than - regimes based on deterrence or rehabilitation. The kind of retributivism I have in mind is of a positive and non-strict variety, the meaning of which will be made clear in later chapters.

Before stating nature of that defence, we should pause to consider some of the central points of contention, or “sticking points”, between the aforementioned theories. Doing so will not only enable us to better appreciate the ways in which these theories differ, but also introduce some of the difficult questions the problem of punishment inevitably poses.

Sticking point 1. The purpose of punishment.

A first and fundamental point of contention concerns the purpose of punishment. What should punishment aim to do? When I have discussed this
question with colleagues and students, they have been taken aback by the fact that I even consider that there might be a retributivist way of answering it. They find it obvious that punishment is about getting the right behavioural results: protect society from crime, transform antisocial offenders into well-functioning and productive citizens, making sure dangerous individuals aren’t free to roam the streets, etc. In other words, they have a more or less consequentialist outlook on punishment. Punishment is a regrettable but necessary device of social control, which in itself is evil. It is only justified if it leads to the right overall consequences.

Retributivism, at least on its more ambitious formulations, takes issue with this. Punishment, according to this view, is about giving rule breakers what they deserve, as a matter of justice. My colleagues and students have tended to hold this idea in extraordinarily low esteem. The notion that punishment can be justified on grounds of moral desert alone is taken to be absurd. Indeed, it is held to be a notion that only masquerades as a rationally defensible philosophical position, but which in reality is merely a rationalization of deeper and darker motivations. For behind the doctrine that punishment is justified by, and should aim at, giving rule breakers what they deserve is namely the desire for revenge, I have been told. This desire, while an undeniable part of human nature, is furthermore something which should be resisted, or at least not be used to morally justify a contentious political practice such as punishment. The desire for revenge can at best explain the origin and shape of punishment as we know it. All in all, the retributive theory is taken to be muddled, politically reactionary, and potentially dangerous.

This sticking-point, then, concerns the basic rationale of punishment. Those who have a broadly consequentialist outlook on punishment hold that punishment should aim at getting the right behavioural results, and that the justifiability of punishment is contingent on its effectiveness in this respect. Some, myself included, hold that there are other things to punishment. According to at least some forms of retributivism, punish-

---

11 The most sophisticated argument to this effect is perhaps Mackie 1982. Mackie here argues that there is a “paradox of retribution” – retribution cannot be rationally defended, but it rests on inescapable emotions of wanting to “bite back.”
ment is instead primarily about making sure that criminal wrongdoing gets what it deserves.\textsuperscript{12}

\textit{Sticking point 2. The appropriate form of punishment.}

A closely related point of contention concerns the appropriate form of punishment. We shall define punishment more exactly in chapter 2, and as we shall see it is often a good idea to be wary about referring to rehabilitative responses as “punishment”, but let us here understand the concept intuitively: punishment is a unifying term for the things that happen to people who break the law and get caught. The question is: What sort of things should happen to the lawbreaker?

Retributivists and deterrentists are in clear agreement here. Punishment should harm rule breakers. It should make them worse off. For retributivists, this follows from the belief that the lawbreaker has (typically) done wrong and that he or she deserves to be subjected to pain or deprivation as a result. Deterrentists hold that punishment must harm if it is to deter, for the only reason why punishment acts as a deterrent is that we do not fancy being subjected to it. To take a crude but clear example: if people in general prefer being locked up inside prisons to being outside them, this form of punishment has no deterrent value whatsoever (indeed, it serves to increase the probability of crime).

Rehabilitationalists, on the other hand, typically argue that punishment should not be about harming rule breakers, but rather about making them, on the whole, better off. This usually leads them to argue that punishment, as conceived by deterrentists and retributivists, is irrational and counterproductive. If rehabilitation is better served by offering rule breakers non-punitive treatment, this is the only defensible thing to do.

Note that for those who entertain a consequentialist outlook on punishment, whether rehabilitation or punishment-as-deterrence is endorsed depends on ultimately factual considerations. Utilitarian-minded people

\textsuperscript{12} As will be made clear in chapter 3, however, some retributivists hold that the purpose of punishment is to reduce crime (or otherwise affect behaviour in a desired direction), but that punishment is only morally justified if the punished person deserves it.
will endorse the form of punishment that on the whole seems to them to have the best consequences. 13 In my experience, whether a person ends up advocating one response rather than the other depends chiefly on their temperament or disposition – their way of perceiving the “facts”. I have made a habit of distinguishing among my student and colleagues between “optimists” and “pessimists” (the latter, however, prefer to see themselves as “realists”). Pessimists are those who believe that crime is a consequence of the underlying brutish selfishness that ultimately governs men, and that the threat of punishment is the only thing that keeps us from reverting into the state of nature. They tend to advocate punishment-as-deterrence. Optimists are those who believe that criminals are corrigeible, if only given the right help, and that it is feasible to address the very motivation of offenders. They tend to advocate rehabilitation.

If utilitarians make the choice between punishment and treatment on essentially factual grounds, retributivists are necessarily committed to the view that punishment must involve harm. No empirical observation about the efficiency of harm can change this, as it is a normative requirement that wrongdoers experience pain and deprivation. Moreover, retributivists hold that punishment must have a certain emotional tone. Though not about mere revenge, retribution must involve something which we may refer to as resentment, moral disapproval, or simply blame. This has to do with the fact that, for retributivists, to punish someone means to hold them responsible for their actions in a particular sense – to address them as people who have done something morally blame-worthy, and who therefore deserve moral censure.

---

13 See e.g. Moran in Conrad & Schneider 1992. As will become obvious, when I in this book speak of consequentialism, I for the most part refer to the aggregative and agent-neutral form of consequentialism known as utilitarianism (Sinnott-Armstrong 2007). While declared dead in some fields of political philosophy (cf. Page 2006: 79), utilitarianism is alive and kicking in the philosophy of punishment. At any rate, determentism and rehabilitationalism works best if understood as utilitarian theories.
**Sticking point 3. Personal responsibility.**

A third point of contention between the theories is a big one. It concerns the concept of *personal responsibility*. As we just saw, retributivists hold that punishment is, or should be, a species of holding people responsible for their actions. Unlike deterrentists and rehabilitationalists, retributivists therefore put a lot of weight into the concept of personal responsibility.

What does “personal responsibility” mean? Answering this question requires that we first address the general concept of responsibility. Responsibility is usually thought of as having to do with questions of authorship, causation or production of things. To be responsible for the proverbial state of affairs $X$ (an action, outcome, state of affairs, event, or whatever) means, roughly, that the occurrence of $X$ is somehow attributable to you, and attributable in such a way which corresponds to your degree of authorship, causation or production of $X$.\(^\text{14}\) Under this general understanding of responsibility, a variety of subspecies of responsibility can be discerned.

A first distinction to be made is between causal and moral responsibility. *Causal* responsibility concerns only causal relationships. One is

---

\(^{14}\) This means that one can be more or less responsible for $X$. A situation where $X$ happens solely because of you is one where you bear full responsibility for $X$ happening, but it may just as well be the case that while $X$ happens partly because of you, there are many others as well which contribute to $X$ happening (e.g. global climate change). In such cases, you are partially responsible for $X$. Note that this conception of responsibility does not exclude that one may be responsible for $X$ in virtue of *omissions*. If it is true that $X$ happened because of your neglect or passivity, it could be said, counterfactually, that *unless* you had remained passive, $X$ would not have happened (e.g. “had you intervened, the child would not have drowned”). It can therefore make sense to treat omissions as a basis for responsibility for $X$. Note that I am here primarily concerned with responsibility as a *backward-looking* concept, not the *forward-looking* responsibility which is sometimes referred to as “treatment” or “remedial” responsibility. For valuable treatments of the concept of responsibility, see for instance Zimmerman 1988; Fischer 1999; Miller 2004.
causally responsible for X if one has part in the fact that X came about (and responsible to the degree that corresponds with one’s degree of causal contribution). Causal responsibility as such has nothing to with questions of praise and blame. We can for example in a causal sense say that “the earthquake was responsible for the destruction of Lisbon”. This does not commit us to somehow finding the earthquake at fault or blameworthy for the havoc it wreaked. By \textit{moral} responsibility, on the other hand, theorists usually intend precisely the property that makes agents deserving of praise and blame.\textsuperscript{15} If someone is morally responsible for X, she stands a relation to X such that X can be attributable to her in a way that makes her praise- or blameworthy.\textsuperscript{16} It is in this familiar sense we can say, e.g., that “Zhukov was responsible for the destruction of Berlin” or that “the students are responsible for the pleasant atmosphere in the classroom”, thereby implying something evaluative about Zhukov and the students, respectively.

Moral responsibility is usually seen as quite the exclusive thing. It is ordinarily assumed that only reasonably rational, sane, and morally competent adult human beings can be morally responsible. Unlike, say, infants, machines, or rocks, these may be said to deserve praise for good things and blame for bad things (Zimmerman 1988, Smilanski 1996). The exact reason why only sane and rational human adults can be morally responsible is contested, however. Standard suggestions are that they have an exclusive capacity for control, rationality, or reason-

\textsuperscript{15} As Pereboom 2001 notes, however, one can be morally responsible for a morally indifferent or neutral thing, which serves to prove that moral responsibility need not always be connected with praise and blame.

\textsuperscript{16} Moral responsibility clearly does not have to coincide with \textit{legal} responsibility, which is the prerequisite for legal punishment. A person may for instance be morally responsible for wrongs that are not legally prohibited and for which he or she consequently cannot be legally responsible. Moreover, some forms of legal punishment, punishments of strict liability, do not require moral responsibility. Crimes of strict liability are crimes which concern pure causal responsibility. An example could be a manufacturer’s obligation to compensate victims of faulty merchandise. Under conditions of strict liability, the manufacturer may be held liable to compensate even though it/he/she did not intend to cause harm, nor could reasonably foresee that it/she/he would cause harm. Causation is enough in these situations, and moral fault is not relevant.
responsiveness (see e.g. Fischer & Ravizza 1998; Fischer 1999; Morse 2000).

Personal responsibility is often used as a synonym for moral responsibility, but it is more correct to see the former as a specification of the latter. The concept of moral responsibility as such makes no assumptions about what kind of entities or things that can be responsible. It is perfectly possible, for instance, to defend the notion of collective moral responsibility, such as when one ascribes praise- or blameworthiness to corporate actors in a way that is not reducible to individuals. Personal responsibility, however, is used to refer specifically to the responsibility of a single individual. If Zhukov is personally responsible for the destruction of Berlin, this means, as I use the term, that he as an individual is morally responsible for it (wholly or partially).

Personal responsibility tends to evoke quite specific connotations of a normative kind. If one puts emphasis on personal responsibility, this tends to mean that one thinks that individuals can be, are most often are, morally responsible for what they do. One will not buy into stories to the effect that people are powerless and blameless parts of greater wholes, such as their culture, society, or zeitgeist. Take as an example the alleged “fat-boom” in modern societies. Someone who emphasizes personal responsibility will not be impressed by claims such as that individuals are victims of a fast-food culture, or that society as a whole must accept collective responsibility for the prevalence of obesity. Instead, she will hold that each and every individual is personally responsible for maintaining healthy eating habits. This idea – that individuals generally are personally responsible for their behaviour – is sometimes referred to as the principle of personal responsibility, the spirit of which is that we should “own up” to the consequences of our actions. “We ought to answer for our own actions”, James Q. Wilson proclaims along these lines, “and not, save for the extraordinary reasons, claim that we were compelled to act badly by forces over which we had little control” (Wilson 1997: 1).

What does this conceptual unpacking have to do with punishment? Well, personal responsibility is a sticking point in the debate between rehabilitation, deterrence and retribution in only too many ways.

Again, a first thing to note is that personal responsibility is a prerequisite for retribution. According to retributivists, it is the fact that the rule breaker deserves punishment that justifies that the state punishes him or her. But unless one is personally responsible for one’s crime, it seems that one cannot deserve punishment – personal responsibility is
thus a prerequisite of desert, and hence also a prerequisite for retributive punishment.17 (We will return to this relationship between punishment, desert and responsibility throughout this book).

A second thing to note is that rehabilitation and deterrence do not seem as reliant, if reliant at all, on personal responsibility as part of their justificatory scheme. It seems to make no decisive difference whether rule breakers are personally responsible and blameworthy for their crimes if the purpose of punishment is to improve offenders or deter potential rule breakers.

A third thing to note concerns the substantive debates that revolve around personal responsibility. Opponents often charge retributivism with being obsessed with the notion of rule breaking as “badness” - that retributivism’s insistence on individual guilt and moral desert only makes sense as long as one adopts a prescientific and obscure view of people as self-made selves; fundamentally free and autonomous agents who engage in wrongdoing simply because they choose to, not because they are being driven to crime by quite natural and understandable circumstances. For such opponents, retributivism’s continued insistence on personal responsibility serves to cloud our vision and forestall the invention of more rational ways of controlling crime.18 Retributivists, on the other hand, flip this reasoning around and charge their opponents with failing to treat rule breakers with respect. They tend to argue that since retribution assumes that rule breakers are personally responsible and thus blameworthy, retribution treats them with the respect that is their due as persons. On this retributive account, it is supposedly far more dignified to suffer punishment because one is seen as a wrongdoer, capable of deserving moral censure, than when one is punished simply to send out the right incentives to others, or when one is deemed in need of treatment.

All manner of examples can be drummed up to illustrate this point of disagreement. Here is my favourite: suppose medical science comes up with a pill that could permanently cure violent aggression. If we gave the pill to people convicted of violent crime, we could be sure that they wouldn’t commit further acts of violence. Suppose also, to keep the ex-

17 For a defence of that responsibility is a prerequisite for desert, see e.g. Smilansky 1996; Corlett 2003.
18 A clear statement of this kind is Clark 2003.
ample clean, that the quality of life of those who take the pill is not made worse – not worse, at any rate, than it would have been during and after years of imprisonment – and that crime rates would not go up if the criminals went unpunished. Should we opt for the pill in such a situation? Should we opt for only the pill?

Those who maintain a utilitarian outlook on punishment, it seems to me, have no reason to forgo using the pill. The pill would get the right results done, and there is nothing beyond the right results which makes punishment important. The medical solution would take care of recidivism without subjecting the violent criminals to unnecessary suffering.

From a retributive perspective, and to the amazement of more pragmatically minded people, there are however reasons to be very cautious about opting for the pill. First of all, the pill may seem an illegitimate interference with the autonomy of the rule breaker. After all, he or she may want to engage in future acts of violence once he or she has paid the price for the present one, and perhaps should be allowed to do so. Furthermore, it may seem unfair that the rule breaker gets off with “only” taking a pill, as it does not match the harm his or her crime has presumably caused the victims. (It would seem particularly unfair if other rule breakers who are guilty of similar crimes are punished instead of given the pill). Finally, and going back to the notion of personal responsibility, the pill may seem inappropriate because it doesn’t seem to be anything even close to holding the rule breaker responsible. When we give rule breakers the pill, we seem to treat their crimes less as expressions of wrongdoing, and more as something which should be understood, treated and cured. In short, a retributivist may reject the pill simply on the grounds that it fails to affirm the rule breaker as a personally responsible person whose choices warrant blame.

---

19 This example of course simplifies things in more than one way. A full utilitarian answer to the question would involve at least the following factors: the effect on general law-abidingness; the possible welfare or utility people derive from seeing rule breakers in distress; the financial costs of the pill as opposed to other measures; and the distress involved in taking the pill as opposed to other measures.
The pill-example above is of course science fiction at this point, and some people object to the very idea of getting points across by using hypothetical examples. (Concocting far-fetched thought experiments is a philosopher’s disease and, alas, one I shall submit to more than once in this book). Fortunately, the very same kind of problems crop up in actual penal practice from time to time. Consider this recent Swedish example:

A couple of years ago, Jens Orback, then minister of equality of the later defeated social democratic government, announced that he planned to create a treatment clinic for men who batter women. The reasons he gave for this were straightforward: male violence against women is a gigantic social problem which needs to be solved. It is not certain that punishing violent men is the best way to go about it, however. Rather, experiences from other countries suggest that the best way of lowering recidivism among violent men is to offer them treatment for their behavioural problems. To offer violent men treatment, moreover, does not mean that we victimize them or that we trivialize their repugnant behaviour. It only means, Orback concluded, that we do what we need to if we are serious about coming to grips with the problem of male violence against women.

These politically cautious statements might seem innocuous. But the idea of a “male clinic” was far from uncontroversial, at least not given the gender-fused debate which raged in Sweden at the time. A controversial 2001 public report had namely argued that male violence against women is endemic, claiming that as many as 46 percent of all Swedish women had experienced male violence since their 15th birthday. This

---

20 Indeed, the minister said that his “long-term ambition” was to establish “crisis centres” for violent men in every city.
21 Svenska Dagbladet, 050801: “Orback vill ha en mansklinik” [“Orback wants a male clinic”].
22 The report in question is Slagen Dam [Captured Queen] (Lundgren et al. 2001), which was written on behalf of The Crime Victim Compensation and Support Authority and co-authored by Eva Lundgren, a controversial sociologist. The
hardly supported, it was argued, that violence against women is perpetrated only by disordered males - given the scope of the problem, violence against women must rather be understood as something which even “normal” men perpetrate. Representatives for a national organization of women’s shelters, in opposition to the idea of the violent man as a disordered and abnormal outlier, held that “Male violence against women should be regarded as a criminal act and judged accordingly”, thus neatly summarizing what they felt was wrong with Orback’s entire idea.23

Similar reactions to Orback’s proposal were expressed in the newspapers. An editorial in a national tabloid criticized Orback for “unreflectively” having declared that the violent males need treatment. To use that word, the editorial said, implies that the offender is ill or disordered; somebody whose behaviour warrants understanding and should be excused. But if the offenders are disordered, how come they’re often so calculating and controlled in their behaviour, the editorial asked? That the offenders are in control over their behaviour, and often well-integrated and functioning men in other respects, suggests that they can hardly be called “ill”.24

Are these reactions hogwash? Some certainly think so. A criminologist with the Stockholm police had the following thing to say:

“All can see what the problem is, but no one acknowledges it […] We send these men in and out of jail, and in between they commit further crimes. The prison sentence doesn’t bite on them. In my opinion, what rules them is a pathological need to control women.”

authors explicitly argued that the problem of male violence against women should be understood in terms of gender, not in terms of individual psychology or socio-economics. The report has been the object of much criticism, in particular concerning its definitions of violence.

23 This claim was made in reaction to a 2002 public report on penal policy and mentally disordered offenders, in which it was claimed that “it is common knowledge that many of those who are convicted of crime, especially violent and sexual crime, are in need of psychiatric treatment”. (Psykisk störning, brott och ansvar [Mental disorder, crime and responsibility], SOU 2002: 3, p. 18). For the response from ROKS, the women’s shelters organization, see Ju 2002/451/L5.

24 Expressen, 050802: “Orback spelar högt” [“Orback takes a risk”].
And further:

“This is not subject to extensive debate. […] Some wants us to believe that the abusive men are just like any men. If we instead accept the view that there is a mental disorder at the heart of things, we would see that there are preventive measures to use. This behaviour cannot be punished away, it has to be treated.”

Like the imaginary case of the crime-curing pill, this skirmish directs our attention to an important sticking-point: whether there is a conflict between punishment and treatment, or, perhaps better, between punishment and understanding. Those who objected to the idea of a “male clinic” did so in part because they felt that treatment would send out the wrong moral messages: that if we refrain from punishing violent men and instead treat them in some way or another, we would fail to express the appropriate blame and resentment. In part, the sticking-point has also to do with how we explain crime. Those who criticized the notion of violent men as mentally disordered instead preferred to explain the behaviour by citing the influence of patriarchal social structures, which serve to make violence against women “normal”. This was held to be a particularly threatening and odious mode of explanation by others.

---

25 Both quotes from Dagens Nyheter, 060523: “’Fängelse biter inte på männern som slår’” [“’Imprisonment does not work on abusive men’”]
26 This was the position of many feminists. To explain male violence against women by citing the mental disorders of a series of individual males, they held, is not only false but also blinds us to the structural nature of the problem. Interestingly, when Orback restated his proposal some year afterwards, he did so in the very gender-theoretical terms used by his critics. The treatment clinic, Orback held, would engage and question the socially constructed notion masculinity, as “the traditional male role contributes to the maintenance and reinstatement of a masculinity which affects the lives of women and girls and restricts the freedom of action of boys and men” (UNT 060510: “Mansrollen ska diskuteras” [“The male role will be debated”]). This idea, talking about gender-roles, at least looks far from the dreaded behavioural engineering a la A Clockwork Orange some originally had in mind.
27 In a parliament debate about ending financial public support to ROKS, the organization above which held that male violence against women “is a criminal act and should be judged accordingly”, a parliamentarian for the Christian Democrats, for example, held that the gender-based explanation makes “the will
The point of contention here goes back to the former one about personal responsibility. If rule breakers are responsible for their actions, it might seem inappropriate to offer them anything less than punishment, as this would perhaps serve to reduce their guilt and portray them as victims rather than perpetrators. As one commentator, reflecting on the fact that many rule breakers come from very deprived and problematic backgrounds, put it:

“On the one hand, it’s clear that persons who have been brought up under bad social circumstances are overrepresented in our prisons. On the other hand, I think that we have to discuss the responsibility a criminal has for his or her actions. I believe there’s always a choice. The man who batters his wife actually chooses to hit her.” (Hellgren cited in Hagberg 2006: 63. My translation.)

What would our theories of punishment say about Orback’s clinic? Deterrentists would prefer punishment, provided that treatment isn’t on the whole a better way of preventing crime. Rehabilitationalists presumably see in the clinic precisely what they have in mind: an attempt to figure out what is wrong and address the root causes of crime, rather than a pointless and vindictive imposition of pain and resentment. Retributivists, finally, would probably be sceptical. Provided that the violent men aren’t so disordered that they lack responsibility for their actions, they deserve punishment and this is what they should get as personally responsible agent. Indeed, as I shall argue, one could make a retributive case for that that even the punished men themselves have an interest in being punished rather than treated.

1.4 WHERE WE ARE TODAY: CONFLICTING TENDENCIES

Having introduced some points of contention between retributivism and its rivals, and before stating the nature of my argument, it is perhaps worthwhile to reflect on the contemporary standing of retributivism; the and responsibility of the individual disappear” and furthermore “enables collectivizing male guilt” (Interpellation 2004/05:648).
theory I will defend in this book. Though my students and colleagues almost uniformly reject retributivism, it is nevertheless the case that the theory seems in better health than for a very long time. It is widely recognized that retributivism, which at the beginning of the 20th century enjoyed precious little support, nowadays has a very strong standing in academia. Moreover, it is also widely held that retributive ideas, starting from around the 1970s and onwards, have exerted a strong influence on Western penal practice. On the basis of these developments, it may seem as if retributivism is in little need of defenders.

There are reasons, however, why an investigation and defence of retributivism is timely and important. First of all, it is wrong to assume that retributivism has a consolidated standing just because its clearest rival - the ideal of rehabilitation - is presently held in low esteem. True, the great visions of a penal regime based on treatment, once so influential, are today few and far between. But new champions of the rehabilitative ideal will emerge - rehabilitation is simply too compelling an idea to be dormant for long. Furthermore, while it seems true that there is a renewed emphasis on punishment in Western penal regimes, claims that this is evidence of retributive influence on policy must be taken with scepticism. There are indeed signs that penal policy has taken a punitive turn. But that turn largely seems of a distinctively anti-retributivist character.

Secondly, and more fundamentally, the trajectory of certain long-term developments seem to be such that we need to ask whether the principle of personal responsibility and its associated practices of deserved praise and blame are tenable at all. For taking a historical perspective, it seems as if confidence in the existence of personal responsibility has steadily declined over time. It can be predicted that developments in sciences of human behaviour will eventually make us question the very notion of personal responsibility, guilt, desert and similar familiar concepts, as these seem to rely on a theory of action which many are increasingly starting to question. Retributivism, with its emphasis on responsibility and desert, must be addressed in light of this development.

---

28 See e.g. Dolinko 1997; Smilansky 2006.
29 See e.g. Dolinko 1991; Tham 1995; Bedau 2003
Let’s begin by taking a brief look at developments in penal policy, in order to undermine somewhat the notion of retributivism’s dominant standing there.

It is often held that the most significant thing about late 20th-century development in penal policy was retributivism’s ascent at the expense of the rehabilitative ideal. The received story is this:

In Western states, the ideal of rehabilitation grew strong during the latter parts of the 19th century and the beginning of the 20th century. Penal policy was designed to break criminal habits and remould antisocial personalities. Thus we find the appearance of “correctional” institutions, the proliferation of forensic psychiatry, and the invention of policies such as indeterminate sentencing, parole, youth prisons, etc. Sweden was certainly no exception to this general pattern. Indeed, dismissal of retributive principles in penal policy in favour of rehabilitative ones was probably more marked here than in most other places.30

Over the last couple of decades, however, Swedish penal practice has become “tougher” (Tham 1995). Convicted rule breakers are more likely to be sentenced to prison, and prison terms are on average longer. Moreover, proportionally fewer mentally disordered offenders are committed to forensic psychiatric institutions instead of regular prisons. In 1975, 70 percent of those found guilty of murder or manslaughter were sentenced to compulsory psychiatric care, compared to only 20 percent in 2004 (Hagberg 2006).31

30 See for instance Qvarsell 1993. A particularly clear example of the dominance the ideal of individual prevention exerted a couple of decades ago is found in the public report Skyddslag [Protective Law] SOU 1956:55, which was considered as a candidate for a new Swedish criminal code. The subsequently accepted code – Brottsbalken, adopted in 1965 and still in place – is widely held to have been a compromise between general prevention and rehabilitation.

31 Hagberg rhetorically asks if murderers are healthier today. It could just as well be asked if they were sicker in 1975.
The same tendency towards “toughness” obtains in other Western states (see e.g. Garland 2002). Many have taken this as evidence of the ascent of retributivism on behalf of the once so influential ideal of rehabilitation, which fell into disrepute due to widespread suspicions that rehabilitation simply does not work. There is probably some truth to saying that the tougher penal policies in part are a result of the increasing influence of retributive thought, even though it is incorrect, as we shall see in chapter 3, to equate toughness in punishment with retributivism. But penal policy in Western states has become more accepting of the paradigm of “just desert” than before (Bedau 2003; Dolinko 1991; Raynor & Robinson 2005).

Based on this relative ascent of retributivism in policy-making, one may indeed think that the theory is in no need of champions. However, some developments serve to qualify that thought.

First of all, while it is beyond doubt that the ideal of rehabilitation is no longer nearly as popular as once was the case, and while it seems true that penal policy in Western states indeed has become more punitive, we must be careful not to infer the influence of retributivism from this. Tougher penal policy is not the same thing as retributive penal policy. In fact, one can just as well interpret the tendency towards “toughness” as evidence of a different, less refined attitude to crime and punishment. That attitude is sometimes referred to as the “Nothing Works” view. According to this view, many criminals are impossible to rehabilitate and largely cannot be deterred from crime. In order to protect itself, society must incapacitate these individuals for as long as needed. It need not pay too close attention to principles of justice in this enterprise.

There are indeed many signs today that the primacy of social protection overrules more principled considerations about desert and fairness. To take just a few recent examples from Sweden: In 2002, a public report proposed the creation of a “public protection measure”, intended to work as a way of keeping dangerous offenders deprived of their liberty even though they might not be liable for their crimes. According to the re-

---

33 The report in question is the aforementioned SOU 2002:3. The public protection measure was intended to close a “gap” in Swedish penal practice – the possibility of a severely mentally disordered offender quickly becoming better as a result of psychiatric treatment, with the consequence that he or she could not be sen-
port, authorities should in the interest of public safety have the opportunity to incarcerate some particularly risky individuals on indefinite basis, even though they may not be accountable for what they did. The present government now entertains a proposal to the effect that some offenders should be sentenced to *regular prison* even though they were not accountable for their crimes. The rationale of this reform, evidently, is that some offenders cannot be “sufficiently” sanctioned otherwise (in plain text: mentally disordered offenders might as a result of psychiatric treatment get healthy too quickly). Along the same lines, two leading politicians have recently taken the art of penal populism a step further: as long as their crimes are severe enough to merit imprisonment, *all* offenders, sane or insane, should be sentenced to regular prison, they have argued.

These are examples of “get tough”-policies which express little regard for retributive principles, and which should not be taken as evidence of a growing focus on “just deserts.” They are rather the result, I believe, of a concoction between the attitude of “Nothing Works” and what is sometimes referred to as penal populism. Individuals in Western societies have become increasingly afraid of crime. This rise in fear is for the most part not motivated by a corresponding increase in real crime, but politicians have been quick to exploit the feeling of insecurity and propose reforms that from a retributive perspective seem highly problematic. Retributivism claims that desert is a necessary condition for punishment. Proposals that boil down to the punishment of undeserving in the name of social protection can scarcely be taken as evidence for “retributive” penal policy.

\[
\text{34 The proposal in question is Ds 2007:5, written on behalf of the Ministry of Justice. What constitutes a “sufficient” sanction is not clear, but a plausible guess is that the standard of sufficiency is provided by populist sentiments.}
\]

\[
\text{35 SVD 050830, “Fängsla psykiskt sjuka brottslingar” [“Imprison mentally disordered offenders”].}
\]

\[
\text{36 That rise of the fear of crime is disproportionate to crime rates is argued by, e.g., Kristjánsson 2006. Analyses of the political logic behind politicians becoming tougher on crime are found in, e.g., Wilson 1997; Garland 2002; Raynor & Robinson 2005.}
\]
*A shrinking sphere of responsibility?*

If there is reason to doubt retributivism’s consolidated standing in contemporary penal policy, there is even greater reason to question its long-term viability. The sphere of beings or entities held to be fit to be punished namely seems to be steadily shirking over time.

A couple of centuries ago, it was not uncommon to punish animals for their crimes, and there are even odd reports of inanimate objects being brought to justice (Glover 1970; Ferry 1995; Pinker 2002). Children were also often held responsible in a way firmly believed to be unacceptable today. These examples now seem like historical follies. Today, the bearer of personal responsibility is taken to be the normally functioning adult human being. But even within this class of people, there are signs that the sphere of responsibility is becoming smaller. 37 Most clearly this is seen in the process some sociologists refer to as the medicalization of wrongdoing.

The medicalization of wrongdoing is the process through which previously wicked, sinful or criminal behaviour is instead seen as indicative of medical problems. It is a process of moving from a paradigm of “badness” to one of “sickness”, as some have nicely put it (Conrad & Schneider 1992). For better or worse, medicalization indeed seems upon us. We find today a host of officially sanctioned medical diagnoses of things that were previously seen as moral wrongs or evidence of a bad character. Perhaps the clearest examples are those of addiction. Those who drink too much are today said to be suffering from a disease – alcoholism – which affects their ability to control their drinking. 38 That the behaviour is now seen as an illness has consequences for our attitudes: drinking too much seems less of a moral flaw than a medical problem; something which is worthy of treatment and understanding rather than blame and disapproval. Alcoholism is by no means the only illness of this kind. It is today suggested that we may similarly be addicted to gambling, smoking, drugs in general, shopping, eating, adrenalin rushes, sex,

37 A similar argument is Morris 1988.
playing computer games, surfing the web, and so on. These diagnoses, some official and some unofficial, all have a tendency to exonerate the addict from ordinary personal responsibility. The person who is addicted to food eats too much – a morally problematic behaviour, let us assume – but cannot control his or her behaviour. Consequently, he or she cannot be ordinarily blamed. Other instances of behaviour previously thought of as moral vices but today increasingly seen as medical problems are bad temper (“Intermittent Explosive Disorder”); youthful pig-headedness (“Oppositional Defiant Disorder”); repetitive petty theft (“Kleptomania”); and callousness (“Antisocial Personality Disorder”).

We simply seem to have far more ways to be disordered today, which must translate into far more ways to be less than fully responsible. Some have even suggested that having a “rotten social background” – not an uncommon trait, sadly - may sometimes qualify as a legal excuse. There thus seems to be a tendency today towards less personal responsibility for blameworthy behaviour, which coexists uneasily with the trend of increasing toughness in penal practice. It is perfectly possible that our descendants, a few centuries from now, will look at our habit of punishing criminals with the same bewilderment as we look at the medieval practice of bringing non-human animals and inanimate objects to justice.

As I see it, retributivism makes two claims of central importance. One is conditional. It is the claim that punishment can only be justified as long as the person being punished deserves it. The other is substantive. It is the claim that people in general are personally responsible and hence capable of deserving praise and blame. This is a factual claim, and one which may turn out to be false. If wrongdoing is more like disease or

---

39 These disorders are official in the sense that they are included in the American Psychiatric Association’s official text, the DSM, which has a global influence over both psychiatry and public policy. See DSM-IV, APA 2000.
40 Delgado 1995, see also Hefferman & Kleinig 2000.
41 Scheffler 1992 has noted that there is a conflicting trend, however, which is the increasing tendency to treat issues of personal health and security as a personal responsibility.
42 Not all retributivist will accept this claim (see e.g. Corlett 2001 and Murphy 1995), but unless one subscribes to it, it is difficult to account for prevalent retributive arguments that punishment respects persons.
disorder than freely chosen acts, the notion of wrongdoers deserving punishment is fundamentally undermined. There are thus reasons of both practical and philosophical nature why an investigation of the merits and flaws of retributivism is timely and worthwhile.

1.5 MY ARGUMENT

Having illustrated some sticking-points in the theory of punishment, and having briefly reflected on the present standing of retributivism in academia and penal policy, it is time to present the argument I’m going to make. Recall that I started out this chapter by posing the following question:

(Q) Which principle or theory (or principles or theories) should serve as the basis for a state’s penal regime?

I shall defend the thesis that the retributive theory of punishment is a better answer to (Q) than rival theories. Retributivism ought to serve as the basis of a state’s penal regime. A penal regime ought to be retributive in character.

What reasons are there to accept this thesis? The reasons, it will become obvious, emanate from the claim that it is important to be given what one deserves. Retributivism, as I understand it, is the theory that desert is a necessary and (often) sufficient condition for justified punishment. Critics have objected to this in largely two ways: that desert is insufficient to establish that punishment is justified, and/or that desert does not exist. I shall refer to these objections, around which my argument will revolve, as:

(O1): The objection from insufficiency
and
(O2): The objection from non-existence.

My argument for answering (Q) along retributivist lines will emanate from an attempt to counter (O1). For to provide an answer to this objec-
tion is in effect also to provide reasons for why the state should design its practice of punishment in accordance with retributivism, i.e. according to principles of desert. It is to answer why a penal practice which gives rule breakers what they deserve indeed is justified. This admittedly depends on that we accept that if it turns out that some variant of the retributive theory is the best answer to (Q), then the rival answers are ultimately wrong. But this seems to follow from the fact that the rival answers are rival.43

I shall argue that there are two significant reasons in favour of the retributive answer to (Q), i.e. in favour of a retributive penal regime.44

1. *The institutional reason.* A retributive penal regime ensures that certain principles of fairness and justice are respected. These principles are: that no one should be punished without deserving it; that like cases should be treated like; and that there should be a proportionality between crime and punishment which is bound both upwards and downwards. These principles serve to exclude unfair inconsistency in punishment, excessive punishment, punishment of the innocent, etc. The rival theories of deterrence or rehabilitation do not seem able to respect this set of principles to the same extent.

2. *The symbolic reason.* A retributive penal regime engages in retributive punishment, or retribution. Retribution, unlike punishment-as-deterrence or rehabilitation, necessarily involves moral censure – it necessarily involves moral emotions of blame and re-

---

43 That is, I will throughout maintain that retributivism, deterrentism and rehabilitationalism present theories that are incompatible with one another: Retributivism differs from its rivals in its emphasis on desert, and rehabilitationalists and deterrentists are in disagreement about the proper way to arrange penal policy. When we say that (Q) should be answered along retributive lines we say that a retributive penal regime is justified. As a consequence of the incompatibility between the rival theories, this either means that the other theories do not represent justified penal regimes, or that they fail to provide as justified regimes as retributivism. My argument should be read these latter lines.

44 My debt to Morris 1968 is here undeniable. In fact, chapters 4-6 of this book are intended as elaborations of the views set out in this article.
sentiment. Why should this matter? Because of the tacit understanding of people inherent in moral censure. We do not blame or resent dogs, natural disasters, machines or infants when they cause harm or distress. We do resent reasonably mature human beings, however. Seen this way, to be blamed or resented actually affirms that you belong to the latter category. It affirms your status as a responsible moral agent; an equal who is held to expectations. If an elephant steps on my foot I shall regret my pain, but I will not hold it against the animal. The elephant could not properly be blamed, since it cannot be held to normative standards of behaviour (such as the norm that one should not step on others’ feet). It consequently does not matter if the elephant stepped on my foot accidentally or not. If you step on my foot, however, it matters a great deal if it was accidental or intentional. If it was intentional, I will resent you for it, and will perhaps believe that you deserve moral criticism for what you did. It is important to note that you should take this criticism of mine as an unequivocal compliment, even though it might be unpleasant. Quite frankly, it means that I do not consider you an elephant.

Retributivists have referred to this second reason as the respect for persons argument. Following the restrictiveness of moral emotions of blame and resentment (or praise and gratitude, for that matter), they have concluded that retribution is the only penal response that respects the criminal as a person. In morally censuring him or her, the criminal’s choices are respected as choices of an autonomous moral agent – an agent that could have and should have acted differently.

Why should we care about (1) and (2)? Why should we take them to justify retributivism as a guiding theory of penal practice? Well, the first reason stands and falls with the importance of aspects of the Rule of Law. If you find such principles as consistency and fairness between individ-

---

45 We could discuss whether or not an animal such as the elephant is capable of intentional action. It does seem obvious however that an elephant may step on your foot without knowing it, but also that it may do it in the process of attacking you. By saying that it does not matter whether the elephant steps on your foot accidentally or not, I intend that it makes no moral difference.
als unimportant, you can safely discard it. For those who find them important, however, it is to retributivism’s merit that it so coherently incorporates them. The second reason is perhaps more elusive. Why is respect for persons important? Such a question cuts to the very heart of a morality which may loosely be called Kantian. Respect for persons is important since it is our due as morally free beings (cf. Murphy 1995). We will have reason to return to this tricky notion at several stages in the next chapters.

Note that (1) or (2) are reasons which in principle anybody can accept. They are reasons even rule breakers can find important. Thus I shall explicitly argue that even those who face a penal response for having broken a rule may prefer retribution to other penal responses. They may prefer it (indirectly) since it is a consequence of a penal regime which in central aspects is superior to alternative regimes based on rehabilitation or deterrence, and they may prefer it (directly) since it respects them as persons. This boils down to the claim that they too would or should prefer a retributive penal regime over the rival regimes.

A further question regards from where the ultimate justification of (1) and (2) derives. Are the institutional and symbolic reasons ultimately important for consequentialist reasons? One prominent retributivist, James Rachels (1997), has argued that they may well be: a penal regime based on desert is ultimately justified simply because it leaves people better off than regimes based on other principles. I shall return to this difficult question throughout this book.46

A fundamental objection

Recall that I just said that there are two main objections to retributivism. One is that desert is insufficient to justify the practice of punishment. I will argue that (1) and (2) suffice to meet that objection. The other is (O2). According this objection, desert does not exist. In particular, as we have seen, some claim that retributivism relies on a “badness” paradigm which a scientific understanding of crime, most notable in the medicalization of wrongdoing, will prove incorrect. Having demonstrated why

46 I do not offer a separate discussion of it until the closing chapter 10, however.
retributivism is normatively more attractive than its rivals, I shall turn to
this question. Is it true that retributivism, and desert-based theories of
justice in general, turn out to be scientifically untenable? This is a factual
question, but it pertains to the normative question in so far that it threat-
en to undercut the very basis of my normative claims.

One can think of it this way: In discussing (O1) I make a positive
case in favour of a retributive answer to (Q). In effect I argue that there
are things – respectfulness towards criminals and procedural justice –
that a penal regime ought to accommodate. Since a retributive penal re-
gime accommodates these things better than its rivals, this counts in fa-
vour – should lead us to take a normative presumption in favour – of
retributivism. But positive normative claims such as this can be chal-
lenged on factual grounds. In the end, the normative claims can turn out
to be empirically untenable. If so, we should drop them. We cannot de-
defend normative theories whose necessary elements turn out to be fig-
ments or illusions. The world is under no obligation to be normatively
pleasant.

(O2) is a challenge to the positive normative claims I make in this
respect, in particular the symbolic reason and the weight I attach to it.
The objection says: “Even if you are right that retributivism in some
normative respects seems superior to rival theories, it matters little if
much of retributivism’s set of assumptions are false. A normative theory
cannot justify something on the basis of desert, for you cannot justify
something with nothing.” If (O2) is true we must reject retributivism, at
least as traditionally understood. The normative reasons for preferring
retributivism over its rivals, or more precisely for preferring a retributive
penal regime over other regimes, are then practically worthless.

A first thing to note here is that retributivism has no problem ac-
commodating excuses. If an offender has a good excuse, the theory on the
contrary implies that he or she should remain unpunished. But it is
unlikely that the theory can withstand any amount of excuses and still
seem worthwhile. The theory arguably relies on the assumption that of-
fenders by and large are personally responsible. Retributivism is not
shaken by the observations that some thieves suffer from a disorder –
kleptomania – which render them undeserving of punishment. It need
not even be shaken by the possibility of wife beaters being mentally dis-
ordered or victims of deeply entrenched gender structures. But if we
become able to explain more and more crime in such seemingly excusing
scientific terms, the theory will eventually break down. It will perhaps
seem, as Marx once said, as correct “in the abstract” but largely inapplicable in practice.\textsuperscript{47}

I shall argue that there is a clear sense in which some explanations of wrongdoing, if true, do exonerate from responsibility. In fact, it is possible that \textit{all} causal explanations have such an effect. This possibility of course presents us with the formidable problem of free will. Do people ever act in such a way that they can rightly be praised or blamed for their actions, i.e. in such a way that they are personally responsible for them? This question lies at the heart of what could be called the scientific objection against retributivism: that it presupposes a prescientific and false conception of human action.

Contrary to the present orthodoxy, I shall not try to solve the free will problem by so called \textit{compatibilist} analyses of such things as “free will” and “responsibility”. Throughout, I will accept an \textit{incompatibilist} view: the thesis of (hard) determinism does genuinely threaten some concepts and practices that are worth wanting, in particular those of praise and blame. Nor shall I “solve” the problem of free will by asserting that determinism is false. Instead, I opt for treating determinism as a thesis which presents us with a decision under uncertainty. Since we do not know whether determinism is true or false, moral and political considerations - the institutional and symbolic reason - counsel us to bet that it is false.

Returning from this ever more radical tour into deep philosophical problems to questions of penal policy, I conclude by sketching the make-up of a humane and just penal regime based on retributive principles.

\textit{The argument in context}

In dealing with questions about punishment and personal responsibility, I am in dialogue with two grand scientific discourses. On the one hand the philosophy of punishment, with its deep connections to general moral and political philosophy. On the other hand what we may loosely refer to as the problem of autonomous man, which concerns such ques-\textsuperscript{47} For Marx’s interesting views on punishment and retributivism, see chapter 10 and Murphy 1995.
tions as the potential truth of determinism and the existence of free will, and asks whether the traditional image we have of ourselves is fundamentally mistaken.48 The amount of work done on these topics – taken separately and in combination – is nothing short of staggering, and to get an overview of the field(s) is therefore very difficult. This reinforces the need to be humble in one’s claims. It is difficult to know when one is original and when one rehashes old arguments; when one adds to the overall body of work and when one merely repeats, however competently, what others have already said. I would nevertheless like to position myself in the debate, and point out some aspect of my argument which seem to me innovative.

With regard to my argumentation for retributivism, or a retributive answer to (Q), I rely on reasons – the “institutional” and “symbolic” as I call them – which by no means are new. However, my overall strategy is somewhat original, and maybe not as clean as the one employed by most retributivists. Whereas retributivists tend to argue that punishment of wrongdoers is an intrinsical good, I instead argue that, when it comes to criminal justice, there are things about being treated as one deserves which are attractive – so attractive, in fact, that all would or should prefer it to other modes of treatment. My argument in favour of retributivism can thus be said to take a partly consensual route – what makes retribution justified is that all would reasonably consent to it. Unlike much retributive thought, this argument does not necessarily rely on desert as a moral primitive. Furthermore, in trying to clarify a way in which libertarian free will can be foundational to both retributive punishment and democracy, I try to spell out what has previously mostly been merely hinted: that there is some connection between democratic rights to inclusion and the capacity to deserve punishment.

The clearest example of an innovative thing about my argument, however, seems to me to be the way I analyze the free will problem, or the problem of autonomous man. Contrary to the prevailing orthodoxy, I do not try to solve the problem by dissolving it through compatibilist analyses. Instead, I take what may be called a pragmatic route: given that determinism presents us with a choice under uncertainty (change society

48 All can intuit what a work in the philosophy of punishment can look like. “The problem of autonomous man” is probably a less clear topic. A particularly representative example of a book in this tradition is Pinker 2002.
or not without knowing whether the changes are factually motivated), I argue that we should bet that it is false. This argument - which by the way is not a call for the maintenance of useful illusions - is to my knowledge somewhat new.

***

But we have gone ahead of ourselves. Let us, in the next chapter, begin at the proverbial beginning by defining the concepts of rules, wrongdoing, and punishment.
The rest of the book is outlined as follows: In chapter 2, I offer some analysis and definitions of central concepts. I define “punishment” and “rules”, as I here understand them, and offer some remarks about the difference between wrongdoing and crime. I also clarify what I mean by “penal regime”.

In chapter 3 I proceed to describe and discuss the three dominant theories of punishment which I have previously mentioned: retributivism, deterrence and rehabilitation. Particular emphasis is placed on retributivism, which is in more need of elucidation than its rivals. I chisel out a non-strict, positive version retributivism, which states that desert constitutes a reason - but not always a conclusive one - to punish. The chapter then continues by asking why acting on desert is important, i.e. can be a reason for punishment. I cast this question in terms of the alleged “gap” of retributivism raised by David Dolinko (1991). I argue that grievance-based versions of retributivism are distorting, and that intrinsic-good retributivism may seem counter-intuitive. The question, then, is why we should care about desert as a principle of punishment; why we should make desert be the guiding principle of a state’s penal regime.

Part II of the book is devoted to crafting an argument for retributivism that manages to traverse Dolinko’s gap. I begin, in chapter 4, by examining Hegel’s (in)famous claim that (retributive) punishment is a criminal’s right. Why should anyone accept this claim? It is easy to see why victims, or law-abiding citizens in general, may want to punish rule breakers. But a right is usually something that the right-bearer would like to have. And an apprehended rule breaker would clearly like to forgo his or her “right” to punishment. I argue, however, that there are two reasons, one direct and one indirect, why all – including the rule breaker – has an interest in being retributively punished. These reasons are developed in the subsequent chapters 5 and 6.

Chapter 5 is devoted to the analysis of the indirect argument in favour of being punished, which I refer to as the institutional reason. The reasoning is as follows: among the penal regimes considered in this book, one based on retribution seems most able to adhere to central principles
of justice. I here follow James Rachels, who has provided a small but compelling test to support precisely that conclusion. The upshot of Rachels’s test is that retribution as a penal aim coheres better with important principles of justice than its rivals – it is the only penal aim (in the test) which at the same time accounts for principles of guilt, proportionality, consistency and excuses.

Why should this count as a reason for the importance of being retributively punished? It is an indirect reason for it: if there are reasons to prefer a retributive penal regime to other regimes, it is only fair that one accepts one’s own punishment under that regime, should one culpably break a rule. One cannot coherently enjoy the benefits of a just penal regime without having to face the consequences of that regime for oneself, should one qualify for them.

In chapter 6, I develop a more direct and controversial reason for the importance of being retributively punished. I call this reason the symbolic reason. The basic idea here is that when one is retributively punished, one is affirmed as an autonomous individual. Being held responsible for a thing (and this encompasses both good and bad things) is a compliment. It means that one is a person - the kind of being whose choices warrant respect, and whose actions are sometimes appropriately praised and blamed. Examining the models of agency underlying the penal theories of retributivism, deterrence and rehabilitation, I argue that only a retributive penal regime is compatible with full respect for persons. So long as one wants that respect, there is a completely clear way in which one may claim to have a Hegelian right to punishment – unless I am blamed for my wrongdoing, I am either seen as a rational agent whose actions need to be changed with the prudent use of incentives or, worse, that my actions are the result of disease or disorder, and largely beyond moral assessment.

So far the argument in favour of a retributive penal regime. In part III of the book, I proceed to address a fundamental objection to the theory. This objection has to do with the existence of desert. What if, as hinted in the present chapter, a scientific conception of crime and criminals is incompatible with a retributive conception of “badness”? In chapter 7 I outline a “theory of excuses” – an account of why most scientific explanations indeed seem to threaten personal responsibility, which is a prerequisite for desert. I argue that since most scientific explanations undermine the notion of alternative possibilities and/or explain by citing factors beyond the individual’s control, there is a clear sense in which a
scientific explanation of crime (if true) indeed seems excusing. I offer some examples of different types of explanations and speculate somewhat about whether it is psychologically realistic to conjecture that an increased ability to explain prima facie blameworthy behaviour, or at least believe that one has such an ability, means that one is more willing to excuse.

Chapter 8 is devoted to the presentation and defence of the position which seems to be the end product of the reasoning in chapter 7: the thesis of hard determinism. If all behaviour can be scientifically explained, and if all scientific explanations are incompatible with personal responsibility, then personal responsibility for behaviour does not exist. Hard determinism is the thesis that advocates such a conclusion. I defend hard determinism against objections from compatibilism, probabilism and explanatory incompatibilism, and try to tease out its moral and political consequences. In particular, hard determinism seems to rule out retributivism.

Having defended hard determinism from the objections that are often levelled against it, I proceed in chapter 9 to offer a rejection of it. Hard determinism, I suggest, is a thesis which suggests that we take certain action. It suggests that we change society in significant respects. But since we do not know whether determinism is true or not, making those changes or not is a decision under uncertainty. I argue that since there are important values to lose, we are justified in betting that determinism is false.

In chapter 10, I offer some conclusions and sketch the outlines of a just penal regime. The thrust of the argument made there is that since we often are epistemologically uncertain about questions of responsibility, the benefits of a retributive regime should counsel us to treat them conservatively. The burden of proof, it is held, should fall on those who argue that some action species of wrongdoing is to be excused, not punished. This is because of the values inherent in (retributive) punishment.
2. Rules, Wrongs and Punishment

"But that is not it I intend to speak of here; my design being not to show what is law here, and there; but what is law; as Plato, Aristotle, Cicero, and divers others have done, without taking upon them the profession of the study of the law.” (Hobbes 1998: 175)

PREVIOUSLY said that this book is intended as an answer to (Q), which was understood the question concerning which principle should serve as the basis of a state’s penal regime. I will argue that there is a version of retributivism that satisfactorily answers that question – answers it better, at any rate, than competing theories of punishment.

The purpose of the present chapter is to provide the theoretical and conceptual background of (Q). I will clarify what I mean by rules, wrongdoing and punishment; address the relationship between law and morality; and describe the concept of a penal regime.

2.1 RULES AND RULE BREAKING

In the previous chapter, I said in passing that this is not an abolitionist book. I presupposed that there should be such a thing as a penal regime, though the nature of its justification is unclear. Why would one presuppose this?

First of all, everyone, with the exception of the odd anarchist, accepts that every society needs rules of behaviour; rules which prohibit or
render obligatory certain action. The legitimacy of such rules as a general phenomenon is, when we take the quickest glance at political theory, normatively overdetermined. Among the thing said to be ensured by rules by various authors are order, trust, moral excellence, calculability, cooperation, happiness, and social cohesion. Whether all of this other things equal are ensured by behavioural rules is questionable. But even if rules only ensure a fraction of the good things political theorists say they ensure, we can be confident that rules can be morally justified. They also seem to be a social given. All societies seem to have rules of some sort.

A “rule” is here understood as a technical term, signifying only a norm, formal or informal, that some action be taken or not taken. The most familiar form of a rule is that of a law, a formal, public and common rule, the breaking of which is associated with legal punishment. But other and less tangible rules, such as informal rules of etiquette, are structurally similar to laws, and as such pose no fundamentally different questions than laws. It is in principle the same thing to say “Do not murder” or “Pay your taxes” as it is to say “Do not burp in public” or “Thank the host after dinner”. Both classes of statements refer to a standard of behaviour supposedly in force, and both naturally suggest that sanctions may follow failure to live up to the standard. Though I shall treat laws in this book, I shall for the most part use the term “rules” instead, since doing so makes argument easier to generalize to other cases of rules (and rule breaking) than public laws and crimes, should one wish to do so.

49 The kind of rules focused on here are of the command and prohibit type, classically referred to by Hart as primary rules (Hart 1961). These are to be distinguished from secondary rules, which concern ways to “introduce new rules of the primary type, extinguish or modify old ones, or in various ways determine their incidence or control their operations” (ibid. 79). A secondary rule, then, might be that an elected majority is conferred a right to make primary rules.

50 Influential treatments of the benefits of common rules are Hobbes 1998; Locke 1998.

51 See Dahl 1989. On a theoretical note, and contrary to Madison’s old notion that angels have no need for government, it is plausible to say that even a society of morally perfect people is in need of government (and rules). See Kavka 1995.

52 A more aesthetic reason is that I like the ring of “rules” and “rule breaking” better than other expressions. It is worth mentioning, however, that some maintain that there are analytical differences between rules and laws. Heywood, for instance, argues that laws differ from other social rules in that they (a) are made
A second reason to presuppose a penal regime is that if there are rules, there will invariably be people who do not follow them. Thus, a society not only has rules of behaviour, and justifiably so. It also faces the problem of those rules being, for whatever reasons, broken. I shall sometimes refer to these instances of rule breaking as “crimes” or “offences” or “lawbreaking”. Those who break the rules shall be referred to as “rule breakers”, “criminals”, “offenders” or “lawbreakers”.

A society has rules which will sometimes be broken. This fact raises the question of how to handle rule breaking and uphold the rules - for if the rules are worthwhile, it is other things equal better if they are complied with. One option is to leave it to each of us, as individuals. This does not seem promising. Political theory instead overwhelmingly recommends that the state – the overarching political authority in a society – should be entrusted the task of upholding the rules. A classical thought-experiment can be conducted in defence of that claim. Imagine individuals in a state of nature, i.e. individuals without a common power. In such a situation – supposing that there can be rules at all - the business of handling rule breaking would be up to each of us. Suppose that A breaks a rule and thereby harms B. In the absence of a common power, anyone and everyone may respond to A’s crime. This includes B. But B, as a harmed party, is likely to exact too severe a punishment on A. This in turn would lead A to get back at B whenever he or she has the opportunity. Spirals of revenge and counter-revenge would be set off.53 So even if individuals could agree on common rules in the state of nature – political theory’s version of the bogeyman – they would do a poor job upholding them. It is better if the state is entrusted the task of handling rule breaking. In fact, one of the great recommendations of a common power is that it, apart from facilitating social life by providing common rules, can uphold the rules in a more disinterested, reliable and restrictive manner.54 It

___

by the state and trumps other rules; (b) are compulsory; (c) are public to all; and finally (d) are binding to all. (Heywood 2004: 153). Note that with the exception of (a), one may well find that other rules meet criteria (b), (c) and (d) – just think of a family which adheres to public, binding and compulsory rules.

53 This theme is most famously developed by Locke 1998.

54 The prime example of this type of argument is Hobbes. According to him, the state is justified by defeating the various problems of collective action that be-devil us in the state of nature. The reasoning above is of course heavily influ-
can ideally promote compliance with the rules without going to the excesses that would bedevil a situation where private individuals are both the upholders of rules and interested parties.

There are basically two ways for an authority such as the state to promote rule following: to reward compliance and/or punish incompliance. Existing states seem to overwhelmingly opt for the latter.

2.2 PUNISHMENT, WHAT?

What is punishment? We can all pretty much intuit an answer to that question. Punishment is the thing that happens to those who break the rules and get caught. Today, its most familiar manifestations are material deprivation (fines) and restriction of freedom (imprisonment). This may seem all there is to it. But we get a better grasp of the concept if we break it down more carefully.

Punishment is, first of all, obviously something unpleasant for the person suffering it, or at least intended as such. Punishment is about imposing pain or deprivation on someone.\(^{55}\) Secondly, punishment is a species of intentionally imposed pain and deprivation. You cannot inadvertently punish someone, for instance; punishment is deliberate – it is intended to produce displeasure. Third, punishment is normally thought to be reserved for some sort of authority. The state punishing citizens is the paradigm example, but one may also speak of employers punishing em-
ployees, parents punishing children, and teachers punishing students. A subject may perfectly well intentionally impose pain or deprivation on an authority, but lacking the authority to make the rules in their relationship, we would hardly call this punishment. Fourthly, and connected to this, punishment is pain and deprivation imposed for some offence against a rule. Suppose you intensely dislike getting up early in the morning and that your employer – an authority, we assume - requires this of you. Even though the discomfort you experience is intentionally caused by an authority it is not punishment, because it is not inflicted for breaking a rule. This “for”-clause does not mean that it is impossible to punish the wrong person, however. In cases where we punish an innocent by mistake, we have simply intentionally inflicted pain or deprivation on a supposed offender rather than the real one.

Taking these clarifications into account, we could define punishment as:

Punishment\textsubscript{det}: pain or deprivation intentionally inflicted on a person by an authority for an offence.

Some things should be noted about this rudimentary definition.\textsuperscript{56} As we shall see in chapter 3, it is questionable if what we refer to as rehabilitation of offenders can be understood as punishment so conceived. According to the theory of rehabilitation, penal responses are intended to better those subjected to them, which puts the entire idea of the penal response as harm in question. Rehabilitation is rather to be seen as good for you, or at least intended as such. Sometimes this goal may be

\textsuperscript{56} Some other definitions of punishment: “an authority’s infliction of a penalty, something intended to cause distress or deprivation, on an offender or someone else found to have committed an offence, an action prohibited by the law” (Honderich 2006: 15); “a penalty inflicted on a person for a crime or offence” (Heywood 2004: 169); “the imposition of something which is intended to be burdensome or painful, on a supposed offender for a supposed crime, by a person or body who claims to have authority to do so” (Duff 2004). One thing left unclear in these definitions, as well as in mine, is whether only individuals can be persons, or if it makes sense to treat collectives (such as corporations) as persons in the eyes of the law. I shall not have anything to say about this (fascinating) subject. I will assume that “person” means “individual human being”.

pursued by imposing pain or deprivation on you, but sometimes it may be pursued without it. The latter cases, although being interferences in your life because of some offence, clearly cannot be conceived as punishment. I shall instead use the term *treatment* for these cases. In the context of penal policy, treatment will be understood in the following way:

**Treatment** \_def: steps taken against a rule breaker, with or without his or her consent and without intending to cause pain or deprivation, to address the disorder(s) that led to or facilitated the offence on the grounds that doing so is for his or her own good.

Treatment thus refers to penal responses which do not intend to cause pain or deprivation and are intended to work to the benefit of the individual.57

Furthermore, a retributive definition of punishment would perhaps add that punishment is the intentional infliction of harm or deprivation on a person for an offence *for which he or she is personally responsible*. According to retributivism, proper punishment necessarily involves moral censure, and moral censure arguably cannot exist without responsibility. Thus, according to a retributive conception of punishment, we for example cannot properly say that we *punish* a dog, even though we may intentionally inflict pain and deprivation on it for some “offence” (e.g. we may correct a dog for wetting the carpet). Punishment, says the retributivist, requires not only that a rule has been broken. It also requires that the offence is an instance of wrongdoing.

These points of contention will be made clear as we go along. The original definition above suffices, for now, to grasp the intuitively quite familiar concept of punishment.

---

57 An example of treatment could be to coercively help an addicted offender to overcome his or her addiction at a clinic. In such cases, the person subjected to the treatment may feel harmed. But the treatment is not intended to harm, and does not necessarily involve pain and deprivation. Note that imprisonment, which often is intended to harm, will not be felt as harm by everyone either. A difficult question is what weight to give the experience of the punished person. Here it suffices to say that the *intentions* of punishment and treatment are quite different.
2.3 CRIME AND WRONGDOING

What kind of actions should society prohibit and, if performed, punish? A plausible answer is that actions that are harmful, offensive, or otherwise damaging to society should be prohibited. Very roughly, we could imagine such actions as criminal because they are morally wrong. This directs our attention to the difficult question concerning the relationship between wrongdoing and crime.

Any plausible account of law and morality will hold that there are salient differences between crime and wrongdoing. First of all, some acts are wrong even in the absence of rules saying that they are wrong. Though some legal positivists will disagree, I find this point obvious and won’t elaborate on it further. Secondly, some acts may be wrong in spite of being legally permitted. We can divide this class into two subclasses: permitted acts that ought to be prohibited and permitted acts that ought not to be.

This leads us to an important, if often overlooked, problem in the theory of punishment and criminal justice: the problem of how to distinguish between actions that should be illegal, and thus liable to be punished, and actions that should be legal. How can we draw the line? Possible answers are that society should outlaw that which is harmful or offensive to others, or that which is morally repugnant. But there are clearly examples of harmful or otherwise morally blameworthy acts that are not illegal, and which we would prefer to keep that way. Take lying to a friend as an example. We could agree that such an act is prima facie blameworthy, and that the friend may well be emotionally harmed by the lie. But we would presumably also agree that it is not an act that should

---

58 Hobbes 1998 is an extreme statement of legal positivism: right and wrong does not exist prior to rules. Such a view seems unacceptable, as it appears to lead to saying that might makes right – an action is morally wrong only as long as someone with sufficient power has defined it as morally wrong. If we follow that line of reasoning to its grizzly conclusion, we seem forced to say that seemingly repugnant political practices (such as slavery) was not morally wrong until it was prohibited, and that it remained right as long as it was legal. The natural law approach espoused in this book is also problematic, but less so than a positive law approach (cf. Heywood 2004: 156-9).
be criminalized and within the scope of punishment.\textsuperscript{59} At the same time, we are very confident that it makes sense to criminalize murder, assault, rape, theft, and the other usual suspects. But to draw a neat line between what should be considered criminal and what should considered be legally permitted wrongdoing is never going to be easy.

Note however that there are also acts that are wrong exactly because of the fact that they are criminal. A fancier way of saying this is that they are \textit{malum prohibitum} offences.\textsuperscript{60} A paradigmatic example is driving on the wrong side of the road. It makes no principled difference if a society chooses to drive on the left or the right of the road. It must, however, choose one of the two. Suppose the state chooses to make driving on the right the norm. Driving on the left now means breaking the rules, and it is a wrong precisely because it is against the rules, not because the act is wrong in itself. A person driving on the left does wrong because he or she fails to respect what was agreed upon or decided.\textsuperscript{61} The act in itself, prior to the decision, is morally neutral.

The most important thing to note here, however, is that many \textit{criminal acts are not wrong at all and should not be punished}. First of all, rules may be inherently unjust or simply obsolete. Breaking such rules does not warrant punishment. We would not say, for instance, that dissidence against racist Apartheid-laws is a wrong that should be punished, or that there is a legitimate interest in upholding such laws. Moreover, there may be cases where it is right to break a rule which in general is legiti-

\textsuperscript{59} What would the punishment be for, say, lying about the beauty of a friend’s sweater?

\textsuperscript{60} The label \textit{malum prohibitum} applies to offences that are immoral exactly because they violate a law. \textit{Malum in se}, on the other hand, applies to acts that would be wrong even in the absence of legal prohibition – murder, for instance (see Dolinko 1991).

\textsuperscript{61} He or she of course also does wrong because breaking the rules here means endangering other motorists: we presumably have a prima facie natural duty not to harm or endanger others. Note, however, that most hold that we have a natural duty to uphold just institutions, and that driving on the wrong side of the road therefore may be construed as a natural wrong even if it does not endanger others. It would be the wrong of not restraining one’s behaviour in light of existing legitimate agreements in place. For a helpful discussion of these matters, see Bird 2006. Of course, it may also be that we should respect a collective decision even in the absence of natural duties to do so.
mate. There may be cases where it is right to, say, steal another’s property even though the laws prohibiting theft in general should be followed. Saying that laws should be obeyed does not mean that they have to be slavishly and invariably obeyed.

2.4 IS CRIME SOCIALLY CONSTRUCTED?

The view set forth so far assumes that there are some actions that are wrong in themselves (wrong even in the absence of rules saying that they are wrong) and which should be, if of sufficient magnitude, criminalized. This assumption, which says that crimes represent violations of natural duties, may be questioned.

Some suggest that crime is socially constructed. What counts as crime is subject to a number of contingent factors. Criminalization of an act normally requires that it is seen as morally wrong or otherwise problematic. But what is seen as wrong or otherwise problematic, they argue, is highly contingent. The upshot of this view is that crime is a volatile and shifting category. What one society, in one era, judges to be criminal and liable to be punished might perfectly well be legal in another.

Inspired by this line of thought, Nils Christie offers the following advice on how to think about crime to fellow social scientists:

“Crime does not exist. Only acts exist, acts often given different meanings within various social frameworks. Acts, and the meaning given them, are our data. Our challenge is to follow the destiny of acts through the universe of meanings. Particularly, what are the social conditions that encourage or prevent giving the acts the meaning of being crime?” (Christie 2004, p. 3)

The idea here is presumably this: since crime depends on labelling (declaring such-and-such acts as legal and such-and-such acts as illegal), we should abandon the belief that there is some natural category of acts that are criminal, and instead focus on the processes of criminalization and
decriminalization. No act is criminal. Acts are only criminal if they have been framed as such.

To focus on such labelling fits well with functionalist analyses of crime, such as Durkheim’s (1964). According to Durkheim, laws express central moral beliefs of a society. The purpose of the practice of punishment, he argues, is to give members of society a valuable opportunity to express outrage at behaviour that fails to conform to these beliefs, and thus to rally around the moral code they believe is correct or appropriate. The use of punishment is therefore not premised on specific actions being done. According to Durkheim, it is not as if crime represents a determinate set of actions, in the absence of which none would be punished. It is not as if a society in which no one stole, murdered, assaulted, slandered, evaded tax, and so forth, would stop punishing people altogether. Since punishment in Durkheim’s opinion exists to promote social cohesion, it is a practice without a stable referent. It is not premised on a specific list of actions, determined by some purportedly objective criterion such as harm (cf. Fischer 2006), being performed. A society that managed to eradicate crime as we know it would instead criminalize things that are presently non-criminal. It is possible, he argues, that a society can react with the same outrage against trivial offences as we do against severe offences. Durkheim writes:

“Imagine a society of saints in an exemplary and perfect monastery. In it crime as such will be unknown, but faults that appear venial to the ordinary person will arouse the same scandal as does normal crime in ordinary consciences. If therefore that community has the power to judge and punish, it will term such acts criminal and deal with them as such” (Durkheim 1982: 100).

---

62 Typically, social constructivist approaches to crime will add that the ability to criminalize and decriminalize is reserved for the powerful in society. See Conrad & Schneider 1992.

63 According to Hacking 2001 the great appeal of social constructivism (or constructionism) is precisely that it suggests that things could have been otherwise; that things are not fixed and immutable. Note, however, that it is a fallacy to assume that one is committed to essentialism with regards to crime just because one believes that there, speaking with Christie, “is” such a thing as a criminal act.
The upshot of Durkheim’s argument is that “we must not say that an action shocks the common conscience because it is criminal, but rather that it is criminal because it shocks the common conscience” (ibid. 81). This suggests that the very concept of crime is highly contingent. In promoting social cohesion a certain amount of crime is **valuable**, and in the unlikely event crime were to disappear, it would have to be “reinvented” - even if this means criminalizing insignificant “offences” such as eating sloppily at a dinner party.

In one sense this “labelling thesis” is trivial. It is obvious that crime can only exist when some actions are defined as criminal, and it is plausible that some crimes represent moral attitudes that may well be contingent. The most blatant examples of behaviours that used to be legal but are now widely considered illegal are perhaps those which concern racism. For much of its colonial past, European discourse on the differences between peoples was to a ludicrous extent steeped in ideas about the hierarchy of races, and Europeans seem to have habitually thought of themselves as biologically and culturally superior to others. In such an intellectual climate, it should not surprise us that no rule against derogatory speech towards other ethnic groups was enacted. Thus, under the entry “Negroes” (Negrer) in the 1913 edition of the widely distributed Swedish encyclopaedia Nordisk Familjebok (“Nordic Family-book”), one could read remarks such as: “In mental respects the negro in general can be placed on level with the child, with little talent and a vivid imagination”, and that he “in general appears shy, suspicious, and lacks every trace of mercy or compassion”. Moreover, from a moral perspective the “negro” could “sooner be said to lack morals than being immoral” (Nordisk familjebok, 1913, volume 19, p. 729). Today these appalling statements would probably be punishable on grounds that they breach laws against incitement of racial hatred, which notably state that it is forbidden to express contempt against ethnic groups.64 And even if the author could get away unpunished with his entry in the encyclopaedia, the discourse has changed so that his views today strike us as very much blameworthy, whereas they were not at the beginning of the 20th century. This suffices to demonstrate that the class of legally punishable or otherwise blameworthy actions need not be constant over time.

---

64 In Sweden, that law is “Lagen om hets mot folkgrupp”, Brottsbalken, chapter 16, §8.
What is the relevance of examples of contingency such as this? Well, they can put the whole notion of justifying punishment into doubt. In colonial times, it would have seemed ludicrous to say that racial slurs merit punishment, or that they should be deterred through punishment, or that racists are in need of rehabilitation. It was a behaviour which was neither seen as wrong, nor criminal. But what if all crime is subject to the same contingency? Chances are that, a hundred years from now, it will seem as ludicrous to punish theft – a practice whose justifiability we take for granted – as it did punishing racial slurs in the early 20th century. A certain relativism is introduced by the labelling thesis, and it can seem to devalue the whole idea of justifying punishment.

We could reply to the labelling thesis in a number of ways. First of all, even if crime is contingent, we are still faced, for any given time and place, with the question how to conceive crime and justify punishment. We need not conclude that anything goes simply because standards of right and wrong may be shifting. Secondly, and more importantly, there is surely a class of actions which not only seem prima facie wrong across time and between societies, but which also seem to have been invariably criminalized. Even Foucault, champion of contingency, has admitted this: “the division between the permitted and the forbidden has preserved a certain constancy from one century to another”, he writes (Foucault 1995: 17). The obvious examples of such “constancy” are crimes against person and property: murder, assault, robbery, theft, etc.

For the purposes of this book, I will assume that there are some actions that it makes sense to criminalize, and that breaking the rules against them constitutes prima facie wrongs. The nature of those rules, I will further assume, is such that all can see that they are worthwhile – including those who break them. In short, I assume that we analyze punishment against the backdrop of societies in which there is at least one rule $R$, such that $R$ is seen as legitimate by all, and that breaking $R$ is seen as a prima facie blameworthy act by all. I do not think that this assumption is terribly problematic. All can agree that a rule against murder is legitimate, for instance, and that it is prima facie wrong to break it. I believe even a murderer can accept this (even though he or she may feel that his or her own particular crime is not blameworthy). For the remainder of this book, when I talk about some punishment possibly being justified for some offence, I assume that it is a type $R$ rule that has been broken.
How many type $R$ rules there are in actual societies is of course up for debate. I will tacitly assume a society or societies in which there is at least one. The reader is free to imagine whichever scope of the set of rules like $R$. To some, it may contain only the essential core - rules against violence and theft, for instance. Others may envision the set to contain virtually all the laws in force in their own particular society, right down to menial rules such as those against jaywalking, illegal parking or disorderly conduct. The amount of type $R$ rules makes no substantial difference to my argument. My justification of the practice of punishment should only be read as pertaining to the punishment for violation of rules which are of the kind above, regardless of how few or how many. If one says that there are very few crimes which the state justifiably can punish, one still has to answer what it is that makes punishment right in the few cases it is right. And this - the nature of the justification of the practice of punishment - is what I am interested in here.

Why analyze punishment against the backdrop of rules that are perceived as legitimate by all? First of all, I would not like to be read as one who tries to justify the punishment of any crime. Some crimes should be encouraged - crimes against inherently unjust laws, for instance - and it would be odd to offer a justification of sanctions carried out against them. Secondly, and more relevantly, separate and very difficult problems are added to the justification of punishment once we allow that different groups or individuals may have fundamentally opposed ideas about permissible and impermissible conduct. In pluralistic societies characterized by genuine disagreement, any justification of penal practice must address ways to handle the fact that fundamentally different conceptions of the good are in place, and the consequences this has for the analysis of authority, legitimacy and so forth (see e.g. Bird 2006: chs 1, 11). I do not suggest that this challenge cannot be met. But it is not the problem I want to discuss here. Therefore, I restrict my argument to rules like $R$, of which I assume there is at least one - an “overlapping consen-
sus”-approach with regards to criminal justice, if you will.65

---

65 For the strategy of pursuing an overlapping consensus to handle genuine disagreement, i.e. disagreement which is not merely semantic, see Rawls 1993.
2.5 PENAL REGIME, WHAT?

So far I have described my views on rules, rule breaking, punishment, the relationship between rule breaking and wrongdoing, and the notion that crime may be socially constructed. Take again (Q):

(Q) Which principle or theory (or principles or theories) should serve as the basis for a state’s penal regime?

Based on what I have said so far, I understand this question to concern punishment of crimes against legal rules that all can agree should exist and be upheld. I assume that there are some actions that legitimately can be criminalized, and that it other things equal is better the fewer of those actions are carried out. Obvious examples would be assault, murder, or theft. It is of course possible that there are particular situations where assault is on the whole praiseworthy, and it is also possible that such acts should go unpunished. But taken as a general thing, all can surely agree that it makes sense for society outlaw assault and desire that as few people as possible are assaulted.

But (Q) concerns specifically a penal regime. What does this term stand for? Theorists of criminal justice usually focus on punishment: they wonder what justification, if any, there is for the practice of inflicting pain or deprivation on people who break legitimate rules. By speaking of a penal regime rather than simply punishment, I intend to widen the focus slightly. A penal regime, as I understand it here, is the whole set of formal practices and rules the state employs with respect to lawbreaking. Punishment, and the reasons for inflicting it, takes centre stage in such a regime. But a penal regime also specifies when not to punish. It includes criteria for when to hold responsible and when to excuse, as well as considerations on the forms and purposes of penal responses. The theories I shall present as potential answers to (Q) in the next chapter are theories of punishment. But they are also theories of penal regimes. They represent different ways for the state to organize the practice of responding to rule breaking – including different ways for the state to intervene or not intervene in the life of the individual offender.

What is the scope of (Q)? It includes a reference to a generic “state”. Do I have any particular state in mind? No. The argument is intended to apply to all states that fulfil the criteria set out above: they must install
and uphold rules that are legitimate, binding to all, and perceived as worthwhile by all. I shall predominantly assume that such states are democratic in character, for the assumption that some rules will be seen as binding for all works best if all, as citizens, may be construed as co-legislators of the rules. And some of my more specific arguments – that the model of rule breaking inherent in retributivism is more consistent with democratic principles than the rival theories’ models, for instance – only makes sense if democratic principles are taken to be valuable.

Other than that, (Q) is not restricted to any particular state. Some of my empirical examples will be taken from Swedish penal policy, and much of the theoretical views from Anglo-American philosophy, but the argument is intended to generalize.

***

Society has worthwhile rules. To break such rules is prima facie morally wrong. Yet for various reasons rules will, even if all can agree that they are worthwhile, be broken. The state is entrusted with the task of acting against such rule breaking. A penal regime is the forms and nature of the state’s responses to rule breaking.

According to which principle or theory should the penal regime be designed? Three theories of punishment (which in effect also represent three different penal regimes) will be entertained in this book: deterrentism, rehabilitationalism and retributivism. The next chapter is devoted to presenting these theories in detail – in particular retributivism, the elusive and often misunderstood theory defended in this book.

---

66 A fundamental tenet of democracy is that one should be given the opportunity to affect the rules which one is expected to follow; see e.g. Dahl 1989. Note that rules which in themselves are legitimate may not be binding if the state enforcing the rules is inherently illegitimate. Take rules against theft in a dictatorship as an example. Though a rule prohibiting theft in general is worthwhile, and one all can see as binding, it may still be that it is illegitimate because the dictatorial state is illegitimate in turn. (In particular, one can question whether such a state has a right to punish). However, there is nothing in principle that rules out that legitimate dictatorships, if there can be such things, may have the right to punish violations of worthwhile rules. Furthermore, it may be that even an illegitimate state retains some right to punish. I will return to these questions in chapter 10.

67 I make this argument in chapters 6 and 8.
3. Three Theories of Punishment

“Today it is impossible to say for certain why people are really punished”
(Nietzsche 1998: 60)

Given that there are legitimate rules, what ways are there to arrange and justify taking penal action against those who break them? As we have seen, most seem convinced that the practice of interfering punitively in the lives of offenders is somehow justified. But the nature of that justification is elusive, and proposals consequently abound.

The justificatory principles of punishment mainly discussed in the literature are retribution, deterrence, and rehabilitation (see e.g. Rachels 1997; Murphy 2003; Heywood 2004). As we have seen, they are also the ones I shall consider in this book. It is important to note that even though these theories are the most important and in my mind compelling ones, they are by no means exhaustive of the literature on the justification of punishment. There are other ideas on what it is that justifies the practice of punishment out there. This fact has consequences for my conclusion.

---

68 Here follows a rough overview of what seems to me the most important other positions in the philosophy of punishment: Sometimes incapacitation is mentioned as an independent justification (punishment is justified since it prevents criminals from committing further crimes), and punishment is sometimes also believed to be about restitution, or the paying back to the victim the criminal’s advantage (Tamburrini 1996). Hampton has advanced a moral education theory of punishment, according to which punishment is justified in that it communicates to the wrongdoer that what he or she did was morally unacceptable (Hampton 1984). Recently, restoration has been suggested as a justification of punishment. According to this idea, punishment should aim at promoting forgiveness and closure, for instance by arranging deliberative meetings between criminals and
By design, I shall only be able to say which of the examined theories that seems on the whole preferable. I will remain ignorant about whether there is an unexamined theory out there, which performs even better.

I shall in this chapter present the theories of deterrence, rehabilitation and retribution more closely. Some general remarks are in order before we go into their details. The aim of this book is to answer a deceptively simple question – according to which principle or theory should a just penal regime be designed? Some may well feel that this question is too imprecise, clumps too many things together. Hart (1968), the great jurisprudent, for instance held that we must ask about the practice of punishment not only what its “general justifying aim” is, but also answer the (purportedly separate) questions about whom to punish, and how.69

More recently Dolinko (1991) has suggested that the “justification” of punishment must be broken down into its rational and moral justification, the former being about what goals we are trying to promote by punishing and the latter about whether or not we are morally right to do so.70 As will become obvious, I shall not make too much of these distinctions. Deterrence, rehabilitation, and retribution in my opinion represents comprehensive candidates for answering (Q): each of the theories seem to

victims (see Duff 2004; Braithwaite 2002). Morris has defended a paternalistic theory of punishment, which holds that punishment should aim at improving offenders without thereby violating their status as persons (Morris 1981). Bedau has sketched the outlines of a liberal theory of punishment, which seems to originate in part from collective action research, in part from ideas on procedural fairness (Bedau 2003). Reform, finally, is sometimes intended as something different from rehabilitation (see Raynor & Robinson 2005). I will use reform and rehabilitation interchangeably, even though it should be noted that reform may be able to incorporate responsibility to a higher degree than rehabilitation, which in a clearer sense implies that the criminal in some sense is disordered.

69 Hart’s own idea is that the first of these questions should be answered along utilitarian lines, whereas retributivism can inform answers on the remaining two. We will return to this “hybrid” or “mixed” theory in chapter 5.

70 The distinction has been suggested for quite specific purposes by Dolinko. He distinguishes between bold and modest retributivism. The latter is the view that while we rationally justify punishment in terms of social protection, it is desert that morally justifies punishment. (Bold retributivism, on the other hand, is the view that desert is both the rational and moral justification of punishment). I will return to whether bold retributivism is a plausible doctrine below.
offer both rational and moral justifications of punishment, and each answers, in its own way, the questions set forth by Hart. The easiest way to think about them is this: the theories are primarily morally justificatory ones. Each theory presents an account of what it is that makes punishment right. But by the nature of that account, each naturally suggests a particular penal regime.

In describing the theories here I shall be quite brief and keep critical argument to a minimum – their strengths and weaknesses will be examined in greater detail in later chapters. An exception on both counts will be made for the oft-misunderstood and controversial theory of retributivism, which is the theory I intend to defend in this book. I offer a lengthier description of retributivism than the other theories, and respond to some criticism against the theory that seems to me misconceived. The chapter closes, after having discarded the ideas of mixing the theories into a composite penal regime, by identifying the weaknesses of retributivism that will be addressed in the remainder of the book.

3.1 DETERRENTISM

Deterrence theory, or “deterrentism”, claims that what justifies the punishment of a rule breaker is its capacity to deter future rule breaking. One usually distinguishes between two forms of deterrence: special and general. Special deterrence stands for the notion that a rule breaker, when being punished, is less likely to break further rules in the future. By imposing some pain or deprivation on the rule breaker, he or she will think twice before running the risk of further punishment. General deterrence is the same logic applied to other people. If you and I can see that rule breaking gets people punished, we will be less inclined to break the rules.

Deterrence has probably always been considered a justificatory aim of punishment. It is hard to imagine anyone considering the usefulness of punishment entirely detached from its capacity to deter crime. Deterrence is today often taken to be a distinctively utilitarian approach to punishment, but it is important to note that one need not be a utilitarian in order to advocate it (Honderich 2006). However, it was not until the emergence of utilitarianism that deterrence was fully worked out as an
independent and coherent justificatory theory of punishment, and I shall here present it in utilitarian terms.

**Utility**

Utilitarianism is the theory which states that an action (or rule, institution, etc.) is right insofar as it leads to greater total utility than any alternative. It is the prime example of a consequentialist moral theory, i.e. a theory which holds that it is only the consequences of X that determines the rightness or wrongness of X. *Utility* – the good utilitarians seeks to maximize - may be understood in various ways. Traditionally, it has been given a hedonistic meaning. Utility and disutility on this understanding refers to mind-states of pleasure and pain. Others have conceived utility as desire fulfilment or preference satisfaction. Today, many have settled for the rather inclusive label "welfare", which is the term I will use interchangeably with utility. Welfare refers to the thing or things which increase a person’s well-being.71 The more welfare a person has, the better off he or she is. A utilitarian will say that a morally right action is one which, among alternatives, leads to the greatest amount of total welfare among morally relevant beings. A utilitarian theory of punishment will consequently say that punishment is justified as long as it leads to greater total utility than the alternatives; more welfare than the alternatives.

Central to all utilitarian theories of punishment is the view that everybody’s welfare counts - the rule breakers’ included. Rule breakers cannot simply by having broken a rule be subjected to any and all punish-

---

71 Kymlicka 1995 offers a nice overview of different understandings of utility. See also Kelman 2005. Though it is not important for my analysis, I interpret welfare as the proper meaning of utility. In the case of punishment, the kinds of things of relevance for welfare are, on the positive side, for instance the increased security of low crime rates, the satisfaction involved in seeing offenders punished, and the possibility of making a rule breaker’s life go better. On the negative side, punishment typically lowers the welfare of the punished and causes distress among the rule breaker’s family and friends. There is also a considerable costs involved in maintaining a criminal justice system.
ment the state may see fit. They are to be treated with equal concern and respect in the initial stages of the so called calculus of utility (the process in which we evaluate alternatives in terms of their aggregate consequences). They are not people towards whom we, in virtue of their crime, no longer owe any obligations or moral concern.

This obviously has ramifications for how we think about punishment. Utilitarianism is an egalitarian theory although it can justify deeply unequal outcomes. In the input stage of the calculus of utility there is no discrimination: every morally relevant being is included, and the utility of one does not in principle count as more important than the utility of another. But the act or policy that, after having given everyone’s interests equal concern, turns out to maximize utility may be deeply unequal in its outcomes. It might be, for instance, that the overall welfare of humanity would be maximized if five percent of the population were enslaved and forced to work for the good of the remaining ninety-five percent. In that case, a utilitarian would say that it would be right to enslave them. But it should be noted that the utilitarian calculus does count the disutility suffered by the slaves. It’s only that the worsening of their lives is outweighed by sufficiently large gains in the overall welfare of the rest.

What the indiscriminate extension of the principle of utility to all means is clearly that punishment, in itself, is a bad thing. It does not matter that a criminal might have performed odious acts; his or her suffering is still, other things equal, something to be avoided. Since the suffering the criminal undergoes when being punished must be considered, it produces a kind of morally relevant disutility that needs to be outweighed by gains in utility. Without such gains, all that would be left is pain and suffering. That would make punishment wrong. As we recall from chap-

---

72 I shall here regard, contrary to the gist of utilitarian thought, that the morally relevant beings are humans living in a particular society. This is problematic, first, since utilitarians typically assume that we have just as weighty obligations to members of other societies (see Singer 2002). (If this becomes unbearable, just imagine that the society we are dealing with here is a global one – and that there are no relevant beings besides us in the universe!) It is problematic, secondly, since it rules out that non-humans are morally relevant beings. Many have argued that all beings that can experience pleasure and pain or can have preferences should be encompassed (again, see Singer 1990).
ter 1, Bentham thus wrote: “All punishment is mischief: all punishment in itself is evil. Upon the principle of utility, if it ought at all to be admitted, it ought only be admitted in as far as it promises to exclude some greater evil” (Bentham 1988: 170). The “greater evil” usually put forth by utilitarians is the widespread suffering and feeling of insecurity that crime causes. In deterring crime, and thus promoting compliance with social rules, the practice of punishment can outweigh the suffering it causes the criminal, and those who care about him or her. 73

It is easy to see why such an attitude to punishment is welcomed by many. First of all, it gives punishment a seemingly rational justification: punishment is not about inflicting pain for pain’s sake, but serves an identifiable and wholly secular purpose – to make people in general better off by securing compliance with the law. Secondly, it recommends a sense of moderation in punishment. Punishment should be as light as possible. One should settle for the level of punishment that is the optimal trade off between the criminal’s suffering and the utility it produces. If the desired deterrence can be obtained by a lighter penalty, opting for a more severe one would be senseless cruelty. On a historical note, this logic meant that the utilitarian theory of punishment was a voice of moderation at a time when punishment was often appallingly gory and grave. 74

Deterrence

Deterrence is about making people refrain from doing certain things out of fear or prudence (Honderich 2006: 75). Punishment is a deterrent

73 Bentham considered general deterrence (he called it “general prevention”) to be the real justification of punishment, although he noted that the punishment of a particular person might serve the valuable goals of intimidating (i.e. special deterrence), reforming, or incapacitating him or her (see Bentham 2005).

74 See Foucault 1995.

75 Deterrence is about making agents refrain from acting in certain ways regardless of whether or not they want to. An oft-discussed example of deterrence is the nuclear politics of the Cold War. Once effective second-strike capabilities were secured by both the U.S. and the USSR, it didn’t matter whether one party might have wanted to attack the other – they kept the peace anyway, out of fear or
since it represents a cost that rational agents will consider. In deliberating whether or not to break a rule, a rational agent will compare the expected gains of the crime with the prospect of being apprehended and punished. If the expected disutility of punishment is weightier than the expected utility of the crime, he or she will be deterred from breaking the rule. Many factors may be relevant to the deterrence of a punishment. Those most easily recognised are severity and certainty.\footnote{Beccaria 1986 added swiftness to this list. He thought that a punishment would deter more the swifter it was.} Severity is the pain involved in punishment. Certainty is a compound function of (a) the probability of apprehension when breaking a rule on the one hand, and (b) the probability of being punished if apprehended on the other. Severity and certainty are communicating vessels: a very painful punishment might deter even if the risks of undergoing it are very small, and a highly probable punishment can afford to be relatively light. As a general rule, however, the deterrent effect of a punishment is assumed to go up when its severity and/or certainty goes up.

Robert Nozick (1974: ch. 4) has formalized the theory of general deterrence as follows: the likelihood of an agent breaking a rule is equivalent to a \((1-p)\) chance of obtaining a gain \(G\), where \(p\) is the likelihood of getting caught. If the agent is caught, he suffers a cost \(C\).\footnote{Nozick here assumes that getting caught is the same thing as suffering punishment.} \(C\) might be of any kind – guilt feelings, sentence severity, compensation to victims, and so forth.\footnote{Nozick himself prefers to label these different costs \(C\), \(D\) and \(E\), but it is of no relevance.} When the expected costs are higher than the expected gains, deterrence is likely to be effective. Thus, deterrence presupposes that:

\[
p \times C > (1-p) \times G
\]
To make deterrence the chief aim of punishment has important implications. First of all, to punish a crime that for some reason is a unique event seems unjustified. This at any rate was Bentham’s conclusion:

“General prevention ought to be the chief end of punishment, as it is its real justification. If we could consider an offence which has been committed as an isolated fact, the like of which would never recur, punishment would be useless. It would be only adding one evil to another. But when we consider that an unpunished crime leaves the path of crime open not only to the same delinquent, but also to all those who may have the same motives and opportunities for entering upon it, we perceive that the punishment inflicted on the individual becomes a source of security to all” (Bentham 2005: 20-1).

Secondly, as we have seen, punishments should be as light or grave as the calculus of utility recommends. If adequate compliance with a rule could be obtained by minor penalties, appeals to more severe punishment in the name justice should, as long as they unsubstantiated from the point of view of utility, be rejected. Thirdly, there is a sense in which crimes that cannot be deterred should not be punished. They should either be left alone – for without an effective deterrence value the penalties serve only to impose pointless suffering – or be motivated from other utilitarian concerns. Fourthly, actions without harmful consequences should not be punished in the first place. Crimes without either victims or bad social consequences should be decriminalized immediately, for otherwise the state would, again, impose pointless suffering. Bentham, for instance, defended the decriminalization of homosexuality on these grounds. Consenting adults who engaged in homosexual acts harm no one, he argued, nor do they threaten reproduction or the institution of

---

79 A crime only committed by those unable of rational thought could not be deterred, for example. But if those who commit the crime in question are incurable, there might be utilitarian grounds for incapacitating them. Consider for instance the case of paedophiles; a class of people sometimes assumed to suffer from an incurable disorder. Punishment, we could imagine, could not serve to deter them, because the disorder is such that they cannot control their actions. But since paedophiles cause enormous pain, utilitarianism might recommend that they be separated from society indefinitely. Note that it is questionable if this policy could appropriately be called “punishment”.
marriage. The criminalization of homosexuality was merely a reflection of an aesthetic sexual morality. From a utilitarian point of view this made it senseless to prohibit it.\textsuperscript{80}

\textit{Problems}

As I said, critical discussion of the theories will be kept relatively short in this chapter. Nevertheless, some problems will be discussed, if only to enable us to see more clearly the content of the theories. The strengths of the deterrence theory should be plain for all to see. The idea of making the production of compliance with the rules the justification of punishment is plausible, and there is little doubt that punishment is a deterrent.\textsuperscript{81} Many potential rule breakers no doubt would commit crime unless they calculated with the cost of punishment. Furthermore, as Honderich has remarked, the threat of punishment probably produces “unreflective law-abidingness” (Honderich 2006: 83-6).

The problems associated with the deterrence view are, however, numerous. Some are practical: do we ever know enough about criminality and the deterring effects of punishment in order for us to establish the appropriate levels of punishment? How could we possibly know how much a criminal suffers from a punishment, let alone the total effects crimes and punishments have on aggregate utility? Such objections have always been levelled at utilitarianism, in all areas of application, but it should be noted that they only concern the practical possibility of implementing deterrence as the primary aim of punishment. It does not challenge the idea that deterrence, if we only knew how to bring it about, \textit{would} justify punishment.

Other problems exclusively concern the nature of deterrentism’s justification of punishment, or more precisely its justice or fairness (or lack thereof). This should come as no surprise: utilitarianism has perennially been criticized for being capable of supporting what some consider highly unfair and immoral outcomes. Consider first the objection that

\textsuperscript{80} Bentham made this case in the unpublished paper “Offences Against One’s Self”. See \texttt{<skeptically.org/utilitarianismtheethicaltheoryforalltimes/id22.html>}.  
\textsuperscript{81} For a valuable survey of research on punishment and deterrence, see von Hirsch et al. 1999.
guilt plays no principled part in a deterrence-based regime of punishment. If maximizing utility is what counts, what is to exclude us from punishing the innocent? If greater deterrence could be produced by punishing an innocent rather than the guilty, would it then be right to punish the innocent? This objection, and utilitarian attempts to meet it, will be treated in chapter 5.82

A second justice-based objection concerns rule breakers who have limited responsibility for their acts. Consider for instance someone whose mental disorder impairs his or her ability to control his or her behaviour. A classic example of this is the kleptomaniac. Suppose a kleptomaniac is caught committing one of his or her petty thefts. Should we refrain from punishing him or her? From deterrentism’s perspective, there is a clear sense in which we should answer this question in the negative – and this even if we accept that kleptomaniacs have no responsibility for acting out of the “irresistible impulses” their disorder causes. The reason for this, again, is that guilt plays no prominent role in deterrence theory. What is important is changing future behaviour, not making backward-looking “justice”. It might be true that kleptomaniacs cannot be deterred by punishment, as they do not respond to rational incentives. But even though kleptomaniacs do not respond to the threat of punishment, other prospective thieves can still be deterred by punishing them. Deterrentists have

82 Bentham, by the way, seems to have cheerfully embraced at least one consequence of dropping guilt and justice as guiding principles in the practice of punishment. Since the punished person’s suffering in itself is evil, it follows that the best punishment is one that looks worse than it really is. Indeed, in The Rationale of Punishment he remarked that if one could hang an effigy of a criminal without thereby losing deterrence value, it would be mere cruelty to hang the real criminal (Bentham 2005). Regardless of whether one finds this conclusion objectionable, it directs our attention to an important idea of deterrence theory: achieving as much deterrence as possible with as little pain as possible. Beccaria, the enlightenment theorist of punishment par excellence, similarly argued that the best punishment was the one which made a great impression on the minds of man while being as painless as possible for the criminal (Beccaria 1986). John Stuart Mill, another great utilitarian, famously defended capital punishment in an 1868 Parliament debate partly on these grounds: death as punishment looks worse than it really is. Life imprisonment, on the other hand, looks better than it is, and is thus particularly bad from a utilitarian point of view. (How Mill divined that death really isn’t that bad is a perfectly legitimate question).
therefore tended to argue that a penal regime where punishment follows without exception will be more effective than one where rational agents can hope for the possibility of being excused or pardoned. Thus Mackie has noted that,

“It is true that only intentional actions can be deterred, and hence that a penalty is deterrently effective only in so far as a rational agent sees it as being attached to some possible intentional action. But it is not true that his seeing it can result only from its being legally connected to actions of just that sort. Though an unintentional killer, say, could not himself have been deterred, it is quite possible that potential intentional killers should be more effectively deterred by the punishing of all killers than by the punishing only of the intentional killers and the excusing of unintentional ones.”
(Mackie 1990: 209-10. Emphasis added)

Such a conclusion has struck many as unjust, since it fails to excuse those without responsibility for their actions. It should be noted, however, that some deterrentists do not accept that we are justified in punishing all rule breakers, regardless of their ability to be deterred (Dennett 1984: 160-1).

A final justice-based objection is that deterrence theory, even if it would restrict punishment only to the guilty and responsible, could support excess severity (or leniency) in punishment. This might be seen as flying in the face of the theory. Recall that the theory’s principal recommendation is that we should punish only as harshly as deterrence requires. Still, “requires” is here an unclear term. What deterrence theory suggests is that punishments should be set at the level where the trade off between the utility and disutility is optimal. At what level the optimum is found is an empirical question – it might be that murder should be punished by five years imprisonment, but it might also be that it should be punished by torture. Many deterrentists have argued that proportionality in punishment serves the aim of deterrence, or that a utilitarian calculus of utility would recommend that punishments stand in some relation of proportionality to the severity of the crimes.83 But this is an

83 Bentham in particular put a lot of thought into rules of proportionality in punishment (Bentham 1988: ch. 14). My simple point, which will be augmented in chapter 5, is that a utilitarian defence of proportionality must be conditional. Proportionality will only be justified as long as it promotes total utility or wel-
empirical claim, which may well turn out to be false. To say that punishment should be as lenient as possible, then, is somewhat of a truism. It rules out pointless cruelty (i.e. pain not outweighed by gains), but it does not exclude very severe punishment indeed. Critics have thus consequently objected that deterrence theory can justify too much.\textsuperscript{84}

\* \* \*

Deterrence theory provides a compelling justification of punishment. The idea that protecting society from further crime is the aim of punishment is plausible. Many believe that the deterrence punishment produces is enough to make many potential rule breakers opt for compliance, and that this fact alone is sufficient to justify punishment. Still, deterrence theory faces weighty justice-based objections, and questions can be raised about deterrence as a practicable principle, as we do not seem able to calculate utility with any degree of exactness.

3.2 REHABILITATIONALISM

Rehabilitation is like deterrence an old justification of punishment. However, rehabilitation as a principal theory of penal policy did not emerge until the birth of modern social and medical science and the proliferation of the prison system. Although the reform of offenders had long been thought of as a purpose of punishment, especially in religious circles where imprisonment was believed to lead to moral atonement and spiritual rebirth, it was not until the 19\textsuperscript{th} century that it aspired to be the main penal policy (Foucault 1995; Qvarsell 1993). With science (perhaps most notably psychiatry) presenting the alleged know-how, and the prison system supplying the necessary institutional structure, voices were raised

\textsuperscript{84} Or, for same reason, that it can justify too little, i.e. defend too lenient punishments.
that punishment could cease to be about inflicting pain in the name of justice and instead seek to eradicate the causes of crime.

The penal theory of rehabilitation, or rehabilitationalism, has at least since the later half of the 19th century exerted a massive influence over Western penal policy. The parole system, indeterminate sentencing, non-custodial sentences, the increased use of forensic psychiatry in the management of crime, are all offspring of the rehabilitation theory. Familiar names also testify to its influence. To call prisons “correctional” institutions is one example. That Sweden’s Prison Board is called Kriminalvården (roughly “Care of Criminals”) and its prisoners “clients” is another. 85

Rehabilitation is simply the idea that the penal practice should aim at changing the factors that caused the crime in the first place. Unlike deterrence, which seeks to achieve the right behaviour without addressing the basic criminal motivation, rehabilitation seeks to change the offender’s very desire or need to commit crime. Hudson offers the following definition:

“Taking away the desire to offend, is the aim of reformist or rehabilitative punishment. The objective of reform or rehabilitation is to reintegrate the offender into society after a period of punishment, and to design the punishment so as to achieve this” (cited in Raynor & Robinson 2005:4).

The mode of rehabilitation can take different forms depending on what factor is believed to have caused the criminal to break the rules. If lack of education and consequent inability to compete at the job-market is believed to be the cause, rehabilitation consists in providing the offender with an education. If a mental disorder is seen as the cause, the state should provide psychiatric treatment or psychotherapy. If addiction is believed to be the cause, the offender should get help overcoming his or her addiction. And so forth.

As we saw in the previous chapter, rehabilitation urges us to be cautious about the way we use the term “punishment”. Punishment can be defined as an authority’s intentional infliction of pain or deprivation

85 Kriminalvården’s official English name is The Prison and Probation Service, a translation that fails to do justice to the Swedish name.
on a person for having broken a rule. We saw that rehabilitation too might involve the intentional infliction of pain and deprivation, but that in treatment-modes of rehabilitation, the intention to inflict pain or deprivation is dropped. What sets rehabilitation apart from most other theories of punishment is its straightforwardly paternalistic element.\(^{86}\) Whereas punishment is typically understood as something that makes the punished worse off then they would otherwise have been (i.e. as something which harms them), rehabilitation is usually conceived as \textit{good} for the person undergoing it. In being helped to overcome the disorders that breed crime, the criminal is believed to be restored to a better state.\(^{87}\)

Having said that, some rehabilitationalists could defend punishment as we know it on the ground that it leads to the improvement of the offender: that by being subjected to pain or deprivation, the offender sees the errors of his or her ways and becomes a better person. Such was the idea of traditional reform theory (Foucault 1995). Modern rehabilitationalists, however, tend to regard the idea of improving rule breakers by punishing them as far-fetched (Honderich 2006). They lean more towards treatment as defined in chapter 2. Unless otherwise stated, this is also what I shall intend by rehabilitation in the remainder of this book: Rehabilitation theory is the view that the penal regime should aim at \textit{treating} offenders.

Now, just as the modes of rehabilitation can differ, so can the nature of its justification. The standard justification is perhaps straightforwardly utilitarian: by rehabilitating offenders we reduce crime, which benefits society. Rehabilitation is justified, a utilitarian could argue, since it is the best way to maximize utility (Duff 2004). But other justifications have been advanced. Some argue that rehabilitation is the \textit{right} of an offender. Since an offender \textit{needs} rehabilitation in order to overcome his or her problems, offering rehabilitation is a humanitarian obligation; a way of giving an offender the best tools available to lead a good life. Such a justification can be disconnected from considerations of social utility - it could be argued that we owe offenders rehabilitation even if there are, from a utilitarian point of view, better ways of promoting utility. Others

\(^{86}\) See however Morris 1981.

\(^{87}\) Admittedly, one could imagine rehabilitation that works only to the benefit of the community, not the rule breaker. More often than not, however, rehabilitation is believed to benefit community and rule breaker alike.
still have defended rehabilitation on the ground that it is the best way of protecting the interests of potential *crime victims*. Since rehabilitation is the most effective means of protecting potential crime victims, to rehabilitate is an obligation we have towards them, some have argued (Raynor & Robinson 2005).

It is important to note that rehabilitation entails a solid focus on the individual rule breaker. It does not focus on the structure of society as such. Thus, to say, as many contemporary politicians have said, that criminal policy should aim at eradicating the causes of crime – understood as, say, poverty, poor education, drug abuse, etc. – is *not* to advocate rehabilitation. All theories of punishment treated in this chapter are compatible with, say, fighting poverty in order to prevent crime. Rehabilitation is instead normally understood as intervening, with or without consent, in the life of a rule breaker in order to restore him or her to some state of “normal” or desired functioning. This entails a preoccupation with the individual offender, not social background conditions, although it of course may be the case that the crime-causing problems have social origins (such as in cases where socio-economic deprivation is believed to produce anti-social personalities).

*Problems*

To its advocates, rehabilitation represents a profound optimism. What rehabilitation assumes is that individuals are *corrigible*; that they can change for the better given the right help. Much Western penal policy during the 20th century expressed this optimism. However, by the final quarter of the century such optimism seemed misplaced. After decades of attempts, it was largely agreed among criminologist and social theorists that the quest for rehabilitation had failed. Evidence of successful rehabilitation was conspicuously scarce. The received wisdom thus became

---

88 Rehabilitation theory has therefore been criticized for failing to see the larger picture, especially by the academic left. It has been argued that by focusing on the disordered individual, rehabilitation theory misses the structural and societal problems that underpin crime.
that, as far as rehabilitation goes, “nothing works”. As a consequence, Western penal policies began to distance themselves from the rehabilitative ideal, which was seen as unattainable and/or inappropriate. Politicians, sensitive to the fact that their electorates were increasingly afraid of crime, adopted policies that were perceived as “tough on crime” (Garland 2002). Public protection by means of incarceration was put forth as the basic rationale of the prison. “Lock ’em up and throw away the key” was the natural conclusion of the “nothing works” paradigm (Raynor & Robinson 2005).

This story, which I can only tell sketchily here, needs to be moderated. The profound pessimism about our ability to rehabilitate offenders is, and has always been, exaggerated, and there are presently signs that the principle of rehabilitation is reasserting itself in Western penal policy, especially in the form of cognitive behavioural therapy (ibid.). Nevertheless, the fact remains that rehabilitative programs are not a successful as one would have liked. As a rule they fail to have a significant negative impact on recidivism, and in some cases can even increase it. Rehabilitation theory is naturally affected by this. Its persuasiveness relies on that fact that rehabilitation is a feasible aim of penal policy. In the absence effective rehabilitation programs, it becomes hard to motivate why punishment should not instead aim at maximizing public safety, or at least at giving the incorrigible offender his or her “just desert”. Rehabilitation advocates can reply that we won’t find effective treatment unless we are willing to go through a long period of trial and error, and they may add that regardless of whether or not rehabilitation is effective, it is our humanitarian obligation to strive for it. But the low success rate of rehabilitation programs is undoubtedly a problem. Critics are always left with the option of saying, “Sure, rehabilitation would be nice if we knew how to do it. But we don’t”.

This objection is purely empirical and contingent. The fact of the matter is that the future may bring wonderfully effective ways of remov-

---

89 “Nothing Works” was the title of an influential article by Robert Martinson in 1974. For a survey of the context of the article and the debates it sparked, see Raynor & Robinson 2005: 64f.
90 Studies have for example shown that to offer social sensitivity training to psychopaths may in fact increase the probability of recidivism, see Dennett 2004: 308.
ing the desire to commit crime. Moran has offered the following prediction, based on the prospects of an increased biomedical ability to reduce recidivism:

“The development of biomedical techniques to control crime is still in the experimental stage. Once such techniques become available for large-scale application, however, it is probable that the rehabilitative idea in corrections will reassert itself. The current antirehabilitation forces are sure to raise their voices in protest, and the ‘nothing works, lock ‘em up’ people are certain to become more adamant. Both groups will be silenced, however, by the pragmatists, who, in pointing to the capability of biotechnics to ‘correct’ the criminal, will argue, ‘but it works’” (in Conrad & Schneider 1992: 223).

Such a development would of course obliterate the objection from unfeasibility. Is resistance against the rehabilitative ideal then dependent on the purely empirical question of how successful rehabilitation is? By no means. Some objections levelled at the ideal are of a very principled nature.

Some have argued, firstly, that whatever the success rate of rehabilitation programs, they are still unjustified in that they all have a tendency to victimize the wrongdoer. Since rehabilitation, whether or not it proclaims to exonerate from responsibility, rests on the assumption that there is something wrong with the offender, it always seems to work in the direction of reducing the offender’s guilt (cf. Heywood 2004). I shall return to this criticism throughout this book.

Associated with this criticism is a second criticism that rehabilitation offsets proportionality and considerations of procedural fairness. The logic of rehabilitation admits of penal responses that may seem problematic: if removing the desire to offend is the purpose of punishment, then it could be that a maladjusted petty thief requires a longer time in rehabilitation than a well-adjusted criminal who is guilty of quite serious crimes, e.g. treason. This will trouble all who believe that a punishment must stand in some proportion to the criminal act itself. Rehabilitation theory, however, tends to focus on the criminal as an individual, not his or her particular crime. It is ultimately the individual person which requires help. A small time offender may be in greater need of rehabilitation and pose a greater risk to society than people who are guilty of more serious crime.
Furthermore, some are troubled by the fact that the logic of rehabilitation can work to remove completely traditional considerations of punishment from the penal regime. Assume, as we did in chapter 1, that a rule breaker could be “cured” through a one-time medication (a pill that fully removes the risk for future crimes). The rehabilitative theory would seem to recommend (a) that he or she be medicated, since this meets the criterion for interfering in the first place, and in addition say (b) that no further measures ought to be taken against him or her. If for instance sexual offending is more effectively rehabilitated by a painless chemical castration than a lengthy time in prison, then, following the assumptions of rehabilitation theory, chemical castration would be the preferred course of action. It would achieve the right result without imposing pain. But some object that when penal policy become disconnected from considerations that figure prominently in traditional accounts of punishment, we risk ending up in a situation where the state engages in irregular and intrusive action against rule breaking. Built into in the rehabilitationalist position is an instrumental rationality which, when coupled with the paternalist attitudes that are inherent in the desire to intervene in someone’s life for his or her own good, can easily lead to that crime/criminal propensity and cure/punishment blend into each other, they argue. In such a situation, there is a risk that the state starts to engage in pre-emptive rehabilitation and other problematic practices.

A further problem of rehabilitationalism is that it tends to regard crime as indicative of disease or disorder, i.e. an expression of something in need of treatment. This outlook seems to have an inappropriate conformity bias. It assumes that to comply with the norms of society is the normal thing to do. This cannot be universally true. Consider for instance crimes of principled dissent. Is it plausible to label those who engage in civil disobedience as deviant? Furthermore, consider petty and widely

91 Of course, it might be that the offender himself refuses the treatment, which may make it impermissible. There is nothing in rehabilitation theory that forces one to directly disallow the importance of consent.
92 I will have more to say about this theme in chapters 5 and 6.
93 Definitions of mental health, however, often emphasize the ability to live contently and in conformity with social norms. This is evident, for instance, in the highest scale-steps of the so called GAF-scale (Global Assessment of Functioning), which is an extensively used scale among psychiatrists in various countries.
committed crimes such as speeding, or crimes against rules that are regarded as inherently unjust or obsolete. Is it plausible to say that every speeder, or every person who fails to comply with a rule they see as inoperative, is disordered and in need of rehabilitation? The answer appears to be no. The rehabilitative theory is most easily applicable to crimes that are widely seen as indicative of behavioural disorders, such as violent and sexual offences. Even so, social constructivists are quick to point out that classifications of deviance are social constructs, and that they serve the interests of some more than others (Christie 2004; Conrad & Schneider 1992).

***

To many, rehabilitationalism is quite obviously the correct justification of “punishment”, or the correct guiding principle of the penal practice. It to them represents an inquisitive, rational and humane position, more interested in understanding and helping offenders than in blaming and harming them. Just like it can be unproductive and possibly inhumane to blame someone who suffers from a physical illness, it can be unproductive and inhumane to blame someone who commits crime, they argue. A decent and rational penal regime is available once we take this to heart and drop the notion of inflicting harm on offenders in the name of justice, rehabilitationalists argue.

Nevertheless, for all its appeal, the theory faces some powerful objections. A widely spread, if exaggerated, objection is that rehabilitation does not work; that we have no idea how to change the behaviour of rule

---

The GAF scale runs from 0 to 100, where 0 is minimal and 100 maximal “functioning” (see DSM-IV, APA 2000). Here are the two top steps of the scale. Scale-step 100-91: “Superior functioning in a wide range of activities, life’s problems never seem to get out of hand, is sought out by others because of his or her many positive qualities. No symptoms.” Scale-step 90-81: “Absent or minimal symptoms (e.g. mild anxiety before an exam), good functioning in all areas, interested and involved in wide range of activities, socially effective, generally satisfied with life, no more than everyday problems and concerns (e.g. an occasional argument with family members).” Critics have objected that accounts such as these lay too much emphasis on social embeddedness and conformity. Some have quipped that if an ability to conform is what signifies mental health, it follows that all liberals in Nazi Germany were mentally ill (see Glover 1970: 104).
breakers. At a more principled level it has been objected that rehabilitation is an ideal that cannot guarantee such things as proportionality and procedural fairness; that it implausibly regards crime as abnormal; and that it invariably works to undermine the responsibility of the wrongdoer.

3.3 RETRIBUTIVISM

Retributivism is a desert-based theory of punishment. It argues that punishment is justified if, when, and to the extent that, it gives criminals their “just deserts”. Retributivism is arguably the most controversial of the three theories discussed in this book, and also notoriously hard to pin down. Cottingham in 1979 claimed to have distinguished between no less than nine different senses of “retributivism” (Cottingham 1979). Jeffrie Murphy, himself a prominent retributivist, has noted at least six different versions of retributivism based on how the central notion of “giving wrongdoers what they deserve” is interpreted (Murphy 2003). Honderich, finally, claims to have identified a whopping sixteen distinct versions of retributivism (Honderich 2006). Subsequent references of mine to the retributive theory should therefore be taken with healthy scepticism. In what follows, I shall describe what I intend by “retributivism”. Since the term is contested, chiselling out the theory will inevitably alienate readers who by “retributivism” intend something else than I. Nevertheless I hope that the version I present is understandable and that it strikes the reader as worthwhile.

Despite retributivism’s contested nature, there are some commonalities between most versions of the theory. A classical definition, which I think most of its proponents would accept, is found in Johan Rawls’s *Two Concepts of Rules*. Rawls here offers the following description:

“[..] the retributive view is that punishment is justified on the grounds that wrongdoing merits punishment. It is morally fitting that a person who does wrong should suffer in proportion to his wrongdoing. That a criminal should be punished follows from his guilt, and the severity of the appropriate punishment depends on the depravity of his act. The state of affairs where a wrongdoer suffers punishment is morally better than the state of affairs where he does not...” (Rawls 2001: 21-2)
Now, a number of things are noteworthy in this passage. Punishment, first of all, is said to be somehow intrinsically justified when administered to those who deserve it. It is “morally fitting” to punish a person who has done wrong. This “fittingness” is normally understood as a matter of justice. When Rawls’s writes “that the state of affairs where a wrongdoer gets punished is morally better than the state of affairs where he does not”, he expresses the retributive belief that is just to punish those who deserve it (and that it is less just, or unjust, to refrain from doing it). A second thing to note is that punishment is required to be “in proportion” to the wrong. Retributivism claims that when one does wrong, one should get a punishment that is somehow commensurate to “the depravity of the act” (whatever this might mean). It rejects excessive leniency and harshness in punishment alike: in order for the punishment to be justified, it must be proportional to the offence. It must be what the offender deserves.

Desert

As noted by Rachels (1997), retributivism is the application of the principle of desert to punishment, and as such could be said to be part of the larger class of desert-based theories of justice. Desert-based theories of justice all claim that people can come to deserve things/modes of treatment in virtue of so called desert-bases - facts about people that render it appropriate to treat them in certain ways. When we say that the virtuous deserve praise, or that the diligent deserve success, we say that there are some facts about these people (their virtue/effort) which make specific treatments or outcomes appropriate (praise/success). According to

94 The seminal work when it comes to the analysis of desert is Joel Feinberg’s “Justice and Personal Desert (reprinted in Feinberg 1970). Here Feinberg launches the idea that every desert-claim takes the form “A deserves X in virtue of P”, where A is an agent, X a mode of treatment, and P the desert-basis. The relevance of the concept of desert for justice is that it is evaluative, and so can serve to inform discussions about distributive and retributive justice. Desert is about the fittingness or appropriateness of a mode of treatment and the evaluation of the agent. For an excellent collection of papers on desert and justice, see Olsaretti 2003a.
desert-based theories of justice, claims such as these have moral weight. Other things equal we should, as a matter of justice, give people what they deserve.

Action is the most natural desert-basis. People may come to deserve good or bad treatment in virtue of the way they act. Those who do good things – those who despite risk to themselves rescue others from harm for example - deserve praise or reward for their commendable action. Those who do bad things – those who make no effort to help others in need, say, or those who intentionally inflict harm on others – deserve blame or punishment. Retributivism contends that what justifies punishment is that it can, when properly administered, treat people in accordance with their desert in this manner.

It is important to note that the kind of (moral) desert retributivism considers is of preinstitutional origin. A distinction is often made between institutional and preinstitutional senses of desert. Institutional desert, or entitlement, is a kind of desert which is strictly derivative of bodies of rules or institutional frameworks, which in turn need not have a basis in the principle of desert. Take the game of football as an example. The rules of football say that if team \(A\) scores at least one more goal than team \(B\) without cheating, then team \(A\) wins. If \(A\) has fulfilled this criterion for winning, it has “legitimate institutional expectations” to be awarded the win. In an institutional sense, therefore, \(A\) deserves it – \(A\) is entitled to the win given the body of rules which governs the game.

Unlike institutional desert, preinstitutional desert is a fundamental moral concept, which is logically prior to any given institution. We could therefore ask for any given institutions or body of rule whether it is in line with preinstitutional desert or not – a question which is nonsensical if desert is taken to be derivative of the institutional framework. For instance, in a preinstitutional sense of desert we could say that team \(B\) deserved to win even though team \(A\) scored one more goal than \(B\) without cheating (for instance because \(B\) played better football or tried harder). This amounts to the claim that the institutional rules of football do not always treat agents in accordance with their (preinstitutional) desert.

Retributivism as a rule understands desert preinstitutionally.95 Crime should be punished since/if it is wrong, not because it is a conse-

---

95 For work on institutional or preinstitutional desert, see Feinberg 1970; Scheffler 1992; Rawls 2000; Scheffler 2001; Olsaretti 2003b. Debates concerning desert in
quence of a particular penal scheme. It is for instance possible to say, according to retributivism, that a legally permitted act ought to be punished since it is morally wrong, which boils down to the claim that the act in a preinstitutional sense deserves punishment.

It should be clear that retributivism differs from utilitarian theories of punishment in a number of respects. As we have seen, utilitarians are committed to the belief that punishment is justified if and only if its positive effects outweigh its negative ones. Its general justifying aim is its capacity to produce utility, for instance by deterring crime or rehabilitating offenders. Retributivism on the other hand holds that justified punishment is justified irrespective of its consequences, though modern retributivists tend to see beneficial consequences of punishment (such as deterrence) as welcome side-effects (see e.g. Armstrong 1961; Corlett 2001). Punishing someone who deserves it is justified in itself. Desert is a necessary and sufficient condition for justified punishment.

Entailed by this difference in terms of justification is clearly a difference in temporal focus. Utilitarians adopt a forward-looking attitude to punishment. Punishment should not be about correcting past injustices; rather, it should be about taking steps to ensure that what happens from here on becomes as beneficial as possible. There is nothing valuable, utilitarians contend, about exacting justice for past wrongs. What is valuable is only the extent to which punishment can serve to deter future crime, enhance a sense of security in society, better the criminal, etc. Retributivists, on the other hand, adopt a backward-looking approach. When a wrong is committed, the state has to look at what transpired when the crime was committed. If an identified guilty party is deserving of punishment, the state should – or is at least permitted to – proceed to punish the guilty in accordance with what his or her action merits. This is what justifies punishment, and acting on desert sets limits to extraneous considerations influencing the punishment. Considerations of the future, e.g. what the 20th century has been animated by Rawls, who in A Theory of Justice (Rawls 2000: 88-9, 273-7) held that preinstitutional desert had no substantive role to play at the level of the basic structure of society, at least when it comes to questions of distributive justice (Rawls seems to have, inconsistently or not, have advocated ideas akin to retributivism when it comes to punishment, see Scheffler 2001). The concept of preinstitutional desert is nowadays treated with more interest, if not more liking, by political philosophers.
would be most instrumental in giving the victims a sense of “closure”, should not influence the punishment. The punishment should solely be a function of the severity of the wrong committed, not something arrived at after having performed a calculus of utility. “The severity of the wrong” is here usually understood as a compound of the harm caused and the level of responsibility. An intentional harm for example merits a harsher punishment than an equal harm that was caused inadvertently.

A prerequisite for retribution is of course that the criminal, in virtue of his or her actions, deserves to be punished. 96 Needless to say, this means that retributivism excludes punishment of innocents. But it also means that the crimes need to have been in some sense wilfully committed: the rule breaker must have acted freely in a way that makes him or her responsible for the act. Retributivism thus excludes punishment of those who are unable to be held responsible for their actions (e.g. very young children, nonhuman animals, accidental rule breakers, severely mentally disordered, etc.). Opinions may diverge on whether crimes, generally taken, are free choices of such kind that can serve as the basis of negative desert. Some retributivists seem to substantively assume that most crimes are freely chosen and most criminals are fit to be held responsible (see Anderson 1997; 1999). Much retributive criticism of other penal models, especially the rehabilitative one, makes little sense unless this is held to be true. Other retributivists, however, are open to the possibility that the nature of crime and society might be such that punishment is never deserved (Corlett 2001; Murphy 1995). Desert then plays solely the role of a conditional: unless a criminal deserves punishment, punishment can never be just. In case no one deserves punishment, no one should be punished.

There are many attempts to differentiate between different versions of retributivism. Some distinguish between negative and positive retributivism (Honderich 2006). Negative retributivism is the view that desert is a necessary condition for punishment – unless punishment is deserved, punishment cannot be justified. According to negative retributivism, desert only establishes the permissibility of punishment, but does not constitute a positive reason to punish. Positive retributivism, on the other hand, is the view that desert not only establishes the permissibility of

96 Note that “action” here probably should be understood so that attempts, omissions and criminal recklessness or negligence can be encompassed.
punishment, but is also a reason for punishment. According to positive retributivism, the fact that a punishment gives a wrongdoer what he or she deserves is something which speaks in favour of punishment. I will understand retributivism along the lines of positive retributivism. Negative retributivism is virtually undistinguishable from a side-constrained utilitarian theory of punishment, and does not seem to capture what retributivism is about.

Positive retributivism can in turn be divided into strict and non-strict versions. Strict retributivism holds that punishing the guilty is a perfect duty. It is obligatory that the wrongdoer be treated as he or she deserves. Kant seems to have advocated strict retributivism:

"The law of punishment is a categorical imperative, and woe to him who crawls through the windings of eudaemonism in order to discover something that releases the criminal from punishment or even reduces its amount by the advantages it promises […] For if justice goes, there is no longer any value in human beings’ living on the earth" (Kant 1998: 105).

Non-strict retributivism, on the other hand, holds that desert is a pro tanto reason to punish, but not always a conclusive one. Both versions thus agree that punishment of the undeserving is always impermissible, and that desert constitutes a reason to punish. But whereas strict retributivism holds that desert must be acted on, non-strict retributivism holds that various considerations may trump the pro tanto reason to punish.

97 There is some debate about how best to characterize Kant’s philosophy of punishment, however. See footnote 123 below.

98 I understand retributivism to be concerned with acts that are of relevance to the criminal law. Retributivism thus need not confront difficult questions that hit a general desert-based moral theory (it needs not, for instance, discuss whether we have an obligation to redistribute an inheritance that gives one beneficiary more than he or she deserves; see Dolinko 1991; Fischer 2006). Furthermore, I here presuppose that desert constitutes a reason to punish those who are apprehended and found guilty. A commitment to retributivism need not entail that we have a perfect obligation to apprehend all who have committed crimes that deserve punishment – something which would require an astronomical increase in the police force’s capabilities.
The kind of retributivism I will defend in this book is of a non-strict (and positive) variety. I have in mind a pluralistic view which says that acting on desert is justified in most cases, but not all. There may be cases when the acting on desert poses too great costs to other values.\textsuperscript{99} We should then refrain from acting on desert. As Feinberg once remarked: “desert is one very important kind of ethical consideration, but it is not the only one” (Feinberg 1970: 94). To this one may add: justice is one very important element of what we ought to do, but not the only one.

Many things still need to be sorted out in order to fully understand what kind of theory retributivism is, and what kind of penal regime it leads to. We lack a more thorough discussion what it means to be responsible for something, how this constitutes desert, and why it is important to treat people according to desert. Before proceeding to these important questions and the problems they present, however, I should first like say something about what retributivism is not.

3.4 SOME MISCONCEPTIONS ABOUT RETRIBUTIVISM

Retributivism is a controversial theory of punishment, but some of the controversies surrounding the theory are based on misconceptions. I will here try to spell out a couple of mistaken objections, which I have frequently encountered in the process of writing this book.

RETRIBUTIVISM CONFUSES WRONGDOING AND CRIME. It is sometimes suggested that retributivism entails, or contingently leads to, the view that those who break rules should be punished quite irrespective of the nature or circumstances of the act. Retributivism is thus said to entail, or lead to, a legalistic and rigid penal regime which confuses crime and wrongdoing. Whether or not retributivism “leads to” such a regime is an

\textsuperscript{99} One could, as Nozick has suggested, imagine cases where “catastrophic moral horror” overrides normally held obligations (see Nozick 1974: 30). Less demanding, one could imagine cases where the reason to forgive trumps the reason to punish.
empirical question. It can safely be said, however, that the theory in itself in no way entails legalism.\textsuperscript{100}

Those who believe that retributivism entails unacceptable legalism usually present a case along the following lines. Suppose two shoplifters: on the one hand an impoverished woman who steals food in order to feed her starving children, and on the other hand an affluent cheap-skate who simply cannot be bothered to pay for his groceries. Both break a legitimate rule. But the circumstances of their actions are clearly different - and a standard objection is that retributivism cannot account for this fact. Instead, the theory recommends that the woman and the cheap-skate should both be punished, and, if they stole similar things, be punished \textit{equally hard}. But this kind of stern and even-handed justice is unacceptable, the objection continues. There are mitigating, if not exonerating circumstances in the case of the impoverished woman which are not present in cases where people steal out of sheer greed.

This certainly seems right. But it misses completely the mark as an objection to retributivism. Those who present cases such as this have completely missed the foundational importance of \textit{desert} to retributivism. Recall that the theory states that punishment is justified if, when and to the extent that it is morally deserved. Given that the mother and the cheap-skate are not equally at fault, which seems plausible, the theory simply requires that we treat the mother more leniently than the cheap-skate. She deserves a lesser punishment; in fact she presumably deserves no punishment at all. For similar reasons it is incorrect to say that a retributivist is committed to the legalistic view that all crimes are to be punished, regardless of the nature of the laws. Consider a society which consistently permits morally vicious and outlaws morally virtuous behaviour. In such a society, a coherent retributivist will conclude that no-one should be punished.\textsuperscript{101} More realistically, no coherent retributivist would consider, e.g., dissidence in an apartheid-state to be deserving of punishment.

\textsuperscript{100}“Legalism” is here used to describe a rigid and strictly even-handed attitude to criminal justice, such as when one says that every crime deserves punishment, and that every crime of a certain \textit{type} deserves exactly the same punishment.

\textsuperscript{101}Unless one invariably considers, implausibly, incompliance with the laws a moral fault.
RETRIBUTIVISM IS NECESSARILY TOUGH ON CRIME. Some suggest that retributivism is necessarily “tough on crime”, which is to say that it is favour of comparatively harsh punishments. Given that the ascent of the theory has coincided with the tougher penal policies of Western states, there might be some empirical grounds to this claim.\(^{102}\) Note, however, that to say that retributivism is “tough” will only count as an objection as long as it is tacitly understood that penal policies ought to be soft. The point can thus be said to be a circular. Probing into the claim that the theory entails tough punishments will allow us to say some things about the difficult and central notion of proportionality, however, so if the reader bears with me, I shall dwell on it at some length here.

What has led critics to concluding that retributivism is essentially “tough on crime”? One need not look far for an answer - the reason in all likelihood is the principle of proportionality (\(PP\)) that retributivism of all sorts endorses. The essence of \(PP\) is that the punishment should “fit the crime”. The idea here is that every crime stands or should stand in a relation to some punishment, such that the punishments match or are commensurate to the seriousness of the crimes, and that the graver the offence is, the harsher the punishment should be.

Now, plausible as \(PP\) might seem at a general level, it leaves us wondering about more specific questions about proportionality. How do we decide the amount of punishment a crime merits according to \(PP\)? The historically most prominent answer to this question, and the one that presumably has led critics to conclude that retributivism is “tough on crime”, is the lex talionis. The lex talionis (Latin for “law of retaliation”) suggests that questions about proportionality can be settled by making crime and punishment stand in a relation of literal sameness to one another.\(^{103}\) The criminal should suffer as much, and in the same way, as his

\(^{102}\) Though one may wonder which is cause and which is effect here: Are penal policies tougher as a result of the increasing influence of retributive ideas, or have retributive ideas gained influence as a result of people’s desire to find a theory that justifies tougher penal policies?

\(^{103}\) Some understand the lex talionis differently than I do here. Davis in an influential piece understands it as simply the idea that the punishment should fit (i.e. be proportional to) the crime. It is always implied, he argues, that this fittingness is one of equivalence and not sameness (Davis 1995). Contrary to Davis, I treat the lex talionis as one interpretation among many of the principle of proportionality.
or her victim(s). The slogan of “an eye for an eye” is the best known expression of this view: taking the eye of another according to the lex talionis requires that the offender’s eye be taken in return, as a matter of justice. This view is expressed in the Old Testament and the Koran as a leading principle of justice in punishment and has exerted an enormous influence over penal policy throughout history.

The lex talionis is often regarded as harsh and unforgiving (although it deserves to be mentioned that it also curbs excessive punishment). It strikes many as the essence of revenge. If retributivism necessarily endorsed the lex talionis, then arguments to the effect that retributivism is always about being “tough on crime” would seem quite warranted. However, lex talionis is a completely unworkable and probably immoral doctrine that has enjoyed precisely little support among retributivists (cf. Armstrong 1961; Corlett 2001; Hegel 2005). It is unworkable for a number of reasons. Most trivially, it can be impossible to “pay back” the literal equivalence of a crime. A trite example is the difficulty involved in poking the eye out of an eye-less eyepoker. Furthermore, and more relevantly, many crimes lack a clear victim and it thus seems hard to say what the criminal deserves on the basis of literal sameness. Taking an eye for an eye is intelligible, but what about illegal parking (a victimless offence the designers of the Mosaic laws might have had a hard time anticipating)? It seems as if there is no clear way in which an illegal parker can be “paid back”. The same thing could be said for many crimes (e.g. tax fraud, treason, speeding, etc.). The lex talionis is immoral, moreover, in that it seems to recommend punishments that most of us would reject. It seems abominable to torture the torturer, or rape the ra unpist (see however Reiman 1995).

It is instructive to note that few, if any, retributivists have endorsed a literal interpretation of retribution. 104 Indeed, not even in the Mosaic Laws - the heartland of exact retributivism - is the lex talionis advocated in its literal sense. A central passage in the Old Testament states: “If anyone injures his neighbour, whatever he has done must be done to him:

104 Even Immanuel Kant, who forcefully advocated the lex talionis, admits exceptions from it. There are for example cases of killing where Kant’s general recommendation to take a life for a life is not operative: one soldier’s killing of another soldier in a duel, and a mother’s killing of her illegitimate child (Kant 1998: 107-9).
fracture for fracture, eye for eye, tooth for tooth. As he has injured the other, so he is to be injured” (Leviticus 24:19-20). This certainly seems to, at least where “injuries” are involved, be straightforwardly in favour of retribution as sameness. But the Mosaic Laws include plenty of examples where this quest for exact retribution is overridden. Almost ludicrously dependent on the death penalty, the laws state, for instance, that adultery, blasphemy, breaking of the Sabbath, and the cursing-of-one’s-parents is punishable by death. None of these “crimes” would seem to motivate death as punishment if one follows lex talionis literally.

Conceiving proportionality between crime and punishment as sameness is neither viable, nor desirable. It should be duly noted, however, that many find the gist of the lex talionis as just about right. When a bully gets bullied, a scam-artist gets scammed, or a wife-batterer gets battered, some feel strongly that this is just. People like that “had it coming” and “gets a taste of their own medicine”. Similarly, we are often upset when rotten people “get away with it” and perhaps wish there was a cosmic order that ensured that each was given his or her due. But although some can experience a certain satisfaction when a wrongdoer undergoes the same hardship as his or her victims, this is an instance of poetic justice and should not serve as the basis for policy. In particular, a retributive scheme of punishment should not pay attention to the feelings of crime victims, or people close to them.

---

105 These examples are mainly found in Leviticus 20. Reportedly, there are no less than 36 offences are punishable by death in the Mosaic Law (see Megivern 1997).

106 Further examples of the lex talionis being displaced include the burning down of another’s crop. This crime, we are told, should not be punished by burning down the offender’s crop in turn. The offender should rather compensate the victim materially.

107 Punishing in accordance with desert requires that considerations about the feelings of crime victims are excluded, at least in the standard sense (it could be argued that having caused ill-feelings is the basis of desert, so that having angered at lot of people greatly through a crime means that the crime is more severe than one which hurts slightly the feelings of few). For instance, victims’ thirst for revenge or capacity for forgiveness should not determine the wrongdoer’s punishment – the consequence of this could clearly be that crimes against forgiving people would be punished too leniently and crimes against vengeful people too harshly. Retributivism, in imagining that giving people their due is an
Retributivism by proportionality thus cannot – or should not - mean sameness. Since sameness won’t work, the theory is left with having to interpret what kind of punishment is proportional to what kind of crime, based on the loose understanding that graver wrongdoings merit more severe punishments. There are two major ways to interpret the notion of proportionality. None of them necessarily leads to “tough” penal attitudes.

Two Interpretations of Proportionality

Proportionality may be understood in two ways (Duff 2001: 133f). It may be seen as an *absolute* concept. This is to say that particular acts of criminal wrongdoing intrinsically merit specific punishments (without saying that the punishments are about the literal sameness between crime and punishment, as the lex talionis does in its caricature form). Or it may be seen as a *relative* concept. This means that proportionality is nothing more than the idea that the greatest crime should merit the harshest punishment in the set of available punishments, and so on in a complete downward ranking, without making any claims about what punishment a wrong in an absolute sense merits outside of the set. Both conceptions of proportionality holds that there should minimally be an ordinal match between crimes and punishment, and preferably also a cardinal one.

108 An *ordinal* ranking concerns only the order of the ranked items, whereas a *cardinal* one in addition brings in considerations of the distance between items. Thus, if there are two crimes and two punishments, an ordinal ranking consists in reserving the more severe punishment for the more severe crime. Cardinaly, if the one crime is twice as bad as the other, we also seek to punish the former twice as severely as the latter (see e.g. Elster 2006). Contrary to some who claims that ordinal ranking is sufficient (e.g. Duff 2001), I would suggest that the cardinal ranking is preferable (though possibly a mirage). Ordinal rankings may be inadequate. Consider a state where the most severe penalty ($P$) is a lot more severe than its second-most severe punishment ($p$). Suppose further that the two most serous crimes are almost equally serious (where crime $C$ is the worst and crime $c$ almost as bad). Ordinally, we would have to say that it would be propor-
The relative interpretation of proportionality, however, holds that crimes should stand in relative proportion to the set of punishments, but does not care whether the set in itself is soft or tough. The absolute interpretation of proportionality, on the other hand, says that each crime merits a specific response, and that a just penal regime will, as far as possible, punish crimes in accordance with what they objectively merit. Contrary to relative proportionality, absolute proportionality thus seeks to answer such questions as where the upper and lower limits of a penal scheme should be placed.

In order to illustrate this distinction, let \( \{A, B, C\} \) be the possible acts the state punishes. Assume further that \( A \) is the greatest and \( C \) the smallest wrong, so that \( A > B > C \), where “>” stands for “is worse than”. Let \( \{\text{Severe, Medium, Mild}\} \) be the three available punishments in the set of punishments, indicating their degree of severity. Assume finally that each punishment applies to only one criminal act, so that if \( C \) is punished with Mild, \( A \) and \( B \) cannot be punished with Mild.

Now, any retributive theory will say that \( PP \) implicates that \( A \) should be more severely punished than \( B \), which should be more severely punished than \( C \). Thus, ordinally we get the penal scheme \( \{A/\text{Severe}, B/\text{Medium}, C/\text{Mild}\} \). This order may never be reversed or otherwise offset. The idea of proportionality sets limits here and this is usually recognised as a merit of retributivism.

Returning to the distinction above, however, an absolute idea of proportionality requires that one judges the specific content of the crimes in question. Suppose that crime \( A \) is twice as bad as \( B \), which is twice as bad as \( C \). Assume further that \( C \) is proportionally punished with Mild, which has a “severity value” of 20.\(^\text{109} \) This means that Medium should have a value of 40, and Severe a value of 80. We are committed to an absolute

\(^{109}\) The values are of course arbitrary. The reader may see the numerals as months in prison, but I stipulate them for the sake of illustration and - hopefully - clarity.
interpretation of proportionality if we believe that 20, 40 and 80 are in themselves appropriate punishments for the respective acts. We may for instance believe that murder is a crime that is to be punished by death. That is, we believe the death penalty is the proportionate punishment for murder – a view that, according to the absolute interpretation, may hold regardless of whether the set of punishments in question contains capital punishment or not. In this sense proportionate punishment means that the punishment is as severe as the crime in question objectively merits. If $C$ ought to be punished with 10, but $C$ is de facto punished by 20, then a retributivist will conclude that the punishment is out of proportion as it is too severe.

The absolute approach, which holds that wrongful acts have a value that should be codified in the state’s penal regime, is probably the more common retributive approach. But one could also think along different, more relative lines. A relative interpretation of proportionality does not care whether a set of punishment is tough or not as long as ordinal (and ideally cardinal) proportionality is upheld. Suppose a state punishes $C$ with Mild(20). Assume furthermore, as we did above, that $A$ is twice as bad as $B$ and $B$ twice as bad as $C$, so that $B$ is punished with Medium(40) and $A$ with Severe(80). This is a scenario consistent with PP. But now assume that we shift the entire set of punishments in a less severe direction, so that Mild is 10, Medium 20, and Severe 40. In a relative sense, we would still obtain cardinal proportionality. Given that $A$ is twice as bad as $B$, which is twice as bad as $C$, and given that we cannot say conclusively for any of the acts ($A$, $B$, $C$) that they objectively deserve such-and-such punishment, the situations below – one called “Soft on crime” and the other “Tough on crime” - may be said to be equally proportionate.

---

110 It is consistent with PP in a formal sense at least. In the scenario above, we might still disagree with the order of the crimes, i.e. disagree with the notion that $A > B > C$. We might also reject the notion that one or more of the punishments are appropriately severe. Finally, we may contest the idea that there is an equi-distance between $A$ and $B$ and $B$ and $C$. 
Table 3.1 Two proportional penal schemes

<table>
<thead>
<tr>
<th>Type of crime</th>
<th>Tough on crime</th>
<th>Soft on crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Severe (80)</td>
<td>Severe (40)</td>
</tr>
<tr>
<td>B</td>
<td>Medium (40)</td>
<td>Medium (20)</td>
</tr>
<tr>
<td>C</td>
<td>Mild (20)</td>
<td>Mild (10)</td>
</tr>
</tbody>
</table>

Comment: The figure assumes that crime A is twice as bad as crime B, which is twice as bad as crime C. Mild, Medium and Severe are the types of punishments. The values in parenthesis arbitrarily indicate degree of severity of the punishment, where bigger numbers are more severe than smaller ones (e.g. months in prison).

If the two penal regimes may be said to be equally proportionate, is not the whole idea of relative proportionality suspect? If a retributivist could choose either one of the regimes and still claim to be proportional, is not the choice essentially arbitrary? The retributivist Corlett has treated this question (2001: 100-101). Suppose we have a choice between two penal codes, which Corlett calls the “Barbaric Code” and the “Humanitarian Code”. Both are proportional in the relative sense outlined above, but the Barbaric Code is much tougher than the Humanitarian Code. How do we know which one is justified? Corlett concedes that the choice between them is arbitrary, but argues that this is not a problem unique to retributivism since the principle of proportionality is important to theories of punishment in general.\[111\]

Corlett’s point is perhaps a bit strong, as one can at least logically imagine a penal code based on deterrence that disregards proportionality completely, while it would be impossible to conceive a retributive penal code not based on proportionality. However, his point serves to illustrate

---

\[111\] Indeed, the principle of proportionality is subject to scrutiny and approval in such classical works on deterrence as Beccaria’s (1986: 14-16) and Bentham’s (1988: 178-188).
why it is wrong to say that retributivism is necessarily tough on crime. If a retributivist employs a relative interpretation of proportionality, taking an interest only in the internal orders of crimes and punishments and making no attempts to specify what a certain crime merits in an absolute sense, then it is obviously incorrect to say that he or she necessarily has tough penal attitudes. He or she will not care about the toughness or softness of the penal code – for instance its upper and lower limits - but only about whether punishment is proportional given the ranking of crimes and the rankings of punishments. If a retributivist wholly or partially employs an absolute interpretation of proportionality, however, he or she will care about whether the penal code is tough or soft. It is possible that the choice between toughness and softness is essentially arbitrary. But if the choice is arbitrary, this means that it is made based on his or her prior notion of proportionality. A “tough” retributivist will opt for the harsher penal code while a “soft” retributivist will opt for milder punishments.

Most likely retributivists are moved by considerations of both absolute and relative nature. The appropriateness of certain punishments for certain crimes – most notably the death penalty for murder – is often debated, and this reveals that there is an interest for what kind of reaction a class of crimes in an absolute sense merits. We often have ideas about whether some particular punishment is too lenient or too severe for its crime which are not based on internal comparisons with other punishments in the set. At the same time, one has to acknowledge that it seems very hard reach any level of precision here. The objective benchmarks or Archimedean points of a punitive regime are not easily found – indeed, they probably cannot be “found” at all but must be postulated. And even if we could somehow establish beyond all doubts that, say, shoplifting normally merits such-and-such punishment, we would still have to wrestle with how much better or worse than shoplifting other crimes are. (Is robbery ten or twenty times worse than shoplifting?) The difficulties involved in penalty fixing notwithstanding, it should be clear that labelling retributivism as necessarily “tough on crime” is incorrect. Even considering a retributivist who gives proportionality an absolute interpretation, it seems clear that whether he or she is tough on crime or not depends on whether his or her prior idea of proportionality is tough or not. He or she may defend a very tough set of punishments, but he or she may also, if this is his or her idea of proportionality, defend a soft one – even a considerably softer one than those most existing states em-
ploy. Lacking an absolute standard of proportionality, these kinds of differences are likely to remain. This is not to deny that retributivists in practice often have defended tougher penal regimes. But there is nothing in retributivism that in itself implicates being “tough on crime”. I have for example often been told that my own standard of proportionality is quite “soft.”

RETRIBUTIVISM AS POINTLESS REVENGE. Many tend to think that retributivism is barbaric and inhumane: that it represents an outdated and arcane morality that is “primitive and unenlightened” (Rachels 1997: 470). When retributivism states that there is something intrinsically valuable about punishing wrongdoers, many feel that this is a mere rationalization of a more basic urge, namely the desire to “bite back” and get revenge. It advocates, according to these critics, the imposition of pain for pain’s sake, and thus flies in the face of such values as forgiveness and mercy. Examples of the bad press retributivism gets are not hard to find. Liberman (2006) defines retributive beliefs as anti-humanitarian. Tännö thinks that retributivism falls on its own “inherent cruelty” (1998: 166). Dolinko, warns us that “adherence to retributivism may well encourage greater vindictiveness and a peculiarly self-righteous and smug indulgence of our society’s most punitive reflexes” (Dolinko 1991: 559). The list could go on.

The upshot of these remarks can perhaps best be summed up in a charge that retributivism is a barbaric and inhumane theory. To label the theory as barbaric is, it should be noted, not an argument against it. It is a question of semantics and does not present us with reasons to reject (or accept) the theory. However, since retributivists have been adamant to rebut the charge of inhumanity – and indeed have argued that retributivism is a humane theory since it “respects persons” – it nevertheless seems to be a case of semantics with some importance. Few would like to advocate a “barbaric and inhumane” theory, so the question of appropriate name-calling is relevant.

Now, if the belief that punishment has some valuable characteristics that are not of the normal utilitarian sort is in itself barbaric and inhumane, then retributivism is indeed (by definition) barbaric and inhumane. But to advocate this view is non-argumentative to the extreme and
few seem to have resorted to it. It is not very charitable to reject a premise of a moral theory on the grounds that it is essentially immoral. What has been questioned is rather whether retribution is the same thing as revenge. On the widespread view that revenge is prima facie wrong this is a question that seems relevant, for if retribution is similar to revenge and revenge is wrong, then retribution must be wrong too. On this note Pereboom has argued that “Perhaps the deepest problem for [retributivism] is that retributivist sentiments are at root vengeful desires, and that and that therefore retribution has little more plausibility than then vengeance as a morally sound policy for action” (Pereboom 2001: 160).

It seems clear, however, that this critique rests on confusion about the respective nature of revenge and retribution. Revenge and retribution are not the same. Nozick (1981) has provided a list of differences between the two concepts, which I find reasonably complete.  

1. “Retribution is done for a wrong, while revenge may be done for any injury or harm or slight and need not be for a wrong” (Nozick 1981: 366). Thus, retribution requires some wrong in order for it to be deserved. It is therefore more naturally connected to justice than revenge, which may be taken against the undeserving.

2. Contrary to retribution, revenge has no internal limit. The revenge taken on an enemy might be wildly out of proportion to the harm suffered, whereas retribution is limited by its nature. For all its bad press, a variety of the retributive proportionality principle – the infamous doctrine of “an eye for an eye” – is actually limited and proba-

---

112 Murphy has refuted this criticism: “Typically, the charge is that infliction of punishment on such grounds [i.e. retributive] is nothing more than pointless vengeance. But what is meant by the claim that the infliction is ‘pointless’? If ‘pointless’ is tacitly being analyzed as ‘disutilitarian’, then the whole question is simply being begged. You cannot refute a retributive theory merely by noting that it is a retributive theory and not a utilitarian theory. This is to confuse rede- scription with refutation and involves an argument whose circularity is not even complicated enough to be interesting” (Murphy 1995: 13).

113 See however Honderich 2006: 180.
bly worked as a conservative principle when it was formulated. After all, it says “an eye for an eye” – not “your head for my eye”.

3. Revenge is personal, while retribution requires no personal ties to the victim. Retribution can be performed by a detached third party (in cases of criminal wrongdoing, the state). Revenge, on the other hand, is typically about the harmed parties harming the harmers back in return.

4. Revenge and retribution differ in “emotional tone”. “Revenge involves (…) pleasure in the suffering of another, while retribution either need involve no emotional tone, or involves another one, namely, the pleasure at justice being done.” (ibid. 367) Retribution, contrary to revenge, can be done dispassionately. One can retribute without being angry or resentful.

5. Revenge requires no generality, whereas retribution does. “Not only is the revenger not committed to revenging any similar acts done to anyone; he is not committed to avenging all done to himself. (…) Whereas the imposer of retribution, inflicting deserved punishment for a wrong, is committed to (the existence of some) general principles (prima facie) mandating punishment in other similar situations” (ibid. 368). In short, since retributivism is a general theory it can aspire to be a theory of justice. Revenge cannot.

This list should caution us not to speak too frivolously about retribution as revenge. Nozick admits that revenge and retribution in some cases will be mixed, and that sometimes the desire for revenge and the desire for retribution can be intertwined. However, revenge and retribution

---

114 It might also be that retributivism has grown out of a deeper desire of “biting back”, which emanates from a desire for revenge deeply entrenched in human nature (see e.g. Mackie 1982). Retaliation, or negative reciprocity, is for instance on Brown’s list of “human universals” – it occurs in all human cultures (see
can be disentangled, and it is therefore incorrect to run them together.\textsuperscript{115} Retributivists have as a rule emphasised that they are \emph{not} advocating revenge, but rather justice.

This is not the end of the story, however. Even granted that revenge and retribution are separate concepts, and that it is question-begging to assume that a justification of punishment must be of a utilitarian variety or else fail, what animates claims that retributivism is about “pointless revenge” is presumably that the theory, even taken on its own account, cannot spell out an intelligible point of punishment. Even if we concede that punishment can coherently be said to be give offenders what they deserve, why should we think that this is important? What \emph{good} does it do? These reflections take us into the realm of valid criticisms of the retributive theory.

\textit{Problems}

Retributivism faces plenty of problems. One problem, which we have sketched above, concerns proportionality, or more specifically how we are to determine which punishment is proportional to which crime. Of course, we can all ballpark proportionality. Life imprisonment is too severe a punishment for shoplifting, and a monetary fine is too lenient a punishment for murder. But how do we arrive at more precise judgements? Settling for a relative conception of proportionality solves some problems, but by no means all of them. Deterrence and rehabilitation as penal aims face similar “practical” problems: how do we know what level of punishment is the optimal trade-off between compliance and suffering, and how do we effectively rehabilitate offenders? But the problems associated with the principle of proportionality suggest that it

\textsuperscript{115} Note, however, that there is nothing in Nozick’s distinctions that precludes that private individuals can engage in retribution; a question I do not treat in this book.
would be incorrect to say that the retributive theory, unlike the competing theories, is free from non-arbitrary elements.

Another problem, which we encountered in chapter 2, concerns which classes of wrongdoing should be legally punished. Some actions are generally taken to be morally wrong yet are not outlawed. How do we draw the line between, say, the least severe legally prohibited wrongdoing and the most severe legally permitted wrongdoing? It seems as if this question, if coherently answerable, involves bringing in factors extraneous to the theory.\footnote{See Fischer 2006, who argues that we can answer questions of what acts to outlaw by bringing in considerations of harm and offensiveness.}

The majority of objections levelled at retributivism have to do with the concept of desert, however. The retributive theory is most easily characterized as the view that punishment is justified if, when, and to the extent that it is (morally) deserved. This idea can be questioned in two distinct ways. On the one hand, one can question the \textit{sufficiency} of desert. Granted that \( X \) deserves to be punished, why should this fact about \( X \) play any justificatory role? Why should the fact that \( X \) deserves pain or deprivation count as a reason for actually imposing pain or deprivation on \( X \)? On the other hand, one could question the \textit{existence} of desert. This is compatible with granting desert potentially great moral significance. It only questions whether rule breakers really are deserving of punishment. Marx, for instance, thought that retributivism is the correct theory of punishment in principle, but argued that since actual societies are marred by deep inequalities, actual societies are rarely justified in punishing (see Murphy 1995). Others argue that desert cannot exist because it requires a particular theory of agency which science has proven false. If \( X \) breaks the rules because of various factors beyond his or her control, such as his or her genes or social background, then it is surely wrong to say that \( X \) deserves punishment. His or her crime is less a wicked choice than a product of causal forces, they conclude.

* * *

Retributivism is perhaps at once the most and least problematic theory of punishment treated in this chapter. One the one hand, retributivism seems to be in accord with many of the things we ordinarily think about
punishment: that one should not be punished unless one has committed a criminal wrong; that those guilty of greater crimes should receive more severe punishment; and that there is an irreducibly backward-looking element to punishment. However, for all these intuitively appealing things, the theory also faces objections of profound nature. Perhaps most pressing is the suspicion that retributivism fails to address the real reason why we are justified in punishing. Utilitarian theories seem to present a more straightforward and plausible answer to both why we punish as a general practice, and why we are (sometimes) justified in punishing actual rule breakers – that doing so makes society and its members better off.

I will return to these questions shortly, when I elaborate on the defence of retributivism I undertake in this book. An idea that we need to address immediately, however, is the idea that we could mix all three theories into a composite justification of punishment, and consequently construct a penal regime which at once seeks to retribute, deter and rehabilitate.

3.5 AN ENDURING CHOICE: COMPARING THE THEORIES

Recall that the aim of this book is to answer (Q), which concerns which principle or theory a penal regime should be based on, i.e. which principle the penal practice should adhere to. The theories of deterrence, rehabilitation and retribution are theories which purport to answer that question. Each theory tells a story about what it is that makes punishment right, and what the practice of punishment should therefore seek to do. According to deterrentism, punishment is justified by making potential rule breakers refrain from acting on criminal desires. This idea leads to a penal regime which attaches sanctions that are sufficiently threatening to acts of rule breaking. Rehabilitationalism holds that it is right to intervene in the lives of offenders since doing so allows us to address their criminal desires. This idea leads to a treatment-based penal regime which is designed to break criminal inclinations. Retributivism, finally, holds that punishment is right if it gives criminal wrongdoers what they deserve. This idea leads to a punishment-based penal regime governed by certain rules (e.g. the principle of proportionality).
My claim will be that retributivism answers (Q): that we have reason to prefer a retributive penal regime over ones based on deterrence or rehabilitation. In response to this claim, it could reasonably be asked why we have to choose between these theories at all, as each of them clearly makes sense. Couldn’t we simply say that all the theories bring something important to the table, and conclude that the practice of punishment ought to simultaneously reform, deter and give people what they deserve?

Such a conclusion is tempting. It seems plausible that a penal regime that incorporates the ideas of all three theories would, other things being equal, be better. Instead of promoting one aim and dropping the others, it would promote all three. Unfortunately, the theories are to a higher or lower degree incompatible with one another, both in terms of the strategy of justification and in terms of the institutional design they suggest. Thus, it is not only ideas about what aspect of punishment that justifies interfering without consent in the lives of individuals that come into conflict. The theories also conflict at the level of fairly concrete political practice.\(^{117}\) In order to see these incompatibilities, consider the following cases.

*Retribution vs. Deterrence.* If deterrence is set forth as the aim of punishment, then some policies would be acceptable which retributivism would

\(^{117}\) Thus, I do not subscribe to the view that one can accept a justification of punishment without this having implications for institutional design. Some have suggested the opposite, for instance that desert is what justifies punishment, but that we may have other rational aims of engaging in punishment (e.g. deterrence). But if one accepts that desert is the moral justification of punishment, this presumably has consequences for how one designs the penal regime (and the same goes for rehabilitation or deterrence). It seems for instance absurd to believe that one could accept rehabilitation as the moral justification of punishment but design the penal regime around the aim of giving wrongdoers what they deserve. Such a view, it seems, would lead to that plenty of offenders are unjustifiably punished - if rehabilitation is what counts, surely we would employ various institutional techniques, such as indeterminate sentencing and treatment programs, which a penal regime based on retribution seems unlikely to include. A special case, however, is the notion that the questions regarding what justifies a rule and what justifies distributions under that rule are fundamentally distinct. I will return to this question in chapter 5.
find unacceptable. Deterrence needs not coincide with proportionality, for instance. The quest for deterrence may lead to punishment that from a retributive point of view would be too lenient or too severe. A real world example of this conflict can be seen in the use of the “three-strikes-and-you’re-out”-scheme and similar crime control programs. The idea behind these programs is that past crimes increase the punishment for future crimes, for instance that the third serious offence leads to life imprisonment. This is believed to reduce recidivism - by raising the severity of punishment the stakes are raised and potential rule breakers will think twice before risking further punishment. From a deterrence point of view this can make sense. But a retributive regime of punishment aims at punishing the criminal in accordance with what the crime deserves. And past criminality arguably doesn’t affect the wrongness of a present action.\textsuperscript{118} Desert seems particularly offset if later crimes become increasingly minor while the punishments become increasingly severe. This is only one example where retribution and deterrence can conflict, both on a philosophical and institutional level.

\textit{Retribution vs. Rehabilitation}. Retribution and rehabilitation can clearly also come into conflict. What is effective from a rehabilitative point of view need not be appropriate from a retributive point of view. The opening chapter suggested just such a case: what should be done about men who batter women? A rehabilitative position would be to answer that question with what works. If the behaviour of abusive men can be changed through treatment, then this is what should be done. From a retributive point of view, we said, it can always be argued that such rehabilitation – whether or not it is effective – is inconsistent with the purpose of punishment, which is to give responsible wrongdoers what they deserve. Thus, even if there were highly effective ways of changing the behaviour of these men in a non-punitive manner, a retributivist could still argue that to opt for these ways would be morally unjustified.

A major conflict between rehabilitation and retribution is the one concerning determinate-indeterminate sentencing. Rehabilitationalists emphasise that punishment, or treatment, should be about changing the \textit{criminal person}. Therefore it is a good idea to sentence people indetermi-
nately (e.g. between five and ten years in prison), since authorities can then delay the release of potentially dangerous and unreformed offenders and offer an early release to those who respond well to treatment. Retributivists as a rule emphasise punishing the *criminal action*, not the person. The criminal should get the penalty his or her action merits. Determinate sentencing is seen as an expression of this endeavour, and also as a bulwark against arbitrariness in sentencing. A person should not receive a harsher sentence for his or her crime than comparable persons, simply because he or she seems, say, less accepting of the community’s majority values.

*Retribution vs. Rehabilitation.* As we have seen, deterrence and rehabilitation share important common ground. They both regard punishment as a necessary evil: there is nothing intrinsically valuable about the practice of punishment or about being punished; punishment can only be justified to the extent it produces valuable consequences for the rule breaker or society at large. Often, but not always, are those consequences understood in terms of the aggregate welfare rule-compliance creates.

But deterrence and rehabilitation theories can also conflict, especially if rehabilitation is construed as a moral obligation we owe those who need help. If we pursue metaphor of crime as illness we will see this clearly. Some would argue that the rehabilitation of criminals follows the same logic as the treatment of the ill: just as we have an obligation to treat those who are ill, we have an obligation to help those who suffer from crime-causing disorders. To punish rather than to offer treatment seems very unjust from this perspective – one does not punish those who suffer from pneumonia or cancer, so why should one punish those who suffer from behavioural disorders or the like? But as we have seen deterrence theory can justify punishing some to the benefit of others in a way that offsets the illness analogy, for even “sick” criminals can be used as messages to potential rule breakers. Rehabilitation and deterrence can also come into conflict concerning the role of pain in punishment. Deterrence sees pain or deprivation as necessary to punishment: in order to count as a reason against breaking the rules, punishment must be seen as unpleasant. Rehabilitation, as we have seen, sees no such need to incorporate pain as a necessary component of punishment (or “punishment”)
– it will incorporate pain to the extent it is effective for rehabilitating offenders.\textsuperscript{119}

If a utilitarian were to choose between rehabilitation and deterrence as guiding theories, he or she would settle, as usual, for the theory that seems to best promote aggregate utility.\textsuperscript{120} In this sense, there is no incompatibility between the two theories. But note that not even a utilitarian can pursue rehabilitation and deterrence simultaneously, since they suggest different institutional designs. That which is likely to deter is not always that which is likely to rehabilitate. For example, just like a retributivist, a deterrentist will advocate determinate sentencing, while a rehabilitationalist will advocate indeterminate sentencing. Moreover, a deterrentist will see the penal regime as a system of pain – uncomfortable penalties which function as foreboding signs for the public to read – whereas a rehabilitationalist will be more willing to try different, non-punitive ways of responding to crime.

* * *

There is some purchase to saying that a penal regime cannot or should not express or build on all three principles at once. Since the theories we are considering here represent different justifications of punishment

\textsuperscript{119} The difference between rehabilitation and special deterrence (deterring the criminal from future crimes by inducing pain) is, as we have seen, that special deterrence does not require that the criminal desires of the offender changes. A criminal that out of fear of punishment refrains from committing crimes he or she would like to commit is effectively speaking not rehabilitated – he or she merely responds to threats in the intended way. A rehabilitated criminal, on the other hand, is normally understood as one who no longer desires, or desires less, to commit crimes.

\textsuperscript{120} Brian Barry has noted, in a funny quote, that the basic empirical nature of utilitarian answers to moral problems means that utilitarianism is an unexciting brand of philosophy: “Adherence to utilitarianism makes for very boring political philosophy, because once the goal has been postulated (some version of Bentham’s ‘greatest happiness principle’), everything else is a matter of arguing about the most efficacious means to that end. Is there a duty to obey the law? It depends on the consequences for aggregate utility. Can governments legitimately be overthrown by force? Same answer. Is economic equality desirable? Same answer again” (Barry 1990: xxxv).
and/or different institutional architectures, a penal regime which tries to accommodate them simultaneously will arguably be contradictory and confusing. These are rival accounts of what a penal regime should do. If we believe that one account is more plausible than the others, we cannot accommodate the others without adhering less to the account we found most attractive in the first place. This is not to deny that states in practice often have penal regimes that compromise between different aims and justifications, taking little interest in whether the regimes would be more coherent or effective if they paid more attention to incompatibilities of the sort discussed above. To a political theorist, however, such mixed regimes are a bit like the bumblebee: they fly even though they in theory shouldn’t be able to. My assumption here, at any rate, is that we are justified in searching for the best answer to (Q), and that the best answer cannot incorporate more than one theory if the theories are rival.

3.6 GOING FORWARD: QUESTIONS OF RETRIBUTIVISM

Since the idea of bringing in all theories into a composite penal regime seems untenable, we are left with having to choose between them. As declared, my intention here is to defend fashioning the state’s penal regime according to retributivism.

Why choose retributivism? What are the arguments in its favour? Answering this question actually coincides with defending retributivism against the first of the two principal objections I mentioned earlier: (O₁) the objection from insufficiency. This objection concerns an alleged “gap” in retributive thought, more specifically that desert is not sufficient to establish that punishment is justified. Chapters 4-6 are to be understood as an attempt to meet that objection, and it is here that my argument in favour of choosing retributivism can be found. I shall argue that a retributive penal regime, in trying to give rule breakers what they deserve, has merits that make it preferable to rival regimes. By identifying the merits of a retributive penal regime, ruled by the principle of desert, we shall have identified something which allows us to traverse the alleged gap.

The other objection, (O₂) the objection from non-existence, is animated by the idea that a scientific conception of crime and criminality will offset the notion that criminals deserve punishment. The objection raises diffi-
cult questions concerning scientific explanations, responsibility, free will and determinism. I shall deal with this objection in chapters 7-9. The rest of this chapter will be devoted to presenting (O1).

The gap of retributivism: the objection from insufficiency

Why should one take a presumption in favour of retributivism as the answer to (Q)? Many, myself included, find that there is something intuitively appealing about the way the theory connects the practice of punishment to such notions as proportionality, guilt, and responsibility. This serves to make the whole idea of letting desert be the guiding principle of criminal justice attractive.

David Dolinko (1991; 1997) has suggested, however, that there is a gap in retributive thought, which concerns exactly desert and its capacity to be such a guiding principle of punishment. The gap, he contends, consists in the distance between two logically separate claims. These claims are:

1. X deserves punishment
2. Punishing X is morally justified

Retributivists hold that punishment is justified when it gives the rule breaker what he or she deserves in virtue of his or her actions. Like desert theorists in general, they hold that if X deserves some treatment or thing A (in virtue of some desert-basis P), this counts as a reason, albeit perhaps not a conclusive one, for giving A to X. Olsaretti has argued that desert-claims in this way have moral force: “To say that someone deserves something is to claim that she ought, other things being equal, to get that thing, or that it would be morally better if she get it” (Olsaretti 2003b: 4). But why should we hold this? Why should someone deserving some blame or punishment (or praise or reward) count as a moral reason to give it to him or her? We can certainly all imagine cases where it is true both that X deserves A and that we should not give X to A. It might be, for instance, that a deranged man, in virtue of such-and-such considerations, deserves to get his gun back. But why should we take this to be a conclusive reason for returning the gun, or even a reason for it at all?
These are the kind of questions Dolinko raises in charging retributivism with having a gap.

Note that Dolinko might be proving a bit too much here. If there is a gap between (1) and (2) above, there seems also to be a gap between the following two claims:

(3) X does not deserve punishment
(4) Punishing X is morally unjustified

Yet here I suspect that many would be willing to say that (4) is entailed, although perhaps not in a strictly logical sense, by (3).\textsuperscript{121} Be that as it may, Dolinko suggests that retributivists must produce some additional reason why we are justified in punishing the deserving. He writes:

“The claim is that punishment – which involves doing to wrongdoers things that we ordinarily think of as violating people’s rights, like incarcerating them against their will for years – is morally permissible because it is what wrongdoers deserve. Yet we do not, in general, believe that treating a person in a way that would otherwise violate his rights is automatically permissible simply because this person deserves this kind of treatment” (Dolinko 1991: 543-44).

Note that unlike Dolinko, I am here considering not only why desert makes punishment permissible, but also why desert is a \textit{pro tanto} reason for punishment. Note also that I am not concerned with questions about when desert can be such a reason and when it cannot. I suspect that for any desert-claim put forth one could raise Dolinko’s gap. Regardless of whether we are discussing punishment or something else, we can ask why somebody deserving whatever thing \(A\) should count as grounds for saying that we would be justified in giving \(A\) to him or her. The gap first

---

\textsuperscript{121} Since (3) is (quasi)factual, it follows from Hume’s law that we need an additional premise of normative kind, i.e. one that states that we are never justified in punishing the innocent, in order to deduce (4). But a retributivist may hold that this intermediate premise is so widely and strongly held that the operation is (almost) automatically performed.
of all questions the very point of acting on desert, not why we should act on desert in one area as opposed to another.122

Do retributivists have to answer Dolinko? Could they not simply say that giving someone what he or she deserves is good in itself, and thus conclude that there is no gap between claims (1) and (2)? Recall that Rawls earlier described retributivism as the view that “The state of affairs where a wrongdoer suffers punishment is morally better than the state of affairs where he does not”. Couldn’t we settle for this view: that we think it is intrinsically good that the guilty suffer punishment, and that the penal regime therefore should be designed to punish the guilty? Is this any more problematic than when we, say, support proportional taxation simply because it counters undeserved differences in wealth (and not because it is good for social cohesion), or when we support giving good student papers high grades simply because they deserve praise (and not because it encourages students to try harder)?123

Such a view is certainly possible, and some retributivists do seem to say that punishing wrongdoers is a good in itself. Kant, as we saw, took his retributive convictions about justice so far as to bombastically declare that if “justice goes, there is no longer any value in human beings’ living on the earth” (Kant 1998: 105). Small wonder, then, that we are right to

122 Dolinko’s (1991: 545) strategy is to present cases where we do not want to act on desert (such as overriding a will that leaves one inheritor better off and another worse off than what they deserve). He then concludes that since we are unwilling to act on desert in such situations, it is unclear why acting on desert should be taken as a reason for punishment. (See Fischer 2006 for a recent response, which argues that there is a gap in retributivism but it is not one that threatens the coherence of the theory.)

123 However, even if we allow that giving someone what he or she deserves is good in itself, we are still left with whether or not this goodness is sufficient to make acting on desert morally proper. That is, we are still left with the difference between the claims (1) A’s receiving the punishment he deserves is intrinsically good and (2) It is morally permissible (or proper) to inflict the deserved punishment on A (see Dolinko 1997: 525). Strict retributivism is the idea that there is no gap between these two claims. I will defend a non-strict version of the theory, and thus grant that (2) does not always follow (1). There are other values, which might lead us to conclude that punishment of the deserving is not morally proper, such as when punishment leads to sufficiently dire results (imagine, for instance, that punishing A would mean the start of World War III).
punish those who deserve it: if the “value in human beings’ living on the earth” depends on it, it’s a no-brainer. Kant even came up a (in)famous example to illustrate the importance he attaches to justice, the so called Island-example. Kant writes:

“Even if a civil society were to be dissolved by the consent of all its members (e.g., if a people inhabiting an island decided to separate and disperse throughout the world), the last murderer remaining in prison would first have to be executed, so that each has done to him what his deeds deserve and blood guilt does not cling to the people for not having insisted upon this punishment” (ibid. 106).

Here it seems as if Kant advocates punishing the last murderer in spite of the fact that, as the society is about to dissolve, there is no longer a society to protect. And we should do so in spite of no tangible benefits to anyone (let’s disregard the part about the “blood guilt” clinging to people and that the murderer may pose a future risk to other societies). This has been taken to mean that we should punish those who deserve it simply because it is intrinsically good or just. Honderich refers to this position as intrinsic-good retributivism (Honderich 2006: 24-9).

Intrinsic-good retributivism rejects Dolinko’s gap. Indeed, on a strict version of that theory, claim (1), if true, is a conclusive reason to punish,

124 I am aware that it might be a mistake to interpret Kant as a bold and strict retributivist. A number of writers have found passages where Kant seems to endorse a less strict form of retributivism. Scheid, for instance, argues that Kant subscribed to a consequentialistic justification of the institution of punishment, and uses the concept of desert only to establish the permissibility of punishment. That is, according to Scheid Kant believes that punishment is justified since it promotes public safety, but that the institution must take desert into account to ensure that people aren’t treated as mere means instead of ends in themselves (e.g. punishing an innocent in order to maximize social utility) (Scheid 1993). Corlett, among others, notes that Kant seems to accept reform and deterrence as subsidiary aims of punishment; if it is established that a person is guilty, then we may proceed to ask how to punish (or not punish) in order to effectively reform or deter (Corlett 2001). But the Island example has in a sense outgrown Kant: whether or not it expresses what Kant himself thought, it nevertheless describes a position that seems in need of analysis.

125 Needless to say, intrinsic-good retributivism is a species of what was previously called positive retributivism.
which certainly means that claim (2) is entailed. (If we have a perfect moral obligation to punish wrongdoers, the claim that punishing X is morally justified strictly follows from the claim that X deserves punishment). But is intrinsic-good retributivism plausible? Honderich, for one, thinks not. He finds it a circular position, which basically states that punishment is right because it is right. To say that intrinsic-good retributivism is circular is a tad question-begging; in particular if one presupposes that a justification of punishment must ultimately be cast in terms of utility or else fail. But in order to see the problems associated with intrinsic-good retributivism, consider the case of The Repentant Wrongdoer.

The Repentant Wrongdoer. A man has broken a legal rule and is liable to be punished. He is morally responsible for the act - he broke the rule willingly and knowingly - and accepts full responsibility for it. In fact, he is very repentant and would consent to his own punishment, would he to be punished. However, he pledges that, whether or not he is punished, he will never break another rule again. Suppose that the members of his community somehow knew that this was true: whether or not the man is punished, he would comply perfectly with the rules for the rest of his life. Suppose further that no member of his community, victims of his crime included, wants to punish him, but instead wants to grant him mercy. Suppose, finally, that the choice between punishing him and not punishing him has no consequences for the general law-abidingness in the community. No one will be more prone to break rules if the community forgoes punishing him.

Should we punish the Repentant Wrongdoer under these circumstances? A strict intrinsic-good retributivist, it seems, would say yes. Note that the way the case is construed, there is no deterrence to be gained from punishing him, nor any need to rehabilitate him. Furthermore, there are no grievances to be satisfied by punishing him, nor will anyone feel unfairly

126 However, see note 123 above
127 He also finds it is a case of moral madness which calls for diagnosis rather than argument; see Honderich 2006: 25f, 192.
treated if he is granted clemency. What sense does it then make to say that we should punish him? Doesn’t this prove that intrinsic-good retributivism is fundamentally mistaken – that it advocates punishment in cases where it seems entirely pointless, indeed wrong?

We could pump all manner of intuitions from The Repentant Wrongdoer. It does seem pointless to punish him. However, suppose that he is punished. What would we then be able to say? I suggest that we can say all manner of things, but not that he has been unjustly treated. To the extent we hold that intuition, it is clear that retributivism retains some force. Based on backward-looking considerations the Repentant Wrongdoer deserves a punishment, and if he is punished, it cannot be said that he has been treated unfairly or unjustly.

Regardless of this, we might simply not buy a theory which says that the Repentant Wrongdoer should be punished. It seems so pointless. What good will it do? As Glover has remarked in relation to Kant’s island-example:

“To impose suffering or deprivation on someone, or to take his life, is something that those of us who are not retributivists think needs very strong justification in terms of benefits, either to the person concerned or to other people. The retributivist has to say either that the claims of justice can make it right to harm someone where no one benefits, or else cite the curiously metaphysical ‘benefits’ of justice being done, such as Kant’s concern that we should have ‘blood-guiltiness’ removed” (Glover 1990: 229).

This is, in a nutshell, Dolinko’s gap.

Retributivists have tried to spell out various ways to traverse the gap which do not boil down to the assertion that acting on criminal desert is simply intrinsically good. The problem is that many, if not all, of the attempts have been found wanting. I will here briefly survey some of the more prominent attempts, and point to why they seem unsuccessful.

A popular idea is that punishment somehow restores the balance upset by the crime. Hegel, for instance, spoke of punishment as crime annulment – a crime represents a false moral claim, and punishment, by

\[\text{\textsuperscript{128}}\text{Dolinko (1991) himself examines and rejects a few attempts to traverse the gap (attempts by Hampton, Sher, Moore, and Morris). The analyses of Hampton and Morris below are indebted to him.}\]
correcting that claim, serves to annul the crime (Hegel 2005; see also J. L. Anderson 1999). This idea has garnered little support, basically because it is difficult to see how a crime, once it has happened, could be annulled at all. There are modern variations of Hegel’s idea, however. Most famous is Morris’s idea that punishment removes the unfair advantage a criminal illegitimately has gained over the law-abiding. On this view, the most important building block of social order is equal compliance with the rules. Society is premised upon a contract which says that all for the sake of the common good should accept a degree of self-restraint, i.e. refrain from doing things contrary to common rules of behaviour. A criminal casts off the burdens of self-restraint accepted by the compliant, and in effect free rides on social cooperation.\textsuperscript{129} Punishment is appropriate since it removes the criminal’s unfair advantage over others, and thus restores the balance (Morris 1968).

Morris’s attempt, while influential, is open to severe criticism. It has been noted that if a criminal’s advantage is his or her freedom from self-restraint, it seems as if the advantage gained is unduly sensitive to the temptations of people in general. A heinous crime which is contemplated by very few would then seem to be a lesser wrong than an everyday crime which tempts many, as one in the former case throws off a “burden” which others do not find difficult to carry (Dolinko 2005). Furthermore, the notion of “unfair advantage” seems to misrepresent what makes (most) crimes wrong. If I assault you, it is counterintuitive to think that it is the fact that I am free-riding on others self-restraint which makes me at fault. It seems more appropriate to say that I am fault because of some more direct quality of my act, e.g. because I harm you.

Jean Hampton has suggested yet another twist on Hegel’s idea of annulment. Punishment, on her view, is justified since it reinstates the equal dignity or worth of the victim. A criminal, Hampton contends, expresses his or her superiority over the victim. He or she (tacitly) says, “I can do this”. Punishment falsifies that claim; it humbles the criminal at the hands of the law (Hampton 1988). But this view seems unpromising.

\textsuperscript{129} A thief, for instance, enjoys the benefits of living in a society where property rights are upheld, but does not accept the burdens of self-restraint for him- or herself. Thereby, the thief gains an unfair advantage over others. He or she enjoys the fruits of others’ cooperation (i.e. does not have to fear that his or her own stuff will be stolen), but does not contribute to the common good.
too. First of all, some crimes are without victims, at least in any direct sense of the word. It is difficult to see who is declared inferior by me possessing illegal narcotics, for instance. Hampton’s theory cannot explain why it might be justified to punish victimless offences. Furthermore, it is not clear why punishment is needed to make the criminal see the errors of his or her ways. Could we not make criminals see the equal worth of their victims by, say, hammering in that point by subjecting them to a lecture-series on egalitarian thought?

So why should desert count as a *pro tanto* reason to punish? Honderich (2006) has suggested a particularly bleak answer. Crime evokes grievances of varying strength. Punishment may serve to satisfy those grievances; to satisfy, that is, desires to see the wrongdoer in distress. Retributivism, in Honderich’s mind, boils down to the view that punishment is right precisely since it has this effect. Interestingly, there is now a clear good that punishment produces. It produces grievance satisfaction, which, Honderich claims, is something we want. Honderich himself thinks that we cannot justify the infliction of pain or deprivation on a person with such immaterial a good as the (vengeful) pleasure of seeing rule breakers in distress. But grievance satisfaction is at least an intelligible attempt to traverse Dolinko’s gap, which turns on a mentalist conception of value.

Note, however, that the notion of grievance satisfaction is clearly of a consequentialist variety. It is not uncommon for people to defend punishment on the ground that it allows a community to express its outrage over a crime, or that it validates people’s conception of justice. Such functions of punishment are among other things believed to foster social cohesion.\(^\text{130}\) Sometimes such views have been referred to as “retributive” (see Cottingham 1979). But these considerations take us a long way from what is classically understood as retributivism. For if social cohesion is what we are after, why not punish innocents, thereby allowing an unwitting public to become more firmly compliant with the prevailing norms? If there is such a thing as a coherent retributive theory, it surely cannot abandon the view that there is something in punishing *the deserving*, and only the deserving, that makes moral sense.

\(^{130}\) See e.g. Durkheim 1964 and Lord Denning’s classical claim that the ultimate justification of any punishment is that it is “the emphatic denunciation by the community of a crime” (cited in Cottingham 1979: 245).
Attempting to bridge the gap

Small wonder, if grievance satisfaction is the basic rationale for retribution, that people want to have criminals punished. They get satisfaction from it. However, this cannot be the base-line defence of a plausible retributive justification of punishment. What needs to be done, if we are going to justify retribution as the basis of a penal regime, is to shift focus from the parties doing the punishing (the state, and in extension the general public) to the party being punished. We need to think about whether retributive punishment is something acceptable or valuable even for the person undergoing it. This move may seem audacious. But only if there is some value in retribution, which even the person being punished can recognise as important, can we begin to traverse Dolinko’s gap in a way that neither appeals to an intrinsic and unanalyzable goodness involved in giving people what they deserve, so unacceptable to some, nor to the idea that punishment simply satisfies the grievances of law-abiding people (an idea which would seem to firmly propel retributivism back into rationalizing revenge).

Recall what we said in chapter one: a justification requires that we demonstrate the value of something “in a rational and impartial way” (Bird 2006: 13). This presents us with an ethical hurdle: when we justify something, we should seek to do so with reasons which all in principle can accept - including those who might be adversely affected by it. In our case, this means that if we are going to justify a retributive penal regime, we must do so in a way which is in principle acceptable even to those who are punished by it.

Note that I am not here suggesting that we play with words. A punishment, properly understood, is unpleasant and painful to those who suffer it. If there is some value in being punished, it is certainly not as directly oriented towards promoting the good of the person being interfered with such as in the case of treatment, which is defined by being intended to make the treated person better rather than worse off. I will not suggest that those undergoing punishment like to do so, nor that they see it as being in their immediate interest. What I do intend to suggest, however, is that there are values inherent in designing the penal regime around the concept of desert, and that there are also values inherent in
being treated in accordance with one’s desert. These values can rationally be recognized even by those who undergo it.

More specifically related to (Q), what I plan to do over the next three chapters is to offer a defence of a retributive penal regime: to show why all have reason to prefer a penal regime built on the basis of retribution over regimes based on deterrence on rehabilitation. I shall begin this task by taking a look at one of history’s most (in)famous philosophical claims: Hegel’s contention that punishment is a criminal’s right.
Part II. A Defence of Retributivism
4. The Strange ‘Right to be Punished’

“...because the primary rules are designed to benefit all and because the punishments prescribed for their violation are publicized [...]
there is some plausibility in the exaggerated claim that in choosing to do an act violative of the rules an individual has chosen to be punished.” (Morris 1968: 479)

GERMAN philosopher Hegel, one of history’s most influential advocates of retributivism, once said that punishment is the right of the criminal. Anthony Quinton has replied to this idea by noting that “it is an odd sort of right whose holders would strenuously resist its recognition” (cited in Tamburrini 1996: 124-25). Quinton’s witticism captures nicely what is likely to be the first reaction to the idea: A right is something one would rather attain or keep. And how bizarre, masochistic an idea that says that criminals want to be punished! (And the amused chuckles are soon followed by scornful comments about “German idealism” and “metaphysics.”)

The basic reason why Hegel’s idea seems preposterous is that it would appear irrational for the criminal to desire that harm be inflicted on him or her. In our present times, when economists have touted the “calculating egoists” assumption of rational choice theory for quite some time, it perhaps appears more so than ever. A crude rational analysis of the interests involved in an act of punishment would namely look like this:

131 Paraphrased from Hegel 2005: 36-37. Hegel’s original (translated) version is as follows: “The injury which the criminal experiences is inherently just because it expresses his own inherent will, is a visible proof of his freedom and is his right.”
Everybody benefits from a rule \( R \) that, say, prohibits assault. The integrity and safety of one’s person, which is a good all desire, is promoted by \( R \). Still, an individual \( S \) might want to assault somebody in spite of \( R \). Suppose \( S \) gambles that he or she won’t get caught and breaks \( R \), but is caught after all. The interest of \( S \) in this situation is clearly to escape punishment (for instance by being excused). Others, by contrast, have an understandable interest in punishing \( S \), hoping that enforcing \( R \) will make \( R \) more respected by \( S \) and others in the future, and making sure that their own compliance with \( R \) is not a “sucker’s payoff”.

Wouldn’t this appear self-evident? If it was up to the rule breaker, the best outcome would be to enjoy the benefits of \( R \) while being able to break \( R \) at will, thus free riding on others’ compliance. In the words of John Rawls: “The absolutely best for any man is that everyone join with him in furthering his conception of the good. Or failing this, that all others are required to act justly but that he is authorized to exempt himself as he pleases” (Rawls 2000: 103. Emphasis added.).

It would appear so, I suppose. Though many actual rule breakers are likely to understand the rationale for why acts of rule breaking should be punished, it is nevertheless probably the case that most of them, if given the choice, would prefer escaping punishment to being punished. The fact of the matter seems to be that it is society – the rest of us – that has an interest in punishing \( S \); the exception perhaps being those criminals who recognise that they need help to come to grips with addictions, mental disorders, and so forth. \( S \) would normally, as Quinton suggests, forgo his or her “right to be punished”.

But it is obvious that we need to invoke a distinction between desires or preferences of a more immediate nature from more abstract ones if we are to see the point behind Hegel’s claim. We may begin this task by briefly studying the disagreement between Beccaria and Kant. In the 18th century, the aforementioned legal theorist and contractarian Cesare Beccaria (1986) argued against the death penalty on the grounds that rational individuals could not be assumed to have granted the state a right to capital punishment. When forming a social contract it is unlikely that they would give the power over life and death – life being the highest of all goods – to society. Rational individuals, Beccaria thought, would secure a right to life and would not consent to any regime that included the death penalty. (Would they consent to life imprisonment? 15 years?)
Kant wrote a fuming denunciation of Beccaria’s argument, calling it “sophistry and juristic trickery” (Kant 1998: 108). Kant argued that:

“No one suffers punishment because he has willed it but because he has willed a punishable action, for it is no punishment if what is done to someone is what he wills, and it is impossible to will to be punished. – Saying that I will to be punished if I murder someone is saying nothing more than that I subject myself together with everyone else to the laws...” (ibid.)

We may debate whether Kant attacks a straw man here, but he clearly has a point.\textsuperscript{132} When people say about a punishment that it was the rule breaker’s own choice, they do not mean that the rule breaker has chosen the punishment per se. Rather, they mean that the rule breaker has chosen to engage in action which, as a foreseeable result, risks incurring a certain punishment. When we say that people might be rational to want punishment, we are therefore not saying that they actually want the actual punishment. Whatever goes through a rule breaker’s mind when he or she is being fined, imprisoned or otherwise harmed, it is not likely to be blissful happiness. His or her synapses won’t churn away an emphatic “yes”.\textsuperscript{133}

\section*{4.1 WOULD WANT AND SHOULD WANT}

Let’s accept, with Kant, that it is impossible to want one’s own punishment. If the basis of the claim that a criminal has a right to be punished is that punishment is what the criminal really desires, then we may safely

\textsuperscript{132} The straw man being to confuse the individual as a contracting party on the one hand and the individual as an actual criminal on the other. It is doubtful if Beccaria committed this fallacy: he seems to argue that contracting parties, not actual criminals, would not consent to capital punishment.

\textsuperscript{133} Indeed, Kant raises the curious question whether the punishment of someone who desires to be so treated can be thought of as punishment at all. (There is a joke about a masochist and a sadist. The masochist says, “Spank me”. The sadist replies, “No”.)
conclude that the claim is false. It is a fairly safe generalization that most people would not want their own punishment. Assuming that the alternatives are not worse, most criminals would opt for escaping punishment if given the choice.134

Or does this depend on the nature of the rule breaker? Would rule breakers who have retributivist beliefs about criminal justice be a different story? Corlett thinks so. Part of the reason why mercy is often regarded as suspect by retributivists, he argues, is that retributivists are people who would not like the state to have mercy even on themselves. Thus, according to Corlett, “the retributivist just is one who herself wants to be punished if she deserves it” (2001: 106). If this is taken to be an actual empirical claim, one suspects that there are very few retributivists around at all. Even criminals who have retributive intuitions about justice are likely to opt for escaping their own punishment, if given the choice.

To say that denying a criminal his or her punishment equals thwarting one of his or her actual desires therefore seems untenable. The reason for saying that punishment is the criminal’s right is instead presumably that there are important things about being punished, which makes punishment somehow in the interest of the criminal. We could perhaps state it as follows: no criminal de facto wants to be punished, but there may be a sense in which they should want it. It is in this sense we might be able to mount a defence of Hegel’s claim that punishment is a “right of the criminal”. It might be the case that everyone – the criminal included – has an interest in rule breaking being retributively punished, and in this sense should want it.

We must be careful here, for the notion of punishment as a “right” easily lends itself to ludicrous scenes (imagine a criminal, devastated by his or her sentence, being told: “Don’t you know we’re doing this in your interest. You should be happy.”).135 If everyone has an interest in rule

134 Thus, we may imagine punishment being preferable to some forms of civil commitment. Kesey’s classic novel One Flew Over the Cuckoo’s Nest (1963) is an argument to this effect, I think.
135 For a similar point, see Murphy 1995: 16. Dolinko further notes some dangers of talking of a criminal’s hypothetical consent to his or her own punishment: “it is difficult to understand the justificatory force of a fictitious or hypothetical
breaking being punished, this interest need not be felt as something valuable or agreeable. But having said that, nor can it be a completely esoteric interest that some philosophers, for unclear reasons, have decided is important. What we need, if we are to defend Hegel’s claim, is an account of why everybody – including, in a particular way, the persons being punished – has reason to prefer retributive punishment to other kinds of responses such as being excused or being subjected to coercive treatment.

Recall that we are here considering possible ways to traverse the gap between “X deserves punishment” and “punishing X is morally justified”. We are searching for ways of moving from the first claim to the second, and doing so out of reasons which even those punished can rationally accept. There are different ways of doing this. One is to argue that a retributive penal regime, which rests on the assumption that desert justifies punishment, is somehow preferable to other penal regimes. If everyone accepts this, one could argue that it is incoherent to refuse one’s own punishment under that regime, should one break the behavioural rules. For one could not reasonably claim to have a right to enjoy the benefits of that regime without also having to accept the consequences of it for oneself.\footnote{Such a way of arguing fits with Scanlon’s influential contractual scheme, which famously holds that right and wrong are judgements about “what would be permitted by principles that could not reasonably be rejected” (Scanlon 1999: 4). One could not say for instance, that the egoist described by Rawls above (he who wanted to exempt himself from rules as he pleases) has reasonable grounds for rejecting punishment. For a recent debate on treating punishment in Scanlonian terms, see Brettschneider 2007a; Lovett 2007; Brettschneider 2007b.} One could, like Murphy, construe this as a contractual choice:

> “On this theory, a man may be said to rationally will X if, and only if, X is called for by a rule that the man would necessarily have adopted in the original position of choice – i.e. in a position of coming together with others to pick rules for the regulation of their mutual affairs […] Thus I can be said to will my own punishment if, in an antecedent position of choice, I and my fellows would have chosen institutions of punishment as the most rational means of dealing with those who ‘consent’. In our case, how could such a fiction legitimate the harsh treatment of a real, and unconsenting, offender?” (Dolinko 2005: 78).
might break the other generally beneficial social rules that had been adopted” (Murphy 1995: 16).

If this is how the choice of a retributive penal regime can be conceived, we could say that the criminal indeed has chosen his or her own punishment, and that it is something which he or she should accept.

One may also argue that retribution as an act is something in itself valuable to all, including the ones suffering it. On this view, retributive punishment, while unpleasant, is something one should want because its nature is such that it is preferable to other responses (such as rehabilitation, punishment-as-deterrence, or an all-out excuse). Hegel himself seems to have had something like this in mind. Retribution according to him acknowledges oneself as a “rational being”. In a key passage of The Philosophy of Right he writes:

“[…] by being punished he [the criminal] is honoured as a rational being. He does not receive this due of honour unless the concept and measure of his punishment are derived from his own act. Still less does he receive it if he is treated either as a harmful animal who has to be made harmless, or with a view to deterring or reforming him” (2005: 37, section 100).

Thus, for Hegel (and some subsequent theorists), retributive punishment is important since it respects the rule breakers as persons. By this is intended, roughly, that retribution communicates an empowering message: if you are retributively punished, this means that you are seen as an autonomous moral agent that has the capacity for personal responsibility and deserved praise and blame. This contention is held to be more attractive than being punished as a means of changing behaviour. It is also held to be more attractive than being offered treatment, which according to retributivists boils down to being seen as a victim of circumstances. This, the story goes, is something which all could recognise as correct.

The underlying idea here is that if one can construe retribution as something even rule breakers value and consent to, one has taken some steps to show why acting on desert can justify the practice of punishment. The logic I propose is the following: There are particular merits to a retributive penal regime; one which aims to give rule breakers what they deserve. On the one hand, it represents an arrangement of the practice of punishment which is in alignment with a number of principles of justice. On the other hand, it treats rule breakers with respect. These mer-
its are such that all can and should value a retributive penal regime, including those punished under it. This means that all who break the rules can and should prefer retributive punishment to other penal responses. For to prefer a retributive penal regime is to prefer retribution as a penal response, even if one happens to be subjected to it. This is so either directly or indirectly. We prefer it directly if we prefer retribution to other penal responses. We prefer it indirectly if we in general prefer a retributive penal regime to other regimes. For one cannot coherently prefer a regime which engages in giving rule breakers what they deserve without also, by implication, prefer that oneself is given what one deserves, should one break the rules under that regime.

The strategy employed here is thus to try to show why all, including the rule breakers, have reason to prefer retributive punishment to other penal responses. It is in this sense one could argue that retributive punishment is the right of criminals, i.e. that it lies in their interest. I take this to have some consequences for Dolinko’s gap. My claim is that there are general merits to a penal regime which aims to give $X$ what he or she deserves. This means that we have reason to value such a regime, and value it in a rational and impartial way. The conclusion of this seems to go in the direction that a retributive penal regime is justified. It will be noted, however, that this reasoning does not traverse Dolinko’s gap in a head-on and complete way. I will defend that there are things about giving rule breakers what they deserve which serve to justify it. This need not mean that we counter Dolinko’s claim that desert as such is insufficient to justify punishment.

The next two chapters are devoted to presenting two merits of a retributive penal regime. In chapter 5, I argue that a retributive penal regime is in better alignment with a number of principles of justice than rival regimes. I shall refer to this as the institutional reason. In chapter 6 I develop the idea that retribution, unlike treatment or punishment-as-deterrence, respects the rule breaker as a person. This boils down to the claim that a retributive penal regime represents a more appropriate mode of treating people than rival regimes. Suitably qualified, I believe this thought is basically correct and that it is politically and socially important. I shall call it the symbolic reason to prefer retributive punishment,
since it claims that something important is being communicated by the act of retribution (i.e. a “symbol”).  

The argument in chapters 5 and 6 are my arguments in favour of answering (Q) along retributive lines. I close chapter 6 by making an assessment of whether the arguments are adequate to close Dolinko’s gap, and to present the fundamental objection to retributivism which is the subject of part III of this book.

137 It will perhaps be noted that I oscillate between retribution being preferable to any alternative response and retribution being preferable to deterrence and rehabilitation. There is surely a difference between the two claims, and the disclaimer from chapter 3 still holds (I only compare retribution to punishment-as-deterrence and rehabilitation). However, while I cannot exclude that there is some theory of punishment which we should prefer to retributivism, we should keep in mind that retributive punishment should be compared to other plausible consequences of breaking the rules. That is, we may assume that something will happen to the rule breaker, and that it won’t be altogether pleasant. We should thus not compare retribution to a situation where, say, our rule breaking gives us a year’s consumption of free mojitos, a hefty chunk of money, and a state-financed carnival held in our honour.
5. The Institutional Reason for Retribution

There are different ways of arranging the state’s penal regime. A merit of arranging it according to retribution, retributivists hold, is that doing so will ensure a fairer, more just regime. In particular, a penal regime that aims at giving rule breakers what they deserve will have merits over ones which aim to maximize utility. This means, as I said in the previous chapter, that all – rule breakers included – can agree that a retributive penal regime is preferable. This institutional reason for retribution (or retributivism) is the subject of the present chapter.

5.1 THE INSTITUTIONAL REASON TO PREFER PUNISHMENT

We may begin by noting that there is a longstanding line of thought which holds that punishment is preferable to more radical expressions of the rehabilitative theory. Some have argued that we need an altogether different approach to criminal policy, replacing the ideas of the “old penology” with a far-reaching treatment model.¹³⁸ We should discard our old ideas about desert, responsibility and the imposition of pain, and instead focus on therapy-based or medical approaches to crime control. Criminals should not be subjected to pain or hardship in the name of justice. Instead, they should be offered effective treatment for their maladies. Against this treatment model, those who defend a penal regime

¹³⁸ Bertrand Russell, whom we quoted in chapter 1, seems to have advocated this. See also Menninger 1968 and Skinner 1971.
based on punishment have argued that a system of punishment (unlike one based on treatment) is foreseeable and procedurally fair. One famous argument to this effect is Morris (1968): in the choice between a system of punishment and a system of treatment we would (or should) choose the former, in part because doing so allows us to control what happens to us under that system. A system of treatment, on the other hand, would open up for irregular and arbitrary exercises of power.

This criticism of a treatment model of crime control is quite common. It has primarily been raised against the backdrop of the significant and historically speaking quite recent influence that forensic psychiatry has gained over states’ penal regimes (for it was not until the advent of modern psychiatry, and its alleged ability to explain, predict and treat behavioural disorders, that a seemingly viable alternative to traditional penal regimes emerged). Along with the increased influence of forensic psychiatry came the criticism that this development threatened the rights of citizens, in particular those connected to the rule of law. This criticism is multifaceted. Some argue that psychiatry is a value-laden discipline, whose definitions of mental disorder are merely expressions of social norms. Psychiatry, according to these critics, is a discipline which rests on social prejudices (e.g. values of conformity) and which functions as a device of social control. Others argue that psychiatry deprives patients of their rights and treats them capriciously. A longstanding criticism of the commitment of criminals to mental hospitals is that such commitment, unlike punishment, puts people in a situation where untarnished paternalism reigns: the patient is at the mercy of politicians or doctors, who have the power to keep patients as long as they see fit. This, it has been argued, subjects the patients to the arbitrary will of others. For these reasons should criminal and public alike prefer traditional punishment to therapeutic treatment, some have concluded (Szasz 1974; 1997; Kittrie 1971; Kesey 1963).

Assuming there is merit to this criticism, it should be noted that it only establishes the superiority of a penal regime based on punishment, generically construed, over one based exclusively on treatment. But this is not sufficient to prove that a retributive arrangement of punishment is superior to other arrangements of punishment. It is thus necessary for

139 For historical accounts of the emergence of forensic psychiatry, see Conrad & Schneider 1992; Qvarsell 1993.
retributivists to defend not only punishment against treatment, but also retributive punishment against others forms of punishment, in order for them to prove that a retributive regime is institutionally superior to other regimes.

Here I will look at a recent attempt to demonstrate why a retributive regime is superior to its alternatives: James Rachels’s essay “Punishment and Desert” (1997), a stringent and elegantly reasoned piece which seems to prove, by means of a simple test, the superiority of a retributive penal regime.

5.2 RACHELS’S TEST

Rachels begins “Punishment and Desert” by an account of what it means to treat people as they deserve, and why it is desirable to do so. He argues that to treat people as they deserve is to treat them in the same way in which they have (voluntarily) treated others. The reason why we should adhere to this principle of desert is that it gives people control over how they are treated by others, encourages reciprocity (without which, Rachels argues, imperfectly benevolent beings would have little reason to behave morally), and produces fair distributions of burdens and benefits. Retributivism, Rachels notes, is the extension of the principle of desert to the special case of criminal punishment: “it is the view that people who commit crimes […] deserves to be punished and that this alone is sufficient to justify punishing them” (Rachels 1997: 474).

Having made the connection between a general desert-based ethics and retributivism, Rachels goes on to present the two main alternative theories of punishment: deterrence and rehabilitation. Deterrence he defines as the view that, in the interest of general welfare, we “ensure compliance with […] rules, [by attaching] sanctions to them” (1997: 475). Rehabilitation is defined as the view that rather than simply punishing rule breakers we should “address the problems that caused their aberrant behaviour” – rule breakers should be “educated, trained and treated, with an eye to making them into ‘productive members of society’ who will not repeat their offences.” (ibid.)

Having presented retribution, deterrence and rehabilitation as the main alternatives, Rachels acknowledges that all three theories are plau-
sible. But he thinks that it is retribution that in the end should determine the design of the penal regime.

“As for the deterrence theory, there is no doubt that sanctions are useful. They ensure massive, if imperfect, compliance with the social rules. That would probably be enough to justify punishment even if there was no other argument available. Moreover, it is hard to deny that rehabilitating criminals would be a good thing, if we only knew how to do it. Yet it can still be argued that the criminal justice system should not be designed to promote deterrence or rehabilitation. Rather, it should be designed along lines suggested by the retributive theory” (Rachels 1997: 476).

What would such an argument look like? Rachels here introduces the idea of a compatibility test, where each theory is tested against four principles of justice. The principles are:

(1) **Guilt** – The principle that only the guilty may be punished.
(2) **Equal treatment** – The principle that “like cases be treated like”. If A and B have committed the same crime they should get the same punishment.
(3) **Proportionality** – The principle that the punishment should be proportional to the crime; that the punishment should be “commensurate with the harms one caused to others” (Corlett 2003: 286).
(4) **Excuses** – The principle that if A has a valid excuse for what he or she did, A should not be punished (or be punished less severely).\(^{140}\)

We may note two things about these principles. The first is that they are not all independent of each other. (1) and (4) rests on the same assumption – that punishment may only be inflicted on the deserving (i.e. the guilty).\(^{141}\) Similarly (1), (3) and (4) all express the idea that moral desert is

---

\(^{140}\) I will go to some lengths to analyse the concept of excuse in chapter 7.

\(^{141}\) Assuming that “guilty” entails some moral element such as when we say “you were morally wrong to do X, therefore you are guilty of X”. But the term “guilty” might be applied to cases where mere causation is required. This we
an important element when punishing. If you have a valid excuse for what you did, then the proportional punishment will be lowered, if not withheld entirely.

A second thing to note is that the principles may appear conspicuously “retributive” in character. Seen as they are chosen as part of a defense of retributivism, this naturally leads to the thought that Rachels is rigging the test so that it will come out in favor of the preferred theory. I will deal with this objection in a short while, but let me here say that just because the test in one sense is “rigged” it doesn’t mean that it proves no substantial point.

Rachels’s plan of attack is to run each theory against principles (1)-(4) and see whether the theories adhere to or violate to them. His general idea is that “Any system of punishment is unjust if it departs from these four principles” (ibid. 477). This is an imprecise way of expressing how the test is going to work. Is violation of one principle sufficient to reject a theory? Are all principles equally weighty, so that we can rank the theories quantitatively, based on their respective number of violations? If two or more theories are able to satisfy each principle, are those theories equally good? The reason why Rachels is not clearer on this point is presumably what he takes to be the conclusive results of the tests: it turns out that retributivism manages to satisfy all four principles while deterrence and rehabilitation violate three each. Here is how the test unfolds:

**DETERRENCE.** Deterrence according to Rachel's violates principles (1), (3) and (4). Concerning (1) the principle of guilt, he writes:

> “There is no reason, if we are concerned only with deterrence to punish only the guilty. As far as deterrence is concerned what matters is not whether the person punished is guilty, but whether he or she is generally believed to be guilty. If people believe she is guilty, the deterrent effect will be the same as if she really were guilty. Moreover, from this point of view, it would be much better to convict an innocent person, if not guilty, might illustrate by the statement “You did X; therefore you are guilty of X”. If this is what Rachels intends by the principle of guilt, then (1) and (4) are independent. (1) would ensure that only those who stand in the relevant causal relation to X should be punished while (2), as a second step, would ensure that among those in the relevant causal relation, only those responsible would be punished.
son (…) than for the crime to go ‘unsolved’, because when crimes are unsolved people get the idea that the law is ineffective and the deterrent effect of the law is diminished” (1997: 477).

That deterrence admits of punishing the innocent is a classical criticism that has been levelled at utilitarianism for quite some time, and it is of course not one that has been accepted as conclusive.\textsuperscript{142} I will return to this question shortly.

As for (3) the principle of proportionality, Rachels argues that there is no guarantee that deterrence is served by making the punishment “fit” the crime. If we are serious about producing compliance, it might even seem recommendable to increase the severity of punishments significantly. Life imprisonment for illegal parking would of course be a more effective deterrent than an insignificant fine, but it would be entirely out of proportion.\textsuperscript{143}

(4) The principle of excuses is also violated. Rachels contends that it might best from a perspective of deterrence to have a “no excuses accepted”-policy. If the would-be criminal knows that the law is rigidly implemented, he or she is more likely to be deterred. If the state excuses some crime, the criminal could perhaps calculate with the possibility of avoiding punishment even if he or she gets caught.\textsuperscript{144}

\textsuperscript{142} For arguments in this debate, see McCloskey 1963; Smart & Williams 1973; Murphy 1995; Rawls 2001; Honderich 2006.

\textsuperscript{143} A utilitarian deterrentist would of course reply that such an outrageous example misses the crucial distinction between utility and deterrence – a utilitarian does not promote deterrence at all costs, but weighs the gains in utility due to deterrence against other costs. In the case of life imprisonment for illegal parking, for instance, he or she will have to include the widespread feelings of insecurity and grief among motorists and the pain experienced by convicted illegal parkers. The weight of these is probably such that life imprisonment for illegal parking is not recommended by the calculus of utility. The point still stands, however, for there is nothing that in principle excludes that utility will be served, via deterrence, by disproportional punishment. This is, as we previously said, an empirical question.

\textsuperscript{144} Dennett, however, argues that there is an optimal point between punishing and excusing here beyond which the system receives diminishing returns, unsubstantiated from a calculus of utility (Dennett 1984: 156-63).
The only principle out of the four that deterrence satisfies, then, is (2) the principle of equal treatment. Rachels here argues in roughly the same manner as when he concludes that deterrence is promoted by not accepting any excuses: if punishments are inconsistent, criminals will calculate with the possibility of receiving a lesser penalty rather than a more severe one.

**REHABILITATION.** Rehabilitation violates principles (1), (2) and (3) according to Rachels. (1) The principle of guilt is violated since if we are interested in removing the causes of crime, it is not necessary relevant that someone has actually broken the law. The aim of rehabilitation, Rachels contends, “is to transform people who are inclined to commit crimes into people who are not inclined to commit crimes” (ibid. 477). This does not mean that we are interested in transforming only those who already have committed crime (although that is a great indicator of criminal “inclination”). It is just as relevant to transform those innocents who have yet to act out their criminal tendencies. Therefore, Rachels concludes, if rehabilitation is the aim, nothing prevents us from taking action against risk-prone people before they actually commit crimes.

(2) The principle of equal treatment is also violated. If rehabilitation is the aim of punishment, the duration of imprisonment will depend on how well the criminal responds to treatment. If A and B are convicted of similar crimes, A may be deprived of his or her liberty much longer than B if A is deemed to respond worse to rehabilitation than B. That the duration and form of “punishment” should be tailored to each individual case is of course an important element in the philosophy of rehabilitation, and its influence is seen in many actual penal practices (e.g. being committed indefinitely to a psychiatric clinic, parole contingent on good behaviour, indeterminate sentencing, etc.).

Rachels argues that (3) the principle of proportionality is violated for similar reasons: the rehabilitative regime’s response will be in accordance with the rehabilitative needs of the criminal, and not the seriousness of the crime. In fact, the idea of rehabilitation can be defined as the attempt to respond to the individual criminal and not to the crime. 145

---

145 This is all in theory, of course. It might thus be that a well-adjusted and sane murderer will require less treatment than a mentally disordered shoplifter who is addicted to drugs, and that a strict application of the rehabilitative idea will
The only principle the theory of rehabilitation does not violate, according to Rachels, is (4) the principle of excuses. For if \( A \) has a valid excuse for having broken a rule, then there is nothing to rehabilitate. It would for example be ridiculous to rehabilitate \( A \) for doing \( X \) if \( X \) was done accidentally.

Rachels is right about this only to the extent we consider excuses that are *not indicative of abnormalities* (e.g. being coerced by others to do \( X \), doing \( X \) by accident). But as we have seen, excuses might be of a different variety. Kleptomania, for instance, might count as an excuse for stealing. But it is the kind of excuse that very much seems to require rehabilitation. The theory of rehabilitation, then, will not opt for non-interference with criminals just because they are not responsible for their crimes.

**RETRIBUTION.** The retributive theory, in Rachels’s mind, fares much better. It violates none of the four principles; indeed the theory “incorporates [them] in the most natural way possible” (ibid. 478). Here is how he reasons.

(1) The principle of guilt is satisfied since punishment according to retributivism depends on desert, and innocent people have done nothing to deserve punishment.

As for the principle of equal treatment (2), Rachels argues that according to any desert-based theory it will be true, that if \( A \) and \( B \) have behaved in the same way they should get the same response. Therefore they are treated equally; like cases are treated alike. This might sound strange for readers who mistakenly think that retributivism necessarily advocates that punishment should be rigidly meted out according to the type of crime. In relation to the previously discussed example of the two shoplifters (the disadvantaged mother and the affluent cheap-skate) we said that some believe that retributivism recommends that both be equally punished – that this is what even-handed retributive justice requires. But we saw that this was incorrect, for if moral desert is the basis make us incarcerate the latter for a longer time than the former. In practice, however, the nature of the crime will be considered as a factor in judging the abnormality of the criminal. It can thus be argued, and indeed has frequently been argued, that “normal people don’t commit murder” and that the risks posed by a murderer compared to a shoplifter warrants that the murderer be incarcerated longer.
of punishment, it trivially follows that a mother of starving children deserves a lesser punishment – if any punishment at all – than the affluent cheap-skate. Two affluent cheap-skates, however, are equally deserving of punishment and therefore will be punished equally. A desert-based conception of morality is not in opposition to, nor invariably in accord with, principles of equal treatment.

(3) The principle of proportionality, which we encountered and questioned in chapter 3, is an unavoidable part of any retributive account. The murderer is more at fault, we normally think, than the shop-lifter and the former should therefore be punished more severely than the latter. A transitivity between crimes and punishments is therefore ensured in the retributive system. The retributive theory, moreover, holds that one should be treated in accordance with one’s desert, no more and no less. This is thus the idea that responses should be proportional to how well or badly one has acted, which in the case of punishment is captured by the slogan of “making the punishment fit the crime”.

Finally, (4) the principle of excuses is also satisfied. Rachels writes that “a retributive system of punishment would have to accept excuses, because what people deserve depends only on their voluntary behaviour. Acceptable excuses show that behaviour was not voluntary; that is why the demonstration that one was coerced, or that it was all an unavoidable accident, and the like, gets one off the hook” (ibid. 478). As I have noted in previous chapters, however, excuses pose a special problem for retributivism. It is no doubt correct to say that retribution admits space for

---

146 Bo Rothstein has suggested to me in conversation, however, that the idea of proportionality based on desert might be all wrong. According to the traditional account, the more harmful or wrong a crime is the more punishment the criminal deserves. But it can be argued that this is mistaken. Instead, it may be true that the more harmful or wrong the crime, the less responsible and thus deserving of punishment is the criminal, since the worse one behaves, the more reason we have to suspect that something is fundamentally wrong – mentally, genetically, or whatever – with the criminal. This clearly offsets the idea of more serious crimes being punished more seriously in the name of proportionality. Take as example epic wrongdoers such as Hitler, Stalin or Pol Pot. Do we believe that sane people could commit heinous crimes such as theirs? If not, are we committed to the belief that they were not responsible or less than fully responsible for them? The suspicion that graver crimes mean less desert is potentially devastating for the idea of proportional punishment.
excuses – if one does not deserve punishment, one should not be punished. Corlett has even noted that retributivism is compatible with a world where none is responsible.\(^{147}\) If none is responsible, the retributive theory simply recommends that no one be punished (Corlett 2001). However, this solution, while logically impeccable, might be a bit too glib. Much of retributive thought simply seems to assume that people by and large are responsible for their actions. It is thus doubtful whether the theory can sustain any amount of excuses. If no one deserves to be punished, then not only is retributivism’s traditional place in the debate as a watchdog over persons’ dignity and autonomy offset. Its recommended course of action – not punishing anyone – also appears problematic.

Back to Rachels’s test. As we have seen, Rachels finds that retributivism violates none of the four principles whereas the alternative theories violate three each. He summarizes this result in table-form (from Rachels 1997: 479):

<table>
<thead>
<tr>
<th></th>
<th>Retribution</th>
<th>Deterrence</th>
<th>Rehabilitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guilt (1)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Equal treatment (2)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Proportionality (3)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Excuses (4)</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Rachels from this concludes that we should design our penal regime along retributive lines since “a system of punishment based on retributive principles is fairer and more just than systems fashioned around after those other ideas” (ibid. 476). Indeed, he argues that “retributivism is the only idea that provides the basis of a just system of punishment. The other ideas do not even come close” (ibid.). To the extent all want the

\(^{147}\) The thesis that none is responsible is most famously advocated by hard determinists. I discuss the thesis of hard determinism in detail in chapter 8.
most just penal regime possible, then, we may argue, if Rachels is correct, that all should prefer a retributive penal regime.

5.3 RACHELS’S TEST – AN APPRAISAL

Rachels’s test seems to provide a powerful and intuitively plausible argument for the institutional benefits of retributivism. There are two main ways in which one can criticize the test. Either one criticizes the choice of principles, and maintains that Rachels’s principles are not the principles against which theories of punishment should be tested, or one criticizes the way Rachels reasons when it comes to whether the theories violate or satisfy the principles.

If one were to follow the first line of criticism, one would be forced to explain why Rachels’s principles are the wrong ones. It thus does not suffice to point to the fact that Rachels never explains why he has picked the four principles he has picked. Nor does it suffice to say that the principles are too “retributive”. It might be objected that Rachels rigs the test so that retributivism will come out on top, but unless one explains what is wrong with the principles Rachels has chosen, that criticism will have little weight.

No doubt an advocate of the deterrence theory would like to introduce other principles that a system of punishment must satisfy. “Capacity to sufficiently reduce crime” or “no imposition of pointless suffering” might be candidates, as they are principles, deterrentists assume, the retributive theory will have problems handling. (The second candidate, however, is probably rhetoric). An advocate of rehabilitation might want to introduce principles such as “ability to decrease recidivism” or “ability to treat criminals humanely”, which similarly are believed to expose the weak points of retributivism. Be that as it may: what is important here are the four principles at hand. Rachels never states that the four principles are the only ones a system of punishment can or must adhere to. What he says is instead that unless a system adheres to these principles, it

148 Ideally, one would also say what principles the theories should be tested against instead.
can never be just. To satisfy the four principles is a necessary (although perhaps not sufficient) condition for a justificatory theory of punishment. And since retributivism is the only theory (in the test) that satisfies the principles, retributivism is the only theory able to fulfil this condition.

Part of the appeal of Rachels’s test is that the principles he chooses to work with are all very plausible. (Running the theories against weak or strange principles would of course render the test worthless). They strike most of us as important aspects of what we loosely refer to as the rule of law. Most would agree, I think, that it seems highly unjust to punish innocents even if society as a whole would benefit from it; that it seems wrong and possibly dangerous to pre-emptively interfere in the lives of individuals who seem likely to become criminals (see however Morse 1999); and that proportionality and consistency in punishment are important aspects of a fair penal regime. If it is correct that rehabilitation and deterrence fails to respect these principles, then the theories indeed seem significantly weakened.

Assuming that we have now accepted Rachels’s choice of principles, let’s move on to the other main way of criticizing the test: the objection that Rachels mishandles the theories vis-à-vis the principles.

There are instances where Rachels’s conclusions about whether a theory violates a principle seem doubtful - the condensed nature of his text has the consequence that not every step in his argument is sufficiently developed - and I will not mention them all here. We have already seen that his analysis of the principle of excuses is incomplete. One might furthermore argue that deterrence does leave room for excuses. 149 It may also be objected that the principle of equal treatment is completely

---

149 The classical way of reaching this conclusion is as follows: suppose criminals of type X cannot be deterred, say because they suffer from a compulsive disorder. This means that a particular criminal of that type could not have been deterred by penal threats, and will not be less prone to commit future crimes simply in virtue of being punished. Furthermore, to punish him or her in order to deter other type X criminals is irrational, for they are also unimpressionable by penal threats. Therefore, it makes sense not to punish type X criminals. But as we saw in chapter 3, the fallacy of this argument is obvious: it might be that we by punishing unimpressionable type X criminals can deter other types of criminals, who are impressionable by penal threats (cf. Mackie 1990: 209-10). Dennett, however, has argued that a deterrence-based system will have utility-based reasons to employ some excuses (see footnote 144).
plastic and without specific commitment to retributivism.\textsuperscript{150} Overall, though, I find little wrong with the way Rachels’s handles the principles.

The gut reaction to Rachels’s test, however, is probably that he interprets the theories too strictly, and that this works to the disadvantage of the theories he wants to attack. An advocate of rehabilitation might for instance agree that the rehabilitative ideal \textit{in theory} admits of pre-emptive incarceration, but nevertheless maintain that this is unlikely to happen \textit{in practice} under rehabilitative regimes. Similarly, a deterrentist might admit that it is possible that the quest for deterrence can justify punishing the innocent, but hold that this is unlikely to be a viable course of action in practice. These objections boil down to the claim that Rachels is too strict in his analysis the principles, and that he produces caricatures of how deterrence and rehabilitation theorist actually think.

When facing objections of this sort, we must always ask whether the consequences we are sticking to others are rejected because the consequences do not follow from their beliefs, or because the consequences are uncomfortable. In the first case we have argued incorrectly and must withdraw our accusations. In the second case we have simply made our opponents feel uneasy, and it is their responsibility to either accept the consequences (“bite the bullet”) or abandon their prior beliefs. Much of what can be objected to Rachels’s test on behalf of deterrence or rehabilitation is, I think, of this second variety. But the objection that his test is too abstract to have anything to do with punitive practice warrants an answer.

The question, then, is whether one must abandon rehabilitation or deterrence as normative ideals of punishment if one accepts Rachels’s four principles and believes that his conclusions about them vis-à-vis the theories are correct. It is a question that must be answered by first determining what weight we should attach to normative ideals and abstract principles. I tend to support the view, that a normative ideal that in theory can support injustice should be rejected \textit{even though it may in practice seldom support or lead to that injustice}. That is, we can ask what different theories support or lead to in principle, and evaluate them accordingly.

\textsuperscript{150} Honderich notes that we could just as well suggest that equal treatment requires that the undeserving should be punished and that the deserving should not. As long as we treat all undeserving equally, we meet the requirements of the principle. Yet we violate completely retributivism. See Honderich 2006: 69-70.
Utilitarians of the deterrentist variety, for instance, have long claimed that even though utilitarianism may *in principle* support punishing the innocent, *in practice* it will never be of utility to do so. Since the general public is likely to come to know about the state’s frivolous attitude towards the principle of guilt, the standard reply goes, their sense of safety will decrease, and as a consequence punishing the innocent will have adverse effects on overall utility. Against this reply I would hold, that even if it is factually true that punishing the innocent is never recommended by the calculus of utility (which I doubt), we may still question the reasonableness of a theory that can support such an injustice in principle. The problem is not that a calculus of utility would or would not support punishing the innocents. The problem is that it is an open moral possibility, whose justifiability depends on factual circumstances. Thus, I tend to think that we may judge the reasonableness of theories or ideals with what they in principle can recommend, and not only what they are likely to produce in practice. And I think it is legitimate to reject the theory that, as it currently stands, can make problematic recommendations (or, since every theory faces its own set of problems, reject the theories that are most problematic).

John Rawls begins *A Theory of Justice* by saying that, “Justice is the first virtue of social institutions, as truth is to systems of thought. A theory however elegant or economical must be rejected if it is untrue; likewise laws and institutions no matter how efficient or well-arranged must be reformed or abolished if they are unjust” (Rawls 2000: 3). If this notion is complemented with the idea that institutions may be evaluated by how their regulative ideals hold up when subjected to theoretical scrutiny, then my conclusion is clear: among social institutions, the institutions of punishment is the one for which it seems most important that it adheres to principles of justice. And since retributivism seems to admit of a penal regime which better than its alternatives adheres to principles of justice, it is legitimate to reject the theories that are most problematic.  

---

151 Utilitarians have also tried to get out of the pickle by “solving” the problem at the level of definition. Punishment, they hold, is pain or deprivation inflicted on someone for an offence. Logically, there can thus be no punishment of the innocents. Such a “solution” leaves much to be wanted. Surely we can do the-kind-of-thing-otherwise-known-as-punishment to an innocent, even deliberatively so. And the question is precisely what utilitarian grounds there are for not doing it (see Armstrong 1961).
it is also the theory upon which the institution of punishment should be based. To the extent citizens prefer ideals that adheres to principles of justice to ideals that do so to a lesser extent, all have reason to prefer retributivism as the foundation of the state’s practice of punishment. This means, indirectly, that they also choose their own punishment, should they violate the laws under a retributive regime. Being punished for rule breaking is a natural consequence under that penal regime, which one has reason to prefer.

5.4 RULE-UTILITARIANISM: RAWLS’S ATTEMPT TO DISSOLVE THE CONFLICT

Some readers may have grown restless by now. I have so far said nothing about the most influential strategy of dissolving the conflict between retributivism and utilitarianism to date; to mix them in a way that savours the strengths of both theories. One influential argument to this effect is found in John Rawls’s “Two Concepts of Rules” (2001). Here Rawls argues that a correct theory of punishment in fact incorporates elements of both retributive and utilitarian character.

Rawls in “Two Concepts of Rules” attempts to save utilitarianism from the objection that it seems to support that it is sometimes right to punish innocents and to break one’s promises. Rawls’s general idea is that rules such as “keep your promises” and “do not punish the innocent” are expressions of practices that are justified in terms of their utility, but that these practices are constitutive of rules according to which particular acts are performed (or not performed). A practice, Rawls argues, carries with it certain behavioural rules: the rules of the game, one might say.

In the case of punishment, and the problem of punishing innocents, this conception works as follows. Rawls admits that retributivism and utilitarianism both seem plausible and that they both bring important things to the table. Retributivism captures the intuition that a state of affairs where the guilty – and only the guilty – are punished seems morally better than one where they are not. Utilitarianism, on the other hand, has a plausible account of the very purpose of punishment: punishment
is “one of the devices of social order” (ibid. 22). Faced with two plausible but conflicting theories, Rawls attempts to reconcile them.

The solution to the conflict between retributivism and utilitarianism, Rawls argues, consists in making a distinction between what justifies punishment as a general practice, and what justifies a particular action under this practice. He illustrates how this distinction works in the case of punishment by a very compelling hypothetical discussion about the practice of punishment between a father and his son. The son asks his father why a particular criminal, $F$, was put in jail yesterday. The father replies that $F$ has robbed a bank and that $F$ was duly tried and found guilty. But if the son were to proceed to ask why people are put in jail, his father would reply “to protect good people from bad people” or the like. Rawls argues that these are different questions:

“We have two very different questions here. One emphasizes the proper name: it asks why $F$ was punished rather than someone else, or it asks what he was punished for. The other question asks why we have the institution of punishment: why do people punish one another rather than, say, always forgiving one another?” (ibid. 22)

Rawls contends that the first question is answered by retributivism and the second question by utilitarianism. Retributivism governs actions falling under the practice of punishment, such as why $F$ and no one else went to jail. Utilitarian considerations answer the question what the practice as such aims to do, and why it is justified in the first place. In this sense, utilitarianism is the more fundamental theory of the two, for it answers the question about the very point of having an institution of punishment.\(^{152}\) The two questions furthermore address different institutional levels, Rawls argues. Legislators, who are interested in social protection and overall welfare, pass laws in order to control crime, but are not concerned with the everyday workings of the criminal justice system. That is the responsibility of the courts. But the courts are ruled by retributivist ideas about desert and proportionality, not utilitarian considerations. The conflict between retributivism and utilitarianism is therefore a superficial

\(^{152}\) Speaking with Hart, we could say that utility provides the “general justifying aim” of punishment (Hart 1968). Speaking with Dolinko, we could say that utility provides its “rational justification” (Dolinko 1991).
one, Rawls concludes. As long as we keep in mind that the theories ask different questions and apply to different levels, we see that there is no need to choose between them. And this also, Rawls contends, suffices to prove why an overall utilitarian justification of punishment need not lead to that we justify punishing the innocent. For the actions taken under the practice of punishment, he argues, are ruled by retributive rules. The idea is very similar to Hart’s idea of “retribution in distribution” (Hart 1968), and is a prime representative of the “mixed” or “hybrid” theories of punishment that seek to incorporate both utilitarian and retributive elements (Dolinko 2005).\(^{153}\)

However, Rawls’s hybrid theory has been incisively criticised (Oftstad 1982), and retributivists have scorned at the allegation that they are unable to provide a general justification of the institution of punishment (cf. Corlett 2001). The basic reason for why Rawls solution is not as strong as it first appears is that it relies on rule-utilitarianism. Unlike act-utilitarianism, which is the theory that states that the rightness of wrongness of actions depends solely on their consequences, rule-utilitarianism holds that the rightness and wrongness of an action depends on the consequences of the rule the action follows. According to rule-utilitarianism our individual actions should not stand in direct relation with the principle of utility. We should instead act according to rules, which we in turn should choose based on what consequences they produce. Thus, while an act-utilitarian is committed to the view that it is right to break a promise when doing so appears to have better consequences than keeping it, a rule-utilitarian would ask what good a general compliance with the rule of keeping one’s promises produces. Presumably finding that people are better off in a world where the obligation to keep promises is upheld, the rule-utilitarian would conclude that we should make keeping one’s promises as a binding rule. In short, we ought to choose the rules that from a utilitarian standpoint are the best, and then act in accordance with those rules (Smart & Williams 1973; Urmson 1969; Goodin 1995).

Rule utilitarianism thus stated seems to solve some problems. In the case of punishment, it seems likely that a penal regime which includes a

\(^{153}\) Influenced by the idea of hybridization, Jonathan Glover has even suggested that it seems “less and less helpful to discuss penal policy in the crude categories of retributivism and utilitarianism” (Glover 1970: 144), thus rejecting the premise of what Feinberg called the “classic debate” (see Dimock 1997).
rule against punishing innocents has better consequences than a regime which does not. In Mackie’s words:

“It is easy to construct imaginary examples, and not impossible to find real concrete ones, where an act utilitarian would have to say that it is right to kill innocent people [...] or to betray a trust. But the rule utilitarian can say that each such individual act is wrong because the general performance of each of these classes would plainly have a very bad effect of the general happiness” (Mackie 1990: 137).

In essence this is what Rawls argues: it is of greater utility when the practice of punishment is curbed by a number of retributive rules than when it is not. Unfortunately, by relying on rule-utilitarianism Rawls is balancing on a knife’s edge. For rule utilitarianism is unstable: it always seems prone to either collapse back into act-utilitarianism, or push in a direction more deontological than utilitarian. Here is why:

Act-utilitarianism has always reserved a place for rules, but they are never of a morally binding nature. Rules are seen as “rules of thumb” – generally wise advice on what to do in situations when one has too little time or too little knowledge to make a reasoned calculus of utility. But when there is time and one has sufficient knowledge, and the principle of utility recommends breaking the rule of thumb, act-utilitarianism recommends breaking it. Rule-utilitarianism, however, grants rules a stronger standing. Once we have a rule which, among alternatives, leads to the best consequences, we are under an obligation to follow it. The rule is morally binding. But this move seems to conflict with the principle of utility, for surely there may be cases (at least logically) where the calculus of utility does support breaking the rule. And if rule-utilitarianism still recommends following the rule, it is not obvious that it is being consistent with its own normative assumption (the principle of utility as sole criterion for right and wrong). Instead, it seems to bestow deontological weight to the rules.\textsuperscript{154}

\textsuperscript{154} I here understand deontology (minimally) as the view that the goodness or badness of things is not fully accounted for by their consequences (Scanlon 1999). More familiarly, a deontological morality tends to hold that there are moral duties to do certain things, such as respect certain rules, which are independent of the consequences of doing so. A classical example of such a duty is the duty to
A rule-utilitarian might object, as Rawls seems to do, that making exceptions to the rules will always produce more disutility than utility. This might be, but it is a factual and contingent claim which still needs to answer what happens if the principle of utility supports making an exception to the rules (for surely we don’t buy that it is logically impossible that an exception would be of greater utility). And the only utilitarian answer, as far as I can see, is to accept that it is right to make an exception (cf. Smart & Williams 1973: 67-73). 155

For this reason, it seems as if we can reject the solution proposed by Rawls. A theory that even in principle excludes punishing the innocent is not utilitarian. And a utilitarian theory of punishment remains capable of supporting injustice.

5.5 CONCLUSIONS

In this chapter I have attempted to argue that a retributive penal regime is to be preferred to other penal regimes by means of presenting and elaborating Rachels’s lucid test. There is a strong case for a penal regime designed along retributive lines. Such a regime, it seems, adheres to principles of justice most of us find important, including central aspects of the rule of law. Regimes based on deterrence or rehabilitation, by contrast, admit violations of these principles. This may manifest itself in actual penal policy based on deterrence and rehabilitation. But even if the problems remain only philosophical, it is still questionable to en-

never punish an innocent. A deontologist could claim that it is wrong to breach this duty even though the consequences of doing so may be wonderful.

155 If, however, the rule utilitarian bites the bullet and concedes that there may be particular cases where an exception to a rule would be of utility but nevertheless maintain that the rule should be followed, a new host of problems present themselves. It is always possible to play around with the relevant rule. One could say, for instance, that the rule should be “do not punish innocents, unless the fate of the world depends on it”. Since rules in this way can be modified so as to ultimately be recommendations about how to act in unique situations, rule-utilitarianism once again risks collapsing into act-utilitarianism. See Juth 2005: 429-30.
dorse, as a basis of policy, theories which support injustices in principle. Normally we take problematic consequences of a theory to indicate that there must be something wrong with its assumptions.

For all the appeal of Rachels’s test, it still rests on the idea of retribution, deterrence and rehabilitation being separate and incompatible penal aims. I closed the chapter by describing Rawls’s composite theory, which seeks to mix utilitarian justifications with retributive principles. I argued that such an attempt, while intuitively appealing, seems to rely on some form of rule-utilitarianism, which doesn’t seem to do either utilitarianism or retributivism justice.

I have not taken odds with Rachels’s contention that the principles of justice he works with are correct. This can of course be questioned, and a fuller treatment of the topics in this chapter would have to address the question of whether it indeed always is wrong to punish innocents or pre-emptively incarcerate and rehabilitate risk-individuals. It is possible that such an investigation would find that such measures are justified in cases of extreme danger, much as we feel confident that the state is right to breach normally held rights and quarantine people when fighting infectious diseases. Nor have I said much about the ultimate justification of a retributive penal regime. That such a regime has institutional benefits is one thing, but we remain more or less in the dark as to what justifies having a penal regime in the first place. One might put it as follows: if one had to choose between a penal regime based on retribution, deterrence or rehabilitation, there is an institutional reason to prefer a retributive one. But the institutional reason, while undeniably in retributivism’s favour, cannot explain why we should have a penal regime at all – it can only tell us according to what principles it should adhere once in place. We will have reason to return to this question.

The institutional reason points to an indirect way in which retribution is valuable: it is valuable because it is a consequence of a system which we should prefer, even when it sanctions our own particular behaviour. The next chapter will concern a direct way in which retribution is valuable. I will argue that we should prefer retribution to other penal responses since it, in expressing blame, affirms us as free and responsible

---

156 Recall that I am here defending a non-strict form of retributivism, according to which we are justified in punishing at some cost to total utility, possibly even at a great cost, but not any cost.
persons. In short, a retributive conception of punishment expresses, like other important social and political practices, *respect for our choices.*
6. The Symbolic Reason for Retribution

“the concept of deserved punishment seems to cohere with the principle that criminals have knowingly, intentionally and voluntarily violated a law and that the offender in so doing accepts punishment as society’s response to the breach of law and disregard of the social contract.”
(Corlett 2001: 84)

“retribution theories have generally been understood by their makers in terms of free will.”
(Honderich 2006: 26)

WHY should we take a retributive penal regime to be in the interest of all? This is the question we are currently considering. One reason, as we saw in the previous chapter, is of institutional nature. Here, however, we shall be concerned with a different, more direct reason. A retributive penal regime engages in retribution, which, I shall argue, embodies and expresses important values, which we should care about. Being treated as one deserves, which in the case of criminal wrongdoing means being punished, symbolizes something all can agree is worthwhile. Defending this claim is a tall order indeed. I shall go about it in a two-step manner.

First, I will probe into the ways in which our three theories of punishment conceive rule breaking. We shall investigate what conceptions of crime and criminality are presupposed by rehabilitation theory, deterrence theory and retributivism. What is the rule breaker like according to them? What explains the fact that he or she breaks the rules? I shall refer to such conceptions of crime and the criminal as models of rule breaking.
These models are foundational to - though not always explicitly stated by - the theories of punishment considered here.

Next, I shall argue that while we may find each of the models appropriate in different instances of rule breaking behaviour, there is a particular appeal to the model of rule breaking presupposed by retributivism. Only according to this model are we truly personally responsible for our actions. Unlike rehabilitation or punishment-as-deterrence, retribution, in being intrinsically about blame and resentment, affirms this. In other words, I shall defend the importance of being given what one deserves on the ground that this is the only penal response that fully accounts for the fact that rule breakers are responsible moral agents. Unlike being offered treatment or being punished in order to produce the right incentives, retribution treats us with the respect that is our due as persons. The fountainhead of modern versions of this type of argument is Herbert Morris’s *Persons and Punishment* (1968). Unlike Morris, I shall be concerned with arguing the symbolic supremacy of retributive punishment specifically.  

6.1 THREE MODELS OF RULE BREAKING

Models of rule breaking are the underpinnings of theories of punishment. Each theory of punishment tacitly or explicitly assumes things about at least the following two, somewhat overlapping questions:

(a) What is the rule breaker like?
(b) Why does he or she break the rules?

One may derive answers to these questions by working backwards from different justifications of punishment. Suppose someone argues that

---

157 As we saw in chapter 5, Morris argues in favour of a system of punishment as opposed to one based on treatment. It is therefore unclear whether his argument is *retributivist* - a conclusion many seem to have jumped to (see Dolinko 1999).

158 It should be noted that the closely related questions (a) and (b) are such that it may be relevant to relate to the general theory of human nature inherent in vari-
punishment is justified since it enables the state to morally educate criminals. Such a view seems to presuppose that some form of ignorance explains criminal action – perhaps the idea is, as Socrates thought, that no one knowingly does wrong. Punishment is appropriate on this account since it allows the state to make criminals see their own ignorance. Thus, we could say that the model of rule breaking here is that (a) rule breakers are well-intended but (b) are ignorant and therefore act wrongly (i.e. break the law). This serves as the basis for the claim that punishment should be designed to morally educate them.

What about the penal aims considered in this book? Retribution, I shall argue, presupposes that crime is a wicked or wrongful choice. Criminal behaviour is a normal, if regrettable, expression of autonomous agency, which deserves moral censure. I shall refer to this model of rule breaking as the Autonomy model.

Rehabilitation assumes what I call the Disorder model. The criminal is here conceived as a person who is in some sense disordered or abnormal. He or she breaks the rules because of some condition, which causally induces crime. Suffering from the effects of this condition, the purpose of rehabilitation is to help the criminal to function better.

Deterrence is much harder to classify. One may argue for a penal regime based on deterrence even if one considers rule breakers along the lines of the Autonomy Model or the Disorder Model. As long as there are would-be rule breakers who are impressionable by penal threats, deterrence is a potential justification. There is, however, one model of rule breaking which is exclusive to (and especially fitting for) deterrence: that rule breakers are rational agents as conceived by Rational Choice Theory. I shall refer to this as the Rationality Model. According to this model, criminals are not necessarily wicked, nor are they abnormal. Instead, they are rational agents who act strategically in order to maximize expected utility. On this view, crimes could well be seen as a result of de-

159 According to this view, everyone seeks to do what they perceive as good. If what they do in actuality happens to be bad, this is evidence of their ignorance – their mistaken beliefs about the good.
fective institutional design. If the state cannot provide rational agents with sufficient incentives to comply with the rules, so that the expected gains of breaking the rules exceeds the costs, then crimes will naturally follow.

I should like to emphasize that these models are not exhaustive; they do not capture every conceivable way of conceptualizing and explaining rule breaking. Criminology is rife with different explanatory theories of crime, some of which might not fall readily under the three models here presented.\(^{160}\) Attempts to explain crime are often too fine-grained to fit my quite general models. What I call the Disorder Model, for instance, applies to some sociological and psychological theories as well as some biological ones, while advocates of these different theories might object to any attempt to construe them as similar (which they often are, however). The three models do not represent the only ways to think about rule breaking. They are merely ideal types: attempts to lay bare the models of rule breaking – in particular the philosophy of agency - underlying the three theories of punishment considered in this book.

6.2 THE DISORDER MODEL

Let us start with the rehabilitation theory. What model of rule breaking underlies this theory? Rehabilitation is the idea that punishment/treatment is justified since, or if, it takes away a criminal’s desire to reoffend.\(^{161}\) Clearly, then, a proponent of rehabilitation must hold that treatment, broadly conceived, can make rule breakers less likely to break further rules in the future. Unless one believes that crime is a behaviour that can be treated and changed, commitment to rehabilitation makes

\(^{160}\) For useful introductions to the interesting empirical field of explanation of criminality, see e.g. Herrnstein & Wilson 1985; Maxim & Whitehead 1998; Brown et al. 2007.

\(^{161}\) This remoulding of a criminal’s motivations is of course instrumental for some intrinsical good, most likely societal welfare. Strictly speaking, then, a rehabilitationist argues that punishment/treatment is justified as it is the best way to protect society, and the interests of its members.
little sense. Moreover, a proponent of rehabilitation is likely to hold that the thing being treated is also something that explains, or helps to explain, why a person broke a rule to begin with. It seems incoherent to say that we should offer treatment to criminals unless we also assume that there is something about them that explains why they commit crime. Let's for the moment focus on the latter belief: that crime is explained by some fact about the criminal, which is to be targeted by rehabilitation. The Oxford English Dictionary defines rehabilitation as “the action of restoring something to a previous (proper) condition or status.” This definition suggests that rehabilitation rests on a normative assumption: rehabilitation involves changing someone or something for the better. In the case of punishment, the criminal is to be restored (if he or she ever was otherwise) to the preferable state of not committing crime (Raynor & Robinson 2005). In assuming that rule breakers can be restored to a better state, proponents of rehabilitation are committed to the belief that there is something wrong (or not good enough) with the rule breaker; something that explains, or helps to explain, the crime (and which treatment is designed to counter). This is why rehabilitation can be said to rest on a Disorder model (DM). Underlying the belief that penal responses should rehabilitate rule breakers is that rule breaking indicative of some form of disorder, disease, or abnormality.

A very strong statement of the DM was made a couple of years ago by the neuroscientist Wolf Singer, head of the Max Planck Institute for Brain Research in Frankfurt. In an essay published by a German newspaper, Singer advanced the controversial claim that anti-social behaviour is indirect proof of brain abnormalities. Since the only plausible hypothesis about human behaviour is that behaviour originates from the brain, and since the brain consists in matter which obeys causal laws of nature, we may conclude that what has historically referred to as “free will” (the ability to do otherwise) is a myth, Singer argued. The brain is an immensely complicated system of amino acids and signalling pathways, capable of processing vast amounts of variables. Even though we subjectively often feel as free agents, in ultimate charge of our own destiny, our

---

162 I suppose one could say that we have an obligation to try to help criminals through treatment whether or not it works. Here it is assumed that one who advocates rehabilitation believes that it is possible to in fact change criminal behaviour through treatment.
behaviour is fully caused by neurological factors, at a level that is mostly subconscious. Our behaviour can be fully accounted for by the architecture of our brain. This means, Singer continued, that we have to reinvent the way we think of rule breaking and punishment. Even though we may never come to know the exact workings of the brain, our best guess is that for every act of rule breaking there is a neurobiological cause: for every behavioural abnormality there is a neurobiological reason. Rule breakers act the way they do as a result of the malfunctioning architecture of their brains. They are thus rather to be seen as victims of their brains than culpable agents. This does not mean that anything goes, however. Society cannot do away with rules and norms. But penal policy should be less about revenge and holding responsible, Singer concluded, and more about trying to change the behaviour of the rule breakers, keeping them isolated from society indefinitely if it turns out that they are incurable (Singer 2004).

It should be stressed that a biological way of conceiving rule breaking, such as Singer’s neurological one, is not the only way of ending up at the DM. What is central to the model is the notion of behavioural disorders: that there is something wrong with an individual that affects the way he or she behaves. Such disorders need not exclusively, or even primarily, have a biological basis. Some proponents of something like the DM for example emphasise socio-economic circumstances. A good example is Robert Owen, the 19th century industrialist and reformer who held that crime (as other instances of bad behaviour) is entirely attributable to a malfunctioning social system, which conditions malleable humans into vice. Others emphasise mental disorders, such as psychopathology or neurosis (though these disorders may have an ultimately biological basis). There are different ways of filling the DM with specific substance. There are also different ways in which the disorder can manifest itself in behaviour. It can affect a person’s ability to control his or her behaviour, such as in various compulsive disorders. It can in different ways compromise a person’s cognition, such as in psychosis. Finally, it can distort

164 A classic statement of this kind is Menninger 1968. See also the views of John Hospers (presented in Stroll 2004).
the value system or moral capacity of an individual, such as in the callousness associated with psychopathology.\textsuperscript{165}

Clearly, explanations such as these rest on the notion that crime indicates that there is something wrong with the rule breaker. It is not a “normal” action.\textsuperscript{166} By implication, these explanations instil a great deal of weight to the existing rules of a society: to follow the rules is the norm; to break them indicates some abnormality.\textsuperscript{167} This means that the DM will work best in cases where the rules are contested to a minimal extent, and embody weighty behavioural norms. Some have argued, for instance, that murder is an act that in itself proves the presence of an abnormality – a popular notion is that “normal people don’t murder.” Underlying this claim is clearly the idea that the rule against murder is so legitimate a law, so acceptable to most, that breaking it in itself poses the question of normality. The DM will work much worse in cases of petty crimes committed by a large proportion of the population. It seems problematic to say that all illegal parkers or speeders are disordered, for instance.

We may now answer questions (a) and (b) above. The DM states that:

(a) The rule breaker is a disordered agent, whose
(b) abnormal condition explains, or helps to explain, why he or she broke the rules.

\textsuperscript{165} The official term of what is in everyday speech associated with psychopathology is, I gather, “Antisocial Personality Disorder”. See DSM-IV, the Diagnostic and Statistical Manual of the American Psychiatric Association APA 2000.

\textsuperscript{166} I here use the term “normal” to signify something that is non-disordered. The term can give birth to misleading connotations, for few rehabilitationalists would argue that crime is abnormal in the sense that it is something surprising or mysterious. On the contrary, most would argue that the occurrence of crime is very understandable – and in that sense of the word, it is also “normal”.

\textsuperscript{167} This idea can of course have problematic consequences, for surely dissidents in oppressive states should not be seen as disordered simply because they don’t comply with the rules. Furthermore, traits which are generally seen as disorders in some states can be seen as appropriate and normal in other states. Under some regimes, for instance, a sadistic personality might be more esteemed than in others.
From such a conception of rule breaking, it should not surprise us that many proponents of rehabilitation reject the view that crimes are “free choices”. Instead, crimes are symptoms of problems that the state should remedy, either as a humanitarian obligation and/or out of concern for social protection. That rule breakers are victims, in the sense that they suffer from problems, suggest a couple of consequences:

First of all, the (criminal) agency in question does not warrant the same respect as does other agency. This is not very controversial. We would not, for instance, treat a contract being entered under pathological delusion in the same way as any other contract – it would not be seen as binding. As a consequence of the disordered nature of criminal agency, the state and others have a greater freedom to act paternalistically towards rule breakers. It is generally held that a disordered person is not entrusted the same autonomy as non-disordered persons: we have less reason to respect the choices of someone whose capacity for autonomy in various ways is compromised. The clearest examples are those of the infant or the non-human animal. It is generally held that they have practically no capacity for autonomy whatsoever – they can do things, even pursue goals, but cannot reflect critically on these goals and cannot be held morally responsible. Quite simply, infants and non-human animals fail to qualify as moral and rational agents. Paternalistic intervention in their lives therefore tends to require very little justification as compared to similar interventions in the lives of human adults, which tend to require substantially more justification.\(^{168}\) The will of a human adult is granted more weight – a greater moral significance - than the will of an infant or a non-human animal. Unlike the latter two, the adult is believed to know what’s best for him or her, and is entrusted with a great deal of freedom in conducting his or her own affairs. Put shortly, he or she is granted extensive autonomy, or self-rule.

\(^{168}\) It is not problematic, for instance, to thwart a child’s desire to have lots of candy before dinner. The child will ruin his or her appetite, so we remove the candy for his or her own good. The same interference with an adult’s desire to have lots of candy before dinner, however, presents a much higher ethical hurdle. We may perhaps ask the adult whether he or she knows that he or she is about to ruin the appetite, but a received wisdom is that it is very contentious to actually hinder the (non-disordered) adult beyond this point.
The same considerations that explain why infants and animals are not granted autonomy explain why some human adults are not granted autonomy either. Some adults suffer from disorders which damages abilities which other adults are believed to have. This has consequences for how we interact with them. Disordered adults may for example be paternalistically treated in a way a non-disordered adult may not. (This is an empirical claim as well as a normative one, I think). They may be educated, treated, and detained in ways non-disordered may not. In supposedly lacking the capacity for autonomy, they are not granted any autonomy – or at least not as extensive autonomy as the non-disordered adult. Put simply, their choices are not treated as as significant as the choices of non-disordered individuals. If the choices are seen as symptoms of disorder or disease, as they are by the DM with respect to rule breaking, then there appears to be less reason to respect them as free and responsible; the agent’s own. The moral and political consequences of this view I shall return to later in this chapter.

***

Underlying the rehabilitative theory is the idea that crime is indicative of disease or disorder. I have chosen to capture this idea under the label the Disorder Model. This model makes a number of contestable assumptions. First of all, the quasi-empirical claim that crime is indicative of something being wrong with the criminal is questionable, especially when we consider petty crimes that are widely committed. Similarly, in claiming that crime is indicative of disease or disorder, proponents of DM seem to have a conformity bias. The claim that breaking a rule indicates disorder requires that the rule in question is perceived as maximally legitimate, clear in content, and uncontested by others.\(^{169}\) Naturally, actual rules may often fall short of meeting these high requirements.

\(^{169}\) A rule I regard as illegitimate will probably exercise less force over my action: outdated or unjust rules will be more readily broken than rules that are seen as justified and relevant. Furthermore, rules that are unclear in the sense that it is unclear what the rule requires and forbids will presumably sometimes be broken unwittingly; an event which removes completely the need to invoke disorder to explain crime. Finally, if a rule is perceived as frequently broken by others, a person might break rules he or she would never break otherwise. For example, in
In spite of these problems, there is obviously still something attractive about the DM. On a psychological level, it promises to explain even very repugnant crimes without resorting to notions such as innate wickedness or evil. The DM’s message is that crime is a behavioural disorder that is caused by some underlying disorder. Science can come to understand these disorders as completely natural, if regrettable, consequences of our genotype, our society, or our mind. And hopes are that when science reaches that level of understanding, it can enable the state to create a penal regime that is humane, rational and effective - less dependent on pain and blame than present practices of punishment and more focused on what works to rehabilitate offenders in need.170

6.3 THE AUTONOMY MODEL

The retributive theory of punishment holds that punishment is justified if, when and to the extent that it is morally deserved.171 If a retributivist claims that there are cases of rule breaking that deserves punishment, what conception of the rule breaker does he or she assume? Unlike the Disorder Model, the retributivists clearly cannot assume that rule breaking is explained by various kinds of behavioural disorders. Such an explanation would seem to undermine the very desert that the retributive theory of punishment takes as central. In trying to get away from regarding crime as indicative of disorder, it instead seems necessary to assume that crime is not something only “abnormal” individuals commit. Crime is something even “normal” individuals can engage in.

the state of nature, it can be hypothesised that even normally peaceful individuals, contrary to their preference, will resort to violence on the count that others resort to violence.

170 Indeed, in some particularly heinous cases it seems as if the notion of a behavioural disorder is the only bearable option (consider the crimes committed by a Hitler or a Stalin). Unless we explain such cases with some abnormality, we are left with the terrifying option of concluding that the crimes were perpetrated by perfectly sane, normal men.

171 Note that this catchy phrase obscures that it is only according to strict retributivism we can say that punishment is (invariably) justified if it is deserved.
When a retributivist claims that rule breaking deserves punishment, it seems evident that the act is supposed to be somehow freely or autonomously chosen, as opposed to coerced or compelled. Whereas a proponent of the DM could speak in terms of pathological agency or pathological choices, a retributivist clearly must hold that the choices in question are somehow sane and responsible ones (responsibility, it will be recalled, is a prerequisite for desert). A deranged gunman might choose between turning the gun on himself or on others. But such a choice could scarcely be the basis of desert, as it is shaped and possibly caused by the gunman’s disorder, which undermines responsibility.

So what model of the rule breaking does retributivism assume? A retributivist would have to assume that there are acts of rule breaking that are what Jani L. Anderson has aptly called “freely committed wrongs” (J. L. Anderson 1999: 365). This is minimally to say that the acts are voluntary ones, carried out without (significant) external or internal coercion, and in adequate knowledge of their nature and consequences. Anderson writes: “If people are autonomous moral agents, crimes are a reflection of a willful deliberation, not merely illicit harms resulting from undeliberated drives or emotions. Nor are crimes signs of a need for therapy, a ‘bad attitude’, or of recalcitrance” (Ibid. 366-7).

In saying that crimes are freely committed wrongs, it is important to note that there are different ways of conceiving an action as “freely” committed. Some use the term “freedom” to denote roughly the same thing as “unimpeded”. This usage equates freedom with voluntariness. To have freedom of action on this account simply means to be unhindered in one’s actions - to say that a crime is freely committed here simply boils down to saying that the criminal wasn’t compelled to commit crime in a way he or she would have been if he or she suffered from a compulsive disorder, acted under posthypnotic suggestion, was forced by others, etc.

Does this voluntarist conception of freedom suffice to say that rule breakers are morally deserving of punishment? Though there is some disagreement here, I think not. Retributivists presuppose a stronger conception of freedom. In order for them to say that a criminal deserves punishment, I submit, they not only have to assume that the crime is freely
committed in the sense of being voluntary or unhindered. They also have to assume that the crime was genuinely chosen.\textsuperscript{172}

A belief that you sometimes hear expressed is that every criminal “had a choice”; that regardless of the criminal’s general state and the situations he or she was in, he or she could always have chosen to comply with the rules rather than breaking them. This is surely too extreme.\textsuperscript{173} There are cases where compliance cannot reasonably be expected, and there are probably disorders so fundamental that the notion of choice becomes inapplicable altogether. But behind the popular belief is an idea I take to be central to the model of rule breaking underlying retributivism: that crimes are sometimes chosen in the fullest sense of the word; that a rule breaker could have chosen to comply with the rules rather than breaking them. Since this option of compliance exists, crime represents a course of action among alternatives that the rule breaker opted to pursue. As a consequence of the freely chosen nature of crime inherent in this account, the retributivist can argue that the rule breaker is deserving of blame or punishment: He or she had the chance of doing right, but chose to do wrong. He or she is therefore morally culpable. (This of course assumes that the crime represents a wrong that is deserving of praise and blame.) Let us refer to this model of rule breaking as the

\textsuperscript{172} I here side with Ted Honderich, who argues that retributivism is committed to the idea of libertarian free will (he refers to it as “freedom as origination”). According to this doctrine, free acts are chosen in the fullest sense of the word. They are “not caused or determined, but could have been otherwise even if everything in the past and at the time had been just as it was” (Honderich 2006: 26). Note however that there are some retributivists who defend retribution without relying on or assuming such freedom of the will (Moore 1998, Morse 1999; Morse 2000).

\textsuperscript{173} Proponents of the AM range from those who seem to argue that it is always possible to do otherwise to those who think that genuine choices are few and far between. Sartre might be an example of the former: his existentialism rests on an almost unlimited freedom, which translates into equally unlimited responsibility (see Wiggins 2003). Kane is an example of a more moderate theorist. He argues that true choices are quite rare. They only occur when an agent is faced with a choice between two or more alternatives, all of which are recommended by strong, but different and conflicting, reasons. (An example he offers is when you have to make a choice between furthering your own interests and helping others). He calls this plural rationality (Kane 1996).
Autonomy Model (AM), as its upshot seems to be that rule breaking is autonomous action, in the fullest sense of the word.

When a retributivist says that a criminal in virtue of his or her action deserves such-and-such a response, then, I submit it is entailed that the act is perceived as one which was not only voluntarily performed, but also genuinely chosen. Another way of saying this is that the AM assumes that rule breakers have alternative possibilities in action. To have alternative possibilities here means to be able to “do otherwise”. If an agent is about to do \( X \) at time \( t \), it is necessary, in order for alternative possibilities to exist, that there exists at least one other possible course of action \( Y \), such that the agent has a possibility to do either \( X \) or \( Y \) at \( t \). The notion of alternative possibilities is tricky, and philosophers have argued over different ways of formulating it.\(^{174}\) Robert Kane has provided the following definition, which seems just about right:

\[
\text{"[An] agent has alternative possibilities (or can do otherwise) with respect to A at [time] t in the sense that, at t, the agent can (has the power or ability to) do A and can (has the power or ability to) do otherwise" (Kane 1996: 33).}
\]

Some take alternative possibilities to be the basis for moral responsibility. They believe that one cannot be responsible for an action one could not but have performed. Retributivists, as I take them, rely on this basic in-

\(^{174}\) A prominent debate has centred on whether alternative possibilities require indeterminism at \( t \). An influential line of reasoning, proposed by G. E. Moore and others, holds that the notion that an agent “could have done otherwise” (the result of alternative possibilities) really stands for “could have done otherwise, if he or she would have so chosen”. This reformulation is compatible with determinism: it merely states that an agent could have done otherwise if he or she, contrary to fact, for some reason would have chosen to do something else instead. On this analysis, I could have chosen refrained from eating a banana at time \( t \) if it is true that, if I had chosen not to eat the banana, I wouldn’t have eaten it (Moore 1994). Champions of the alternative possibilities requirement such as the quoted Robert Kane have protested, rightly in my mind, that this reformulation obscures what is intended by alternative possibilities. We want to know whether an agent had a genuine ability not to eat the banana at \( t \). I will return to this so called compatibilism/incompatibilism-debate mainly in chapter 8.
tuition: it is the fact that people can do otherwise which explains why they can deserve blame or punishment for breaking the rules.

To clarify again the difference between freedom as voluntariness, and freedom as voluntariness plus alternative possibilities: many have argued that what is central for establishing whether a rule breaker is responsible and deserving of punishment for his or her action is largely a matter of rationality and control (e.g. Fischer & Ravizza 1998). If the rule breaker is irrational or has little control over his or her actions, such as when brainwashed, coerced, or mentally disordered, he or she is not responsible. If on the other hand the rule breaker has what Pettit (2001) has called rational and volitional control – acts under the rational guidance of his or her beliefs and desires, which in turn are such that he or she can identify with them – then the rule breaker is responsible. But such considerations seem independent of the question of alternative possibilities: one can have rational and volitional control over the choice to do X at t even if no other possibility exists. According to the AM, however, such control is a necessary but not sufficient condition for free choice. The choice must be unimpeded in the sense above. But alternative possibilities are also necessary. For without an ability to do Y, then for all the control in the world, doing X is the only possible course of action at t. And such a “choice” can hardly be called genuinely free, let alone a basis for moral desert, many argue.

In order for choices to be free in the sense required by the retributive theory of punishment, then, they must thus be what philosophers call undetermined.175 By this is intended roughly that the outcomes of choices are not prefigured by the causal history leading up to them: a

175 Or, more in detail, they must be undetermined at some point in time. Moral theory has long recognized that there are cases where an agent freely relinquishes subsequent freedom of choice. Consider the case where a man chooses to take drugs, and subsequently becomes so addicted that his power not to take drugs is lost. If the choice to take drugs was free in the first place, however, it could be argued that the addict is still responsible for his addiction in a way he would not be if the choice to try drugs was beyond his control too. Aristotle elegantly likens cases such as these with having the freedom to let go of a stone or not - when the stone is dropped, it might be too late to retrieve it, but that doesn’t override the fact that it was freely dropped in the first place (Aristotle 1998: 61).
choice is undetermined if the agent has a genuine ability to choose either X or Y, and the outcome of the choice is not given in advance but is up to the agent. Here is how O’Connor, in a lengthy passage, describes it:

“When I decide, say, to go for a walk on a cool autumn evening, I am conscious of various factors at work (some consciously articulated, some not) motivating me either to do so or to do something else instead. And there are some courses of action which, while it is conceivable that I might choose to follow them, are such that they do not represent ‘genuine’ possibilities for me at the time, given my current mood, particular desires and beliefs, and, in some cases, long-standing intentions of a general sort. But within the framework of possibilities (and perhaps even relative likelihoods) that these present conative and cognitive factors set, it seems for all the world to be up to me to decide which particular action I will undertake. The decision I make is no mere vector sum of internal and external forces acting upon me during the process of deliberation (if, indeed, I deliberate at all). Rather, I bring it about – directly, you might say – in response to the various considerations: I am the source of my own activity, not merely in a relative sense as the most proximate and salient locus of an unbroken chain of causal transactions leading up to this event, but fundamentally, in a way that is not prefigured by what has gone before” (O’Connor 2003: 257).

This reasoning may seem innocuous, but the basic argument is as contested as they come in philosophy. The AM suggests that crimes are free choices in the sense that the rule breaker could have chosen to comply with the rules just as well as he or she could have broken them. But if the criminal is free to choose between breaking a rule and complying with it, what explains that the criminal actually broke the rule? The tautological observation that “the criminal chose it” clearly won’t get us anywhere. We want to know why the criminal chose to break the rule in the first place. But since the choice is undetermined (up to the agent), it appears as if we cannot come to learn this, for it would violate the AM’s requirement that choices aren’t prefigured. Aren’t we then forced to say that the crime “just happened”?

To clarify the problem here, consider the way the AM explains rule breaking. Clearly, proponents of the AM will attach great weight to the intentions of the rule breaker. In social science, we say that they thereby aspire to give intentional explanations of why rules are broken. Intentional
explanations are explanations in terms of the purposeful behaviour of agents. In providing an intentional explanation of an action, the trick is to show that the action was a means to further the agent’s desires, given his or her beliefs. “Desires” and “beliefs” are here understood as subjective concepts: an intentional explanation seeks to show that an action was a reasonable one – perhaps even the most reasonable one – given what the agent subjectively values and how he or she perceives the world (Elster 1990). Many acts of rule breaking are surely reasonable in this respect, and can thus be explained along these lines. But so can rule following. In order to tease out the explanatory consequences of the AM’s theory of action, consider the case of *Adam and the Two Worlds*:

*Adam and the Two Worlds*. Adam is at a supermarket, and considers whether to steal a candy bar or not. Adam wants a candy bar but has no money on him. He regards the risk of detection as very small. However, Adam finds stealing morally problematic in itself, and fears the embarrassment of being caught. Now consider two possible worlds. In world one, Adam chooses to steal the candy bar. In world two, he chooses not to.

Now, regardless of which world is realized, we can intentionally explain Adam’s action. In world one, Adam steals the candy bar. We explain this with Adam’s desire to have a candy bar, his lack of money, and his estimation of the risk of being caught. In world two, Adam chooses to refrain from stealing. We explain this by citing the fact that theft breaches Adam’s sense of moral duty, and (less loftily) that Adam fears the embarrassment of being detected. So both actions have reasons and can be given an acceptable intentional explanation. But if Adam finds himself in an undetermined choice-situation, we cannot say *why one set of reasons was acted on rather than another*. We cannot say why one world was realized rather than the other.

It should be stressed that the AM need not be clueless when it comes to explaining crime. The usual suspects can be invoked – greed, selfishness, aggressiveness, amorality, impulsivity, etc. But the strange thing is that the AM requires that even if a rule breaker chooses to break a rule, it is often true that he or she could have chosen compliance instead. In the case of Adam, we have two intentional explanations of two different actions. But how could we say, as the AM does, that both actions were genuinely possible; that Adam could have chosen to shoplift (after which
the first intentional explanation would kick in) or not to shoplift (after which the second would kick in)? How could we explain whether the first or second is eventually Adam’s choice?\textsuperscript{176}

Proponents of the AM have no satisfactory answer to these questions. Indeed, they would argue that the questions are wrongfully posed. If people are autonomous agents, capable of free choice, then we simply have to live with the fact that the will of the agent can be\textit{ inclined} by different factors without being\textit{ necessitated} by them.\textsuperscript{177} I shall have reason to return to this difficult problem in chapters 7 and 8. Let us for the moment accept this conclusion and proceed with the model.

If rule breakers are autonomous agents, why do they choose to break the rules? It is important to note that a premier assumption for the AM is that individuals have a capacity for both good and bad action. They may choose to behave amorally or antisocially, perhaps even consistently so. But since they retain an ability to do or be otherwise, their amorality is, so to speak, a moral one. Unlike automatons, their harmful behaviour is attributed to them as responsible beings. I may selfishly choose to invariably further my own interests, never considering the welfare of others for a single second. But since this behaviour is chosen, I am responsible for it and it is possible to hold me accountable. Unlike the automaton, I am capable of deserving blame. Conversely, I am capable of deserving praise if, e.g., I choose to respect legitimate rules and further the welfare of others. Unless there are rule breakers that are autonomous

\textsuperscript{176} As we shall see when discussing the Rationality Model, it is also difficult to understand why people sometimes act against their best interest. If we say that people always act in their own best interest, it appears as if the kind of indeterminism\textit{ vis-à-vis} the choice never surfaces – people simply make a ranking of the alternatives and proceed to choose the one deemed best. Such a position coexists uneasily with the AM, which states that people genuinely “can do otherwise”. But to say that people sometimes willingly do what is not on top of their ranking of alternatives is also puzzling. Kant, who wrestled with the problem of why people often act contrary to the dictates of reason, wrote: “We can also see that freedom can never be located in a rational subject’s being able to choose in opposition to his (lawgiving) reason, even though experience proves often enough that this happens (though we still cannot comprehend how this is possible)” (Kant 1998: 18).

\textsuperscript{177} See for instance Searle 2001.
in this respect, the retributive theory of punishment would not work. It could only advise us never to punish anyone (Corlett 2001).

We can now answer the questions above. The AM states that:

(a) The rule breaker is an autonomous moral agent capable of undetermined choice, who

(b) sometimes chooses, for whatever reason, to break the rules.

***

The Autonomy Model is less concerned with the specific explanation of rule breaking than with outlining a theory of action that is compatible with us holding each other personally responsible. In this respect the model is very plastic – it admits a plethora of reasons for why people choose to break the rules, as well as why they comply with them. Its basic assumption is that rule breakers, even when heavily inclined towards breaking the rules, retains an ability to choose to follow the rules instead. Even when an act of rule breaking seems very understandable, indeed perhaps even natural, the possibility of acting differently means that the crime was nevertheless a free choice. Unless this is held, it seems difficult to see why criminals should be taken to deserve blame and punishment.

It has been argued that the theory of action underlying the AM is in concord with “ordinary morality” (Wiggins 2003; Stroll 2004). Whether this is a strength or a weakness is up for debate. Some argue that the AM rests on a “metaphysically inflated” notion of choice, which in addition tends to blind us to the effects of structural injustice and fuel moralistic reasoning (Scheffler 2005). Also, we have seen that the AM poses curious questions concerning how it could be explained that an agent chose to form one intention rather than another, and critics will consequently point out that the model is high on morals, but low on explanation. Nevertheless, as we shall see in the second part of this chapter, it can be argued that the AM alone can provide full justification for the normative weight we attach to choice. But it faces formidable challenges from those who argue that the model, whether normatively attractive or not, is empirically implausible.
6.4 THE RATIONALITY MODEL

Compared to the previous two theories of punishment, the deterrence theory is decidedly harder to classify. Whereas rehabilitation and retribution appears to rely on quite a specific model of agency in order to be coherent, the deterrence theory seems compatible with a greater variety of models. Deterrence is the aim of making individuals “less likely to offend again or offend a first time because of fear or prudence” (Honderich 2006: 75). As long as (some) would-be rule breakers are impressionable by penal threats, nothing rules out that deterrence is compatible with the view that rule breakers, by and large, are disordered. And even if one sees rule breakers as autonomous moral agents, deterrence might still look like a plausible justification for punishment. I want to be completely clear on this point: deterrence as penal aim is compatible with all three models of rule breaking presented in this chapter.

With this being said, there is a model of rule breaking which seems exclusive to deterrence theory, and which seems inconsistent with both retribution and rehabilitation as penal aims. It emanates from Hobbes’s theory of action, and is today found in neoclassical economics and decision theory, or Rational Choice Theory (RCT). The simplest way of describing this theory of choice is that people are assumed to be self-interested utility-maximizers. They act in order to maximize their expected utility (or minimize their expected disutility). Choices are said to be rational when they are instrumentally effective, perhaps even optimal, in satisfying the desires of the agent, given his or her beliefs. There is an instructive distinction in the literature between explanatory and normative RCT. Explanatory RCT assumes that rational agents are instrumentally rational, and uses the notion of utility-maximization to explain individual behaviour, as well as collective outcomes. Normative RCT makes no comparable assumption, but employs the notion of instrumental rationality to delineate what a perfectly rational agent would or should do (Elster 1986). As foundational to the deterrence theory of punishment, it is clearly explanatory RCT that is here in focus.

Let us call this the Rationality Model (RM). According to RM, people in general act so as to maximize expected gains over losses in utility. The subclass of criminal behaviour does not require a different analysis: rational agents break rules if and when it is instrumentally rational to do
so. One of the most straightforward formulations of this explanation of crime is given by Jeremy Bentham, the great utilitarian and deterrentist:

“The profit of the crime is the force which urges a man to delinquency: the pain of punishment is the force employed to restrain him from it. If the first of these forces be the greater the crime will be committed: if the second, the crime will not be committed” (Jeremy Bentham, quoted in Tamburrini 1996: 16).

We might wonder at this point what the ‘profit’ of crime is. What is this ‘utility’ rational agents are believed to maximize? Bentham, as is well known, thought in terms of pleasure. The standard motivational assumption of contemporary RCT holds that rational agents seek to maximize material self-interest – quite tangible goods such as material resources or power. The exact nature of the motivations is, as we shall see, not as important as the nature of rationality inherent in the model. Let us settle with ‘utility’ as a primitive, denoting the stuff the rational agent seeks to maximize.

Let’s go back to Bentham’s statement above. He there makes an extraordinarily strong claim. If a rational agent, whatever his or her motivation, finds that the gains of crime exceeds the pain of punishment, then he or she will commit the crime. Given that the choice situation is perceived as one where rule breaking is rational, rational agents invariably will, Bentham seems to say, break the rules. Later proponents of the RM has watered down this claim, saying that when perceived gains are greater than perceived losses, then the probability of crime goes up (Herrnstein & Wilson 1985).

Faced with rational rule breakers, the purpose of punishment is to ensure that rule breaking is perceived as less rational than it would otherwise be. As we saw in chapter 3, deterrence is simply the idea that raising the expected cost of crime will make impressionable individuals less likely to break the rules when they make choices under uncertainty.179

178 This might well be too restrictive. Adding additional motivational assumptions might present more problems for RCT than one may think, however (see Eriksson 2005).

179 Clearly the deterrence theory requires that the rule breaker doesn’t know whether or not he or she would in fact be punished if he or she broke a rule. Even a regime that punished 99 percent of all offences would still be useless as a
Just as raising the cost of a product will make a rational consumer less willing to buy it, raising the expected cost of a crime will make rational would-be criminals more inclined to comply with the rules. The standard justification of deterrence, such as the one we find in Hobbes, holds that since individuals’ quest for utility might often result in collectively disastrous outcomes, a body of rules defended “by the sword” is necessary to promote peaceful coexistence (Hobbes 1998).

Clearly the RM has a tendency to “normalize” crime. Without the threat of costs, whether from the state or from one’s peers, most or all of us would break rules, the model seems to assume. Faced with a particular set of incentives, a rational agent can commit any and all crimes. Explaining rule breaking in this way makes rule breaking seem very “normal” or understandable: people simply cease the opportunities they have to further their own interests, and in this sense the question of blame and moral disapproval becomes auxiliary to the model at best. Criminals are not a breed apart. As Gary Becker in a classic article has described his RM-approach to crime and punishment:

“The approach taken here follows the economists’ usual analysis of choice and assumes that a person commits an offense if the expected utility to him exceeds the utility he could get by using his time and other resources on other activities. Some persons become ‘criminals’, therefore, not because their basic motivation differs from that of other persons, but because their benefits and costs” (Becker 1968: 176, Italics added).

The idea here is that it is the circumstances of criminals which differ from law-abiding people, not their more fundamental nature as persons. Edward Luttwak has applies this mode of thought to underclass criminality. Against those who treat crime as a pathology of the poor Luttwak emphasises the *rationality* of crime for those who have little access to deterrent against an individual who knew for sure, that the risk of him or her being caught and punished was zero. Recall also what I said in chapter 3: the risk of being punished must be a compound function of (a) the likelihood of being caught and (b) the likelihood of being punished if caught.

180 This is to say that criminal and non-criminal action can both be accounted for by an economical analysis. It could well be, however, that criminals are less risk-averse than law-abiding people, which in some sense does represent a motivational difference.
other job-markets. Commenting on a study of the income per capita of drug dealers in Washington DC, he writes that “those who entered the drug trade were making a rational choice based on correct information, and could not have been otherwise directed by any competent management consultant” (Luttwak 1998: 100). Indeed, the impression one gets from reading the views of Becker and others is that crime is a question of (in the words of popular culture) “not hating the player, but hating the game”.\(^{181}\) If rational agents break rules, this is because of a faulty system of incentives. The solution to the state of nature must always be sought in institutional design, not in the character of men.\(^{182}\) As H.L.A. Hart aptly put it, punishment on this account is “an economy of threats” (Hart 1968).

We can now answer the questions above. The RM states that:

(a) The rule breaker is a rational agent who
(b) will break the rules when doing so yields greater expected utility than following them.

The Rationality model and alternative possibilities

We should pause at this point and reflect on the ways in which the RM and the AM differ. As I said earlier, the AM can explain crimes in much the same way as the RM: many crimes are rational choices made by self-interested utility-maximizers. But while the RM assumes that people are fundamentally hardwired to be self-interested utility-maximizers, the AM suggests that people can choose to be like that – but also choose to be otherwise. The AM admits of situations where a rational calculus strongly recommends breaking the rules but where moral reason – exercising an independent force – trumps the gains involved in rule breaking. The rational man, as depicted by RCT, would treat moral reasons as just

\(^{181}\) Imagine finding the Ricki Lake show useful for one’s dissertation!
\(^{182}\) See Binder 2002: 334f.
as any other factor in the calculus of utility, if indeed he or she can take independent impression by moral reasons at all.\textsuperscript{183}

The most dramatic difference between the models emerges, however, when we consider the notion of alternative possibilities. As we have seen, the AM retains the alternative possibilities condition as fundamental. The notion that people “could have done otherwise” is taken as central for the appropriateness of the retributive conception of punishment. The RM, on the other hand, either rules out alternative possibilities or deems them entirely irrelevant. Here’s why:

According to RCT, rationality minimally requires that we (a) rank feasible alternative courses of action, and then (b) choose the highest-ranking one. We do not have to assume, as RC theorists tend to do, that rational agents have transitive, reflexive and complete preferences, nor that they are able to correctly assign probabilities to different outcomes of alternatives. The relevant part is that rational agents, regardless of their criterion or criteria for ranking, are able to rank the alternatives they consider on an ordinal scale, ranging from the best to worst choice. Rationality then requires that the best alternative be chosen. RC theorists call the highest-ranked alternative or alternatives the \textit{maximal set} (McCarty & Meirowitz 2007). The maximal set can contain more than one alternative but that implies that the agent is strictly indifferent between the elements of the maximal set. Rationality then requires that the decision between the alternatives is made with some random decision-making process. If there is one alternative that is strictly preferred to the others, there is a single element in the maximal set. A rational agent will opt for this alternative.

The notion of choosing the alternative in the maximal set is fully compatible with the voluntarist conception of freedom as defined above, but it either renders irrelevant or rules out the alternative possibilities condition. Colin Hay has therefore argued that RCT, far from being a theory of free agency, in fact is a deterministic theory. He writes:

\textsuperscript{183} Thus, having a guilty conscience or the intrinsic disutility involved in breaking one’s own moral standards would be costs just as any costs, capable of being overridden by sufficiently large gains. Another way of saying this is that every value should be translatable to a master value (cf. Scanlon 1988).
“For, within any rationalist model, we know one thing above all: that the actor will behave rationally, maximising his or her personal utility. Moreover, we know that there is, by definition, only one optimal course of action [sic] by which the actor’s utility might be maximised. It follows, logically, that a rational actor in a given context will always choose precisely the same course of action. So much for voluntarism. What this implies is that the agent’s choice (in fact absence of choice) is rendered entirely predictable given the context” (Hay 2002: 52-3).

Hay here relies on a definition of choice akin to the AM’s, and his equation between predictability and absence of choice can be contested. But the upshot of Hay’s argument is clear – if rational agents always choose (or “choose”, as Hay would have it) the alternative in the maximal set, then there is an obvious sense in which RCT seems to threaten the alternative possibilities condition. If rational agents always opt for the highest ranked alternative, perhaps they could not have “done otherwise”.

This interpretation of RCT suggests, as Bentham thought, that individuals will always break the rules in situations where the expected utility of doing so is in the maximal set. This might well be too strong. First of all, RC theorists acknowledge the importance of *akrasia*. Akrasia, or weakness of will, obtains in situations where a rational agent fails to opt for the element in the maximal set – it may be that I realize that shoplifting is the most rational alternative in a situation, but I might be unable to act on this realization out of weakness of will (see Davidson 2001). Furthermore, we need not assume that the reason why the majority complies with rules is that they all suffer from *akrasia*, or that they comply simply out of fear for punishment. What the RM *does* assume, however, is that

---

185 Note, however, that RCT does not have to say that the AP-condition is nonexistent. Assume that a rational agent is always able to rank alternatives such that there is at least one element in the maximal set, and that a rational agent will always go for the alternative(s) in the maximal set. Such a conception is fully compatible with a genuine ability to do otherwise. Perhaps it is somehow metaphysically possible to choose a lower-ranking alternative. But RCT states that, if agents are rational, that possibility will never be realized since the ability to do otherwise is never exercised. On this analysis rational agents “have, if you like, a nominal choice between rationality and irrationality but […] always opt for the former” (Hay 2002: 53).
those who in fact have broken the rules have done so since it was (perceived by them as) rational.\textsuperscript{186}

In treating the alternative possibilities condition so differently, it follows naturally that proponents of the RM have tended to emphasise blame and resentment to a decidedly lesser extent than proponents of the AM. Crime has more to do with the context of action, and the institutions of society, than with the wickedness of the individual criminal. Rational individuals will always do what further their own interests, and there is little sense in blaming them for trying to maximize their own utility. It seems understandable against this backdrop why some have argued that punishment should be more forward-looking and centred on changing behaviour by means of general and special deterrence, than backward-looking and centred on exacting justice. Since people are determined to opt for what they perceive as in their maximal set, or at any rate cannot be expected to act counter to their instrumental rationality, the questions of ultimate responsibility and desert might appear more misplaced.\textsuperscript{187}

What should be done instead is to design institutions so that fewer people will perceive it as rational to break the rules in the future.

\* \* \*

According to the \textit{Rationality Model} it is not surprising that rules are broken. Rule breaking need not be a sign of disorder, nor innate wickedness. (Since we often have rational reasons to break the rules, the fact that we comply with them is perhaps more puzzling.) Given a sufficiently unfortunate incentive structure, the RM says, normal people may commit murder. And this is exactly what many find undignified, if not empiri-\textsuperscript{186} One might ask whether it makes sense to regard all rule breakers as rational. Different schools of thought can be predicted. Some will emphasise the rationality involved in even the most deranged actions (Boström 1996). Others will say that some individuals are so disordered that the apparatus of RCT doesn’t apply to them.

\textsuperscript{187} A classical argument in this respect is found in Smart 1961. In fact, the same author is famous for having claimed that the notion of ultimate responsibility for an outcome is a “piece of metaphysical nonsense”. According to Smart, the relevant question of blame is not “Who is deserving of it?” but instead “Whom would it be useful to blame?” (Smart & Williams 1973: 54).
cally false, in the RM and the utilitarian conception of punishment. A moral agent ought not, some would argue, obey rules conditionally (when they serve his or her interests). The decision to refrain from shoplifting ought not to be based on a cost-benefit calculation, but on moral conviction. An agent can perhaps choose not to shoplift simply because he or she believes it to be wrong in itself, and this may hold even when a rational calculus overwhelmingly recommends shoplifting. One ought to – and perhaps does – act according to a logic of appropriateness rather than the logic of consequentiality operating within utilitarianism, thus heeding to other voices than that of instrumental, self-interested reason. The way utilitarianism treats rule breaking might be seen as quite undignified, as will be made clear when I describe how retributivists find the utilitarian philosophy of punishment as one of dogs, not men. I will leave this thread of the argument temporarily.

6.5 COMPARING THE MODELS: THE SYMBOLIC SUPERIORITY OF RETRIBUTIVISM

So far the models of rule breaking. I said at the start of this chapter that they are ideal types underlying the respective theory of punishment. There will be situations where a given model seems more appropriate than the others. When crimes seem like cool and calculated gambles, the RM may seem fitting. Especially bizarre, meaningless or heinous crimes are likely to invoke the need to employ the DM. Those who are inclined to think that many crimes are simply wicked or immoral choices will tend to think that the AM applies. Yet, even if the models can apply to different situations, it is still the case that they represent different ways of conceiving rule breaking in general. This means that they are in competition.

One obvious debate about the models concerns whether they are plausible as (quasi) empirical theories. Is it really true that rule breakers

188 See Peters 1999 for a discussion on how these concepts have become central in institutional theory, as they represent different ways to “shape and constrain” behaviour.
in general suffer from some sort of behavioural disorder, as the DM states? Does the AM’s ability to do otherwise exist or is it an instance of “panicky metaphysics”? (Strawson 2003: 93). Is RCT a truthful descriptive and explanatory theory of human action, criminal action included? These questions must be approached with humility: at this point we strictly speaking do not know which model best captures the general nature of rule breaking, and there are of course many reasons to analyze different kinds of rule breaking on different terms (murder and speeding – both breaches of worthwhile rules – simply do not seem to evoke the same kind of questions).

I will return to the question factual plausibility in later chapters, when I probe into the plausibility of the AM. Here we are concerned with reasons to prefer a retributive penal regime, and the retribution it engages in. And I believe there are separate *symbolic* reasons for preferring retribution, as it stands for and symbolizes valuable things in a way other penal responses do not.189 The rest of this chapter is devoted to validating this claim.

Consider first rehabilitation, deterrence and retribution as acts of communication; acts that send out certain messages or “symbols”.190 The models of rule breaking, which we have arduously chiselled out above, are relevant here. Retribution, which assumes rule breaking to be immoral exercise of free or autonomous choice, roughly seems to communicate: “you chose to do the wrong thing, and deserve moral censure for it”. Deterrence, at least as long as it is underpinned by the RM, communicates: “you didn’t respond as hoped to our system of incentives, and in order for you and others to see the irrationality of your ways, we punish you”. Rehabilitation, finally, with its emphasis on behavioural disorders, communicates: “there is something wrong with you, and we will help you set it right”. Given that these are the messages being expressed, one could make a case for that the one communicated by retribution is the most attractive. Such a case is usually framed in terms of that retribution, unlike the other messages, is about “respecting persons” (Rachels 1999).

---

189 A separate, if self-evident, point is that I also believe that if I am correct in making this claim, then this is something that counts in favour of the retributive theory of punishment.

190 To analyze punishment as a communicative act or enterprise has become increasingly popular. See e.g. Duff 2001.
6.6 RESPECT FOR PERSONS AND THE SIGNIFICANCE OF CHOICE

Moral and political theory generally assumes that choices carry great significance. A state of affairs that is chosen is analysed differently than one that is not chosen (Scanlon 1988). Luck egalitarianism, to take one influential example, holds that inequalities deriving from unchosen circumstances are unjust, but that there is nothing comparably unjust about inequalities deriving from choice. The entire liberal model of society also builds on the significance of choice. Its basic idea is that individuals, as long as they do not harm others, should enjoy freedom of choice; that they should be able to do what they think best without having their choices overruled or otherwise interfered with by others (Mill 1998). Underlying this idea is the belief that individuals are somehow entitled to enjoy autonomy, or self rule; that they should be able to choose for themselves. This belief in turn presupposes that there is something about individuals and their choices that has a claim on our respect: choices are morally significant only to the extent the people doing the choices meet certain criteria of autonomous or responsible agency.

Democratic rights of inclusion, participation and voice can be seen as relying on the same reasoning. Democracy could plausibly be seen as the extension of the significance of individuals’ choices into the realm of political and collective self-determination. Seen that way, democracy ultimately makes sense to the extent individuals are such that their choices are worthy of respect. We would not support a democracy of non-human animals or infants, for instance. Apart from the obvious practical difficulties of setting up such democracies, this follows from the simple fact that they are not autonomous moral agents whose choices warrant respect. But if an alien civilization were to colonize Earth and enslave the human race, we would presumably say: “This is not right. We are able to choose for ourselves.” We would see it as a fundamental violation of our rights.192

191 Important works on luck egalitarianism include Cohen 1989; Anderson 1999; Arneson 2000; Dworkin 2002; Scheffler 2005.
192 I will return to the significance of choice and the way it pertains to democracy in some detail in chapter 8.
The idea of letting individuals choose for themselves without interference by others is sometimes referred to as the key aspect of respecting them as persons. To respect individuals as persons in this sense requires that we (a) allow them to choose for themselves in various contexts, and (b) grant their choices significance simply in virtue of being their choices. In Morris’s words: “We treat a human being as a person provided: first, we permit the person to make the choices that will determine what happens to him and second, when our responses to the person are responses respecting the person’s choices” (Morris 1968: 492). Note that both (a) and (b) are necessary for respecting someone as a person. We do not respect someone as a person if we simply remove all alternatives but the one we would prefer him or her to choose (such as when a dictatorship only launches one political candidate in an election and makes voting mandatory), nor if we proceed to interfere or overrule a choice after it has been made (such as when elections are re-held because the result was “wrong”). To respect someone as a person in large part means to allow that his or her choices have moral significance. This requires that we leave him or her with a genuine choice, and that we do not overrule what he or she subsequently chooses.

When advanced by retributivists in the context of punishment, the respect for persons argument suggests that the same respect we employ towards action or choices in general should be employed towards criminal action or choices too (J. L. Anderson 1999). When people choose to vote for a particular party, for instance, or chooses to pursue an uncertain career, we do not override their choices (or are not entitled to). Nor do we presume that that they are out of their mind, or that that they do things unwittingly. Instead, we presume that they do what they do knowingly and willingly. Most of us believe that there is a defensible rationale for this presumption. Retributivists add that we should presume the same thing when it comes to criminal action, or criminal choices. We should not presume that criminals suffer from behavioural disorders. We should presume, as we otherwise do, that their rule breaking is freely chosen and should be treated as such. If there is a defensible rationale behind respecting choices in general, there is no reason why criminal choices should not be analysed in the same way.

Now, to what extent are the models of rule breaking considered here, and the theories of punishment they underpin, compatible with such respect? The DM, it will be recalled, would have to be said to score low here. The upshot of the model is precisely that criminals are suffer-
ing from a disorder that undermines their capacity for autonomous choice. On this account, a criminal can scarcely say: “I took a chance and broke the rules. I was caught and accept the penalty that was declared in advance, but once I have done my time, I have paid my debt to society and expect to be duly released.” Since the crime is indicative of an underlying behavioural disorder, the criminal simply is someone whose actions should not be granted that kind of significance. This stance relies on the highly plausible idea that choices or actions of a disordered individual do not command the same respect as non-disordered choices or actions. Consider for example the difference it makes if we construe smokers as suffering from a disease rather than people who willingly uphold a bad habit. If smokers are diseased, in the sense that they are addicted to an extent that they can’t control their behaviour, it is easier to justify quite draconian anti-smoking policies. If they are seen as people who actively and willingly uphold a habit, the same policies become decidedly harder to justify – smokers make a choice to risk their health, and apart from engaging in “soft paternalism” and preventing harm to others (for instance by banning smoking in public places), there seems to be little moral room left in which the state can manoeuvre. 193

To be treated as one who is disordered, then, seems to conflict sharply with the notion of having one’s choices respected. Note that the point here is not that being labelled as disordered admits of paternalistic exercises of state power, though the notion of disorder is obviously part of what justifies such practices. Nor is the point that a rehabilitative penal regime can treat you in an arbitrary and irregular manner, as was argued in the previous chapter. The point is rather that rehabilitation is inconsistent with you being seen as a person who has freely chosen to break the rules. As held by the DM, what accounts for your offence are instead the crime-causing conditions you suffer from. This equals being seen as a victim of circumstances, or a less than fully responsible person, retributivists argue.

193 An interesting discussion about paternalism and smoking is found in Goodin 1997. The distinction between soft paternalism, which states that it is allowed to make sure that individuals are well informed about their choices but not to coerce them against their informed will, and hard paternalism, which states that is allowed to coerce even fully informed and rational individuals for their own good, is laid out nicely by Dixon 2001.
What are we to make of this last claim? I believe that it is basically correct to say that rehabilitation conflicts with respecting the choices of persons. I also believe that each of us wants, other things equal, to have our choices respected, and that such respect is an important part of our predominant modes of moral, social, and political organization – the value of the presumption of autonomy, if you will. It is of course questionable if we can isolate the importance of this value from factual considerations: if criminals by and large are disordered, a rehabilitationalist could well argue that treatment, not retribution, is the respectful thing to do. It is respectful since treatment would then be about *restoring* the persons’ capacity for autonomous choice, which could well be framed as treating people as “ends in themselves”\(^{194}\). Retributivists would agree: it’s just that they find the notion that most rule breakers are disordered bogus, and that to treat them as such means to treat them in an inappropriate (not respectful) way.

Let’s at this point assume that respect for persons is an important value, and that it is detrimental to rehabilitation, which construes rule breakers as disordered, that it seems unable to account for it. What about deterrence and retribution? Here things become more difficult. Both deterrence and retribution are penal aims associated with punishment. Both

---

\(^{194}\) As the reader surely will know, an important aspect of the literature on respect and personhood is the Kantian idea that persons, in virtue of being the only things of intrinsic, unconditional, incomparable, and objective value, should always be treated not as only “means to an end” but as “ends in themselves” (Kant 1998: 186-7). Since persons are the only beings capable of freedom and rationality, they have a dignity which precludes us from using them in various ways. (Note that while persons are typically assumed to be human beings, most theorists accept that other life-forms, such as extraterrestrial aliens, may qualify for personhood). Unlike what is the case for non-persons, we may not use persons simply for our own gain – Kantian ethics forbids us to perform coercive medical experiments on persons, for instance, even if the experiments would lead to marvelous ways of curing diseases (see e.g. Dillon 2007 for an overview). A rehabilitationalist could argue, in relation to this, that to offer disordered criminals treatment is precisely to treat them as an “ends in themselves”, since what we are trying to do is to restore them to their proper way of functioning. It is an open question whether even coercive and manipulative treatment is consistent with responding to criminals as “ends in themselves” as long as it aims to promote their functioning.
also agree that the punishment for various crimes should be declared in advance, and that punishment should normally follow in case the crimes are actually committed. It thus seems like a deterrence-based and a retributive penal regime are equally respecting of persons’ choices: they both seem to agree (a) that people have a choice between following or breaking the rules, and (b) that those who break the rules should be punished (which seems to be in line with not overriding the choices once they have been made). However, retributivists have argued that deterrence, simply in being about creating compliance through the use of penal threats, fails to express respect for persons. I will refute this ill-conceived objection next, before turning to more valid reasons for saying that retribution is more respectful than deterrence.

6.7 THREATS AND AUTONOMY

Some have argued that what sets retribution apart from deterrence is the mode in which it addresses people. The basic reason for this is that a deterrent system of punishment is taken to be based on prudential or amoral incentives, whereas a retributive system of punishment addresses would-be rule breakers as moral agents. Punishment-as-deterrence, then, is seen as nothing more than orders backed by the threat of naked force. It has been argued that deterrence therefore is a penal aim which fails to respect those subjected under the penal regime as persons. Hegel - always a great source for gusty quotes - presumably had this in mind when he wrote: “In this [i.e. the deterrence] view of punishment it is much the same as when one raises a cane against a dog; a man is not treated in accordance with his dignity and honour” (Hegel 2005: 36). The ideal here seems to be that people should comply with the rules out of their own free will and for the right, moral reasons - not because they are swayed by external threats.

Duff has described this argument by making a comparison between a government that erects a deterrent penal regime and a bank robber:

“The gunman realises he can offer the bank staff no relevant reason to give him the money: so he creates for them, and imposes on them, an irrelevant and coercive prudential reason for obeying his demand […]"
The government realises that too few of its citizens will be sufficiently moved by the relevant moral reasons which justify the law’s claim on their obedience […] so it creates for them, and imposes on them, an irrelevant and coercive reason for obedience by threatening them with punishment if they disobey. Gunman and governments do not only use those whom they actually shoot or punish: they manipulate, and thus use as means, all those whose obedience they try to secure by threats of death or of punishment” (Duff 1985: 180).

Whether punishment-as-deterrence is a violation of the duty to treat persons as ends in themselves, which Duff seems to suggest, is a separate question that I will return to below. The relevant thing here is Duff’s contention that the threat of punishment, in failing to give would-be rule breakers the “relevant moral reasons” for obedience, imposes an “irrelevant and coercive prudential reason”. The upshot seems to be that there is something about the threat of punishment as such which is undignified – if punishment addresses prudential rationality only, it fails to address people in the right way.¹⁹⁵

This argument is peculiar to say the least. Two considerations seem to supply us with decisive reasons to reject it:

First, since a retributive penal regime entails that the state attaches and makes public the penalties for criminal actions, it cannot really control whether people feel threatened by the penalties. Nor can it control whether it is only their prudential reason that is being addressed. The fact that a retributive penal regime and regime based on deterrence may look quite different in their design does not change the fact that the prospect of being punished will act as a deterrent – as Jean Hampton has noted, all punishment is like a “fence” that says “do not cross this fence, or else” (Hampton 1984).

It may be argued at this point that the motivations people have for following the rules are important. Many would say, for example, that it is somehow better if people willingly comply with the rules because they see them as just, than if they comply simply out of fear of punishment.

¹⁹⁵ Duff goes on to argue, more plausibly but still unconvincingly, that the difference between the law and the gunman is that the former should present moral reasons for obedience (but still threaten that punishment will follow if the moral reason falls on deaf ears). The difference thus conceived is that the law tries to justify its threat, whereas the gunman relies on brute force alone (Duff 1985: 182).
This is true (although one could argue that unwilling compliance still is better than no compliance at all). But even if one allows that there is something undignified about following rules out of fear, it is still the case that a retributive penal cannot control whether people feel treated (in the words of Hegel) “as dogs” or not. There may be large differences between the rationales of retribution and deterrence, but in practice both defend penal regimes that may threaten equally much.

Secondly, it seems incorrect to say that the mere existence of a penal threat is inconsistent with treating people as autonomous moral agents, or respecting them as persons. It is true that abolishing all rules and all punishment, as in the Hobbesian state of nature, might in one (weird) sense be to maximize autonomy - people can do whatever they like whenever they like it, without being hindered by the threat of punishment. But this is clearly not what retributivists have in mind. Their idea is rather that when people are under a threat, they are forced to act out of prudential rather than moral concern. They are swayed into obedience for the wrong reasons. This is allegedly inconsistent with respecting their autonomy. Note, however, that a threat does not rob a person of a choice (like proper compulsion would). It merely raises the stakes of the choice, making one alternative more costly. A robber who says “your money or your life” does not take away your choice. He or she merely makes one alternative (decidedly) less attractive (Tamburrini 1996; Pettit 2001). Similarly, the existence of a penal threat does not take away your ability to break the law. Nor does it preclude you from complying with the law for the right reasons.

The argument that deterrence necessarily conflicts with respecting the autonomy of individuals thus faces insurmountable problems. If retributivists want to argue that their theory is symbolically superior to the deterrence view, the mere fact that deterrence allegedly treats persons “as dogs” since it relies on an economy of threats will not suffice. Threats as such are not necessarily equivalent with dog-like treatment, and a retributive penal regime will also, if inadvertently, “raise a cane” against potential rule breakers.

196 Following Hobbes’s (1998) suggestive treatment, however, there will certainly be a “net” loss of autonomy in the state of nature.
6.8 THE IMPORTANCE OF BEING HELD RESPONSIBLE

The argument from threats is unpromising, but there is a much better way of arguing the symbolic superiority of retribution. Retributive punishment, with its reliance on the AM, is the only penal response which is necessarily and robustly about holding rule breakers personally responsible — it is the only response which is necessarily and robustly connected to punishment as blame. I shall argue that holding responsible (and being held responsible) embodies important values, and that retribution as a penal response in virtue of this seems preferable to its alternatives. More specifically, I shall make the following argument:

P1: Retribution, unlike its alternatives, is necessarily and robustly about holding rule breakers responsible for their actions.

P2: Holding people responsible for their actions has a particular value.

P3: The value of holding people responsible for their actions is a value even those who are being held responsible can agree is worthwhile.

C: Retribution has a value even those being held responsible can agree is worthwhile.

We can see that the argument as it stands is insufficient to prove that the value of retribution mentioned here is decisive. One could accept the argument in full, but argue that retribution still should be shunned for some other reason. What I hope to show, however, is that there is a value in retribution which has to do with the practice of holding responsible, and which counts in favour of preferring retribution to other penal responses. Needless to say, this in turn counts in favour of preferring a retributive penal regime to other regimes (which do not engage in retribution).

Let’s begin by analyzing the first premise: that retribution, unlike rehabilitation or deterrence, is necessarily and robustly about holding rule breakers responsible. Consider again the messages or symbols communicated by the different penal aims. Rehabilitation communicates that there is something wrong with the rule breaker, which treatment should address. This seems to conflict in a straightforward way with the idea of the rule breaker being personally responsible and thus blameworthy, for
reasons given above: the point of offering rule breakers rehabilitation to begin with is presumably that they suffer from a disorder which warrants understanding and help, not resentment and blame (cf. Heywood 2004). Deterrence communicates that actual and potential rule breakers need to have their incentives changed through the use of punishment. As we saw in the description of the RM above, conceiving punishment this way can scarcely be thought of as holding responsible. Punishment (and possibly blame) is rather used for instrumental reasons – it is a way for the state to make sure that future states of affairs will be as good as possible. This does not seem true to the proper meaning of being held responsible. 197

What, then, is the proper meaning of being held responsible? I submit that however we ultimately define the act of holding someone responsible, the definition will have to include some backward-looking element of blame (or praise). 198 Blame is here understood as a moral concept, expressing various kind of moral disapproval. When we blame someone we hold that he or she in some sense has disappointed or done wrong: that he or she has done less than what could have been expected of him or her, or done the wrong thing altogether. These kinds of judgments cannot be transformed into wholly instrumental, forward-looking ones. If we blame a criminal, for example, this cannot be fully accounted for in terms of us trying to change his or her future behaviour for the better. There is a backward-looking component to blame, which has to do with the fact that we believe the criminal acted wrongly, and that he or she therefore is blameworthy. Blame has to do with disapproval. We blame someone in virtue of something regrettable which has already happened. 199

197 For a statement of the instrumental view of holding responsible, see Smart 1961; Arneson 2003. For a criticism of the kind espoused by me, see Morse 2000; Strawson 2003.
198 The concept of responsibility has a “negative slant” Persson 2005: 377– it is more common to discuss it in relation to blame or punishment than praise and reward. Still, I believe that praise and blame are symmetrical concepts, and that praise just as well as blame therefore can be seen as an instance of respect.
199 I return to blame and other so-called reactive attitudes in the next chapter, when I discuss the objection against retributivism from the non-existence of desert.
Retribution is clearly about expressing blame. Retributivists' central claim is that punishment is justified if it is morally deserved. A crime is morally deserved, furthermore, if it is a freely chosen wrong, as stated by the AM. This seems to boil down to that criminals, in virtue of their wrongful choices, are blameworthy. Like blame, retributive punishment is a backward-looking activity which cannot be transformed into an instrumental view of punishment; and like blame, retribution is supposed to be appropriately based on past evaluations of behaviour.

If the connection between blame and holding responsible is a sound one, which I believe it is, then the following conclusion can be drawn: in being about expressing blame, retribution, unlike rehabilitation or deterrence, is about holding rule breakers responsible. Moreover, retribution is about holding responsible in a necessary and robust sense. It is necessarily about holding responsible since we cannot subject someone to retribution unless we also (try to) hold him or her responsible. Furthermore, retribution is about holding responsible in a robust sense since it is something which is not wholly transformable to forward-looking, instrumental considerations (such as blaming innocents on the ground that doing so will change their future behaviour for the better). If I blame you for doing something, there is an irreducible element of backward-looking disproval from my part which has to do to with the simple fact of me taking odds with your past actions.

So far the first premise: retribution, unlike its alternatives, is about holding rule breakers responsible. What about the second and third? Defending them will require, first, that we say something about why it is valuable to hold people responsible in the robust sense, and, secondly, that we show why this value is one which all can agree is worthwhile, those being held responsible included. Only after showing this can we argue that retribution, as a species of being held responsible, is something all should prefer.

The importance of being held responsible: the moral mechanism
Let us begin this task by analyzing why being held responsible can be preferable to the opposite reaction of being excused. I here deal with excuses which emanate from some supposed lack of autonomy.  

Based on what we have said above about the proper meaning of holding someone responsible, we could say that to be held responsible for whatever thing \( X \) means that one is subjected to blame or punishment for one’s part in the production of \( X \) (assuming that \( X \) is something blameworthy). To be excused means that while one is seen as having part in the production of \( X \), there are circumstances that make others less inclined to blame or punish.  

Now, the core point of saying that there is a particular value involved in being held responsible for one’s blameworthy actions is that one should somehow prefer blame and punishment to being excused (for blame and punishment are species of being held responsible). Since blame and punishment are unpleasant, this seems highly counterintuitive. But there is a fine-grained moral mechanism here, however, which suggests otherwise. Though blame and punishment are most often unpleasant, the unpleasantness involved is also an affirmation of a specific moral status: that of a person whose choices warrant respect. When you are being blamed, you are being affirmed as someone who could and should have known or done better. I believe there is some value, if masochistic, involved in this. Consider the following two (everyday) illustrations:

*The Two Academics.* A couple of years ago, a debate took place between Tiina Rosenberg and Johan Tralau, two Swedish academics. The debate began when Tralau wrote a magazine article in which he accused Rosenberg of plagiarism and sloppy use of references in one of her books. Rosenberg responded to the charges, calling them unfounded. Tralau persisted in his criticism, however, and the two

---

200 Thus, I do not consider the kind of excuses that concern accident, honest mistake, etc. Like I said in chapter 5, these in no way seem to threaten an individual’s status as a person.

201 Excuses, and the way they differ from justifications and pardons, will be examined in detail in the next chapter.

202 Axess 2005: 8. ”Tralau har inte kontrollerat källan: referenser finns”
spent some time sniping at each other in the media. Now, the debate took an interesting turn after Rosenberg had appeared on a radio program, once again discrediting Tralau’s claims. After the radio appearance, which reportedly included some pretty flagrant ad hominem arguments from Rosenberg’s part, Tralau wrote a rebuttal in a large newspaper. He had now changed strategy. Instead continuing the heated polemic with Rosenberg, he criticized the people in charge of the radio show for having let Rosenberg slander him without giving him a chance of rebuttal. Rosenberg, however, couldn’t be blamed since she, according to Tralau, “isn’t accountable”. Therefore, he concluded, “one cannot be angry with Rosenberg”.

The Unruly Youngsters. While reading my local newspaper one morning, I happened to read a letter about an incident that took place on one of Gothenburg’s many trams. The author of the letter, who was an eyewitness to the event, reported that at one stop, a couple of unruly youngsters had gotten on the tram. Wearing allegedly outrageous clothes, and playing loud music on a portable stereo, the youngsters had taken up seats reserved for the handicapped. The reactions of the fellow commuters got increasingly hostile. Irritated looks and scornful comments were exchanged. Suddenly, however - and to the joy and admiration of the author of the letter - an elderly gentleman arose from his seat. He cleared his throat and exclaimed: “It’s not their fault. We have raised them!”

Both these humble examples involve the withdrawal of blame. In The Two Academics, Tralau preferred to blame the people at the radio for having allowed the “unaccountable” Rosenberg speak her mind, rather than Rosenberg herself. In The Unruly Youngsters, the elderly gentleman on the tram preferred to blame the fellow commuters rather than the unruly youngsters, no doubt implying that the youngsters’ behaviour was the result of poor collective parenting. Rosenberg and the youngsters were thus excused; they were not held responsible for their actions. But I pro-

---

204 Expressen 060103, “Sluta sprida lögner, Tiina”. “Not accountable” is my translation of the Swedish term “otillräknelig”
pose there is a perfectly clear sense in which they would have preferred blame to the kind of elevated understanding that was directed towards them (particularly the youngsters, who no doubt intended to be annoying). For in both cases, the withdrawal of blame depended on a demeaning attitude towards them. Rosenberg was portrayed as a regrettable figure that, while a nuisance, could not inspire anger. The youngsters were portrayed as the end-product of lax parenting, social symptoms of their time, rather than persons whose annoying behaviour deserves to be taken seriously enough to warrant blame.

This is, at a very minute level, the moral mechanism involved in being excused rather than being held responsible. When Tralau refused to “get angry” with Rosenberg, he in effect ceased to see her as an equal; as a reasonable deliberative subject whom one can and should hold to normative standards of behaviour. He ceased to see her as a responsible person. In a quite straightforward manner, he thereby lessened her as a human being: when opting not to resent, what was left was an elevated state of calm disapproval.\footnote{205} It is perfectly understandable that Rosenberg might have preferred Tralau to “get angry” with her, for that would have meant that she and her views were taken seriously.

In all simplicity, this is the value of holding and being held responsible. The act of holding responsible expresses a particular view of persons, which seems worthwhile: a person who does wrong should not be seen as someone who is too incompetent, indoctrinated, unaccountable, ignorant, or uncontrolled to be held responsible for his or her actions.\footnote{206} To be held responsible, then, is a compliment. When blamed by others, such as when an opponent gets angry with you in a debate, you are treated as someone who can be held to standards, and who may be blamed if the standards are breached. Being excused, on the other hand, can send out the opposite message – you may be excused since you are not fit to be held responsible.

\footnote{205 Something that in all honesty probably was designed to discredit Rosenberg further.}

\footnote{206 Most hold that praise and blame are symmetrical notions in this respect (Morse 2000, see however Wolf 1980). If a person does something praiseworthy, to respect him or her equally means to see her as someone whose behaviour attributable to her.}
We might say that the notion of personal responsibility is a third aspect of the respect for persons mentioned above: to respect persons requires not only that we allow persons to choose between alternatives, and that we regard their choices as significant. Connected to this last point, it also requires that we hold them personally responsible for the results of their choices. Thus seen, to regard people as responsible for their behaviour is one of the primary social principles which serve to express and reinforce that they are persons, not non-human animals, machines or toddlers. I suspect that there is some deep conceptual linkage here between blame and punishment and the wider set of practices that are choice-sensitive. The paradigmatic example of such a practice is democratic rule. In chapter 8 I shall speculate that there might be some common ground between the capacity to be blamed and the fitness to be part of democratic decision-making: that both presume something akin to the autonomous man as portrayed by the AM.

The case of punishment

To sum up: so far I have argued that to be held responsible is a compliment which serves affirm your status as a person, and that retribution, unlike rehabilitation and deterrence, is about holding rule breakers responsible. The conclusion from this is that retribution too is a compliment and something which respects rule breakers as persons. This in my mind is the best way to argue the symbolic superiority of retribution. In expressing blame, retribution addresses, as the AM states, the rule breaker as someone who could and should have acted differently. It also has the benefit of treating criminal choices in much the same way as we treat choices in other contexts. Retributivists have been right, I think, to stress that the practice of punishment should be analyzed along the lines of other moral, social and political principles (Corlett 2001), though they have not always been clear on why. There is no immediate reason to think that criminal choices should be seen as fundamentally different from other choices. And if there is some justification to the general significance we attach to choice, as surely there is, then there is reason to think that criminal choices ought to enjoy the same significance.

To go over the alternative models yet again, the DM, which is foundational of rehabilitation theory, holds that offenders suffer from disor-
ders. This seems to clearly conflict with the assumption that the choices of persons are worthy of respect. If we regard crimes as free choices, made by rational and autonomous agents, it will seem inappropriate to engage in rehabilitative efforts to change their character. Punishment, not treatment, is what affirms the personal responsibility of the rule breaker. It is blame which confers on the rule breaker the status of a wrongdoer, rather than a disordered, abnormal, or diseased individual who in part must be seen as a victim. (Again, it may be that this status should not be conferred on factual grounds, but that is the subject of the next chapters).

The RM, which underpins purer forms of deterrence theory at least, depicts people as rational agents who will break the rules if and when doing so leads to greater expected utility. This model leads to a carrot-and-whip type scheme of positive and negative incentives, the upshot of which may seem demeaning. As noted above, however, there is nothing about threats as such which undermines your status as a person. But the instrumental view of punishment is disconnected from blame, and this is probably what underpins statements such as Hegel’s that a deterrent system of punishment treats men “as dogs”. A dog can have its behaviour modified by punishment; it can be subjected to pain or deprivation because of some violation of a standard of proper behaviour. But it cannot be blamed. This means, following the analysis in this chapter, that dogs are not held responsible, and thus not affirmed as persons, even when they are punished. It is perhaps the intentions behind punishment that decides whether it is respectful or not.

This view relies on the claim that deterrence is disrespectful of people because it disconnects punishment from blame. Deterrence has no necessary connection to blame, nor to responsibility. Indeed, it is compatible with seeing rule breakers as people who blamelessly cease on the opportunities the structure of the game presents them, and cannot be expected to do otherwise.

There is an additional, more familiar argument here that bolsters the claim that retribution, unlike deterrence, respects the choices of persons. This argument starts out from the observation that under a deterrence-based penal regime, what penal response a rule breaker is subjected to depends on what others do. When I am punished in order to produce general deterrence, I am being punished because there are influenceable agents who the state believes should be influenced in certain ways. This is not respecting my choice to break the rules. In fact, if none is influenceable, myself included, it follows from the theory of deterrence that I
should not be punished at all. Since the punishment of A is justified by
the influence it has on B, C and the others, retributivists have argued the
basic duty of not using people as means to an end is violated. By con-
trast, retributivists hold that a retributive penal regime is respectful of
my choices, since under such a regime I will only be as severely punished
as my choices deserve. The punishment I receive will not depend on
what others do, and it will not fluctuate with the needs of social protec-
tion.

6.9 THE SYMBOLIC REASON FOR RETRIBUTION

The symbolic reason for retribution, then, holds that all, rule breakers
included, would or should prefer this penal response to other responses.
We should do so since retribution, in relying on what I have called the
Autonomy Model, expresses respect for us as persons. Retribution blames
us, and thus communicates an empowering and attractive message, or
“symbol”. That message is that we are autonomous moral agents, whose
choices warrant respect. Since we are free to do good or bad, our good
choices deserves praise, and our bad choices blame.

Note that this reason, as we said in chapter 4, is a direct one. It is to
the effect that we should prefer retribution to other penal responses, even
if it involves ourselves - not just that retribution is a consequence of a re-
tributive penal regime, the virtues of which are such that we may prefer
it. Before taking stock of the way the objection from insufficiency has
been handled in the previous chapters, two objections to the symbolic
reason for retribution will briefly be discussed.

---

207 By contrast, if the aim is to extend proper blame to me, it does not matter if
my crime was the only crime in the history of the world. I should still have been
blamed for the simple fact that I chose to do something wrong, and since respect-
ing me as a person entails blaming me when I choose wrongly.
208 This may be just a little too neat. For an argument about the possible compara-
tive element even in desert-based punishment, see Duus-Otterström 2007.
(1) One could wonder whether it makes sense to extrapolate the value of being held responsible from everyday illustrations such as those offered above. It may seem as stretching it to argue that someone should see the value of legal punishment simply because it is offensive to be excused in other (less serious) circumstances. Even though everyday excuses can be structurally similar to legal excuses – being unaccountable can be an excuse in both legal and everyday contexts, for instance – it is questionable whether we can move from this observation to the conclusion that even criminals have reason to prefer retribution over excuse. After all, there is a big difference between being blamed by fellow commuters and being imprisoned by the state, sometimes even for several years.

This directs our attention to an important problem: Does holding responsible in the robust sense require legal punishment as we now know it? It certainly seems as if doesn’t. It is one thing to say that all of us want to be respected as persons, and that being held responsible for our wrongdoing is part of that respect. It is another thing to say that this requires punishment as we know it. Why couldn’t the appropriate blame be expressed ways which do not involve fines, imprisonment, etc? A criminal could be publicly scorned in newspaper articles, for instance, or be forced to endure a lecture series on Kantian ethics. This would still respect him or her as a wrongdoer – it would address the criminal as a responsible moral agent - but would not involve as much pain or deprivation. Surely a criminal would prefer this kind of response to retributive punishment?

Note, however, that being lectured to or publicly shamed for one’s wrongdoing under present penal regimes would communicate something else than being fully held responsible. We are used to that criminal wrongdoers are fined or imprisoned. Consequently, anything less than this would be interpreted as going easy on the wrongdoers, which in turn risks sending out the wrong moral message (i.e. one of excuse). Note also that a retributivist is not committed to punishment as we know it. As long as alternative sanctions are robustly about blame and perceived as sufficient to give the wrongdoer what he or she deserves, a retributivist could well support a penal regime based on lecturing (it seems implied, though, that the professors would have to be boring!). The objection that we cannot arrive at the value of retributive punishment from the premise that it is valuable to be held responsible thus fails. On the one hand, given our present understanding, punishment as we now know it is the
way criminal wrongdoers are held responsible. On the other hand, retributivists need not assume that punishment as we know it is the only way to hold criminal wrongdoers responsible, from now to eternity.

(2) It could be objected that the symbolic reason’s emphasis on intentions is problematic. I have argued that punishment is valuable if it is intended to blame you for what you did. But we could easily imagine a penal regime, run by clever deterrentists, that only pretends to aim at giving wrongdoers their just desert, but in reality is designed to promote deterrence. Such a regime, let’s assume, will not give rule breakers what they deserve. But it will give the impression that it does exactly this. If successful, how could the rule breakers under that penal regime know that they are not held responsible in the robust sense? More poignantly: how could one know whether the punishment one receives is intended to blame or deter?

The straightforward reply to this question is that it indeed seems impossible to know this. If we believe that we are treated as persons by the penal regime, that value would really seem upheld to us. As long as one defends the view that it is only the things we are aware of can be said to improve or worsen our life (“what you don’t know can’t hurt you”), this serves to disconnect the symbolic reason from actually holding responsible. That our lives cannot be made worse by things we are unaware of is a questionable idea, however - it may seem better, from some vantage point (but which?), to live under a penal regime that truly holds us responsible than to live under one which manipulatively only pretends to do so. But if we are impressed with the idea that intentions are irrelevant, we must defend the superiority of retributivism on other grounds than the symbolic one.

---

209 Thus, as stated in chapter 4, we need to compare retribution to other plausible penal responses, such as punishment-as-deterrence or rehabilitation. We should not compare retribution to fantastical scenarios.

210 For a criticism of the idea that the things we don’t know can’t harm us, see Kymlicka 1995: 25-6.
We are now back where we began in chapter 4. The purpose of this book is to offer an answer to (Q), which is the question concerning upon which theory or principle the penal regime should be based. I defend the thesis that we should answer (Q) along retributivist lines. A penal regime should aim at giving rule breakers what they deserve.

My strategy has been to present general reasons for this conclusion; reasons which in principle even rule breakers, as separate persons, can accept. I have argued that a retributive penal regime has two merits, which I have referred to as the institutional and symbolic reason. It adheres to important principles of justice and it treats rule breakers with respect. These reasons, I have argued, should convince us that retribution – the thing a retributive penal regime engages in - is the most justified penal action out there. Even those who suffer retribution should recognize this, for retribution represents a respectful mode of treatment, and it is in better alignment with central principles of justice. Provocatively speaking, this means that all should consent to their own punishment under a retributive penal regime. It is also in this sense we might speak of retribution as a criminal’s right. It represents values it would be wrong to refuse the criminal.

We started this inquiry, at the end of chapter 3, by noting that it is unclear why the fact that someone deserves punishment should be taken to mean that he or she should be punished. This I referred to as the (O1) objection from insufficiency, and I presented it in terms of a “gap” between the claims “X deserves punishment” to “Punishing X is morally justified.” My attempt to close the gap has consisted in presenting merits about being treated as one deserves in the context of criminal justice, for this would seem to enable us to move from the first claim to the second. But a couple of things should be noted. The first is that since I do not defend a strict version of retributivism here, I do not suggest that we are always justified in moving from the first claim to the second. A second thing to note is that the arguments offered in chapters 5 and 6 do not counter Dolinko’s gap in a head-on and complete way. Dolinko wonders why desert should be taken to justify punishment. I have argued that there are things about being given what one deserves which counts in favour of doing so. This is not precisely the same thing as saying that it is...
desert as such which does the justificatory work. But this is the price one
must pay for not settling for intrinsic-good retributivism.

A third thing to note is that it may seem as if I am confusing different
levels of analysis here. Whereas Dolinko wondered why desert
should count as a reason for the act of punishment, I have argued that
desert gives birth to a penal regime which is superior to other regimes.
Does this mean that I do not have anything to say about the most fundamental
question of all in the philosophy of punishment: what, if anything, justifies punishment? No. While the institutional reason is obviously not an argument for punishment as such – it is only an argument for a particular kind of penal regime once we have one - the symbolic reason is a direct argument for punishment. It states, as we have just seen, that the act of retribution respects rule breakers as persons, which is a good thing. This, then, counts in favour of a singular act of retributive punishment just as it underscores the claim that a retributive penal regime is symbolically superior to rival regimes.

Suppose that the institutional and symbolic reasons are sound, and
that they count in favour of a retributive answer to (Q). A final thing to
note is that it is unclear what the ultimate justification of that answer is.
Retributivism has of course traditionally been understood as a deontological theory (Dolinko 1997). If my argument is read as a deontological one, we would say that a retributive penal regime is good because it treats people fairly and with respect, and thereby adheres to obligations we have to each other. But this interpretation seems to be silent on a fundamental question: why we have a practice of punishment in the first place. It is one thing to argue that the institutional and symbolic reason count in favour of preferring a retributive penal regime over ones based on deterrence and rehabilitation. But can fair and respectful treatment really explain why we are right in punishing rule breakers in the first place? This not only questions whether retributivism can supply us with an answer as to why we have a practice of punishment. More to the point, it also questions whether the reasons in favour of a retributive penal regime are really deontological. It might just as well be that they are of a consequentialist, indeed even straightforwardly utilitarian sort: it might be that we have reason to prefer a regime which gives us what we deserve because it would leave us on aggregate better off than other regimes.

I will return to these difficult questions concerning the ultimate justifi-
cation of a retributive penal regime in the concluding chapter. Here it
suffices to say that the institutional and symbolic reasons count in favour of a penal regime based on desert, regardless of from where they derive their ultimate justification.

6.11 GOING FORWARD: THE OBJECTION FROM NON-EXISTENCE

A restless reader may long have objected that the symbolic reason for retribution must be supremely worthless to rule breakers who do not meet the criteria for the kind autonomous and undetermined agents the AM assumes. If someone suffers from a behavioural disorder that results in crime, surely it is nonsense to say that he or she has an interest in having his or her choice “respected” by being punished. Some could perhaps argue that we should treat criminals as if they are responsible even when they are not, because we thereby uphold a useful illusion (cf. Dennett 1984). But most would say that it is not right to hold responsible those who are not responsible.

Retributivists would agree. They rest their argument in favour of punishment on the substantive assumption that rule breakers by and large are responsible for their actions, in the robust sense associated with being able to do otherwise. Rule breakers should be respected because they are in fact worthy of such respect. Thus, retributivists do not reject rehabilitationalism’s claim that disordered rule breakers should be treated rather than blamed and punished. Retributivists could fully agree with that. It is only that they believe that rule breakers in general are not disordered, and that it would be inappropriate to treat them as if they were.

This claim is of course highly contestable. Contemporary penal regimes usually acknowledge quite a few disorders that mitigate, if not fully exonerate, offenders’ responsibility. In Sweden, for instance, the kind of disorders or diseases that can result in a full legal exoneration include: psychosis; dementia; suicidal depression; severe personality disorders (such as debilitating neuroses); some forms of drug-related withdrawal symptoms; intellectual impairment; compulsive disorders
such as pyromania and kleptomania; and paedophilia. These disorders, when sufficiently grave, presumably explain acts of rule breaking in such a way that they undermine precisely the notion of the offenders being personally responsible and blameworthy for them.

What if lists such as this one become longer and longer? The disorders that can exonerate Swedish offenders seem based on scientific understanding. They involve behaviour the explanation of which at some point in time, prior to scientific “discovery”, presumably were unknown, and which consequently were subject to ordinary punishment. And there is no reason to think that science won’t find further disorders in the future. There is thus no reason to think that we won’t expand the list of the legally exempted even further. Indeed, as the behavioural sciences marches on, becoming able to explain ever more acts of rule breaking in scientific terms, chances are that it may turn out that something akin to the Disorder Model best describes the nature of rule breaking after all.

Those who advocate a scientific approach to crime often suppose something like this. As scientific understanding becomes wider and deeper, they argue, the very notions associated with “badness” will become increasingly displaced. Once we understand the causal etiology of crime or classes of crime, we will discard the notions of rule breakers as fundamentally free and blameworthy characters as little more than primitive myths, moral just so-stories, or figments of our imagination. The symbolic reason for retribution, which relies on the notion of the rule breaker as a fundamentally free agent who is capable of deserving blame in virtue of his choices, would then appear worthless.

There are as many ways of arguing this as there are definitions of scientific progress. B.F. Skinner, the great behaviourist whom we will meet later, held that scientific understanding would unravel the all-encompassing environmental influence over human behaviour:

“In the traditional view, a person is free. He is autonomous in the sense that his behaviour is uncaused. He can therefore be held respon-

---

211 SOU 2002:3. Sweden is somewhat of an outlier when it comes to criminal liability, however. Offenders who are deemed to suffer from severe mental disorder of the kind just mentioned cannot be imprisoned, but may nevertheless be “sanctioned” by the courts to undergo treatment in an institution of forensic psychiatry. They are thus not exempted from legal guilt.
sible for what he does and justly punished if he offends. That view, together with its associated practices, must be re-examined when a scientific analysis reveals unsuspected controlling relations between behaviour and environment” (Skinner 1971: 19-20).

Greene and Cohen instead place their bet on cognitive neuroscience:

“Cognitive neuroscience, by identifying the specific mechanisms responsible for behaviour, will vividly illustrate what until now could only be appreciated through esoteric theorizing: that there is something fishy about our ordinary conceptions of human action and responsibility, and that, as a result, the legal principles we have devised to reflect these conceptions may be flawed” (Greene & Cohen 2004: 1775).

Less fashionable today (and less based in science) but nonetheless in the same spirit, the aforementioned Robert Owen declared that the social system fully conditions man:

“THE WILL OF MAN HAS NO POWER WHATEVER OVER HIS OPINIONS; HE MUST ... BELIEVE WHAT HAS BEEN ... IMPRESSED ON HIS MIND BY HIS PREDECESSORS AND THE CIRCUMSTANCES WHICH SURROUND HIM. It becomes therefore the essence of irrationality to suppose that any human being ... could deserve praise or blame, reward or punishment” (Owen, cited in Roberts & Sutch 2004: 146).

There are innumerable statements of this kind, but the general idea should be clear enough: There are various things that account for human behaviour, and when we understand them properly, we will see that we do not form our own character or actions out of our own free will. Therefore, it is wrong/irrational/unfair to blame people when they act in ways perceived as wrong. (It need not be held that the things which account for criminal behaviour amount to a disorder – one could make the same hidden nuts and bolts behind behaviour exonerate even “sane” criminals – but often something like this is added.)

The upshot of these statements could most concisely be captured by the following claim: we are not personally responsible for our actions, at least not in the robust sense of being blameworthy for them.

This is a stupendous claim, but as we shall see one which many have defended. It naturally conflicts with retributivism’s substantive
assumption about rule breakers as capable of deserving punishment. Responsibility is a prerequisite of desert, so unless we are responsible for our actions (in the deep sense required by the AM), we can never be deserving of blame or punishment.

It will perhaps be recalled that I earlier spoke of a second objection to retributivism. I referred to this objection as the \((O_2)\) objection from non-existence. According to this objection desert - the concept which takes centre stage in the retributive theory - simply does not exist. Nobody really deserves anything. It follows that we cannot justify retribution on the ground that it is an appropriate mode of treating us. For if we are such that we cannot deserve blame (or praise), how could it possibly be the case that we are entitled to a treatment which is premised precisely on our non-existent ability to deserve things? If the concept of desert is a myth, much of what has been said in favour of retributivism so far invariably falls. To be treated with punitive respect is only valuable if we are in fact worthy of that respect.

This is the objection we are now facing. I shall deal with it in the following three chapters, starting out in the next chapter by offering an account of why scientific explanations of blameworthy actions indeed seem to be excusing.
Part III. An Objection to Retributivism
7. A Theory of Excuses

“To understand all is to forgive all.”
(Madame de Staël)

“It is a mistake to let moral condemnation interfere with trying to understand – but it would be a bigger mistake to let that understanding, once it has been attained, interfere with moral condemnation.”
(Baumeister 1997: 387)

The objection we shall deal with over the next three chapters holds, then, that desert, while potentially of great moral significance, does not exist. Retributivism attaches great weight to desert, so the theory is naturally threatened by this objection. But the objection from non-existence of course threatens the plausibility of all positions that substantively rely on desert. A discussion of this objection will therefore not only be of relevance for retributivists and the question of justice in punishment. It will also be of relevance for a much wider set of concerns, such as the standing of desert in general in theories of justice. And as we saw, the implications of the view are such that we will eventually have to question the idea of autonomous agency altogether.

Why would one think that desert does not exist? As we saw at the end of the previous chapter, one way of reaching this conclusion is as follows: Desert requires personal responsibility. In order for us to be personally responsible we must meet certain criteria, such as being genuinely able to choose between alternatives and/or acting out of factors that do not lie beyond our control. But these criteria are incompatible with a scientific understanding of human behaviour. Therefore, a scientific understanding of human behaviour is incompatible with desert. The

212 Of course, it is also of relevance for theories and practices that are responsibility-sensitive (see Smilansky 1996; Olsaretti 2003b).
more we become able to understand human behaviour, the more will we see that people are never deserving of anything, at least not in virtue of their actions.

The central part of this reasoning is the notion that scientific explanations undermine the notion of us being personally responsible for our actions. I shall in this chapter try to validate this claim by offering a “theory of excuses”, the upshot of which is that all (or most) scientific explanations of actions, if true, rule out responsibility (and hence desert) for those actions. Thus, I shall attempt to defend the following claim (C):

(C): a scientific explanation of a prima facie blameworthy action (such as a crime) is, if true, also an excuse for that action.

We will be more specific about what it is about scientific explanations that are excusing in a moment. First, however, we should note that the dominant view is that (C) is false.

7.1 THE RECEIVED VIEW: THE DIFFERENCE BETWEEN EXPLANATIONS AND EXCUSES

It is somewhat of a received wisdom among philosophers, legal theorists and scientists that there is a salient difference between explanations and excuses. Explanations, defined as attempts to answer “why”-questions (“Why did X occur, rather than Y or Z?”), concern empirical questions of causation. Excuses, on the other hand, boil down to a question of morals (“Should we hold somebody responsible for X or not?”) Those who defend the difference between explanation and excuses argue that we can never derive a standard of responsibility from empirical facts: we cannot infer from the observations relevant to the first question an answer to the second. James Q. Wilson offers a good illustration of this line of reason-

---

213 In the social psychology, however, the concept of excuse tends to be treated as a subspecies of explanations, just as the concept of justification. See Shaw et al. 2003.
ing. On the “profound difference between causation and responsibility” he begins by writing:

“One could concede – indeed, if one is an especially ambitious social scientist, one will proclaim – that all human behaviour is caused. That is, if one knew enough about the antecedent conditions of a given act, one could completely explain that act. [...] If we knew enough about my genetic endowment and childhood socialisation, the opportunities and incentives available to me and the beliefs and principles by which I evaluate them, we could fully explain why I am sitting here writing this chapter. But suppose we did fully explain my writing of this chapter [...] Would we then say that I [was] not responsible for [my] actions? Clearly not. The reason is not simply that we have no tested theory of human action that predicts these behaviours with much precision. Some day, in fact, we might have such a theory” (Wilson 1997: 38-9).

Why wouldn’t we say that Wilson is without responsibility if we could fully explain his actions? Because “a cause is not an excuse except insofar as the cause deprives the person of reason or forces the person to act”, Wilson argues (ibid. 42). There is a line to be drawn between explanations and excuses: explanations are only excusing if they are of a certain kind.

According to Wilson, then, all behaviour may be fully caused (and thus fully explainable) and yet compatible with responsibility. Responsibility, we are often told, depends on whether the behaviour is voluntary controlled and performed by a “reason-responsive agent” (Fischer & Ravizza 1998). And not all explanations explain in ways which undermine people’s control or ability to reason.

There is a deeper twist to saying that explanations are different from excuses. Suppose someone says that all explanations are excusing. He or she is then tacitly relying on some standard of responsibility, which explanations are believed to violate. But that standard of responsibility is clearly a normative or at least partly non-empirical one, which needs separate justification. We cannot simply infer from the claim that all behaviour is caused that all behaviour is to be excused. The first claim is empirical whereas the second is not. There is, then, always logical room for the question, “X caused S to do A, but is X an excuse?”
The view that there is a difference between explanations and excuses is very prevalent. Mackie claims that “It is a factual, psychological, question whether an action is intentional or voluntary, but it is a moral or legal question whether or in what ways an agent is to be held responsible” (1990: 208), thereby suggesting that the factual and moral worlds have some degree of independence from each other. In a similar vein, legal theorist Morse remarks that to confuse explanation and excuse is the “fundamental psycholegal error” (Morse 1999: 275). Like Wilson, Morse finds it obvious that all actions have causes. But this does not mean that every action should be excused: we can cite a thousand factors in explaining why a criminal commits crime, but unless we have some essentially moral idea about where responsibility ends and excuse begins, we can never judge whether he or she should be excused for his or her actions. In legal systems, Morse notes, we entrust legislators with coming up with that moral idea: “Responsibility is a normative concept, and we empower legislators, as our representatives, to create normative standards trough legal rules” (ibid.). Neu, finally, recommends that “we should recognize that there is always a story to explain how things have come to be as they are and why a person did whatever they [sic] have done. But not all stories excuse” (2004: 18).

Two claims are being made here. The first is that we cannot settle the question where responsibility ends and excuse begins by empirical inquiry alone. We cannot from “X suffers from disorder such-and-such” infer “X is not responsible”, since the latter is a normative statement. The second claim is that while some explanations may reasonably be taken as excuses, there are others that may not. My aim here is to offer an account of why scientific explanations, if true, almost invariably can be said to be excusing. As will become obvious, this does not mean that I challenge the former of the above claims. I shall not challenge the idea that there is at least in a logical sense a difference between normative questions of excuses and empirical questions of explanations. I shall, however, argue that given some widely held ideas about responsibility nearly all scientific

---

214 It is commonplace to quote Austin’s reaction to the ideas that explanations excuse here. Austin, when presented with the idea that to understand all is to forgive all, is reported to have said: “That’s quite wrong. Understanding might just add contempt to hatred” (cited in Dennett 1984: 32).
In order to accomplish this, we need first to analyse the concept of excuse, and relate it briefly to the adjacent concepts of justifications and mercy.

### 7.2 EXCUSES, JUSTIFICATIONS, AND MERCY

To excuse somebody, $S$, consists of two legs: first an acknowledgment of the wrongness or lamentableness of $S$'s behaviour, and second a reduced or non-existent willingness to punish $S$ on the grounds that $S$ is deemed to be less than fully responsible/not responsible at all for the behaviour. Excuses should be held distinct from two adjacent concepts, justifications and mercy. Excuses, justifications and mercy all concern withdrawal of blame or punishment that seems prima facie appropriate. But whereas excuses admit the wrongness of an act but deny responsibility for it, justification and mercy both accept responsibility. Let us start with the distinction between excuses and justifications, as treated by J.L. Austin (1970).

Let’s say a doctor performs euthanasia (which we assume is illegal) on a patient and is arrested and tried in court. The doctor is unwilling to accept blame or punishment for her action. A strategy of justification would be for the doctor to challenge the notion that killing the patient was wrong. Maybe the doctor claims that the patient lived in unbearable pain and really wanted to die, and that, as a physician, she could not with good

---

215 An example of such a widely held belief is that we should not be held responsible for things we cannot control. That this belief has strong intuitive support is evident from, say, the fact that we are unwilling to hold people responsible for what they do under hypnosis, where one stands under direct control of another. But note that, in a logical sense, there is nothing that commits us to saying that the hypnotized person is without responsibility. One could argue that one is responsible for anything one’s limbs do or produce, which means that you remain responsible for, e.g., breaking a vase under hypnosis. I shall argue that while it is possible to employ such a standard of responsibility, in practice we are unlikely to do so.
conscience deny the patient the right to die. Applying this strategy, the
doctor in no way challenges her part in the patient's death. She admits
that the killing was intentional. The doctor instead challenges the notion
that the act was wrong. In essence she says: “Yes I did it – but I was right
to do it, and I would do it again!” If the strategy is successful she has
justified her action. It no longer appears blameworthy and shouldn’t be
punished.216

Another strategy open to the doctor is that of excuses. *Excuses consist
in claiming that the thing deemed blameworthy isn’t ones fault.* For example,
the doctor might say that the killing was accidental, or that she suffered
from a mental disorder at the time, or that she didn’t know that euthana-
sia was outlawed. In essence, the doctor then says: “it wasn’t my fault.”
Mistake, ignorance, externally coerced action, mental disorder and bad
luck are all catchwords of excuses (see Glover 1970: 57-61).

The distinction between justifications and excuses is not watertight.
Consider for instance a man who performs (generally) blameworthy
Is he entitled to a justification (“I did it, but it was the lesser evil”) or
an excuse (“I had to do it, there was no other choice”)? Or consider a man
who assaults another out of great provocation. Is he entitled to a justifica-
tion (“he deserved it”) or an excuse (“he got on my nerves, I couldn’t
control myself”)? Leaving these difficulties aside, we may note that both
justifications and excuses serve the same basic function - to claim that an
agent does not deserve blame or punishment for what he or she did.

To ask for *mercy*, finally, consists in accepting both responsibility
and wrongness but requesting that one escape blame or punishment
anyway (Murphy 2003: 13-4). Responsibility and wrongness are both
necessary conditions of mercy, for we cannot be merciful towards people
who could not help doing what they did (i.e. those up for excuses), nor
towards people who in the end were right to do what they did (i.e. those
up for justifications). Let’s say our doctor acknowledges that it was
wrong to kill the patient, and furthermore does not try to dodge respon-
sibility for what happened. If the doctor wishes to avoid blame or pun-
ishment under these circumstances, what is left is to ask for mercy - to

216 Or be punished more leniently than otherwise would have been the case.
ask for leniency from the people who would be justified in holding you responsible, should they choose to. 217

From this conceptual analysis we get the following taxonomy: 218

*Excuse*: denying responsibility, accepting wrongness

*Justification*: accepting responsibility, denying wrongness

*Mercy*: accepting responsibility, accepting wrongness

Justifications and mercy do not deny responsibility, so they do not constitute a direct threat to the substantive assumption of retributivism that rule breakers are capable of deserving blame. They are not incompatible with what I have referred to as blame as expressing respect for rule breakers. For our present purposes, therefore, only *excuses* are relevant. A working definition of an excuse, based on what we said above, would be a withdrawal of blame on the basis that an act, disposition or state of affairs, while bad, isn’t something for which the party originally eligible for the blame is responsible. The definition is for the sake of clarity very rigid. We could imagine degrees of responsibility and excuse. We could partially excuse S on the basis of one or another mitigating factor, which serve to lessen the responsibility of S without fully exonerating him or her.

---

217 Mercy lies very close to *forgiveness* but there are important differences between these two concepts. Like mercy, forgiveness assumes that the wrongdoer is responsible and at fault. But forgiveness is, as noted by Murphy, essentially about an *emotional disposition* whereas mercy is about whether, and to what extent, to blame or punish. We may forgive without showing mercy, and we may show mercy without forgiving – forgiveness is about giving up feelings of resentment, and we may still punish those who are forgiven (See Murphy 2003: 13-14).

218 Note that there is no position for “denying responsibility, denying wrongness.” This is because the situations where the three positions above become actualized are such that it is prima facie relevant to blame somebody, and it that it seems to be overkill to deny both authorship and wrongness (it would be awkward to say, for example, “I am not responsible for having started the riot, but in the case I was responsible, I don’t see what’s wrong with riots”).
7.3 EXPLANATIONS AS EXCUSES

As we saw earlier, the received view is that there is a tenable difference between explaining something and excusing it. Explanations operate in the empirical realm. When we ask why $S$ broke the rules, we may cite many different factors that, in isolation or in combination with one another, produced the outcome: $S$’s desperate social situation, low IQ, biological disposition, etc. Never, it seems, do we among these factors find evidence to the effect that we can say whether $S$ is responsible for his or her crimes or if he or she is to be excused. This is because the latter is a normative question which cannot be answered by empirical observations, it is usually assumed. My aim here, however, is to offer an account of why this reasoning might be incorrect, or at least exaggerated. Put short, I aim to offer a theory of why explanations are excuses.

It should be immediately noted that such a theory does not hold for every type of explanation. Many of the accounts we accept as explanations in everyday life do not qualify as excuses. “I overslept” is not an excusing explanation of why I am late, nor is “I missed the bus”. Although we may accept them as explanations in an everyday sense, they are not such that we become less inclined to blame when we accept them at face value. (You may perfectly well believe that I am blameworthy for oversleeping or missing the bus). Nor do explanations of a more spiritual kind count as excuses. Suppose we explain why $S$ commits crimes with $S$’s wickedness – his or her evil exercise of the free choice inherent in every human being because of our likeness to God. Responsibility, we may argue, is built into the premises of that kind of explanation.219

Scientific explanations are a different story, however. They differ from everyday explanations220 and religious explanations in that they seem less capable of incorporating moral elements such as responsibility. Indeed, scientific explanations of human behaviour – for it is those that are relevant when discussing responsibility and excuses – even appear to

219 At least if evil is envisioned as an internal, voluntarily induced state, such as when one stresses man’s ability to corrupt himself through his choices. For a statement of this view, see della Mirandola 1985.

220 With intentional explanations differing the least to be sure. See Rosenberg 1995, especially the passages about folk psychology.
be hostile to responsibility. There are two basic and largely overlapping reasons why scientific explanations have a tendency to excuse.

On the one hand, scientific explanations tend to explain by citing factors beyond the individual’s control. This is because science paradigmatically advances causal explanations. Causal explanations build on the idea that we can explain occurrences by citing their (prior) causes; that we explain something by citing the things which led to or produced it (Humphreys 1989). But the causes we use to explain an occurrence of course in turn have causes too. So in order to achieve scientific understanding, we need to look at and understand causes which lie ever earlier in time. Scientific explanations by their very logic are therefore susceptible to regresses backwards along causal chains. In the case of action, these regresses seem to end up in factors well beyond the individual’s control, the paradigmatic illustrations being an individual’s genotype or social background. We may call this reason for why explanations are excusing their tendency towards regress:

The tendency towards regress. Explanations have a tendency to cite causes that lie ever earlier in time. This means that they tend to cite causes that lie beyond the individual’s control.

On the other hand, the logic of explanation is such that the notion of alternative possibilities seems suspect. The more we can explain why something happened, the less plausible it will seem that something else could have happened instead. A successful explanation shows that the explained event “was to be expected”; that it is no wonder that it happened given the explaining factors.221

221 That explanations show that the explained event “was to expected” is Hempel’s classical way of putting it (Hempel 1965: 337). Hempel is famous for his models of scientific explanations, most famous of which is the Deductive-Nomological (D-N) model. The D-N model states that a successful explanation takes the form of a sound deductive argument, where at least one of the premises is a law or lawlike generalization. Individual events, on this account, are explained by stating initial conditions and subsuming them under at least one covering law. The D-N model lends itself particularly easily to the notion that explanation can be excuses – given the initial conditions and the covering laws, this model assumes, nothing but the occurrence could have happened. As will be-
In more technical words, an explanation is successful if the *explanans* (the thing(s) that explain) renders the *explanandum* (the explained thing) expectable. An explanation simply makes us think that the *explanandum* event is not surprising, mysterious, or unintelligible. There are different degrees of strength of this expectedness. Explanations minimally render the *explanandum* non-mysterious. 222 Maximally, they state that the *explanandum* was determined to happen – that nothing but the *explanandum* could have happened given the explaining factors. But explanations all have a tendency to make us believe that it was *not strange* that such-and-such happened, which often – but not always – leads us to conclude that it was unlikely something else could have happened instead. 223 This means that explanations – at least as long as they do not include alternative possibilities among their premises – may by their very logic be at odds with the notion of alternative possibilities. In case of blameworthy behaviour, explanations might serve to make us think that the behaviour in question was similarly to be expected. Let’s call this reason for why explanations are excusing their tendency towards determinism.

The tendency towards determinism. Explanations are at odds with the notion of alternative possibilities, since the better we can explain why X happened, the less plausible will it seem that not-X could have happened.

How do these tendencies of scientific explanations relate to concept of responsibility? How could one say that explanations in virtue of these tendencies are excusing? Well, the tendencies threaten what seem like two plausible requirements of responsibility. The tendency towards re-

---

222 Mystery is the antithesis to a successful explanation. That which is explained ceases to be mysterious, and the inherently mysterious cannot be explained.

223 With the exception being probabilistic causation. In objectively probabilistic processes, we understand the more and less likely outcomes equally well. Let’s say there is a 0.2 probability of a child being born with blue eyes, and a 0.8 probability of being born with brown eyes. Given that the process is well-understood but inherently random, we can explain either outcome equally well, even though we admit that there were alternative possibilities (see Salmon 1989: 62f).
gress threatens what we may call the *ultimate control*-requirement, which is the idea that in order for me to be responsible for an action, I must have a kind of buck-stopping control over it. The action should not be (fully) attributable to things which I in no way could affect.\textsuperscript{224} But, because of the tendency towards regress, explanations of human action often cite, among other things, cultural context, socio-economic background, early socialization, genetic makeup, internalized norms, institutional setting, or psychological profiles. These are often straightforwardly beyond the individual’s control – one can scarcely control what genes one was born with or the social circumstances of one’s childhood, for instance. Speaking with Rawls (2000), they are brute outcomes of the natural and social lotteries and as such nothing we can be held responsible for. So if one can explain why $S$ breaks the rules by citing such factors, then it seems as if these explanations do undermine responsibility. They undermine the notion that in order for us to be responsible for an action, it must not be (fully) explainable by things beyond our control.

The tendency towards determinism threatens the *alternative possibilities*-requirement, which we introduced in the previous chapter. According to this requirement, in order for me to be responsible for doing something, it must also have been possible for me to refrain from doing it. Unless I could have done otherwise, it seems wrong to hold me responsible for the action.\textsuperscript{225}

Note that on most accounts, these requirements interrelate in a specific way: we require that actions have alternative possibilities *in order* to secure that we can be ultimately responsible. Unless we have alternative possibilities, the causal chain of our lives prods us down a singular possible path. This means, taking the causal regress into account, that the

\textsuperscript{224} More specifically, *ultimate control* means that in order for us to be responsible, sufficient causes of our actions should trace back to factors within our control. In order to have buck-stopping ultimate responsibility, then, we must have the power to originate causal sequences. See Kane 1996 and Strawson 1994.

\textsuperscript{225} Alternative possibilities was defined in chapter 6 as follows: “The agent has alternative possibilities (or can do otherwise) with respects to A at t in the sense that, at t, the agent can (has the power or ability to) do A and can (has the power or ability to) do otherwise” (Kane 1996: 33).
explanation of how our life unfolds and how we act must fully be sought in factors we in no way could control.\textsuperscript{226}

Not all will agree that responsibility requires alternative possibilities or ultimate control. According to currently popular \textit{compatibilist} analyses of responsibility, neither condition is necessary – indeed, they are both positively untenable. I deal with compatibilism in the next chapter. The account offered in this chapter assumes that responsibility requires alternative possibilities and ultimate control.

Since scientific explanations, in virtue of their tendencies towards regress and determinism, seem to undermine responsibility (and hence excuse), I will here postulate that an explanation, when believed to be true, should also make us excuse a prima facie blameworthy action. We will question the practical plausibility of explanations being excusing in short, but let us here settle for the following claim: if there is reason to excuse an agent (because of such-and-such an explanation), then we should modify our attitudes towards that agent. We should not only excuse in theory, as it were, but also in practice.

Note that excusing \(S\)’s behaviour does not mean \textit{tolerating} or \textit{condoning} it.\textsuperscript{227} We do not rejoice at the thought of having our cars stolen by \(S\) just because \(S\) cannot be held responsible for his or her criminal behaviour. Nor do we have to refrain from taking steps to stop \(S\): even if we believe \(S\) to be completely without responsibility and thus undeserving of punishment, we might still want to incarcerate or treat \(S\) in the name of social protection. Thus, the theory of excuses we are discussing here does not imply that we feel good about what a wrongdoer is doing just because we believe the wrongdoer to be without responsibility. Nor does

\textsuperscript{226} The best statement of the view I rely on here is van Inwagen 2003.

\textsuperscript{227} This is very important, and a point that is often misunderstood. I am not claiming that explaining crime means that we take crime any more lightly. Explaining crime does not mean that crimes are “natural” or “unavoidable”. On the contrary, what moves many scientists to inquire into the causes of things is precisely that they want to stop them. A medical scientist doesn’t tolerate cancer just because he or she seeks to explain it, and an international relations scholar doesn’t tolerate war just because he or she seeks to uncover its causes. They are both interested in the very opposite – curing cancer and promoting peace. A social scientist seeking to unearth the causes of crime may have the same attitude: to understand the phenomenon in order to prevent it.
it imply that we should remain passive. What the theory does say, however, is that believing an explanation will preclude reactive attitudes.

“Reactive attitudes” is a term I borrow from P.F. Strawson, whose work I describe at some length below. The term signifies a collection of moral emotions – resentment and gratitude paradigmatically - that are central to truly interpersonal relationships. To have (negative) reactive attitudes towards S is different from merely regretting what S did, or being angered by it. We resent S only in so far as we, roughly, regard S as a responsible moral agent; an equal whom we rightly hold to standards of behaviour. The clearest contrast to this moral emotion is what we feel when natural disasters strike. An earthquake can in a purely causal sense be said to be responsible for levelling a village. But it makes no sense whatsoever to say that the earthquake was responsible in a moral sense (unless we trace the question of why the earthquake occurred back to some agent). Earthquakes “just happen”. We cannot resent them.

Analogously, we might think of S’s behaviour as we think about earthquakes – as something bad that should be prevented if possible, but which it makes little sense to resent. The madman, the infant, the rabid dog; all of these are more or less similar to earthquakes in this respect. Though we might detest them, we cannot resent them. This means that we cannot, strictly speaking, blame them. But if an explanation of an act of rule breaking is such that responsibility is undermined, the rule breaker must be seen as on par with other regrettable things which we do not blame.

The move from a scientific explanation of a blameworthy behaviour to making the wrongdoer in a moral respect resemble an earthquake is surely a very long one. The theory of excuses I am trying to sketch here does not claim that scientific explanations makes rule breaking out to be exactly like natural disasters (for one thing, our ability to stop rule breaking seems greater then our ability to stop natural disasters). What I have tried to sketch, however, is an argument to the effect that given some prevalent requirements of responsibility, such as that a rule breaker

\[\text{228 Some natural disasters of course do depend on human actions, and thus falls under the jurisdiction of responsibility. A case in point would be human induced climate change. Note also that if natural disasters can be traced back to some kind of deity, such in the case of the Theodicy-problem, then moral responsibility becomes relevant.}\]
could have done otherwise and does not act out of factors beyond his or her control, there is a clear sense in which scientific explanations, if true, are excusing. This is the idea behind the claim that retributivism’s conception of rule breaking is unscientific, lost in the confused realm of “badness.” Underlying the claim that no one can deserve anything is the idea that a scientific understanding of (the causes of) crime will prove that excuse, not responsibility, is the appropriate reaction to criminals.

7.4 THE OBJECTION FROM THE NON-DETERMINATE USES OF EXPLANATIONS

The argument pursued so far is that explanations, when they are of a certain kind and are believed to be correct, have a determinate moral reaction. I have namely argued that we ought to excuse those whose actions we can causally explain.

A critic may reply, impatiently, that this rests on a massive confusion of empirical explanations on the one hand, and moral reactions on the other. What I have previously called the “received wisdom” - the notion that the questions of causation and blame are to be held separate – is nothing short of Hume’s law, which famously says that we cannot derive an “ought” from a “is” (cf. Hume 1966). Yet, my claim is that empirical explanations do have determinate moral consequences. This certainly warrants some closer inspection.

One way of criticizing my argument is to point to the fact that explanations are often put to very different use. Two people can agree on what it is that explains something and yet draw different conclusions in terms of what reaction would be appropriate. James Robert Brown has offered a very compelling example of this (2001: 108-12). He notes that people with wildly different attitudes towards a phenomenon may use the same explanations in support of their views. As an example Brown discusses explanations of homosexuality.

Brown suggests that there are two major approaches to explaining homosexuality. One explanation holds that homosexuality is an immutable trait that is either biologically determined or cemented at a very early age. He calls this position essentialism. The other explanation is that homosexuality is a choice. According to this view people construct their
own sexuality (under the influence of the social context in which they live), and can construct it in different ways. Brown names this explanation constructivism. Brown notes that both "pro-gay" and "anti-gay" groups can use either explanation for their purposes.\(^\text{229}\)

7. Brown’s illustration of the non-determinate use of explanations

<table>
<thead>
<tr>
<th>Pro-gay</th>
<th>Anti-gay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Essentialism</td>
<td>Being gay is biologically fixed. Therefore, it is natural and gays should enjoy legal protection as others do on the basis of gender or race</td>
</tr>
<tr>
<td>Constructivism</td>
<td>Being gay is a free choice, one that should be tolerated</td>
</tr>
</tbody>
</table>

Brown is right, of course. Even though the distinction between “is” and “ought” is under heavy attack in some circles (see Bhaskar 1998), I believe it to be basically correct. Consider this argument:

\[
P_1: \text{Homosexuality is an inherited trait} \\
C: \text{Homosexuality is to be tolerated}
\]

This argument clearly violates the is/ought-distinction. We need to supplement it with an additional, normative premise:

\(^{229}\) Figure taken from Brown 2001: 110.
P1: Homosexuality is an inherited trait
P2: Inherited traits are to be tolerated
C: Homosexuality is to be tolerated

The argument is now valid. But as Brown notes, a normative premise for the repugnance of homosexuality could just as well be added:

P1: Homosexuality is an inherited trait.
P2: Some inherited traits, such as homosexuality, are objectionable
C: Homosexuality is objectionable.

This seems to support that explanations do not have determinate moral consequences. We are not forced to move from an explanation to a specific moral reaction. What is missing in Brown’s example, however, is that not every explanation is equally compatible with the notion of responsibility. This will arguably affect the normative reactions to them. For instance, it seems that what Brown calls the constructivist position more readily can incorporate responsibility than the essentialist one. If becoming homosexual is a “free choice”, perhaps one can be held accountable for it. At the same time it would be bizarre to say that homosexuals are responsible for being gay if homosexuality is an inherited trait – one’s genotype being the paradigm example of a “factor beyond the individual’s control.” Few would argue that:

P1: Homosexuality is an inherited trait
P2: One is only blameworthy for the things one is responsible for

It will be compatible with responsibility as long as it assumes freedom of the will or something similar. My argument here is that what appears to be a free choice of an agent might, when facing a regress, be just as incompatible with responsibility. For an clear statement of this view, see Strawson 1994.
P3: One is responsible for some inherited traits such as homosexuality
C: Homosexuality is blameworthy

The reason why this argument looks so unsound is clearly that we believe that P3 is false. It can only be defended, it seems, by adopting a worldview where people are reincarnated and thus may be born with certain traits as a reward or punishment for actions in past lives. But most reject P3 without hesitation.

Now we can see why the independence between explanations and excuses is highly overstated: some explanations, it seems, rule out responsibility, and thus render blame inappropriate. This conclusion no doubt rests on a further belief, i.e. (P3) that we are only blameworthy for things for which we are responsible. But that thought seems perfectly reasonable. In sum: there is a logical difference between explanations and excuses. But the point still stands that given widely held ideas about responsibility, certain explanations do commit us to excuse.

An empirical illustration: the case of decriminalization of homosexuality in Sweden

To believe that homosexuality is explained in a way which exonerates homosexuals doesn’t mean that we have to be “pro-gay”, however, and so there is no need to disagree with Brown in this respect. Brown’s argument is valid as long as we’re talking about attitudes to traits or behaviour. He is right in that a negative attitude to homosexuality may very well survive conclusive results that it is biologically determined. But I

---

231 A position which risks getting caught in a vicious regress, but does not have to.
232 Also: while the distinction between facts and values is correct in a logical sense, to understand the political consequences of scientific explanations requires that we have a clue about how people are likely to use the explanations. In the case of “genetic” homosexuality, I suspect the first argument to be adopted – the non-blameworthiness of inherited traits - but I believe many parents would opt for genetically tailored heterosexuality as well, if given the opportunity. Falke-
propose that the possible biological determination of homosexuality rules out that people can be responsible for being homosexuals, since one cannot conceivably be said to choose one’s genotype. And if my connection between an explanation and diminished or no responsibility – excuse – is correct, the same will hold for all explained behaviour (as long as it is explained in a certain way, of course).

We may illustrate this by taking a look at the history of the decriminalization of homosexuality in Sweden, which supports nicely what I have said here. For a long time homosexuality was regarded as a wicked choice, but when the phenomenon began to be explained along biological lines, the idea that homosexuals were culpable faded away. I want to emphasize that I in no way here construes homosexual behaviour as rule breaking in the sense discussed so far (i.e. illegal behaviour which constitutes a prima facie case of wrongdoing). We will concern ourselves with homosexuality in order to drive home the general point that explanations are incompatible with responsibility, given widely held beliefs about what the latter concept requires.

In 17th century Sweden, the Mosaic laws regarding male on male sexual encounters were enforced to the letter.233 Homosexuality, which was known as sodomy, was a crime against nature and habitually punished by death. The harshness of the punishments had no doubt to do with the way homosexuality was explained. Devoted Christians believed that homosexuality was a sin; a wicked choice. Since homosexual acts were freely undertaken, it made perfect sense to hold homosexuals responsible. They had chosen to perform “unnatural” acts, the sinfulness of which merited very harsh punishments indeed (Rydström 2003).

After a period of relaxed policies against homosexuals during the 18th century, following in the wake of the Enlightenment, there was a shift towards stricter standards in Sweden during the 19th century. In 1865 “indecency” (“otukt”) was outlawed. Indecency was an umbrella concept for “unnatural” acts between persons and bestiality. Under the new legislation, the maximum penalty for indecency was two years of forced manual labour. The law represented not only society’s moral dis-

mark refers to such consequences as “pragmatic” rather than “logical” (Falkemark 1978).

233 These state that male homosexual acts are punishable by death. See Leviticus 1:12 and 20:13. For similar sentiments in the New Testament, see Romans 1:27.
approval of the behaviour, but emphasised that homosexual acts were *criminal*. This meant that they were the result of choice. They were also *harmful*; something which society needed to protect itself from (see Silverstolpe et al. 1999: 70).

The terms “homosexuality” and “homosexuals” did not become widely used in Sweden until the early 20th century (ibid: 246-66). The terms were of medical origin. Physicians and scientists tended to conceive homosexuality as an inherited trait, though some were believed to have acquired the disposition out of choice or habit. As one would expect, this “medicalization” or “biologization” of homosexuality meant that the policies against homosexuals changed. Homosexuals were regarded as people who suffered from a disease that for the most part was inborn and thus beyond the individual’s control. This meant that it became difficult to see how they could be responsible for their disease.

Homosexuality had thus gone from being seen as an unnatural and sinful choice to an inherited disease. A natural consequence was that homosexuality became decriminalized: as of 1944 it was no longer punishable to be a homosexual in Sweden.234 In the government proposal that preceded the decision to decriminalize, the biological paradigm of the times was clearly expressed. The following passage, which stems from a public report the 1944 proposal extensively quotes, is quite representative. It is hard to do the language of the times justice, but a rough translation is:

“The medical research of previous decades has shown, that the genuine homosexuality is caused by certain, probably inherited, deviations from the normal, and that because of exactly this, many homosexuals naturally seek to satisfy their urges and are driven towards homosexual acts without being deterred by threat of punishment. Since punishment therefore cannot - even if it were to be effectively implemented – suppress homosexuality, and thus cannot serve its purpose, its continued existence can only be motivated by the dangers, that homosexuality acts could be said to produce. From this point of view, there is in principle no warrant for making these acts punishable” (Prop. 1944: 13, p. 8. My translation).

---

234 It should be noted that the law against homosexuality applied only to men, as homosexuality among women was not believed to exist.
The proposal adds that homosexual acts between adults in general are socially harmless, and also emphasises the unfairness of punishing homosexuals, since punishment augments the homosexuals’ “severe personal suffering, which the mere awareness of their abnormality produces” (ibid.). It is interesting to note that homosexuality is here “naturalized” but by no means “normalized.” It is treated as a biological disposition that cannot reasonably be suppressed. But it is not normal – again and again words such as “deviant” (“avvikande från det normala”) and “abnormal” (“abnormitet”) are used. The homosexual blamelessly suffered from a lamentable condition or disease. Thus, a principal reason for decriminalization of homosexuality was that it was just:

“[…] the medical science’s discoveries about the nature and causes of homosexuality lead to that the punishment of the constitutionally homosexual must be seen as a harshness, and it is incompatible with the demands of humanity and justice that they be punished for the expressions of a natural drive, the direction of which they cannot be held to account.” (Prop 1944: 13, p. 15)

However, this did not mean that homosexuality was seen as something positive. Fears that it might be “contagious” (“smittorisk”) if it was increasingly practised in public were raised. A government committee, Strafflagsberedningen, suggested that all homosexual criminals – for there were still criminal offences, such as when an adult committed homosexual acts on a boy younger than 21 - be forced to undergo a mandatory medical examination. Thereby the homosexual would “get good advice regarding the treatment of his sexual abnormality, something that in this committee’s mind could have a significant impact on the elimination of recidivism” (ibid. 22).

The reasoning above suggests a thoroughgoing “medicalization” of homosexuality, or a move from “badness” to “sickness” (cf. Conrad & Schneider 1992: ch. 7). The authors of the proposal often quote that medical science has proved the heredity of homosexuality. Central for us is

---

235 It appears to have been an accepted belief that homosexuality came in degrees. Some were believed to be homosexual by nature; hence the term “constitutionally”. Others were seen as bisexuels. A significant subgroup of these was those who engaged in homosexual contacts when deprived the company of women (paradigmatically sailors and inmates).
that this belief is then seen as precluding attributions of responsibility. The explanation served as an excuse. The process was not nearly as neat as I here suggest, however. The badness-paradigm didn’t become replaced by a complete disease-paradigm over night. Rather, as noted by Eman,

“The breakthrough for this “medical” term [i.e. homosexuality] of course also meant the breakthrough for its implicit theoretical assumption: that homosexuality was inherited and therefore should be de-criminalized. However, it should be emphasised that this paradigm change did not happen automatically. Homosexuality could in one and the same article be described as a result of an individual’s nature (genetically determined), while at the same time the actions were described as immoral and hence objectionable.” (Eman in Silverstolpe et. al. 1999: 164. My translation)

So, there seemed to have been conceptual confusion going on. Note, however, that Brown’s contention about the independence of explanations and attitudes to homosexuality is fully supported by the Swedish case. A negative attitude to homosexuality was very much retained. Following the decriminalization of 1944 homosexuality was included on the official list over mental disorders, from where it wasn’t to be removed until 1979.

The Swedish case illustrates how explanations can be excuses. Following a strong tendency to see homosexuality as hereditary, legislators decided that since homosexuals are not responsible for their disposition it would be unfair to punish them. This was decided in spite of that the general attitude to homosexuality was overwhelmingly negative. Again, the theory of excuses sketched here does not preclude negative attitudes, as Brown’s argument show, only reactive ones.

236 One way of making coherent the combination of believing in a biologically fixed homosexuality and viewing homosexuality as immoral is to stress, as the Catholic Church stresses, the difference between a disposition and an action. So one could believe that while homosexuality is an inherited disposition it is also a disposition that one does not have to act on. It is difficult to see how agent and actions could be separated like this (do we not act as we do as a consequence of who we are?), but this at least is a logical possibility. But it is also one that makes the theory of excuses vastly more complicated, and I shall not dwell upon it here.
What is the present view of homosexuality? It is interesting to note that little efforts seem to be made to explain homosexuality. The sexual orientation is increasingly regarded as normal or unproblematic. This illustrates a final important aspect of the explanations/excuses-discussion. It is only as long as a behaviour is regarded as prima facie blameworthy that excuses become relevant. A biological explanation of homosexuality, were this to be pursued, would still explain in a way that exclude responsibility, but it is only as long as homosexuality is seen as morally problematic that the need to excuse emerges.

7.5 VARIETIES OF EXPLANATIONS: WHY DIFFERENT EXPLANATIONS ARE EQUALLY EXCUSING

One could argue that the case just mentioned is way too glib. Small wonder, a critic could object, that explanations are excusing if we consider biological explanations. It is not difficult to accept that explanations are excusing in theory, and perhaps also excusing in practice, if we consider physical explanations such as neurological, genetic, or biological ones. The notion that homosexuals are not responsible for their disposition if it is inscribed in their genes or the architecture of their brain is trivially true. But, the critic could continue, not all types of explanations are biological or physical. There are other kinds of explanations, and many of them seem less excusing. In fact, Brown mentions such explanation: the idea that homosexuality is constructed.

There is some truth to this claim, although I in general find it to be overstated. Some explanations do not threaten responsibility. The clearest example is explanations that include freedom of the kind required for responsibility among their premises. Brown’s “constructivism”, with

---

237 Conrad and Schneider, for instance, in *Deviance and Medicalization* argue that unlike the “positivist” approaches to deviance, their “interactionist” approach assumes that “the behaviours called deviant are by and large voluntary and that people exercise some degree of ‘free will’ in their lives” (1992: 2) Clearly, such an assumption rules out that “interactionist” explanations threaten responsibility. The authors have already assumed that the explained behaviour is “freely” un-
its emphasis on homosexuality being a “free choice”, might be such an explanation. However, social scientists tend to exaggerate the special nature of biological explanations.\textsuperscript{238} Many non-biological explanations, if true, appear to be excusing too. Just as biological ones, they tend to explain by citing factors beyond the individual’s control; and just as biological ones, they render the explanandum expectable.\textsuperscript{239}

I have no idea about how to validate this claim in a general way. Instead, I here intend to briefly analyze particular explanations of prima facie blameworthy behaviour of different varieties. I will try to show why these explanations, even so-called agential ones, have an excusing logic built into them. We shall briefly look at good examples of types of explanations that seem very different from their biological counterparts. First, as an example of a \textit{cultural} explanation, we will analyze Jonah Goldhagen’s well-known explanation of the Holocaust. (This explanation, it seems to me, could just as well be called constructivist). Then we will look at Philipson and Posner’s rational choice-explanation of the AIDS-epidemic. I will argue that both explanations, if true, undermine responsibility and are excusing; yet none of them need to include any references to biology (or physical causes, broadly conceived).

\textit{A cultural explanation: Goldhagen’s explanation of the Holocaust.}

Goldhagen’s aim in \textit{Hitler’s Willing Executioners} is to explain “why the Holocaust occurred, how it could occur” (1996: 5). He notes that previous explanations have focused too much on the operations of the ruling elite and too little on the actual individuals (mainly Germans) who carried out the Holocaust. Goldhagen rightly notes that unless so many people took part in the genocide, it is inconceivable that the Holocaust would have dertaken. Such a view has been referred to as “intentionalist” or “voluntarist” by Hay 2002.

\textsuperscript{238} Even stranger is the sometimes related idea that a biological explanations threaten certain political views especially much, i.e. leftist ones (see Singer 1999; Pinker 2002).

\textsuperscript{239} A case for the equally excusing implications of biological and environmental explanations is nicely made by Richards 2000.
happened – at least to such an abysmal extent. He therefore suggests that,

“It is not sufficient to treat the institution of killings collectively or singly as internally uncomplicated instruments of the Nazi leadership’s will, as well-lubricated machines that the regime activated, as if by the flick of a switch, to do its bidding, whatever it might have been. The study of the men and women who [...] peopled the institutions of genocidal killing must be set at the focus of scholarship on the Holocaust and become as central to investigations of the genocide as they were to its commission” (ibid. 6).

Nor will it do, when one focuses on the individuals that carried out the genocide, to treat them as people who were myopic, coerced, or simply obeying orders. The fact of the matter, Goldhagen argues, is that they (as the title of the book suggests) willingly took part in the Holocaust. They were not people who were unable to say “no”, nor unaware of what they were doing. Rather they did not want to say “no”. They chose to say “yes” to the Holocaust (ibid. 381).

So far this looks pretty much like what social scientists call an agential explanation, or an actor-centred explanation. Goldhagen seems to suggest that the *explanandum* was the collective result of many voluntary individual actions. But the next question is of course *why* people voluntarily wanted to take part in the genocide. What was their motivation and where did it come from? This is where Goldhagen’s explanation stops looking like an agential one.

Goldhagen suggests that the only factor capable of explaining why the Holocaust happened is what he calls the “eliminationist antisemitism” of German culture. Germany, he contends, was steeped in a particularly virulent hatred of the Jews. This hatred was the necessary and sufficient cause of the Holocaust - it could not have happened unless such an antisemitism was in force, and the antisemitism was all it took for the events to happen. Towards the end of the book, Goldhagen suggests that this fundamental insight can account for both why ordinary Germans willingly partook in the killings, and why the Nazi regime could claim power. Both events are explained by the demonization of the Jew in German culture:

“The virulent, racial antisemitism, in motivating Germans to push the eliminationist program forward on the macro, meso, and micro levels,
must be understood to have been motivating people who were operating under constraints, both external ones and those created by competing goals. This was true of Hitler and the lowliest guard in a “work” camp. These ameliorating circumstances notwithstanding, the eliminationist antisemitism was powerful enough to have set Hitler and the German nation on an exterminationist course, powerful enough to have overridden economic rationality so thoroughly, powerful enough to have produced in so many people such individual voluntary, zeal, and cruelty. The eliminationist antisemitism […] resided ultimately in the heart of German political culture, in German society itself” (ibid. 428).

This, then, is clearly a cultural explanation. Individual Germans took part in the genocide because they believed it to be justified, indeed necessary. But this belief was a result of elements in the German culture.

Now, to what extent is this explanation, if correct, an excuse? This question boils down to whether the Germans could be held responsible for their cultural beliefs. Goldhagen suggests that the process of enculturation is very powerful, so powerful that it was in some sense natural that ordinary Germans did what they did. In reality, they could not have done otherwise. But were they responsible for the content of their culture?

Here Goldhagen seems to say that the culture indeed was a factor beyond the individual’s control. The origin of the hateful ideas that caused the Holocaust must be sought earlier in time. The eliminationist antisemitism of German culture was not something which was collectively adopted at an instant. Goldhagen notes that already in 19th century, Germany was pervaded by a racially motivated antisemitism (which in turn had even older religious origins). By the end of the century, there was a stable discourse which depicted the Jews as a symbol of “decomposition, malignancy, and wilful evil” (p. 80). These ideas were prevalent in German culture long before World War One, the Weimar republic and the subsequent onset of Nazism. Buck-stopping responsibility, if it is to be found at all, cannot be placed among the enculturated Germans who, as it were, simply happened to be born into a society where the historical ascent of antisemitism was to culminate.

---

Thus, if Goldhagen is right in that the ideas of German culture were such that they determined the beliefs of ordinary Germans, and that the origins of that culture in addition must be sought earlier in time, there is a clear sense in which Goldhagen’s explanation actually works as an excuse.\textsuperscript{241}

\hspace{1cm}A rational choice explanation: Philipson and Posner’s explanation of the AIDS-epidemic

Perhaps the critical test for the theory of excuses sketched here is the now widespread rational choice explanation. Rational choice theory (RCT) seems to be least capable among explanations to work as an excuse for blameworthy behaviour. RCT holds that various outcomes should be explained by the action of rational individuals who seek to further their expected utility. Game theory is the study of individual rational action in situations of interaction – RCT with more than one player, one could say.

Why should we take rational choice theoretic or game theoretic explanations to be less excusing than other explanations? The simple answer is that they are highly actor-centred or agential. They explain by citing the purposeful, conscious and strategic action of individuals. This seems far from reducing people to cultural dopes, discourse pipes, or other structural epiphenomena.

There is a massive misunderstanding here, however, which has to do with the infamous agent-structure problem. The agent-structure problem concerns how to reconcile the notions of purposeful human action (agency) with the shaping and constraining forces of structures such as power distributions, culture, socio-economic circumstances, institutions, etc. When we seek to explain, say, why working class people tend to vote for parties to the political left, are we to explain this in terms of the social

\textsuperscript{241} Goldhagen himself is aware of this. Although he is adamant about the voluntarist or “chosen” aspect of the Holocaust, and of course writes in understandable disgust, one can certainly question whether he believes that the ordinary German, given his or her beliefs, could have but said “yes” to the killings. Goldhagen begins his book by quoting Tocqueville: “No man can struggle with advantage against the spirit of his age and country.”
cleavages that gave birth to the left-right dimension and the structural interests of the working class as a collective, or are we to reduce the outcome to just a long line of individually rational choices? The agent-structure problem is animated by the difficulties involved in such questions. On the one hand, the interests of the individual seem to have structural origin. But on the other hand, the structural or institutional setting of social life seems ultimately created and upheld by individual action.

The agent-structure problem plays a big part in social science, and one which should be embraced. But social scientists have an annoying tendency to treat agential explanations as different from structural explanations when it comes to what interests us here: their tendency to excuse. Put simply, social scientists tend to think that agential explanations are compatible with responsibility whereas structural ones are not. This rests on a massive misunderstanding. Agential explanations can be just as excusing as structural ones. In order to show why, let’s as an example look at Philipson and Posner’s rational choice treatment of the AIDS-epidemic.

In Private Choices and Public Health (1993), Philipson and Posner set out to analyze the AIDS-epidemic from an economic perspective rather than the usual epidemiological or ethical ones. Their intention is to analyze sexual behaviour as a species of rational choice, and to evaluate various anti-AIDS policies accordingly. Central to their approach is that sexual behaviour is to be thought of and explained as market-behaviour. Philipson and Posner’s central idea is that,

“The AIDS epidemic is primarily, although not exclusively, the consequence of voluntary decisions by persons of different HIV infection status to have unshielded sexual intercourse with each other. […] Risky sexual trades […] are analytically similar to conventional market transactions that take place under uncertainty concerning the reliability of the transactors or the quality of the good or service being sold, as in credit markets or the sale of potentially defective goods” (p. 31).

242 A clear example of this fallacy is Miller et al. 1999, where it is argued that “dispositional” explanations (i.e. explanations that explain action by citing dispositions or characteristics of individuals) are compatible with responsibility whereas “situational” are not (situational explanations explain action by citing the context of the action).
Philipson and Posner assume all sexual acts to be trades. As in other areas, rational actors will engage in sexual trades with the intention to maximize expected utility. To use a condom here counts as a significant disutility, because it reduces the pleasure involved. This disutility is weighed against the risk of not practising safe sex – the probability and cost of catching a disease, for instance. Given this risk, and given that they prefer not to use a condom, rational actors might be willing to take a chance – they do not know whether their would-be partner is infected with HIV, but as long as the risk is perceived as sufficiently small, they will take a chance and have unprotected sex. Philipson and Posner hold, as indeed most RC theorists would, that there is some point where it is rational to engage in risky sexual transactions (i.e., where the expected utility outweighs the expected disutility in the rational calculus).

AIDS thus ultimately spreads as a consequence of rational sexual behaviour.243 A crucial factor for the epidemic in the U.S., Philipson and Posner argue, was that public authorities supplied rational agents with reliable HIV tests. For since you can only get infected once, there is nothing that stops infected individuals from having unprotected sex. They have already suffered the cost, and stand in this respect only to gain from engaging in further unprotected sexual trades.244

Now, to what extent could you say that Philipson and Posner’s individuals are without responsibility? They certainly appear controlled, conscious, and active. They are simply doing what they want to do – Philipson and Posner are adamant that the AIDS epidemic is ultimately to be explained by people’s voluntary decisions. This seems to mean that they are responsible for their actions. However, as we noted in the previ-

243 However, the authors note that the early and rapid spread of AIDS can’t be explained by citing risky but rational choices, for in the beginning the disease was unknown. But continuing infections can in large part be explained by the notion of risky sexual trades, Philipson and Posner argue. For after a couple of years the disease was well-known, but people took deliberate risks anyway (1993: 114).

244 Although the authors relax the initial assumption in the model that infecting others is not a direct cost to an agent, they also note that “Persons who are or suspect that they may be HIV positive will often have an incentive to conceal the information from their sexual partners in orders to induce the latter to engage in risky sex” (1993: 114).
ous chapter, there is a perfectly clear sense in which RCT also has a tendency towards determinism, which erodes responsibility. It might be worth to go over why rational choice explanations can be deterministic one more time, as the point is quite technical. Take again the central notion of maximizing expected utility among alternatives in the feasible set. It has been proven that if preferences are transitive, complete and reflexive, there must be at least one element in the maximal set (see McCarty & Meirowitz 2007: ch. 2). This is a fancy way of saying that for any given ranking of alternatives, there must be at least one highest ranked alternative or outcome.) Rationality is to choose that element. According to some, the implication of this is that once one comes to learn the beliefs, desires and risk proneness of a rational individual, it should be possible to fully predict his or her actions (Hay 2002; Hollis 2002). If rational agents always choose the highest ranked alternative, then there is a sense in which they could not have chosen otherwise. This reflects the tendency to determinism. In Hay’s words: “While actors are free too choose, they will always choose the optimal strategy; consequently, their behaviour is entirely predictable” (Hay 2002: 53). He goes on to claim: “what sense does it make to speak of a rational actor’s choice in a context which is supposed to provide only one rational option?” (ibid. 104)

So, how could one use Philipson and Posner’s reasoning to absolve individuals from responsibility? Well, if the economic explanation of the spread of the AIDS epidemic is correct, then the spread is ultimately to be explained by individual rationality. Given people’s beliefs and desires – in particular their preference for unprotected sex – they did the rational thing by engaging in risky sexual trades. One could question individuals’ responsibility for the origin of their beliefs and desires, of course, just as we did in the case of Goldhagen. This would open up for the tendency

245 Note that while \( x \) might be higher ranked than \( z \), it is possible that \( x \) and \( y \) are ranked equally high (i.e. \( xPz \) and \( yPz \), but \( xLy \) where \( x, y, z \) are the elements in the feasible set). As I said in chapter 6, the maximal set then contains two alternatives and a rational way to decide between them is to flip a coin.

246 The preferences are often treated as “exogenous” in rational choice theoretic and game theoretic work. This means that their origin are not considered (or modelled) – the preferences are simply taken as a given. Recently, however, a particular interest has been taken in evolutionary approaches to rational choice,
towards regress. But above all, we can question whether it is reasonable to hold individuals accountable for being rational; a trait, one presumes, which we cannot choose to be without (or, as the present case suggests, a trait we are doomed to possess).

7.6 DOES IT MATTER IN PRACTICE? P.F. STRAWSON ON THE PSYCHOLOGICAL IMPOSSIBILITY OF THE OBJECTIVE ATTITUDE

Assuming that explanations are often excusing if they are true, and assuming that this holds for different varieties of explanations and not just biological ones, it is another question whether explanations are excusing in practice. Besides the case of the decriminalization of homosexuality, we previously just postulated that explanations would have such an effect. But one could perfectly well argue that while there is an excusing logic built into the theory of scientific explanation, this matters little for social and political practice unless people who believe that a prima facie blameworthy behaviour can be explained will be more willing to excuse as a result. Perhaps we are prone to blame and hold each other responsible despite the fact that we believe actions can be explained scientifically. One who has defended such a view is P.F. Strawson.

I earlier referred to the notion of reactive attitudes - the feelings of love, gratitude, and resentment that a scientific conception of action may undermine. Reactive attitudes occupy the centre stage in P.F. Strawson’s seminal article “Freedom and Resentment.”247 In this piece, Strawson’s makes an important distinction between “participant” and “objective” attitudes. Participant attitudes are the attitudes we have when we regard each other as responsible moral agents and hold each other to normative standards of behaviour. The objective attitude is more of an elevated and detached attitude, which precludes blame and resentment, or holding

---

232

where evolutionary psychology is used to explain why people have the preferences they have. For a critical overview, see Dupré 2001.

247 “Freedom and Resentment” was first published in 1963. I will refer to the reprinting in Watson 2003.
responsible in any appropriate sense of the word. The central claim of the theory of excuses offered so far in this chapter can perhaps best be characterized as the claim that *a scientific understanding of crime will lead to an objective attitude to crime*, centred on understanding rather than blame. But Strawson objects to this claim. In his mind, we are hardwired to have participant reactive attitudes most of the time; hardwired, that is, to regard each other as responsible agents. We could say that if Strawson is right, the notion of us becoming more willing to excuse in the light of scientific advances is simply psychologically unrealistic.

**Strawson’s theory**

According to Strawson, once we realize the fundamentally inter-personal nature of responsibility and its associated practices, we can avoid the cardinal mistake of most treatments of responsibility: the tendency to over-intellectualize the topic. Contrary to other treatments, which have analyzed the practice of holding responsible only after having first established an objective set of philosophical criteria of being responsible, we should start with the actual practice of holding each other responsible, as it appears in everyday life. Strawson suggest that what we thereby see is the importance of reactive attitudes to practices of holding responsible. To blame someone does not consist in coolly calculating whether he or she fits a more or less abstract set of criteria of when it is either just or prudent to do so. It rather consists in feeling resentful towards a person whom we believe has not granted us (or somebody else) sufficient good will or consideration.

The key aspect of Strawson’s theory is the notion of a *participant* perspective on praise and blame. When somebody close to us forgets our birthday, for instance, we tend to resent them for it. Arguments to the effect that given the state of the universe at time \( t \) and the laws of nature, it was unavoidable that he or she forgot my birthday will not impress me. That is because we are involved in a “participant” relationship where such considerations do not apply.\(^{248}\)

\(^{248}\) A recent, even stronger statement of this kind is Habermas 2007. Habermas argues that there is an “epistemic dualism” between a participant perspective.
Strawson believes that the participant perspective, and the reactive attitudes that accompany it, are natural conditions of man: “What I have called the participant reactive attitudes are essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in their attitudes and reactions” (2003: 80). Hurt feelings, resentment, gratitude, love – these are all natural aspects of being a participant member in human relationships. Those feelings are not dispensed (or not dispensed) from such a detached perspective as the literature on responsibility suggest, Strawson argues. Nor can we rationally justify having them, at least not fully.

There is, then, a strong presumption in favour of holding responsible when somebody fails to meet standards of acceptable conduct, according to Strawson. It is a natural reaction among members of our species. But there are a couple of instructive exceptions. First of all we have what I will call the “locally” non-responsible. They are the fully normal agents who are excused for doing something wrong. When we for instance claim that someone “didn’t mean to do it”, or that “it was an accident”, or that “there was no other option”, we are still in the realm of ordinary inter-personal attitudes. Never do we, Strawson notes, suspend our view of them as normal, participant members of human relationships. The second group of exceptions from the sphere of responsibility, however, is what we may call the “globally non-responsible”. Here we have the cases where the agent is, “warped or deranged, neurotic or just

(which he, following the late Wittgenstein, refers to as “the language game of responsible agency”) and an objective, naturalistic one. Habermas’s point seems to be that we cannot account for human agency by subsuming it under an ontologically monistic natural scientific outlook, such as the kind of eliminative materialism of the mind advocated by the aforementioned Wolf Singer. This is because even scientific discourse about human action presupposes a particular point of view - that of the situated and bounded human mind. A science of human action, then, is the human mind trying to learn about its own operations - a situation which ensures that attempts to eliminate the participant aspect of existence is something similar to an impossible act of self-mutilation (and a fully “objective” account could not tell us anything about what it is like to be such-and-such an entity, Nagel 1989).

249 I use this term instead of “irresponsible” since this connotes a character trait or behavioural disposition rather than what is here intended: absence of deserved praise or blame.
a child” (ibid. 79). These are cases where the agent does not qualify as a normal participating member. Contrary to what holds in the case of the locally non-responsible, we do not react to the globally non-responsible as equals. Our reactive attitudes are strongly modified: we leave the participant perspective and adopt what Strawson calls the objective attitude. Facing the abnormal wrongdoer we tend to switch from resentment and anger to more sober or detached feelings:

“To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject of for what, in a wide range or sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained […] The objective attitude may be emotionally toned in many ways, but not in all ways, it may include repulsion and fear, it may include pity or even love, though not all kinds of love. But it cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in interpersonal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love two adults can sometimes be said to feel reciprocally, for each other” (ibid. 79).

It is important to note that Strawson does not see the participant and the objective attitude as mutually exclusive. Sometimes the two attitudes are blended, such as in the case of the maturing child towards whom we react with an increasingly participant attitude. More importantly, nothing stops us from adopting the objective attitude to normal behaviour and normal agents as well, so that our assessment of them become as detached as our assessments of the behaviour of animals, infants, or the mentally disordered. But Strawson thinks that we cannot do so for long: it would be “practically inconceivable” (though not self-contradictory) to maintain such a view: “A sustained objectivity of inter-personal attitude, 250 It should be noted that Strawson partitions the field in a slightly different way. He argues that there are two kinds of cases where we suspend reactive attitudes: those who are rather permanently non-responsible, and those who are only temporally so (e.g. as when we say about someone that “he/she wasn’t herself”). Since Strawson suggests that it is the first group – the group I call the “globally non-responsible” – that is most interesting, I choose to drop the second group from the analysis here.
and the human isolation which that would entail, does not seem to be something of which human beings would be capable, even if some general truth were a theoretical ground for it” (ibid. 81).

So, while we may oscillate between the participant and objective perspective, Strawson believes that human affairs will always find their equilibrium in the former. This enables him to deliver his main message: we should regard holding each other responsible as a natural and unavoidable human condition. It is inconceivable that the notion of personal responsibility, and the practices of praise and blame that go with it, could vanish as a result of the objective attitude’s gaze. The reactive practices, among which we count punishment, are a given in our societies. They might be modified but not abolished. The notion that an increased scientific ability to explain will lead to the decline of responsibility is psychologically implausible.

Analysis of Strawson’s theory

Strawson’s argument makes a lot of sense. It puts the finger – perhaps even too starkly – on the fact that our attitudes towards people vary greatly depending on how we judge their capacity for normal participation in human relationships. The existence of an “objective attitude” is a good way to explain the fact that a child, an animal, or a mentally disordered isn’t treated in the same way as a normal adult. The most profound message of “Freedom and Resentment” is also worth taking to heart: we should not fear advances in the science of action, for we will never have the stamina to think of, as the objective attitude prescribes, our friends and families, our enemies and colleagues as mere vector sums of the factors that shape their behaviour. It seems likely that we will always resent

251 This oscillation is reminiscent of Nagel’s The View From Nowhere, where it is argued that we can oscillate between (something like) the objective and participant attitude towards ourselves (Nagel 1989).

252 Can we not feel gratitude towards a trusty dog, for instance, without saying that it has the capacity for responsible agency as we know it?
our partners a little if they forget our birthdays. If Strawson is correct, all of the worries expressed about the decline of personal responsibility may simply be put to rest. If we are hard-wired to regard others (and ourselves) as responsible, then the main concern of this book – defending the notions of personal responsibility and desert – is superfluous.

But Strawson’s argument is flawed in important respects, mainly because it is overly confident about the resilience of the reactive attitudes. Here are some reasons why Strawson seems too optimistic:

First of all, it is not obvious that the sphere of “normal participant attitudes” will remain as large as it is today, even if some portion of it will always exist (say, between members of a ruling elite). As we saw in chapter 1, the sphere of the “globally” responsible has certainly become smaller over time. We shudder today at the thought of punishing animals or ten-year-olds. But it is possible that we will regard the punishment of those with rotten social backgrounds, or those who have been dealt a poor biological hand, as just as unfair a hundred years from now. Strawson simply overestimates the stability of our reactive attitudes.

Secondly and closely related, Strawson tacitly seems to assume that the categories “normal” and “abnormal” are furnished independently of scientific progress. He writes as if there are people towards whom we tend to apply the participant attitude and people towards whom we tend to apply the objective attitude; and while our subjective standpoint might be oscillating between a participating and objective gaze, the categories in themselves are fixed. But this could be false. It is possible that we abandon the reactive attitudes held towards an agent when we learn about his or her rotten social background, biological abnormalities, and the like. Scientific progress might simply be the thing that serves to usher in the era of the objective attitude; the thing which proves to us that many of the people we previously thought were fit to be held responsible in fact were not. In Berlin’s words, it might be:

“That the more we find out about a prima facie morally culpable act, the more we are likely to realise that the agent, given the particular cir-

---

253 And if we do excuse them, we shall do so mainly by granting them “local” non-responsibility, not global. We shall excuse them as equals that made an honest mistake, not as inferiors who are incapable of making commitments or planning their future.
Let’s say a shamed Japanese mother tries to take the life of both herself and her children.\textsuperscript{254} A language of responsibility and blame assumes that she could have refrained from doing what she did, and that she thus freely chose to act as she did. Probing deeper into the action, however, we realize that she did what she following her cultural beliefs thought was necessary to save face. This serves to make the action more explicable. Our initial reaction of blame may be eroded by this deeper understanding of the situation: a thorough grasp of the mind-state of the women, her cultural beliefs, and the context of the action may make us inclined to say that it is understandable that she did what she did, perhaps even that it was \textit{unavoidable} that she did what she did. Resentment then seems out of place.

The question is, then, whether Strawson’s empirical claim that we are hardwired to treat most of our fellow human beings as participant members is correct. Do people become more willing to excuse behaviour when they believe it to be scientifically explainable? I have argued that they do. The fuller our understanding, the less inclined we will be to blame, I have suggested. Studies support this to some degree, but by no means unequivocally. Iyengar (1988) has shown that when people are given a fuller causal story of “bad behaviour” such as crime or terrorism, attribution of responsibility tends to become less individualistic and more structural or collective. For instance, when crime is reported against the backdrop of the problem of poverty, respondents become less willing to blame individual criminals for their behaviour and more willing to attribute responsibility to the structure of society. Furthermore, experimental studies of American jurors have indicated that scientific testimony can affect attribution of responsibility. In one test, having a psychiatrist witness in favour of the defendant reduced the penalty meted out by the jurors’ from thirty years imprisonment to five; something which may indicate that “a juror’s desire for explanation often

\textsuperscript{254} This example is discussed in Tunick 2004.
overrides the citizen’s desire for punishment” (Wilson 1997: 90). Miller et. al. (1999) have found, however, that while explanations often affect the way we perceive the one providing the explanation, they do not make us less inclined to blame or hold responsible the alleged wrongdoer. Other studies have shown that explanations reduce the willingness to retaliate, but the results are not conclusive.

The literature is to my knowledge too sparse to say whether scientific explanations, when they are believed, lead to lessened reactive attitudes. There is a third reason to doubt Strawson’s argument, however. Strawson commits a common mistake of the moral philosopher. He seems to think that what holds, and is justifiable, on the level of the individual is exactly what holds, and is justifiable, on a societal level. As an individual it might be impossible for me to maintain an objective attitude to the person that, say, assaults me. If somebody intentionally punches me in the face, I will tend to resent him or her for it. But as a member of an expert committee debating criminal policy, it might be perfectly pos-
sible to adopt something like the objective attitude. Indeed, there are institutional settings where individuals are encouraged or required to do so. The power to shape policy is placed in such settings, not in an individual who feels hatred towards the assailant or pity with the victim. In short: as everyday people we may be hardwired to react to people in certain ways. But that doesn’t mean that policy making must be a reflection of these ways. Private morality and public policy might not adhere to the same mechanisms. And it is the prospect of public policy becoming excusing which is here in focus.

In sum, Strawson argues that science cannot usher in the era of the objective attitude since we are hardwired to treat each other in accordance with the participant perspective. We can take the objective attitude to all persons and all action, and not just the deranged, but the prospect of doing so for long is “practically inconceivable”, he claims. I have suggested that our sphere of normal participating attitudes seems to have changed in various ways throughout history, and that scientific explanations might be just the input that accounts for such changes. And even if we in our everyday life cannot sustain an objective attitude (assuming this is what we want) there is nothing that precludes policymakers on crime to make decisions from an objective standpoint. I for one find nothing inconceivable about a penal regime that tries to eradicate reactive attitudes – in fact, Sweden in the middle of the 20th century was on the verge of adopting such a regime (SOU 1956:55). Strawson’s theory is very valuable, but it is flawed in important respects.

7.7 NORMAL AND ABNORMAL ACTION

A final objection to the notion of explanations being excuses I want to discuss is the objection that we interpret normal action but causally explain abnormal or disordered action.

As we saw in chapter 6, some argue that crime is indicative of some abnormality or disorder in the offender. Recall that on the Disorder Model, which is foundational to many rehabilitationalist views, the offender is also a victim. He or she suffers from some crime-causing condition that rehabilitative programs should address. Following the notion that criminal actions are performed by disordered people, one might argue that
while disordered action is explained in ways which erode responsibility, “normal” or “sane” action is not. The idea here is that in explaining disordered action we give fully causal explanations, whereas we explain normal or sane action in a different way. We might say of a kleptomaniac that his or her disorder causes him or her to steal. His or her consciousness is at best an intermediate variable in the explanation – he or she does not in any normal sense of the words choose to steal. The disorder is causally sufficient to produce the behaviour. A “normal” thief - and I am now assuming that there are crimes that are not indicative of disorder - by comparison may be said to freely choose to steal. We attach greater significance to the deliberation of this thief – we often say that his or her act of rule breaking warrants an intentional explanation, or a reason explanation.259 These boil down to interpreting action rather than causally explaining them, some argue (see Rosenberg 1995).

The upshot of this distinction is that it is only causal explanations that have a tendency to be excusing, and that it is only disordered action that we explain in a causal way. The bulk of criminals may be assumed to be non-disordered, however, so it may seem to follow that most crimes warrant intentional explanations, which are not excusing. Put shortly: we explain and excuse disordered rule breakers but understand and blame normal ones.

This position assumes that there is a salient difference between the way we explain normal and abnormal action. There is a methodological counterpart to this assumption, called the A-rationality principle. According to this principle, we should only look for fully causal explanations when reasons are lacking. Where an action or belief is backed by good arguments, evidence or rational reasons, we should let these be sufficient to explain the action or belief. Where no such arguments, evidence or reasons exists, however, we are entitled to look for “mere” causes of the behaviour.260 For example, if someone steals money in order to live comfortably without having to work, he or she steals out of reasoning that

259 I give a fuller treatment of reason explanations, and the way they relate to causal explanations, in the next chapter.

260 It is here assumed that reasons are not causes. I relax this assumption in the next chapter. Many believe that reasons explanations are a subspecies of causal explanation; see especially Davidson 1963. For a statement of the A-rationality principle, see Brown 2001.
seems rational and normal. We take that reasoning at face value, and blame the thief accordingly. If however someone steals in order to finance a global underground resistance to an imminent alien invasion of Earth, the reasoning seems bizarre. It is not something we take at face value. We feel inclined to ask what mental disorders or brainwashing have resulted in that belief. The belief must be explained by causal factors such as the presence of psychosis.

To what extent do these considerations count as an objection to the theory of excuses? As we have already seen in the rational choice example above, there is a sense in which even highly agent-centred explanations may be excusing. They too undermine the notion of alternative possibilities and may cite factors beyond the agent’s control. Moreover, the notion of sane individuals being fundamentally different from disordered individuals can be put in doubt. Consider again Wolf Singer’s idea that criminal behaviour is the result of brain abnormalities. Rule breakers are people whose brains do not function normally: the architecture of their brains is neurobiologically malfunctioning. We may question the notion that rule breaking is the result of brain damage. But underlying Singer’s ideas is the notion that all action, whether criminal or lawful, normal or abnormal, emanates from the brain. This implies a strong sense of symmetry between normal and abnormal action. By symmetry I here intend that we do not or should not use different explanatory tools to explain normal and abnormal action. If, following Singer, there is a neurobiological cause for being abnormal, there should be a neurobiological cause for being normal as well.

The assumption of symmetry seems reasonable, and we do not have to be convinced by Singer’s neurological views in order to see this. Suppose that we explain an abnormal action, $A$, with the conjunction of the causes $X$, $Y$ and $Z$. These generic factors can be of any kind so long as they cause the action – people who suffer from $X$, $Y$ and $Z$ commit $A$. How do we account for those people who do not commit $A$? If $X$, $Y$ and $Z$ are sufficient for $A$, we must minimally state that not-$A$ is explained by the absence of $X$ or $Y$ or $Z$. Thus, not-$X$ or not-$Y$ or not-$Z$ must hold if not-$A$ is to be explained. This boils down to saying that if abnormality has causes, normality must have causes too – at least the absence of the abnormality-producing causes.

Put plainly, if a rotten social background, abnormal species functioning, mental disorder, or whatever explains offending, then decent social background, normal species functioning, or mental sanity should
explain non-offending. There is no reason to assume that normal behaviour is any less causally explainable than abnormal behaviour. It follows that normal action may be explained in ways that threaten responsibility just as abnormal action.

7.8 CONCLUSIONS

In this chapter, I have sketched the outlines of a theory of excuses, the essence of which is that scientific explanations erode responsibility. It is this basic idea which founds the idea that retributivism is “unscientific.” The notion that rule breakers deserve punishment rests on a “badness” paradigm which may not be tenable under a scientific worldview. The more we can explain a prima facie blameworthy behaviour - the more we come to understand it - the less appropriate it will seem to blame it. For scientific explanations explain by citing causes beyond the individual’s control and render the notion of alternative possibilities suspect. This means that scientific explanations are at odds with the idea of personal responsibility.

Or so I have argued. I want to emphasise that I am supremely unsure about to what extent the ideas expressed in this chapter are correct. I do not know whether the claim that nearly all scientific explanations are excusing if they are true is tenable. The purpose here, however, has been to challenge the received view that some explanations may excuse whereas others do not, for this is the basis of a potentially very powerful criticism of retributivism. When people say that retributivism rests on an outdated conception of human behaviour, it is presumably something along the lines of the theory of excuses they have in mind. That “ordinary morality”, which states that people are fundamentally free too choose in various ways, is mistaken. That the scientific project will clash with entrenched and arcane moral concepts.

The next chapter assumes, for the sake of argument, that all criminal action - indeed, all human action - can be scientifically explained. What would that mean for the practice of punishment and other moral and political issues? The next chapter thus deals with the thesis of determinism.
8. The Thesis of Hard Determinism

“But since we are all more or less imperfect, will there be anyone left to be responsible after we have excused all those with good excuses?”
(Dennett 1984: 157)

A prominent critique of retributivism, as we have seen, is to question the concept of desert. Desert, we said, requires responsibility – in order for someone to deserve something, he or she must in a requisite sense be responsible for it. So if one questions the responsibility for an outcome, one also questions desert: a criminal can scarcely be said to deserve punishment unless he or she is personally responsible for his or her crime. And if explanations, given some widely held beliefs about responsibility, explain in ways that undermine responsibility, clearly the explanations are at odds with the notion that people deserve such-and-such treatment.

The implications of explanations being excuses are easy to see. It is plausible to assume that our ability to scientifically explain criminal behaviour will increase in the future. This means, if the theory of excuses offered is correct, that more and more criminals will be excused for their actions. For a retributivist, this is as it should be: if offenders have valid excuses for what they did, they should not be punished. They do not deserve it. However, if ever more crime can be adequately explained by science, the notion of criminal personal responsibility starts to look less and less plausible altogether. It undermines the notion of freely chosen crime that retributivism, in my opinion, substantively relies on.

This might not be a problem if the only thing being eroded is the notion of freely chosen crime. If, as some maintain, crime in general is caused by abnormalities or disorders, it could be argued that criminal action is causally explainable while non-criminal action (of non-
disordered kind) is not causally explainable. But this seems implausible: it makes too much out of the questionable distinction between normal and abnormal behaviour. So under the assumption that normal behaviour is just as causally explainable as abnormal behaviour, an increased ability to explain criminal behaviour probably means an increased ability to explain non-criminal behaviour as well; indeed behaviour in general.  

Put shortly, if the notion of the autonomous criminal becomes increasingly undermined by the progress of science, the notion of autonomous man becomes increasingly undermined too.

There is much to object to this reasoning. But it is clear what danger that now looms on the horizon: the spectre of determinism and the associated problem of free will.

8.1 DETERMINISM

Varieties of determinism abound! Here are some of the versions of determinism I have encountered over the years:  

\[ \text{Causal (Fischer 1999)} \]
\[ \text{Environmental (Dennett 2004)} \]
\[ \text{Genetic (Dennett 2004)} \]
\[ \text{Historical (Berlin 2002)} \]
\[ \text{Human (Honderich 1990)} \]
\[ \text{Logical (Kane 1996)} \]
\[ \text{Material (Delgado & Stefancic 2001)} \]
\[ \text{Neurobiological (Searle 2001)} \]
\[ \text{Physical (Kane 1996)} \]
\[ \text{Psychological (Wolf 1980)} \]
\[ \text{Scientific (Kane 1996)} \]
\[ \text{Social (Berlin 2002)} \]

\[ \text{261 I expressed it thus: if factors } X, Y \text{ and } Z \text{ explain why someone commits a crime, there should be other factors (minimally not-}X, \text{ not-}Y \text{ or not-}Y \text{) that explain why someone does not commit a crime.} \]

\[ \text{262 I shamelessly borrow this stylistic move from Hacking 2001: 1.} \]
This plurality of prefixes can lead one to think that there are lots of different ways to understand determinism. There aren’t. Determinism is, for all its intricacies, a basically straightforward idea. Take first of all what it means for something to be determined to happen. We say that $X$ was determined to happen (at time $t$) if we believe that nothing but $X$ could have happened (at time $t$). Things that are determined to happen in this sense can be (truthfully) deterministically explained. We offer a deterministic explanation of an occurrence if we say that given the explanans nothing but the occurrence – the explanandum - could have occurred. In familiar $X$ and $Y$ terminology, we say that $X$ deterministically explains $Y$ if the probability of $Y$ given $X$ is 1 (i.e. in Bayesian terms: $Pr(Y|X) = 1$).

Determinism is simply the idea that every event is determined to happen, and can thus be (truthfully) deterministically explained.263

Now, what about all the prefixes above? They represent different ways to reach the same destination. For the varieties of determinism all boil down to a single idea: the idea that everything happens by necessity; that all apparent forks in the road of the past are illusory; that nothing could have been different than it was given the determining conditions. Regardless of if one believes that it is culture, genes, socio-economic structures, or whatever that does the determining (or any combination of them), this is the basic idea behind determinism.

The central notion of determinism is thus that everything is the product of sufficient causing conditions. Usually the sufficient causes are conceived as a compound of antecedent conditions and causal laws. The

---

263 It is important to have these distinctions in mind. It is perfectly coherent to believe that every human action is deterministically explainable, while being an indeterminist about the behaviour of subatomic particles. To adopt such a view is to disqualify universal determinism, but believe that determinism governs the subfield of human action. We shall here only treat whether human action (and thought, feeling, etc.) are fully deterministically explainable. As treated below, a determinist is one that believes that all human action is deterministically explainable. Of course, all human action on this view is deterministically explainable since determinism reigns over human actions.
idea is that if there is an antecedent condition \( X \), supplied with a law which says that \( X \rightarrow Y \), we can say that \( X \) is sufficient for the occurrence of \( Y \). Note that \( Y \) may occur for other reasons than \( X \). But we can say that given \( X \), \( Y \) necessarily follows.\(^{264}\)

Let us refer to the thesis that every event, including every action, is the result of sufficient causes in the above sense as determinism (or causal determinism, as I shall sometimes call it). We might state it as follows:

\[
\text{Determinism}_{\text{def}}: \text{the state of the world at any time } t \text{ together with the causal laws necessitates the state of the world at } t+1 \text{ (makes other states of the world impossible at } t+1). \]

Can people plausibly be said to freely choose their actions under such circumstances? If determinism (D) is true it seems like we never have alternative possibilities in action, for this would mean that there are more than one possible future state of the world.\(^{265}\) This is precisely what the truth of D rules out. Thus, unless we are able to falsify the laws of nature, or are able to change events in the past, it seems as if we can never do otherwise (see van Inwagen 2003). Can responsibility find a place in such a world? This is the so-called free will problem.

Theorists have tended to take one of the following stances to this problem: First we have the incompatibilists, who are convinced that D and responsibility are incompatible with one another (they cannot co-exist; if one holds, the other must go). Incompatibilists come in two versions: libertarians (not to be confused with political libertarians) and hard determinists.

\(^{264}\) Kane defines determinism thusly: “Any event (including a choice or action) is determined […] just in case there are conditions (such as the decrees of fate, antecedent physical causes plus laws of nature, or foreordaining acts of God) whose joint occurrence is (logically) sufficient for the occurrence of the event. In other words, it must be the case that, if these determining conditions obtain (e.g., physical causes plus laws of nature), then the determined event occurs” (Kane 1996: 8).

\(^{265}\) That is, if we are able to choose between \( X \) and \( Y \) at time \( t \), this means that there are two possible states of the world at \( t+1 \) – one where \( X \) is chosen and one where \( Y \) is chosen.
1a) **Libertarianism.** If D were true, we would never be able to act otherwise. We would not be morally responsible for our actions, since this requires that there are alternative possibilities genuinely open to us. But D is not true. So it is possible for people to be morally responsible (and it is added that people in general are morally responsible).

This position thus affirms the incompatibility between freedom of the will and determinism but rejects determinism: people have a contra-causally free will. Historically, libertarianism has been advocated by such philosophers as Thomas Reid and, with some qualifications, Immanuel Kant. Recent statements of libertarianism include van Inwagen (1983), Kane (1996), Searle (2001), Berlin (2002), Chisholm (2003), and O’Connor (2003).

1b) **Hard determinism.** If D were true, we would never be able to act otherwise. We would not be morally responsible for our actions, since this requires that there are alternative possibilities genuinely open to us. D is true. So it is impossible for us to be morally responsible.

This position, then, accepts that freedom of the will and determinism are incompatible, but holds that determinism is true. Historically, hard determinism has been advocated by such philosophers as Spinoza and Robert Owen. Recent statements of hard de-

---

266 It should be noted that there are two types of libertarians, who are in deep disagreement about where to place indeterminism. **Agent-causal libertarians** equip the agent with the power to freely initiate causal sequences: indeterminism exists since free agent may choose to do otherwise. **Event-causal libertarians** argue that actions are caused by states or events, but maintain that indeterminacy in the way the states or events come about is sufficient for free will and moral responsibility. Event-causal accounts thus do not employ an agent or substance with the power of contra-causal choice. Among recent agent-causal theorists we find Chisholm, Clarke, O’Connor, and Taylor. Kane and Ginet are examples of recent event-causal libertarians. For useful introductions, see Kane 1996; Pereboom 2001; Fischer 1999.

267 It should be noted that what I here and below refer to as “recent” statements are quite old by some standards. The free will problem, however, is so old that an article written in, say, the 1950s easily qualifies as “recent”.
terminism, or positions close to it, include Skinner (1971), Galen Strawson (1994), and Pereboom (2001).

Then we have those sober debunkers who believe that there is no opposition between D and free will; that they are compatible. These are the **compatibilists**, sometimes also referred to as **soft determinists**.268

2) **Compatibilism.** “Freedom” or “responsibility” does not require that people can choose between alternative possibilities. It rather requires that one’s actions are rationally controlled and unimpeded.269 People can meet such criteria for responsibility without having alternative possibilities: the difference between performing actions voluntarily or not is independent of the truth of D.

The compatibilist can choose one of two basic strategies:

2a) Freedom is compatible with D. D might be true or false: it has no consequences for our freedom or responsibility.

2b) Freedom **presupposes** D. Freedom requires that our actions are rationally directed towards satisfying our desires. D enables the kind of control that rational action requires. Unless our actions are determined by our beliefs and desires they seem subject to indeterminism (i.e. chance). Randomness in action is a far greater threat to freedom and responsibility than determinism.

---

268 They are called soft determinists if they hold determinism to be true. It is of course possible be a compatibilist and believe that determinism is false. One then thinks that although determinism is false, it would be compatible with free will and moral responsibility if it was true.

269 David Hume famously defined “liberty” as “a power of acting or not acting, according to the determinations of the will” (Hume 1966: 95, italics in original). Thus, liberty doesn’t require the ability to choose between alternative possibilities; it requires only that our will is causally effective and that we can do what we want. Pettit has referred to freedom in this sense as rational control. The core component of this kind of control is that one acts under the rational direction of one’s beliefs and desires, i.e. that “depending on how his or her beliefs and desires go, the agent will act, now in this way, now in that way” Pet 2001: 35.
in action. In short, D is a prerequisite for having a rational and causally effective will.

Compatibilism is by far the most frequently inhabited position today (mostly in the 2a version, with Mill (2002) being the most famous champion of 2b). It has a very impressive intellectual pedigree, historically being advocated by among others Hobbes, Hume, and Moritz Schlick. Recent statements include Smart (1961), P.F. Strawson (1974), Frankfurt (1988), Fischer and Ravizza (1998), and Dennett (1984; 2004).

Logically, one position remains to be chiselled out: those who are indeterminists and reject the notion of free and responsible agency. Since D has been widely rejected as an universal account of nature by the dominating indeterministic interpretation of quantum mechanics, there appears to be some room – be it a rather pointless and counterintuitive room – for those who see the human mind as a product of random firing of neurons or the like, and consequently rule out that the kind of control required for responsibility exists:

3) **Random indeterminism.** D is false, and as a consequence of indeterminism, we are not responsible for our actions. Since human behaviour is the result of irreducibly random processes, our behaviour is not under our control.

As far as I know, no one occupies this position. Alexander Wendt (2006) has recently pioneered attempts to create a “quantum social science”, but he does so with the explicit aim of providing room for free will – a room he argues is nonexistent in social science’s classical worldview.

**A very brief history of the free will problem**

The free will problem has haunted humanity at least since Antiquity. It was seriously introduced by the Stoics, who wanted to know how to cope with a world that was strictly governed by causality (Arrington 1998). It
has been continually and hotly debated, but it is widely acknowledged that it remains to be “solved”. Some conclude that the problem is unsolvable (Stroll 2004), while others argue that it is a pseudo-problem (Double 1994).

Compatibilism is the preferred position among contemporary philosophers (Strawson 1986). For the major part of recollected history, however, libertarianism seems to have been the dominant view. The free will problem – by which I here mean the question whether humans are free in such a way that they may appropriately be seen as responsible for their actions, i.e. whether they can deserve praise or blame – was solved by the appeal to the special, indeed divine, place man occupies in creation. In the work of many religious thinkers, the solution lies in the fact that man was created in the image of God. Since God was a “prime mover unmoved”, capable, as the so called cosmological proof of God’s existence taught, of ultimate causation, then it followed that man too was endowed with such a power. That power, commonly referred to as “free will”, resided in the soul and ensured that man was free to choose between alternatives. He was therefore responsible for them, despite the omnipotence of God.

270 I use citation marks here since it is unclear what the problem is – there are many different problems associated with the covering umbrella referred to as “the free will problem”. Part of the reason why the free will problem is difficult is precisely that there is widespread disagreement about what the question is. It is safe to say, however, that most if not all specific and fairly well-defined problems associated with the free will problem remain unsolved, at least in the sense that they remain debated.

271 Prominent examples of this way of arguing are della Mirandola 1985 and Boethius 2002.

272 The cosmological proof came in many versions. A particularly influential one consisted in tracing causation back in time until logically arriving at the ultimate cause, or first cause. This cause, Aquinas among others argued, was God.

273 Difficult questions abound here, however. If man can choose freely between alternatives, what are we to do with God’s omniscience, the attribute of being all-knowing? If God knows the actions of man in advance, then they can scarcely be said to be free. But if God does not know their actions in advance, the actions seem free but how can God then be said to be omniscient? Many theologians settled for that God knows in advance the choices of men, and in this sense they are not free.
for both good and evil in the first place was a separate problem. Theologians settled for the view that God preferred the worship of servants who had a choice to stray over the automatic obedience of servants incapable of doing wrong. 274

So influential was the idea that man was endowed with a special dignity, a divine stamp of approval, that it resisted the emergence of modern science. Galileo’s and Newton’s discoveries posed a threat to free will in the sense that they not only were unprecedently successful accounts of nature, but also of a deeply mechanical nature. The explanations the new science advanced were strictly causal: events were the necessary result of antecedent condition and laws of nature. With the best science of the day being mechanical and deterministic, philosophers faced the dilemma of either discarding good science or denying free will. The solution, with Descartes being the prime example, was a far-reaching dualism. By separating mind and body, Descartes could grant that nature, including our bodies, consisted of matter that was ruled by causal laws, but still maintain that the operations of the mind was a different story. The human soul was not material and as such did not belong to the physical universe. Human behaviour was not caused, it was chosen. Responsibility remained intact.

With dualism in decline, many argue today that modern science provides ample evidence that libertarian free will is a myth. 275 Pinker (2002) names four areas of science that have generated such evidence:

---

274 Plantinga has made this case thusly: A world where there are creatures with considerable freedom is more valuable than a world where there are no free creatures at all. A benevolent creator will therefore opt for a world of free creatures but cannot, as a matter of definition, guarantee that they will choose good rather than evil (if so, the creatures would no longer be free in the sense of having alternate possibilities control). The creator makes morally praiseworthy acts possible only by making morally blameworthy ones possible. Plantinga’s position obviously assumes that praise and blame are only possible under incompatibilist freedom. According to compatibilism, there is nothing contradictory in being responsible while under the sway of determinism, as long as “choices” are made the right way. Thus conceived, God may create creatures that always do right, without thereby ruining their responsibility. For a presentation of Plantinga’s view, in addition to an overview of other arguments here, see Peterson et al. 1997: 110-3.

275 An influential attack on dualism was delivered by Gilbert Ryle Ryle 1984.
cognitive science, neuroscience, behavioural genetics, and evolutionary psychology. These disciplines - and others - explain action in ways that make no reference to autonomous man, divine souls, or transcendental selves. The notion of an agent capable of buck-stopping contra-causal agency is increasingly being regarded as obscure. Among philosophers there seems to be a consensus, albeit weak, for compatibilism. Though quantum physics, the dominant theory of nature, is believed to be inherently indeterministic, the ability to do otherwise is certainly out of intellectual vogue. Yet philosophers contend that most “ordinary people” retain a belief in libertarian freedom. I have, however, never seen the survey to support that claim...  

8.2 HARD DETERMINISM

“If some truth is disturbing, this is no reason not to believe it.”
(Parfit 1987: 324)

If scientific explanations of human behaviour threaten responsibility, as was argued in chapter 7, and if we could at least in principle explain all behaviour scientifically, then it seems as if we are propelled towards hard determinism. As we just saw, determinism is the view that every event, including human thoughts and action, is the result of sufficient causes. The consequence of this is that every subsequent state is the necessary result of prior causes. This seems to mean that people can never do otherwise - there is no time t when an agent is genuinely free to choose between alternative courses of action, for that would mean that some subsequent states of the world are not necessitated by prior causes. Hard determinism is the view that concludes from this that we are never responsible for what we do.

276 Results from experimental philosophy indicate, however, that people tend to accept that moral responsibility is compatible with determinism, at least as long as they are asked to assess concrete events Knobe 2007.

277 I write this noting the absurdity of a “posterior cause.”
The argument for hard determinism (HD) argument basically takes the shape of a *modus ponens*.

\[\begin{align*}
    P_1 & : \text{If determinism is true, then our actions are never free and we are never responsible for them} \\
    P_2 & : \text{Determinism is true} \\
    C & : \text{Our actions are never free and we are never morally responsible for them}
\end{align*}\]

This is pretty much as radical a philosophical theory can get. How could one arrive at it? I will here briefly describe two authoritative statements of HD: Galen Strawson’s and Derk Pereboom’s.

Galen Strawson (1994) has advanced the “Basic Argument” to defend that we are never free in our actions and thus never truly morally responsible. Though it can be extensively elaborated, the Basic Argument is very simple, and probably represents what most think of when they consider the consequences of determinism. Here is the simplest version of the Basic Argument.

\[\text{278 Note that libertarianism and compatibilism also can be cast in terms of simple, valid arguments. For example:}
\]

**Libertarianism**

\[\begin{align*}
    P_1 & : \text{If some actions are free, determinism is false} \\
    P_2 & : \text{Some actions are free} \\
    C & : \text{Determinism is false}
\end{align*}\]

**Compatibilism:**

\[\begin{align*}
    P_1 & : \text{All actions are determined} \\
    P_2 & : \text{Some actions are free} \\
    C & : \text{Some determined actions are free}
\end{align*}\]

279 A caveat: Neither Strawson nor Pereboom are pure hard determinists, however. The arguments they make are intended to challenge the notion of free will and moral responsibility, but both claim that their arguments would work even if determinism turns out to be false. The reason for this, they contend, is that unless actions are governed by determinism, they must be subject to randomness. This is equally bad for free will and moral responsibility. Pereboom dubs this position *hard incompatibilism*. Strawson seeks to argue the conceptual impossibility of free will, based on the assumption that nothing can be the cause of itself. His argument has been referred to as a “no-free-will-either-way”-strategy; see Pereboom 2001: xviii.
P₁: Nothing can be *causa sui* – nothing can be the cause of itself
P₂: In order to be truly morally responsible for one’s actions one would have to be *causa sui*, at least in certain crucial mental respects
C: Therefore nothing can be truly morally responsible

What is the basis of the argument? Strawson exploits the regresses that chains of occurrences are susceptible to, even chains of rational action. Suppose I have a choice between voting on party Left or party Right. Whether I vote Left or Right will depend, Strawson contends, on the kind of person I am at the time of voting. If I am a Leftist person I will vote Left. If I am Rightist one I will vote Right. Suppose I vote Left. Is this a free choice? No, Strawson argues, since I cannot be a *causa sui*. At the time of choosing, I choose according to the kind of person I am (what preferences I have, etc.). But the kind of person I am now depends on the kind of person I was yesterday, which in turn depends on the kind of person I was the day before yesterday, and so on. Just as we saw in the previous chapter, a regress is mounting. It seems to end up in factors well beyond my control, such as my genotype or early socialization. These are not things I am responsible for. But if I am not responsible for where the regress stops, I cannot be responsible for what I subsequently do or have done.

It is of course true that a person may intentionally decide to change him- or herself. A Leftist may desire to be a Rightist person, and make a conscious decision to transform him- or herself to become Rightist. Over time, that decision may produce a genuine change in the person, such as when we intentionally change taste in food (I am presently trying to learn to appreciate olives in this way, but, alas, with little success). But even though we can change the kind of persons we are out of acts of will, we cannot be said to have free will. For the desire to change oneself at \( t \) depends on an earlier state of oneself, which in turn depends on an even earlier state, and so forth.

The conclusion of the Basic Argument is that true self-determination is impossible, and, consequently, that responsibility is too. For Strawson, this means that “it is exactly as just to punish or reward people for their actions as it is to punish or reward them for the (natural) colour of their hair or the (natural) shape of their faces” (Strawson 1994: 16).

Derk Pereboom reasons in similar ways in chiselling out what he calls *hard incompatibilism*. Like Galen Strawson, Pereboom thinks that we are never morally responsible; but unlike Strawson, he believes that there
is nothing logically incoherent about libertarianism. It is conceivable that we have something like libertarian freedom, which enables us to exercise contra-causal free choice. The problem is that such freedom is empirically untenable. “Our best scientific theories”, Pereboom proclaims, “indeed have the consequence that we are not morally responsible for our actions” (Pereboom 2001: xiii). This is so because science explains actions by citing factors beyond the individual’s control. First consider deterministic explanations. They clearly undermine responsibility, Pereboom argues, since they are susceptible to the deterministic regress (travelling ever further back along a causal sequence that ends up in genotype and early socialization). But now consider indeterministic explanations, such as quantum mechanical ones (which Pereboom takes to be the best scientific theory out there). Even if it turns out that quantum indeterminacy governs human action, it is clear to him that moral responsibility would be equally out of the picture. For while indeterminacy in nature may provide grounds for alternative possibilities, it cannot sufficiently bestow control over actions. It is just as bad if I randomly “just happen” to perform an action as it is if I was determined by sufficient causes to perform it. (I will return to this topic in greater detail below.)

Free will and moral responsibility thus find no purchase in a scientific understanding of the world, whether indeterministic or deterministic. Here is how Pereboom understands (part) of the moral significance of hard determinism:280

“The hard determinist position implies that the appalling actions of persons are much more similar to earthquakes and epidemics than they are according to views that hold persons morally responsible. The justification we assume for regarding especially wrongful actions of persons as deeply different from natural disasters is that persons are typically responsible for their actions. But according to hard determinism, because a person’s actions are the result of processes over which he has no control, we cannot consider him responsible for them, just as we cannot hold earthquakes and epidemics responsible for their effects.” (Pereboom 1995: 33. Emphasis added.)

280 Pereboom sometimes uses the term “hard determinism” even though his preferred position is “hard incompatibilism”, or the idea that the kind of free will required for moral responsibility cannot exist even if determinism is false.
Hard determinism holds, then, that no one is ever really responsible for what he or she does. This is no doubt very radical. It rules out praiseworthy and blameworthy behaviour equally. Just as Mother Theresa is not really to be praised for her actions, so is Hitler not really to be blamed for his. They were both determined to become the kind of people that they became, and had no choice but do act the way they acted. In this sense, they were both like “earthquakes.”

The first reaction to hard determinism is likely to be that it is an implausible thesis, and one with outrageous moral and political consequences. I will return to the consequences of the thesis below, but we should first concern ourselves with its plausibility. I believe that hard determinism is a stronger position than one initially may think. Objections levelled at it are often found wanting. I shall here try to defend HD against three such counterarguments: the prominent objections from compatibilism, probabilism and explanatory incompatibilism.

8. 3 OBJECTION 1: COMPATIBILISM

The most popular way to refute HD is to employ reasoning that is compatibilistic in kind. As I said before, compatibilism is the preferred position among philosophers today, though its standing in no way is hegemonic. Less than a unified doctrine, it is a disparate and often impressive collection of ideas that rest on the idea that the “threat” of determinism, as classically conceived, is mistaken.

Recall the argument advanced by incompatibilists: if determinism is true, then no one could ever do otherwise. Choices are always determined by the conjunction of antecedent conditions and causal laws. This clearly seems to mount a regress that excludes that we are ultimate responsible for them. So if determinism is true, freedom and responsibility seem impossible. Compatibilists take objection with this argument.

Compatibilists usually begin by analyzing some common choices of words, and show why they load the issues in a wrong way. People for example often say that if determinism is true, then things are “fated” to occur and are “inevitable.” To say this feeds into the idea that determinism is something that threatens to enslave people and make their agency futile. Not so, compatibilists rightly note. For even if determinism is true,
it is still the case that some things will happen only if we make them happen. Take the mundane example of ordering lunch. It is true that ordering lunch may be a deterministic process. It may be that you are determined to pick the dish you eventually pick. But this is certainly not to say that your input was futile. When you choose what to eat for lunch, the lunch you eventually eat crucially depends on what you pick. Indeed, had you not chosen the dish you chose, you wouldn’t have eaten it all. Determinism does not imply that there is some woeful lunch waiting for you in the future, which you will be forced to eat no matter what you do. This is, in a nutshell, what distinguishes determinism from fatalism.\(^{281}\)

This seems right, but the truth of determinism still seems to rule out the possibility of doing otherwise. Even though what you eat for lunch depends on what you choose, the outcome of that choice is still determined by antecedent causes if D is true. Under more ambitious formulations of determinism, this means that the outcome of your choice has been determined for generations, maybe even eons, since everything happens as a consequence of causal laws and prior conditions. And if the choice of lunch is too innocuous an example, this means that beastly crimes have similarly been determined to happen for generations. How could we then say that a criminal is responsible for the crime, and thus deserving of punishment?

Compatibilists have responded to this line of thought primarily by challenging the importance of alternative possibilities, or the notion that one in acting “could have done otherwise.” Sometimes they try to reinterpret “could have done otherwise” into “could have done otherwise, if we would have so chosen.”\(^{282}\) Let’s say that you face a choice between X and Y, and choose X. You could have done otherwise, according to this reinterpretation, if it is true that if you had chosen to go for Y instead, you would have gotten Y. Take the lunch-example again. I could have ordered a different dish than meatloaf (ordered otherwise) if it is true that, if I had ordered spaghetti rather than meatloaf, I would have eaten

---

\(^{281}\) Fatalism is about the irrelevance or futility of agency. A classical illustration of fatalism is the plot of Sophocles’s *Oedipus the King*, where Oedipus, fated to kill his father and marry his mother, fails to escape this horrid fate despite efforts to the contrary.

\(^{282}\) Moore famously made such a reinterpretation in *Principa Ethica* (Moore 1994). See footnote 174.
spaghetti for lunch. By contrast, I could \textit{not} have done otherwise if I was compelled to choose meatloaf, say because of the paternalist attitudes of the restaurant staff. Normally, however, restaurants have a habit of giving guests the kind of meal they actually pick. Therefore, in this situation as in many others, it is true that I could have done otherwise even if D’s true. For to be able to do otherwise simply means to have the opportunity to do different things, should one choose to.

This reformulation of “could have done otherwise” strikes me as highly misleading. It doesn’t do justice to what incompatibilists are worried about. What we want to know is not whether we would have ordered such-and-such a meal \textit{if} we had so chosen. That seems trivially true. We want to know whether we could have chosen differently in the first place. The debate is thus not about the casual effectiveness of the will. It is about the direction of the will. The compatibilistic reformulation, in obscuring the topic of the actual debate, feeds into claims such as Searle’s that he “cannot think of any interesting problem of free will to which compatibilism provides a substantive answer” (Searle 2001: 278).

Compatibilism is at its best when it squarely allows that the truth of D implies that there is a perfectly intelligible sense in which one could not have done otherwise and instead focuses on the differences between “free” and “unfree” actions that seem to withstand the truth of D. That is, compatibilism is at its best when it develops what Fischer has called “actual sequence” requirements of freedom and responsibility (Fischer 1999). Actual sequence requirements are requirements which do not depend on alternative possibilities, but instead hold that it is the quality of the action which determines whether it counts as free and responsible or not. The reasoning goes something like this: if D is true, which we may assume it is, then every action, like every event, is the necessary product of antecedent causes. But though they are all causally determined, there are some actions that are properly thought of as free and responsible and some that are not.

Take the classical case of a kleptomaniac and a “normal” shoplifter, who both engages in petty theft at time $t$. Given that D is true neither thief could have done otherwise. They were both determined to steal. However, compatibilists argue, it is still perfectly plausible to hold the normal shoplifter responsible and excuse the kleptomaniac. The kleptomaniac suffers from a debilitating compulsive disorder that infringes on the freedom of action in a number of ways. He or she cannot, let’s assume, refrain from stealing even though he or she might desperately
want to do so. Moreover, he or she might genuinely want to be cured from the kleptomania, and thus not act on preferences or desires that he or she in turn prefers or desires.\textsuperscript{283} The normal shoplifter, by contrast, does not suffer from any comparable disorder. He or she is reason-responsive, has a causally effective will, and presumably steals since this is what he or she genuinely wants to do. This means, compatibilists argue, that he or she is fit to be held responsible whereas the kleptomaniac is not. The “actual sequence” of the normal shoplifter’s actions simply meets the qualifications of free and responsible agency whereas the kleptomaniac’s does not.

There seems to be some truth to this: Even if we imagine D to be true, there still seems to be a salient difference between the two shoplifters. Similar things would seem to hold for compelled vs. unimpeded choices; choices that were performed under hypnosis vs. those that were not; choices that reflects ignorance or immaturity vs. those that reflect knowledge or maturity; and so forth. Even if D is true, we still seem able to differentiate between different actions in terms of their quality or nature.

But compatibilists have yet another ace up their sleeve, which bolsters the claim that only actual sequences are of relevance: the so-called Frankfurt-style objection to the alternative possibilities requirement for responsibility. Incompatibilists as a rule emphasise that in order for an action to be free and morally responsible, it must not only meet requirements of control, but the alternative possibilities requirement as well. Thus, in order for me to be responsible for \textit{X}, in addition to me having control over choosing \textit{X}, there also has to be at least one alternative

\textsuperscript{283} The idea of locating responsibility in self-identity has been developed by Harry Frankfurt. The idea here is that one is responsible if one acts on desires one approves of and recognizes as one’s own. Frankfurt speaks in terms of first and second-order desires. The kleptomaniac might be said to have a first order desire to steal, at least in the sense that his or her limbs do not move as a result of external forces but out of his or her own acts of volition. But the kleptomaniac might have a second-order desire not to have the first order desire to steal, but be unable to act on that desire. If so, he or she lacks free will. The Frankfurtian emphasis on motivation has the important benefit that it allows us to focus on internal constraints and not only external ones, which was a flaw of early compatibilism (see Frankfurt 1988).
course of action Y, such that it was genuinely possible for me to choose X or Y at time \( t \).\textsuperscript{284} Attributed to Harry Frankfurt (1969), the Frankfurtian objection is intended to demonstrate that one may be responsible without alternative possibilities. Here is how it works:

Suppose a state \( A \), which counts as a moral agent. State \( A \) is involved in a border dispute with a neighbouring state and considers whether or not to invade it. The thing is, however, that there is a third party, superpower-state \( B \), which very much wants \( A \) to invade its neighbour for “divide-and-conquer”-reasons. If \( A \) opts to invade, \( B \) will stay passive. If \( A \) opts not to invade, however, \( B \) will threaten \( A \) with such overwhelming force that \( A \) has no choice but to invade the neighbouring state, thus averting a conflict with \( B \). But \( A \) is not aware of the fact that \( B \) will intervene if \( A \) chooses not to invade. Here it seems that \( A \) has no alternative possibilities. Either \( A \) chooses to invade voluntarily, or \( A \) is forced to invade by \( B \). To refrain from invading is not a possibility. Does this mean that \( A \) is not morally responsible for its invasion? Well, here is the snag. If \( A \) invades as a consequence of \( B \) having bullied \( A \) around, then we seem willing to excuse \( A \). But if \( A \) voluntarily chooses to invade, which means that \( B \) remains passive, it seems as if \( A \) is morally responsible for that decision. \( A \) is merely doing what \( A \) has chosen to do. But if \( A \) is morally responsible for the decision to invade, then \( A \) is morally responsible despite the fact that there was no alternative possibility. Ergo, alternative possibilities are not necessary for freedom or responsibility.

The Frankfurtian case feeds into the actual sequence requirement of compatibilism: what matters is not whether one could have done otherwise, but whether one acts voluntarily or not.

\textsuperscript{284} As we have seen, this condition, which is foundational to all libertarianism, is of course what is intended by saying that an agent “could (or could not) have done otherwise.” Even theorists such as Robert Kane, who emphasise ultimate responsibility – the intuition that an agent has to be the buck-stopping originator of his or her choices – as the core libertarian intuition, have to treat the alternative possibilities as a necessary condition of responsibility. See Kane 1996: 60.
Compatibilism questions the first premise of HD: that determinism is incompatible with freedom and responsibility. Its central claim is that some determined actions are “free” in the requisite sense. The defence of this claim is often ingenious, and the deterministic world compatibilists describe is a world of deliberation, creativity, and causally effective agency - far from the gloomy apathy of fatalism. But I doubt whether compatibilism suffices to reject HD. For all its sophistication, it cannot overcome the problem that attracts people to HD in the first place: the problem of responsibility in a deterministic world. If determinism is true for human actions, then there is no decision in which it makes sense to say that the agent could have genuinely done otherwise. The kleptomaniac and the “normal” shoplifter are just as determined to engage in petty theft when they do. Neither could have done otherwise. This means that it would be unfair to blame or punish either of them: the actions of both are susceptible to the kind of regress ending up in factors beyond their control that determinism leads to. The kleptomaniac of course shouldn’t be punished. But neither should the normal shoplifter. For if “ought implies can”, how could we say that the normal shoplifter ought not to have stolen, when the priors of his or her choice are such that he or she cannot but shoplift?

What about the Frankfurtian objection? Doesn’t it show that alternative possibilities are irrelevant to responsibility? No. The objection works wonderfully when we consider a choice which is in some sense isolated in time: A can choose X or attempt to choose Y (but be compelled to do X anyway). But now consider that A’s choice is also susceptible to the deterministic regress. Even if A voluntarily chooses X at t, it is important to note that this choice - just like any other choice if D is true - is then determined by prior circumstances, which in turn is determined by prior circumstances, and so on. According to proponents of HD, what rules out responsibility is the absence of free will; an absence which implies that there is no regress-stopping point in time where buck-stopping responsibility can ultimately lie or originate from. The Frankfurtian objection does not alleviate this concern. For even if B does not interfere with A’s
choice, A is not responsible for voluntarily choosing to invade if it is true that it is the necessary outcome of ever earlier events.

* * *

Compatibilism ultimately relies on us accepting that questions of freedom and responsibility really concern the quality of determined choices and not whether the choices could have been different. Lots of intelligent people apparently accept just this. I have intuitions to the contrary. I do accept that there is a difference between the kind of agency a kleptomaniac and a normal shoplifter is capable of. But I cannot see how this difference matters; how punishing the normal shoplifter could be any less unjust than punishing the kleptomaniac, if D is true.

8. 4 OBJECTION 2: PROBABILISM

A second argument against HD is the argument from probabilism. The discussion thus far has relied on the idea that explanations are, or should be, of a deterministic variety. But this may seem outdated. Quantum mechanics is believed to be inherently indeterministic, which means that the even the dominant physical theory has departed from determinism. And many would say that the science of human action, in particular the social science, certainly does not aspire to explain deterministically. To say that social scientific explanations might erode the concept of personal responsibility is thus to misrepresent the nature of the explanations the social sciences offer, which are probabilistic rather than deterministic in form.

Recall that an explanation is deterministic if it holds that the explanans renders the explanandum certain, other things being equal.285 “If X

285 I add this “other things being equal”-qualification since it may very well be true that X invariably leads to Y unless some other force Z overrides the causal influence of X. An example: being employed in the Public sector (X) causes one to vote for parties that endorse increased spending on the Public sector (Y), unless one owns stock in private companies (Z).
then $Y$” is an example of a deterministic explanation.$^{286}$ It states that $X$ is sufficient for $Y$; if $X$ occurs $Y$ will invariably follow. An explanation is probabilistic if it states that the occurrence of $X$ renders $Y$ more or less likely. Formally, a deterministic explanation states that $\Pr(Y|X)=1$, whereas a probabilistic one states that $0<\Pr(Y|X)<1$. $^{287}$

How could probabilism be of relevance to the plausibility of HD? Well, it may be invoked to question its second premise. Deterministic explanations depend on what is sometimes called the constant conjunction requirement: if it is proposed that “if $X$ then $Y$”, then we had better observe that $Y$ invariably follows $X$. But actual science of human action cannot discover such regularities. Social science in particular seems unable to identify any causal laws of this kind. Goodin and Klingeman for instance, in an authoritative introduction to political science, states: “The truths of political science, systematic though they may be, are and seem inevitably destined to remain essentially probabilistic in form. The ‘always’ and ‘never’ of (...) covering laws find no purchase in a political

---

$^{286}$ It is important to note that “if $X$ then $Y$” doesn’t qualify as an explanation if it is merely a correlation. We know, for instance, that tall and thin men are more likely than others to suffer a spontaneous pneumothorax (lung collapse without external trauma). But this observation, even if it was invariably true that all tall thin men get the disease, does not count as an explanation unless one also provides the causal mechanism, i.e. an account of why tall and thin men suffer the disease - what it is about being tall and thin and male that produces the outcome. To think in terms of causal mechanisms is useful, I think, even if one doesn’t accept Elster’s famous contention that mechanisms bring in an element of indeterminacy. (In Elster’s, mind causal mechanism “are triggered under generally unknown conditions or with indeterminate consequences” (Elster 1998: 45).)

$^{287}$ Lieberson 1991: 309 has wrongly defined a probabilistic explanation as one where the presence of $X$ increases the likelihood or frequency of $Y$. It is important to note that a probabilistic explanation only requires that the presence of $X$ affects the likelihood of $Y$. It is a fallacy to assume, as for example Hempel did in his *Inductive-Statistical* model of explanation, that high probability is necessary for a successful probabilistic explanation. That assumption, incisively criticized by among others Wesley Salmon, misses that it is statistical relevance rather than certainty that should be the sought. In probabilistic causation, we may understand highly probable events just as well as ones with low probability (see Salmon 1989).
world, where things are only ever ‘more or less likely’ to happen” (Goodin & Klingemann 2000: 9).

Claims such as this might lead us to reject the theory of excuses offered in chapter 7 as something utterly unrealistic. After all, we would only accept that an explanation of criminal behaviour is excusing if it explains crime with a high degree of confidence. If we could observe, say, that all socially disadvantaged young males with the XYY-syndrome engage in violent crime, then it would seem reasonable to excuse their behaviour on the ground that it was caused by biological and environmental factors beyond their control. But actual science cannot provide such explanations. In reality, the explanations are always probabilistic in form. For even though some explanations explain (or “predict”, as it is sometimes called) a large portion of the variance in behaviour in any given population, there always seems to remain a higher or lower degree of exceptions to the expected patterns. Many young, socially disadvantaged XYY-males grow up to be perfectly law-abiding citizens and do not commit a single violent crime.

Observations such as these naturally suggest that the combination of youth, a rotten social background and an extra Y-chromosome cannot count as an excuse. For since some do not commit violent crime, it may seem as if those who do commit such crime in effect have chosen to do so. And since all crime, and all action in general, are probabilistically explained, it seems as if this case could always be made (save, perhaps, for when we can explain behaviour with very high likelihood).288 If an explanatory factor doesn’t work across the board, it cannot be taken to excuse.

On the basis of this it could be argued that HD is a bit like the bogeyman: a fictional thing fit to scare children with, but with little observational warrant to it. For the truth is that every proposed explanation is marred by lots of anomalies – in fact, one single universal causal law of human action is yet to be discovered. Therefore the theory of excuses I

---

288 Steven Pinker in *The Blank Slate* thus argues that probabilism “doesn’t succeed in allaying the fear that science is eroding the concept of free will and personal responsibility. It is cold comfort to be told that a man’s genes (or his brain or evolutionary history) made him 99 percent likely to kill his landlady as opposed to 100 percent” (Pinker 2002: 177).
Why Probabilism isn’t the Answer

The prevalence of appeals to probabilism in social science is only matched by their unclarity. We can all agree on the definition of a probabilistic explanation: \( X \) probabilistically explains \( Y \) if the occurrence of \( X \) affects the likelihood of \( Y \) occurring. What is less clear is the nature of the probability invoked.

Philosophers of science distinguish between subjective and objective conceptions of probability (Popper 2004: ch. 8). Subjective probability is basically an epistemic concept. When we know too little about something, our lack of knowledge may lead us to say that an event is such-and-such probable, but not certain. The better we become at explaining the event, however, the less probabilistic it will appear. The assumption here is that the world is ultimately not of a probabilistic nature; it is only our incomplete knowledge that leads us to infer that it is. If we were able to see all the nuts and bolts that govern the world, we would exchange probability for certainty.

Objective probability, on the other hand, is an ontological concept. It has to do with the world as it is, not our limitations as scientists. An objectively probabilistic process is irreducibly probabilistic – it is, in itself, governed by randomness. A correct explanation of such a process cannot transcend the degrees of probability that govern it. To ask for certainty in such explanations is a mistake.

Let’s illustrate this important distinction. Suppose that socially disadvantaged males with XYY-syndrome have a 0.6 probability of committing at least one violent crime before they turn 25. This of course means that there is a 0.4 chance of no such crime being committed. These probabilities are arrived at after having studied the population as a whole. Based on them we are inclined to say, were we to make a guess about whether a random individual in the population has committed a crime, that this is a reasonable guess. For any random disadvantaged XYY-male, chances are higher that he has committed a crime than not.

Now, are the probabilities referred to here subjective or objective? It depends on our assumptions about the world. If we believe that the offered can be said to rely on an almost ludicrous exaggeration of science’s ability to explain actions.
criminality or law-abidingness of any given member of the population can be deterministically explained, they are clearly subjective. Our observations on the aggregate level are not detailed enough, accurate enough, or complete enough, to show what is really going on – that every individual outcome is the product of sufficient causes. In this state of relative ignorance, it may seem as if there are elements of uncertainty in the *explanandum*. But useful as probabilistic explanations might be, complete explanations dispel the illusion of probabilism and are deterministic.

Suppose, however, that there cannot be – even in principle – explanations that dispel elements of uncertainty in the *explanandum*. Suppose that the onset of violent criminality among socially disadvantaged XYY males is a process with irreducibly random elements – whether a particular individual commits violent crime depends in part on genuinely random events in, say, the formation of their personality. If this is true, explanations of their behaviour are, and must remain, probabilistic in form. Compare with explaining the outcomes of rolling a random six-sided die. Since the die is governed by random processes, we cannot go beyond specifying the probabilities of different outcomes happening. Suppose we roll the die and it comes up five. This outcome cannot be explained beyond the fact that there was a one in six chance of the die coming up five - since the die is truly random, it is impossible to deterministically explain why it came up five and not something else. If violent crime is subject to the same objective probabilism as die-rolling, we could similarly not deterministically explain why a particular individual commits violent crime. We would have to settle for saying that there was a 0.6 risk of this occurring, and that while there was a genuine 0.4 chance to the contrary, this alternative possibility did not obtain.

With respects to using probabilism to reject HD, it should be clear that the subjective conception of probability will not do. Subjective probabilism in explanations is fully consistent with a deterministic worldview. Though assumptions about subjective and objective probabilism is observationally equivalent – on the observational level it often makes no

---

289 Thus, every individual who commits a crime does so with Pr = 1.0, and every that individual who does not commit a crime does so with Pr = 1.0.

290 I do not pretend to know whether actual dice are governed by objectively random processes or not.
difference if anomalies are genuinely random occurrences or simply due to the incompleteness of our knowledge (see King, Keohane & Verba 1994: 59-63) – one obviously cannot use incomplete knowledge to reject D. In particular, one cannot use observations about sets of occurrences to infer probabilities of individual occurrences. One cannot say that since the probability of \( X \) in a population is 0.6, the probability of an individual member of the population being \( X \) is 0.6 too. This invalid inference obscures frequencies observed at the aggregate level with the probability of an individual element in the set being \( X \) or not-\( X \).

Then again, a longstanding insight in moral philosophy is that objective probabilism will not do either. Suppose some, but not all, socially disadvantaged XYY-males commit at least one violent crime before the age of 25. Suppose further that whether they do so or not is subject to objectively random processes. This means that we cannot fully explain why someone in the population, say Bob, commits crime rather than complying with the rules. We can only cite the probabilities for one thing happening rather than another. But to explain Bob’s crime with the fact that it was, say, 60 percent likely seems awfully close to saying that he just happened to commit crime – not in the sense that the crime was without motives, out of the blue, or bizarre; but in the sense that we cannot conclusively say why Bob broke the rules rather than following them. We can only say that while it was more probable that Bob broke the rules, it was also possible that he would have followed them. But this latter possibility was not realized.

Note that this problem of explaining genuinely random events is not alleviated by raising the probability of crime. If Bob was 90 percent likely to commit crime, we still would have to wrestle with what to do with the comparably improbable event that he did not commit crime. If we could put Bob’s life in repeat, we would see that he breaks the rules nine times out of ten. But whether he does so or not in each lifetime is still random, regardless of the degrees of probability.

On a moral note, such randomness is, as we saw above, just as threatening to personal responsibility as determination is. One’s control seems just as undermined by actions just happening as they are by determinism. Pinker has succinctly summarized this conclusion: “Either our actions are determined, in which case we are not responsible for them, or they are the result of random events, in which case we are not responsible for them” (2002: 178). Here is Mackie’s wordier but eloquent way of saying the same thing:
“Such responsibility would evaporate if we tried to attach it to purely random occurrences just as clearly as it would disappear to infinity along causal chains. It requires for its resting point a contra-causally free and yet determinate and active self, the concept of which is so hard to render coherent. Equally, the notion of an absolutely open choice is as incompatible with an Epicurean physics as with strict determinism: the metaphysical self would be left idle and imprisoned by chance no less than by law” (Mackie 1990, p. 226).

In response to Pinker and Mackie, we could object to making determinism and randomness an exhausting dichotomy. Perhaps there is a middle ground of some sort, neither governed by absolute necessity of pure chance (cf. Popper 1972: ch. 6). After all, those who employ probabilism to reject HD are not looking to exchange determinism with randomness. Rather, they advocate something akin to Leibniz’s classical dictum that “causes may influence without necessitating.”291 The idea is here presumably that causes influence action in various ways, but that people somehow retain the ability to do otherwise. In the flux of all forces affecting an agent, he or she retains some leeway in which to freely act. In O’Connor’s phrase, people are “not wholly moved movers” (2003: 258). Their decisions, though influenced by various factors, are “no mere vector sum of internal and external forces acting upon [them] in the process of deliberation” (ibid. 257).

The idea of non-determining causal influence clearly has some appeal. It allows us to preserve the point of studying regularities and yet live with that there are exceptions to every pattern. Perhaps more fundamentally, it also allows us to preserve room for free will and moral responsibility within a scientific understanding of the world. Some suspect that it is precisely because it enables scientists to paint a morally and politically optimistic image of man that the notion of non-determining causal influence has appeal.292

292 Based on a survey of the sociological theories of Parsons and Giddens, both of whom attach great weight to the concept of “agency”, Loyal and Barnes have recently argued that the question of agency is a “red herring” in social theory. “There is no fact of the matter” they claim “no evidence, however tentative or questionable, that will serve adequately to identify actions as ‘chosen’ or ‘determined’ for the purposes of sociological theory” (Loyal & Barnes 2001: 508). Loyal
Social scientist who deal with prima facie blameworthy behaviour sooner or later tend to consider the question of moral responsibility - they seek to explain, so it is natural that they reflect on the consequences their proposed explanations has for agency. When considering this question, they often pounce on the fact that their results are inconclusive, and proceed to argue that their explanations, if true, therefore do not rob people of their free will. Take as a paradigmatic illustration of this a recent book by the anthropologist Alison Dundes Renteln.

In The Cultural Defence (2004), Renteln investigates the ‘cultural defences’ invoked in legal systems (mainly the U.S. legal system). Renteln’s main argument is that legal systems fail to take the powerful process of enculturation – the process through which individuals internalizes the elements of their culture – into account. In their pursuit for even-handed justice legal systems are blinded to the fact that they are biased toward the norms of the majority culture. Had they taken enculturation seriously, they would have seen the unfairness of imposing a monolithic standard of behaviour on individuals who think, feel and act in fundamentally different ways as a result of their cultural belonging.

With Renteln’s emphasis on enculturation, it is hardly surprising that she finds the need to discuss whether an individual’s internalizing the elements of his or her culture threatens free will. (Since Renteln in studying court cases works with instances of prima facie blameworthy actions, this need is even more understandable). For if one places heavy emphasis on the idea that individuals are conditioned by their culture, then it may seem as if the free will and moral responsibility of individuals in some respect is threatened, because cultural factors may hardwire people to think, act and feel in certain ways. Not so, Renteln asserts, and proceeds to provide a typical reply to any such worries:

“This [that culture strongly influences human thought and action] is not to say that individuals are programmed by their culture in such a way that their behaviour is predetermined. If this were true, there would be no notion of guilt or innocence, as individuals who lack free will cannot make moral choices. (...) Cultural conditioning, in and of

...and Barnes ascribe Giddens’s and Parsons’s emphasis of voluntary choice their desire to craft theories that are not only are “sociologically realistic” but also “politically optimistic” (Ibid. 519).
The point that “cultural conditioning” does not render humans incapable of making moral judgements is trivially true (the question is what judgements they make and why), and need not concern us here. The relevant part is instead Renteln’s idea that a person’s culture can “strongly influence” but not “determine” his or her actions. What is meant by this claim? As far as I can see, there are two ways in which causal factors may be said to influence choices without determining them.

(1) Causal factors may, first of all, delimit the alternative courses of action an agent genuinely considers. Right now, for instance, I am considering whether to go home or stay at work and push on writing. I am not considering joining the Hare Krishna or the Foreign Legion. Causal factors such as norms, personality, culture, and so on no doubt explain why I don’t seriously consider the latter courses of (possible) action. But left with the choice between staying at work and going home, I do feel that this choice is up to me, and that I can just as well choose one as the other.

(2) Causal factors may not only delimit the alternative courses of actions an agent considers, but may also assign different probabilities to the different remaining alternatives. For example: if I am hungry right now (which I am), this raises the probability of me going home and lowers the probability of me staying at work. Hunger makes me going home, say, 0.7 probable, and me staying at work 0.3 probable. But since there is a remaining possibility that I stay at work and push on writing this chapter, I cannot be said to have been determined to go home, should this be what I eventually choose. If this is the case then causes indeed seem to, as Leibniz had it, “influence without necessitating”.

Note however that this idea of causal influence does not solve any of the problems associated with probabilism, either epistemologically or morally. First of all, to say that it was likely, but not certain, that I went home has a perfectly clear observational warrant, but that warrant is to be found in comparing my present behaviour with behaviour of the past. If I in similar situations (hungry at work, thinking about whether to go home) have chosen to go home seven times out of ten, there is a clear
sense in which I was 0.7 likely to go home. But this conclusion is drawn after making a number of observations of individual choices. And it is clearly possible for me to be determined to go home seven times out of ten and determined to stay at work the remaining three times: the probability that reigns over the series of my choices is then straightforwardly subjective. \(^{293}\) This cannot be what a sceptic of determinism is after.

Suppose, then, that the 0.7 probability of me going home applies to an individual choice and that this probability is not subjective. We are then still left with the problem of objective randomness, which we identified above: how can we account for whether one possibility happened rather than another. Again, the problem is not that my actions are pointless or bizarre. It is a situation of “plural rationality”: regardless of whether I choose to go home or to stay at work, I do so for reasons that are understandable and explanatorily adequate (e.g. “I went home because I was hungry”; “I pushed on at work since I was inspired”). \(^{294}\) The problem is rather that the choice between two rational courses of actions cannot be adequately explained in turn. Just as I said above, this seems to boil down to one alternative, though perhaps more likely than another, just happening. \(^{295}\) This is problematic from both a moral and epistemological perspective. For randomness is not the stuff responsibility is made of, and to apply objective probabilism to individual human action may seem awkward in itself.

What animates claims that causes may influence without necessitating is presumably that science about human behaviour fails to identify any universal laws. Speaking with Renteln: some processes of enculturation seem very powerful, but since there are individuals who are not

\(^{293}\) In particular, one cannot infer from the fact that I choose the less probable course of action that I exercised free will.

\(^{294}\) As we saw in chapter 6, Kane has effectively countered the objection that libertarian free will renders action irrational, out of the blue, or something that just happens by launching the notion of “plural rationality”. Loosely speaking, plural rationality obtains when a choice is rational and the agent’s own whichever way it goes; see Kane 1996: 107-15.

\(^{295}\) It does not help to say that the reasons are made effective by being acted on by the agent, since this still leaves us with the question of why the agent acted on reason rather than the other. There is, as Searle has argued, a “gap” here, which cannot be filled if one wants to defend libertarian free will; see Searle 2001.
fully enculturated, or since enculturation does not explain all relevant cases, enculturation does not determine members of a culture to think and act in certain ways. The fallacy of this move is dual-headed. First, it moves from observations about a set of occurrences to conclusions about individual occurrences: we observe, say, that not all ancient Greeks believed in the naturalness of slavery (the prevailing attitudes to slavery in ancient Greece being a compelling example of enculturation), and conclude, unjustifiably, that whether or not some particular ancient Greek believed so or not is subject to some uncertainty. Second, it makes exceptions to expected patterns indicative of free will. That is, it takes the non-perfect causal influence of some factor indicative of individuals having the power to resist the causal influence. But this assumes, also unjustifiably, that those who do not fall under the expected patterns are less influenced by causal factors than those who do. If almost all ancient Greeks was enculturated into believing that slavery was natural, the minority who did not share that belief might be just as determined to believe what they believed by other factors.

***

The argument from probabilism poses many interesting questions and dealing with it is not easy. It is highly questionable, however, if it succeeds in rejecting hard determinism. In its subjective sense, probabilism does not challenge determinism, so this cannot be used as an argument against HD. But this leaves us with objective probabilism, which, despite being recently championed by Wendt (2006), seems ill-equipped to defend lofty moral and political concepts such as responsibility. Furthermore, for all the emphasis placed on the non-lawlike nature of our observations of human behaviour, to draw conclusions about non-determining causal influence in individual actions on the basis of inconclusive correlations over a set of actions is fallacious. Probabilism as here conceived therefore does not suffice to defeat the second premise of HD.
8.5 OBJECTION 3: THE DIFFERENCE BETWEEN REASONS AND CAUSES

Some hold that there is an important difference between actions and events. Events are states of the physical world: planets moving, arms rising, temperatures dropping. They are the result of prior events and laws, and may be given causal explanations - they happen because of some prior event(s) or state(s) of the world. Actions, on the other hand, are different from events, and entail concepts not applicable to events. When a person raises an arm, he may intend to vote in favour of a proposal; something which in turn depends on raising the arm having that meaning (at least by the person doing the raising). And in raising the arm, the person does something for purposes in the future (“I do this in order to vote yes to a proposal and improve its chances to be accepted”), which seems to lend a teleological element to the action not treatable in terms of antecedent causes. These differences, some suggest, mean that actions and events must be explained in different ways. Events should be explained causally in terms of antecedent conditions and laws. Actions, however, should be given intentional explanations, or reason explanations.

To illustrate, compare the event that my arm rises with the action that I raise my arm. Let’s say that my arm rises because of some muscular problem. In that case it is an event, and a purely causal explanation in terms of the physical properties of my muscles would be appropriate. It would be pointless to ask me why I raised the arm, for that brings in a language of action that does not apply. I didn’t raise the arm – it was not intentional or purposive – but something that, as it were, happened without me doing anything. By contrast, if I raised my arm in order to vote yes to a proposal it was an action. It is now appropriate to give an intentional explanation of my action. This means to interpret my action in such a way that the action seems rational or reasonable, at least from my subjective standpoint. For those who see the distinction between actions and events as important, there is a big difference between the explana-

296 For a classical statement of this view, see von Wright 1971. See also Rosenberg 1995.
tory enterprises involved in asking (1) what caused X and (2) for what rea-
son did you do X?

Why should these reflections count as a counterargument to HD? Well, some maintain what Watson has called “explanatory incompati-
bilism”, or the idea that “explanations in terms of reasons are incompati-
ble with explanations in terms of causes” (Kane 1996: 119). The underly-
ing idea here is that only causal explanations threaten free will and moral
responsibility, but since explanations of actions are mostly reason expla-
nations, and reason explanations are not causal, it makes no sense to say
that explanations of actions are excusing.

Thus, if explanatory incompatibilism is correct, reason explanations
(or intentional explanations) would scarcely be excusing. Actions would
not be explained in ways that lend themselves to determinism. The objec-
tion thus attacks the second premise of the argument for HD. We encoun-
tered something similar in the previous chapter, when we discussed the
appropriateness of explaining normal and abnormal action in a symmet-
rical way. Some, it will be recalled, would say that normal (sane, rational)
action requires reason explanations, whereas abnormal (insane, irra-
tional) action requires causal ones.

The problem for explanatory incompatibilism is that plenty of peo-
ple have defended the view that reason explanations are a subspecies of
causal explanations. According to them, reasons can be causes. There con-
sequently is no salient difference between reason explanations and causal
explanations.297 If I am at a meeting where I desire to vote yes to a pro-
sposal, and I believe that raising my arm is the way to realize that desire,
then I have a reason to raise the arm. And it is not at all clear why that
reason cannot be said to be the cause of my action (raising the arm in
order to vote yes) in the same way that a thrown rock is the cause of a
window breaking. The fact that action has a meaning only conferred by
the social context doesn’t seem to change anything about this. The teleo-
logical element in reason explanations, moreover, seems to be a matter of
description. It is true that future states may be what rational action aims
to bring about, such as when we go to the store now in order to buy food
later. But the desire to get food and the belief that going to the store is a

297 The best known defence of this view is perhaps Donald Davidson’s “Action,
Reasons, and Causes” (reprinted in Davidson 2001). Furthermore, see Brown
good way to get food can just as well be thought of as causes of me going – propellers in the back of my action, so to speak, rather than suction devices in front of it. Finally, it is not at all obvious that reasons cannot be given a physical basis, and in that way be similar to events. Materialists about the mind typically assume that having a reason (subjectively), like every mental state, is a pattern of neurons firing in the brain. Accordingly to some accounts, it should therefore in principle be possible to say that having such-and-such neurons firing means having such-and-such reasons (which in turn cause such-and-such actions).

This seems to undermine the force of explanatory incompatibilism somewhat. There at least is nothing obviously absurd or incoherent about claiming that reasons can be causes. However, some argue that the problem is not that reasons can be causes, but rather that reason explanations cannot yield the same kind of regularities as causal explanations of events can. Reason explanations are about processes of practical reason or deliberation, and it is a mistake to think that practical reasoning is a mechanical process. John Searle (2001) has recently defended this view.

Hume famously stated that the connection between motives and actions is “as regular and uniform as that between cause and effect in any part of nature” (Hume 1966: 88). Underlying this claim is the idea that actions are determined by motives, and that once we come to learn a person’s motives fully, we can fully explain and predict his or her actions. Searle refers to this idea as the “Classical Model of Rationality”, and he contests it on the grounds that it portrays human rationality as basically an extension of ape rationality. A particularly strong claim of Searle’s against the Classical Model is his claim that acting on reasons presupposes freedom of the will. On the Classical account, rationality consists in

Note that such an account does not cover what we may call non-subjective reasons. A subjective reason is a reason a person actually has – if somebody believes that black cats are bad luck, to spit three times when a black cat comes in the way is perfectly reasonable given that belief. However, if we say that the same person has reason to believe that such spitting is irrational, as it is a piece of superstition not supported by science, we are dealing with a non-subjective reason. Non-subjective reasons need not be employed by anybody – they are what ought to be done or thought. Only subjective reasons explain actual action.

See also Chisholm 2003: 35; McIntyre 1981: 88-108; Rosenberg 1995: 90.

This is the upshot of the way I have previously analyzed RCT.
choosing the optimal means to an end, which in turn is beyond rational consideration. An agent, $A$, desiring to get end $E$ and believing that means $M$ is the best way of getting $E$, will (invariably) do $M$. Searle thinks that such an account fails to do justice to the processes of practical reason, which are not subject to the same deductive logic as theoretical reason.\footnote{Practical reason concerns questions about what to do, whereas theoretical reason concerns what to believe. It is commonly supposed that theoretical reason is deductive: if I believe $p$, and believe that $q$ is entailed by $p$, then I am committed to believing $q$ by extension. Searle rejects the notion of a similar logic in processes of practical reason, for reasons given below.} When $A$ chooses how to act, $A$ has to decide between different means to different ends, and in addition choose between ends which in turn are often conflicting. This means, Searle claims, that rationality and practical reason presupposes free choice. Searle writes:

“[R]ationality applies only where there is free choice, because rationality must be able to make a difference. If my actions are really completely caused by my beliefs and desires, so that I can’t really help myself, then I have no choice and rationality can make no difference at all to my behaviour. If I am in the grip of causally sufficient conditions, there is no room for me to operate and my actions fall outside the scope of rational assessment” (Searle 2001: 142).

At a psychological level, Searle surely has a point. When we engage in practical reasoning, we assume that the outcome of the reasoning is not given beforehand. We act, so to speak, under the idea of freedom – when we think about doing $X$ rather than $Y$, we have to assume that it is possible for us to do either $X$ or $Y$.\footnote{“Acting under the idea of freedom” is the phrase Dennett uses to capture the notion that we have to assume that we are not governed by determinism when we act; see Dennett 1984: ch. 5. I return to this assumption in the next chapter.} But the psychological impression of freedom might be an illusion, as Searle also admits. But if it is an illusion, it seems false to say that actions are thereby rendered beyond rational assessment. If D is true, outcomes of practical deliberation are determined to happen: I may struggle to decide between shoplifting or not, suffer bouts of paralysis and indecision, have changes of heart, etc. - but in the end the choice will be determined to be as it is. Yet this would not in the least threaten my rationality. We could all, as compatibilists will be quick...
to point out, identify the difference between rational and irrational action even if D were true. A compulsive person being unable to refrain from shoplifting despite desires to the contrary would be irrational, while a normal person who eventually decides what he or she prefers and then sets out to do it would be rational. Searle’s notion that rationality presupposes free will seems false, and one does not have to buy into the very strict Classical Model of rationality to see this. 303

There is thus no reason to assume that there is a salient difference between causal and reason explanations, at least not in the respects that are in focus here. Reason explanations can too threaten personal responsibility. The claim that human action appears less regular and lawful than many other events is true, but we can not use that observation as an objection o HD, for reasons given in the analysis of the objection from probabilism.

8.6 MORAL AND POLITICAL CONSEQUENCES OF HARD DETERMINISM

"Every student of political science is taught that political ideologies are based on theories of human nature. Why must they be based on theories that are three hundred years out of date?"

(Pinker 2002: 305)

The thesis of hard determinism can thus withstand some objections levelled at both its premises. It is a more plausible position than one may think. The next question is whether we should care. Does hard determinism make a practical difference? What are the moral and political consequences of it?

To see the alleged consequences of HD, one may begin by examining why free will – the concept challenged by HD – is important. And

303 Though I do not agree that rationality presupposes free will, I want to emphasise that Searle raises a host of valid points against the Classical Model of Rationality.
that turns out to be quite a mouthful. Kane (1996: ch. 6), for instance, claims that unless we have free will, we cannot be genuinely creative or autonomous; we cannot be truly deserving of praise and blame, nor ultimately morally responsible, and so we cannot be fit candidates for reactive attitudes such as gratitude or resentment; we would not have genuine dignity or self-worth, nor a genuine uniqueness as persons; we would not enjoy life-hopes of the kind only an open future could bestow, and we could not say that we, in the fullest sense of the word, act out of our free will. Perhaps most disturbingly, freely given love or friendship would also be out of the picture. ("My wife married me because of sufficient causal conditions tracing back to before she was born").

Double provides a similar but perhaps a more manageable list of things that makes free will significant:

“Free will is supposed to ontologically warrant the ascription of: (1) our having and expressing reactive attitudes (2) deserved moral praise and moral censure (3) deserved reward and punishment (4) a special degree of dignity that makes us participating members of moral communities [...] (5) a special type of autonomy or self-governance that we do not impute to non-free beings” (Double 1994: 164-5).

Now, if determinism obtains, (1) – (5) seems hard to justify. And since (1) - (5) are very dear to most of us we desperately want to justify them. Hence the fear of determinism. Were it valid reasoning to refute questions of facts with moral arguments, many of us would do so in a heartbeat. But it is not valid. The possible moral unacceptability of determinism does not count as an argument against its truth. So when someone scorns the moral and political consequences of determinism, it is important to note that this is not an argument against determinism as such (unless one assumes, implausibly, that undesirable consequences of a theory counts as evidence against its factual truth). When Berlin famously wrote that,

“There are some terms which, if we took determinism seriously, we would no longer use, or use in only in some peculiar sense, as we speak of witches or the Olympian Gods. Such notions as justice, equity, desert, fairness would certainly have to be re-examined if they were to be kept alive at all and not relegated to the role of discarded figments [...] If determinism is valid, this is a price that we must pay” (Berlin 2002: 15).
it is crucial to see that even if we like the idea of a society of “justice, equity, desert, and fairness”, a determinist could simply bite the bullet and say: “Well, since determinism is valid, this is a price we must pay.”

The consequences of determinism are, as one would suspect, subject to extensive debate. Compatibilists are committed to the view that once we realize the proper meaning of terms such as “responsibility” and “freedom”, determinism leaves our moral and political landscape pretty much as it is. Hard determinists are more divided on the issue, but it is instructive to note that few are willing to espouse radical changes of society. Yes, the truth of determinism means that we are never morally responsible for our actions. But a society without moral responsibility need not be such an alien or scary place, they argue. 304

There are a lot of values being thrown around here, so let’s focus more systematically on certain questions. Most relevant here of course is what the consequences of hard determinism are for the practice of punishment. But it is also interesting to reflect on what the consequences of are for democracy. 305 In chapter 6 I connected retribution to respect for persons. Such respect, I held, presupposes a significance of choice, which we can clearly see expressed in claims that people have a right to autonomy. Given this, it is interesting to ask whether the idea of democracy is incompatible with the truth of hard determinism. I will argue that there is a sense in which democracy is incompatible with HD, and that it is a sense which also, not coincidentally, serves to connect retributive and democratic thought.

Punishment

“Retributivists have reason for dismay”
(Honderich 2006: 160)

There is a wide consensus among hard determinists and others that if hard determinism is true, retributivism must be rejected. This is because the truth of determinism rules out desert – if somebody breaks the rules,

305 We shall thus consider mainly items (3) and (5) on Double’s list above.
this is the necessary result of antecedents beyond the rule breaker’s control. Since the rule breaker could not have done otherwise, it seems plainly wrong to say that he or she deserves punishment in virtue of his or her actions (Pereboom 2001: 159-61).

Honderich (1990) puts a twist on this straightforward answer. As we saw in chapter 3, he thinks that desert-claims in the context of punishment are grievance claims. Somebody does you wrong, and so you want him or her to suffer or pay. Typically this grievance is one of resentment: you resent an offender since you believe that he or she could have chosen not to do wrong, but nevertheless chose to do so. What determinism – or rather the belief in determinism – does is that it undermines the basis for such grievances. You come to realize that the offender could not have but harmed you; that it isn’t really his or her fault. The emotional basis for retributivism is thus tugged away – the reactive attitudes that underlie grievance claims are proven inappropriate. For once we realize that free will is a myth, we indeed, like Pereboom noted, begin to see wrongdoers as similar to earthquakes or epidemics. Hard determinism simply makes the world a more excusing place. “Tout comprendre, c’est tout pardonner” as Madame de Staël said – although in a technical sense she must have intended that to understand all is to excuse all, since to forgive presupposes responsibility.

Retribution is thus out of the picture if HD is true. But what about penal policy in general? How should the state design its penal regime? Should there be a penal regime at all? An extreme answer, which a retributivist who believes in the truth of determinism could perhaps endorse, is that punishment should be done away with entirely. Since no one deserves punishment, and since desert is a necessary condition for just punishment, no one should be punished. Few seem willing to endorse such a conclusion, however. First of all, even if HD is true rules seem still necessary. Hard determinism certainly does not entail political anarchism. But if rules are in place, and if it is better if all complied with them than if they were broken, it seems unavoidable that there should be some sort of response to rule breaking. The state is not obligated to tolerate rule breaking just because HD is true. Some sort of penal regime would still appear necessary.

306 There are certainly a few abolitionists out there who want to do away with the practice of legal punishment altogether, but they are typically not retributivists.
A rehabilitationalist penal regime is straightforwardly compatible with HD, since rehabilitation doesn’t rely on the notion of personal responsibility to begin with. Consider the following from Bertrand Russell, whom we met earlier:

“When a man is suffering from an infectious disease, he is a danger to the community, and it is necessary to restrict his liberty of movement. But no one associates any idea of guilt with such a situation. On the contrary, he is an object of commiseration to his friends. Such steps as science recommends are taken to cure him of his disease, and he submits as a rule without reluctance to the curtailment of liberty involved meanwhile. The same method in spirit ought to be shown in the treatment of what is called ‘crime’” (Russell 1918: 135).

Such a therapeutic approach to crime is obviously compatible with the idea that rule breakers are determined to break the rules and thus without guilt. Their actions are like natural disasters. And although one doesn’t blame natural disasters for the harm they cause, it is perfectly reasonable to take steps to manage and mitigate them. Pereboom along these lines argues that while hard determinism rules out retribution, it is fully compatible with trying to rehabilitate criminals. But if all such efforts to rehabilitate fail, and the risks posed by the criminal are severe, social protection requires that the criminal be quarantined for life, although with minimal discomfort:

“Just as society has a duty to attempt to cure those who are quarantined for its protection, so it has a duty to attempt to morally educate or cure the criminals it detains for its protection. When this is not possible, and a criminal must be confined indefinitely, his life should not be made unnecessary unpleasant” (Pereboom 2001: 186).

Recall that Wolf Singer, the neurologist whom we met in chapter 6, reasons in similar ways: rule breakers are people who suffer from brain abnormalities. They are not responsible for their abnormalities, and so they are not responsible for their crimes. However, since society needs to uphold its rules, it is reasonable to detain rule breakers, treating those who can be treated and placing the dangerous and untreatable in quarantine indefinitely. Rehabilitation, and if need be indefinite incarceration, are
thus responses to rule breaking that are compatible with the truth of HD. They do not presuppose responsibility as retribution does.\footnote{307}

What about deterrence? This penal aim also seems compatible with the truth of HD. As long as people are rational in the sense that they respond to threats, the fact that rule breakers are not morally responsible for their crimes is unimportant. It is true that if HD is correct, the rule breakers who in fact break the rules could not have done otherwise – they were causally determined to do so. It follows that there can be no sense in blaming them in virtue of fact that they deserve it. But deterrence as a penal aim has never been about blaming for the sake of blaming, but about imposing pain or deprivation in order to deter potential rule breakers from future rule breaking. And as long as some potential rule breakers are impressionable in this respect, punishment of the undeserving can be justified on utilitarian grounds. Indeed, even if we assume that there are people who will commit crime quite regardless of each and any penal threat, and who thus could never be deterred, one could still punish the undeterrible in order to deter the deterrable (see Mackie 1990: ch. 10). A deterrentist penal regime, like a rehabilitationalist one, is thus fully compatible with HD.

\textit{Democracy}

I suggested earlier that there is a deep conceptual connection between retributivism and democracy. Both emphasise autonomy, and both essentially rest, or claim to rest, on the idea that people’s choices are worthy of respect. If the thesis of hard determinism challenges retributivism, does it also challenge democracy?

\footnote{307 It is instructive to note that hard determinism and some retributivists here come full circle and coincide. Pereboom actually shares common ground with the retributivists when he claims that HD invalidates retribution since it invalidates moral desert. A retributivist who treats desert as purely a conditional for justified punishment will readily agree. Such a retributivist does not assume that rule breakers in general deserve punishment. But if they do not deserve punishment, they cannot be justly punished – and the truth retributivism explains why. Again, see Corlett 2001.}
This question may seem unwarranted. After all, the question of determinism deals with the “inner” side of freedom – whether we are free to think and act otherwise. But democracy concerns largely the “outer” side of freedom, its social or political side. It concerns what we have a right to do without others interfering, and how we reach political decisions. My freedom to, say, surf the web is threatened if the government restricts and censures access to the internet. But it is not in any understandable way threatened by the truth of determinism. This suggests that the free will problem is of no political relevance. As Scanlon has argued, however, there is a perfectly clear political dimension to the free will problem, which has to do with the “significance of choice as a legitimating condition” (Scanlon 1988: 155). We often say, for instance, that voluntarily entered contracts are binding, that voluntarily taken risks do not warrant compensation, and that freely given consent establishes legitimacy. If free will is a myth, one could argue that no one voluntarily chooses anything – everything is the result of prior sufficient causes. Determinism simply threatens to “deprive choice of its moral significance because (…) the agent’s action is the result of outside causes” (ibid.).

Furthermore, Philip Pettit (2001) has argued the need for comprehensive theories of freedom, like those of Hobbes or Kant. Comprehensive theories of freedom deal with both “inner” and “outer” freedom. In particular, “outer” freedom has to be based on an account of “inner” freedom; how we organize our political life must be coherent with the model of human nature we employ. Hobbes’s defence of the Leviathan and Kant’s defence of a Kingdom of Ends were both ultimately based on their respective analyses of “inner freedom”, Pettit claims.

There is thus a clearly political side to the free will problem. But to investigate the political consequences of HD is more difficult than to investigate its consequences for punishment. True, there moral theories and political ideologies that seem threatened by the thesis – most clearly those that depend on moral desert.308 But what can be said about democracy itself; the idea of collective self-rule? I propose there are two differ-

---

308 Honderich has for example argued that conservatism, which traditionally emphasizes individuals’ moral desert in questions of justice, “is peculiarly vulnerable to determinism” (1990: 244). The same vulnerability would seem to hold for other political philosophies that are desert or responsibility-sensitive, such as luck egalitarianism Scheffler 2005.
ent basic justifications of democracy that need to be discussed here. On the one hand democracy as preference aggregation, which is entirely compatible with the truth of D. On the other hand democracy as the exercise of free choice, which seems incompatible with the truth of D (but is very controversial).

**DEMOCRACY AS PREFERENCE AGGREGATION.** There is an immediate sense of democracy that is not affected in any way by the truth of HD. If democracy is taken to mean – very roughly – the aggregation of individual preferences into a collective outcome, the idea of the preferences being deterministically explainable does not matter. Let us for the sake of simplicity imagine a choice situation where three individuals – A, B and C – choose between alternatives X and Y. (We may imagine that X and Y are policy outcomes, such as in a referendum, or parties or leaders, such as in a representative election). Suppose that A and B vote X and C votes Y. The majority principle holds that X is what the collective {A, B, C} decides. The preferences of the individuals aggregate into a collective outcome.

Now, if this kind of preference aggregation is the basic function and justification of democracy, news that A, B and C were determined to vote the way they did, and could not have voted differently, does nothing to undermine the idea of democracy. What is important in this conception of democracy is its ability to aggregate preferences into collective choices, not that the individual members are genuinely able to vote in different ways.309

If preference aggregation is too slender a justification of democracy, note that much of what we take to be distinctive and right about democracy is also compatible with the truth of D. Democracy is often taken to mean the exercise of free choice. In a (liberal) democracy, people are free to decide who should rule, what to watch on TV, how to raise their children, etc. To grant an individual such freedom of choice is, roughly speaking, to allow the individual to do as she chooses, because she so chooses.310 An individual is granted freedom of choice (in a particular

309 Needless to say, it is of no relevance if the aggregated preferences are the result of a deliberative process or pre-figured.

310 An adjacent concept is that of autonomy, or self-rule. Autonomy shares many things with the concept of freedom of choice, especially if the freedom involved
context) if her choice determines what she does, gets, etc. (within boundaries that all are under an obligation to respect). Thus, I have freedom of choice in voting if I am allowed to vote X rather than Y simply out of the fact that I choose to vote X. By contrast, I do not have freedom of choice if the choice is overruled, for instance because someone thinks that I fail to realize that voting Y is really in my best interest. Proper freedom of choice is thus incompatible with hard paternalism.

In our society it is, loosely speaking, extended to adult, mentally sane and rational human beings. It is not extended to small children, animals, or severely mentally disordered individuals. There are grey areas here, of course: youngsters and mildly disordered are examples of people who are granted some, but not full freedom of choice.

What's important about freedom of choice? Why is choice significant? Scanlon has suggested that choice is significant for three reasons. To choose has a (1) demonstrative value. By choosing, I can demonstrate that I care. It can for example be better to pick out a gift for your spouse yourself, rather than to let someone else pick it out for you. This is because in making the decision yourself, and quite irrespective of the quality of the gift, you show that you have made an effort. Furthermore,

is taken to include aspects which following Berlin are referred to as “positive” freedoms (Berlin 2002). But autonomy brings in a component of authenticity or identification into the picture (see Juth 2005). An autonomous individual is one who governs herself as she chooses, and acts out of considerations that in some sense are authentically her own (as compared to considerations that stem from indoctrination, hypnosis, mental disorder, etc.). I shall return to the concept of autonomy in short.

Thus, we would not say that an individual is free to harm others, nor could we be said to infringe on her freedom by making sure that she does not harm others. To outlaw murder is thus not an infringement on the legitimate autonomy of an individual, although strictly speaking the law in one sense interferes with her freedom of choice.

As we saw in chapter 6, hard paternalism is overriding the choices of those who act knowingly and voluntarily. Soft paternalism is making sure that those who choose are acting knowingly and voluntarily. Hard paternalism constitutes a greater problem, especially for liberal political philosophy, but it should be mentioned that there are some clear cut examples where hard paternalism actually seems to enhance the autonomy of the person being interfered with (see Dixon 2001).
choice is significant since it confers a (2) _symbolic_ message. In being allowed to choose, you are acknowledged as competent – “I may value having a choice because my not having it would reflect a judgment on my own or somebody else’s part that I fell below the standard of competence” (Scanlon 1988: 180). Finally, and most straightforwardly, choice is significant since it has (3) _instrumental_ value. If you are allowed to choose for yourself, chances are greater that you get what you want than if somebody else does the choosing for you.313

Democratic thought usually recognizes the third value of choice as particularly important. A cornerstone in democratic thought, Robert Dahl argues, is namely that individuals are the best judges of their own interests. When it comes to deciding what an individual desires, no one is better equipped to do this than the individual him- or herself.314 All others have to speculate from the outside and are in an epistemologically “disadvantaged position” (1989: 101-2). To grant individuals freedom of choice is therefore the best guarantee that their interests are protected, furthered, or taken into account.315

The same reasoning that explains why individuals should be free to choose also explains why democracy, or the idea of collective self-determination, is a good idea. To let individual preferences aggregate into a collective outcome is the best guarantee that the interest of the _demos_ is similarly protected, furthered, or taken into account. As Dahl notes, democracy at least maximizes the chance that you live under rules you consent to or have chosen (ibid. ch. 10). A collective’s freedom of

---

313 This is not invariably true, Scanlon admits. When you are at a restaurant, it is normally a good idea to choose the meal for yourself – it maximizes the chances that you get something you like. However, “When I go to an exotic restaurant with my sophisticated friends, the chances of getting a meal that accords with my preferences may be increased if someone else does the ordering” (1988: 179).

314 Whereas I prefer to speak of “freedom of choice” to capture this, it should be noted that Dahl uses the concept “personal autonomy.” Personal autonomy for Dahl means, loosely speaking, to be able to make binding decisions about oneself, and being recognised as the best judge of one’s own interests (see Dahl 1989: ch. 7).

315 Mill noted in _Considerations on Representative Government_ that “the rights and interests of every or any person are only secure from being disregarded, when the person interested is himself able, and habitually disposed, to stand up for them” (Mill 1998: 245).
choice can thus be motivated in much the same way as an individual’s freedom of choice: it has, first and foremost, instrumental value.

Note, however, that we have yet to identify a single aspect of democratic thought that is incompatible with HD. Even if people are causally determined to act, think and feel as they do, it could still be advisable to grant them freedom of choice, individually or collectively, since allowing them to choose for themselves maximizes the chances that their interests are furthered. Many are convinced that the significance of choice therefore does not presuppose or require libertarian free will.316

DEMOCRACY AS RESPECT FOR PERSONS. But there might be a different sense of why democracy is important, which indeed is threatened by HD. This sense of democracy has not - at least not only - to do with prudent aggregation of preferences, nor with the instrumental value of choice. It has rather to do with the exercise of faculties that are distinctive to mankind. The reasoning behind this sense of democracy can be described as follows:

Human beings are unique in nature in that they are not at the mercy of their circumstances or instincts. Human beings are capable of genuine autonomy; "the power of agents to be the ultimate creators and sustainers of their own ends and purposes" (Kane 1996: 4). This capacity, which is central to Kantian accounts of autonomy and arguably requires that some version of libertarian free will exists, means that we have a specific dignity or moral worth. Often this dignity or worth is expressed, as in chapter 6, by saying that human beings are persons.

The notion of personhood takes a central but bewildering role in political philosophy. To be a person, it is usually argued, is to occupy a particularly exalted position in the world. Only persons are valued as ends in themselves. Whereas non-persons (non-human animals, notably) may be treated as means to an end, persons are not to be used in a similar manner; whereas non-persons may be treated with what Strawson refers to as the objective attitude, persons should not be similarly treated. Persons, it is often argued, is worthy of respect. They may not be manipulated, controlled, deceived or otherwise used.

To treat persons with respect means, as we have seen, to respect their choices. In Morris’s words: "We treat a human being as a person

316 For explicit arguments to this effect, see Scanlon 1988; Scheffler 2005.
provided: first, we permit the person to make the choices that will determine what happens to him and second, when our responses to the person are responses respecting the person’s choices” (Morris 1968: 492). To respect someone’s choice here means to not only refrain from interfering with them, but also to take the choice seriously — to recognise it as being undertaken by a free and rational equal.317

Now, in the respect for persons-tradition, the special moral status of persons depends on capacities that are specific to them. It is often held that the basis for treating persons respectfully is their capacity for genuine autonomy; their unique status in nature. As Buss writes:

“If we really do have a moral status that distinguishes us from all other animals, then on what can this status depend if not on the fact that we alone are not mere pawns of nature, at the mercy of our instincts? If we are especially worthy of respect, then mustn’t this be because we alone can and do decide for ourselves whether and how our inclinations are really worth satisfying? If there are limitations on how we can justifiably treat each other, then isn’t this, in part, because, independent of our more particular concerns, it is a good thing for each of us to determine our own will?” (Buss 2005: 196)

Suppose, then, that human beings have “a moral status that distinguishes us from all other animals” since we alone “can and do decide for ourselves whether and how our inclinations are really worth satisfying.” Suppose also that this accounts for why, say, a democracy of nonhuman animals or ten year-olds strikes us as preposterous. If this is true, then the ultimate warrant for granting an individual autonomy is precisely that he or she is capable of genuine autonomy. To grant him or her autonomy is the only thing that fits his or her nature. By extension we could say that the ultimate warrant for extending autonomy to a collective is that the collective is composed by such autonomous individuals. Besides all prudential reasons for democracy as a rational and peaceful way of reaching collective decisions, then, there is a deep moral justification of democracy according to this line of thought. Democracy allows the exercise of the

317 This is why, we said, we cannot be said to retributively punish non-persons. A dog wetting the carpet may be grounds for some punitive intervention on our part, but we cannot hold the dog responsible (i.e. blame it). The dog’s “offence” is not one we take seriously: it is not a freely willed flouting of legitimate rules.
uniquely human capacity for autonomy; our ability to “determine our own will”. This conception of democracy does seem threatened by the truth of D because D flatly invalidates that human beings have such a distinctive ability. The notion that human beings have a capacity for free choice – a capacity to choose between alternatives in such a way that the eventual outcome is not prefigured by the circumstances of the choice – is an illusion if D is true. The deep moral justification of democracy we are now considering would then be untenable. There would be no special dignity involved in human agency which requires the exercise of autonomy, either individual or collective. There would at least be no special dignity in the sense that human choice is less governed by causal factors than the behaviour of other creatures. There are different ways to illustrate this line of thought:

First, if your choice between X and Y at t is not genuinely free, but prefigured by the causal etiology of choice, then it is no longer necessary that you’re the one who’s actually doing the choosing. Suppose that a scientist could fully explain and predict your every choice. (This is a really far-fetched assumption of course, but it serves to illustrate the principles involved). The scientist would then be able to choose for you, as you would have chosen. Recall that Scanlon argues that choice is significant in part since it has instrumental value – by allowing you to choose, chances are greater that you get what you want. But if your choices are governed by deterministic processes, it is in principle possible to achieve that instrumental value despite letting others choose for you. 318

To illustrate: if social science were able to accurately predict the outcome of a democratic election (taking into account the possible effects the prediction itself has on voting behaviour), it could provide the outcome of the election without the election actually taking place. The outcome would be the same, and to actually hold the election would merely be an expensive way of going through the motions. By contrast, if voters have free will, the outcome of the election will irreducibly be beyond (full) prediction. If we want to see how the election turns out we have to

318 However, letting others choose for you would rule out the intrinsic pleasure you might derive from the act of choosing. (It would also, speaking with some of the gloomier existentialists, rule out its intrinsic displeasure).
hold it: if we have free will, choice is not a potentially unnecessary intermediate step between deliberation and outcome.

A second reason why the truth of D could threaten the significance of choice is that it could serve to disvalue or demystify choice. Part of what makes choice appear important is that humans are not, in Buss’s words, believed to be “mere pawns of nature.” There might be a mechanism here, though it is very hard to write about in concrete terms, which boils down to this: the dignity of human choice decreases as our ability to scientifically explain it increases, at least for some modes of scientific explanation. Suppose I choose to vote Left in an election. The dignity of that choice may seem threatened if it is proposed that I chose Left because of my parent having socialized me to be Leftist, or that I am genetically predisposed to Leftist values, or that voting Left is an expression of underlying neurosis, or that voting Left was in my material self-interest and I always do what is in my material self-interest. Such explanations seem to reveal the hidden nuts and bolts behind my decision, and it may serve to disvalue it in some respects. It no longer appears to be an exercise of free choice in the sense that I “determine my own will.”

Much, if not all of this is highly contentious. Determinists could object that all choices are explainable in one way or another, and that the “dignity” of one’s choice depends solely on the nature of the causation involved. Unless the claim that we “determine our own will” means that we choose for reasons we accept as our own, rather than out of propaganda, brainwashing, peer pressure or the like, the claim that we “can and do decide for ourselves whether and how our inclinations are really worth satisfying” may seem hopelessly muddled. To require that choices are undetermined in order to have full significance is nothing more than to require that they are unpredictable, random - impossible to explain beyond the useless truism that somebody chose X because he or she chose to choose X. It is basically to treat choices as something magical or inherently mysterious, determinists argue. Why should the moral significance of choice - or democracy for that matter - be taken to rely on such dubious foundations?
The unifying account of B.F. Skinner

The consequences of HD for democracy are very difficult to ascertain. There is at least one sense of democracy where the truth of HD is completely immaterial: that democracy is simply the extension of freedom of choice to a *demos*, motivated instrumentally by the fact that it aggregates preferences into political outcomes. I have hinted, however, that there might be a deep conceptual linkage between free will and democracy when democracy is conceived as the only political system compatible with respecting our autonomy. This linkage also serves to connect the question of punishment with that of democracy. I have suggested that democracy and retributive punishment share common ground in that both practices essentially are about respecting choices. But the potential truth of D seems to undermine the significance of choice. If this is true, to question the person assumed by retributivism – the creature of the Autonomy Model - is also to question the person assumed by some democratic theory.

Support for this contention comes from surprising places. The great behaviourist B. F. Skinner, sadly underrated as a political theorist, thinks in ways similar to mine - if only to discard the notion of what he calls the “autonomous man” entirely. We shall here take a brief look at Skinner’s ideas, which convey a compelling, unified account of the moral and political consequences of hard determinism.

B. F. Skinner is famous for his radical behaviourism: the view that science can explain human behaviour without making references to internal mental states. Echoing such theorists as Hobbes and Bentham, Skinner envisioned human beings – indeed, all organisms – to be ruled by pleasure and pain. They are attracted to pleasant stimuli and repelled by painful stimuli. Through his experiments on the associative learning of organisms (mainly pigeons and rats), Skinner was convinced that one could shape and change behaviour of human beings by controlling the environment – indeed, Skinner claimed that “the variables of which human behaviour is a function lie in the environment” (Skinner 1977: 1). He called such environmental control *operant conditioning* (a term he overtook from Edward Thorndike). Operant conditioning is simply the idea...
of using reinforcers and punishments to increase or decrease the occurrence of behaviours.\textsuperscript{319}

What is interesting about Skinner is that he tried to draw moral, social and political conclusions from his scientific views. Two books of his are especially interesting: Beyond Freedom and Dignity, in which Skinner spells out his thoughts about mankind and society and attacks the doctrine of autonomous man, and Walden Two, his fictional work about a non-democratic utopia designed on the basis of scientific principles.

Skinner begins Beyond Freedom and Dignity (1971) by lamenting the fact that the thing man understands the least is man himself. Greek theorizing about nature is today only of historical interest; it has given way to massive developments in natural science. When it comes to human behaviour, however, Skinner notes with embarrassment and disgust that we still use many of the ideas of ancient Greece. There has been no development in the science of human behaviour comparable to that of the natural sciences.

Skinner attributes the underdevelopment of the social sciences to their reliance on what he calls the doctrine of “inner” or “autonomous” man. This doctrine expresses the idea that people have a self that is active, free and capable of origination. For Skinner, this idea, which as we can see resembles the rule breaker postulated by the Autonomy Model, is an unwarranted assumption. It rests on ignorance and stands in the way of a proper science of man. Here is how Skinner describes the genesis and function of the autonomous man:

“Unable to understand how or why the person we see behaves as he does, we attribute his behaviour to a person which we cannot see, whose behaviour we cannot explain either but about whom we are not inclined to ask questions. We probably adopt this strategy not so much because of any lack of interest or power but because of a longstanding conviction that for much of human behaviour there are no relevant antecedents. The function of inner man is to provide an explanation which will not be explained in turn. Explanations stop with him. He is not a mediator between past history and current behaviour, he is the center from which behaviour emanates. He initiates, originates, and

\textsuperscript{319} Skinner allowed for the importance of biological traits and genetic predispositions, but emphasized mainly environmental factors as an explanation of human behaviour.
creates, and in doing so he remains, as he was for the Greeks, divine. We say that he is autonomous – and, so far as a science of behaviour is concerned, that means miraculous.” (ibid. 14)

The more we come to learn about human action, Skinner contends, the less reason there will be for assuming something like the autonomous man: “Autonomous man serves only to explain the things we are not yet able to explain in other ways. His existence depends upon our ignorance, and he naturally loses status as we come to know more about behaviour.” (ibid.)

Commitment to the “autonomous man” not only hinders scientific progress, Skinner argues. It also stands in the way of more effective ways of dealing with the problems of society. Inner man is employed to secure the freedom and dignity of human beings, but in cementing notions such as praise and blame, it is no longer a progressive force but an unwelcome one. Whereas the doctrines of human freedom and dignity once served as an impetus in the fight against the aversive control of tyrants, they today serve to keep the mythical notions of autonomous choice and moral responsibility, and its associated ill-conceived practices, alive (Skinner was particularly sceptical of the practice of punishment, which he saw as inferior to positive reinforcement). There are better ways to organize society, Skinner contends, and in Walden Two he sets out to describe a utopian society based on a scientific conception of man.

Walden Two (1976) is a fictional story, the details of which need not concern us. The main characters are professors Burris and Castle, who visit the community of Walden Two, where Frazier, a former academic, has created a small rural community based on the principles of behaviourism. As the story unfolds Castle, a philosopher, remains sceptical of Frazier’s creation, while Burris is gradually taken in by the small-scale and self-sufficient community, with its apparently happy inhabitants, four-hour workday and rewarding leisure time. As the story ends, Burris

---

321 About where Skinner’s sympathies with the characters lie: “Burris”, a psychologist, is very close to “Burrhus” – Skinner’s first name. It is safe to say, however, that Frazier represents Skinner’s ideas. Castle serves as the humanist-moralist antagonist to Frazier in the novel, and is basically there to be proven wrong by Frazier.
returns home but realizes Walden Two’s supremacy over modern society, and goes back to live there.

What is interesting about Walden Two – apart from it being a surprisingly enjoyable read - is that it is a classical Platonic argument made in a slightly new way. Skinner shares Plato’s disregard for democracy but substitutes the Philosopher-King with the Psychologist-King – what constitutes the best ruler is not an affinity to the world of ideas, but an expert command of the scientific facts of human nature. Like Plato, Skinner is profoundly collectivistic – indeed, even organic - in his thinking. The purpose of a society is to strive for the common good. The good of the community as a whole comes before the good of this or that group or individual. Fostering such collectivism requires a number of institutional arrangements: children are raised communally in order to overcome the importance of biological parenthood; strict economic equality is pursued through a “labour credit” system for work; and competition is downplayed (indeed, only Skinner allows two games in Walden Two – chess and tennis - but these are never played in tournaments). All of this is intended invoke cooperative “community first” attitudes in the members of the community. The strategies are very similar to those of Plato.

The political structure consists of Planners, Managers and the Rest. The Planners are the highest decision-makers and are responsible for both the election of new Managers (people in charge of different sections of work, such as Dairy or Health) and the election of new Planners. The Planners are thus not democratically elected, nor are they democratically accountable. Although Skinner emphasizes that the Planners listen and are sensitive to the people, there is strictly speaking no democratic way for the Rest of replacing them. Why would people accept such a non-democratic society? Because they would “feel free” in a society where they get what they want. Like Plato, Skinner thus believes that people would consent to the rule of the competent once they see its benefits.

322 Walden Two is also interesting since it has spawned real-life epigones. Los Horcones, Mexico, is a community based on Skinner’s ideas. Visit them at their website: [http://www.losborcones.org.mx/](http://www.losborcones.org.mx/)

323 Those who are neither Planners nor Managers are referred to as Scientists or Workers (Skinner 1976: 48-50).
Of direct relevance for our purposes is Skinner’s treatment of freedom, punishment and democracy. He offers a unifying account where democracy and punishment are both done away with for the same reason: the scientific invalidity of autonomous man. Skinner argues that a good society is one where some basic criteria of the good life are met, but adds that democracy is not the best conceivable system in this respect.324 Frazier, the protagonist in *Walden Two*, at one point declares:

> “But the triumph of democracy doesn’t mean that it’s the best government. It was merely the better in a contest with a conspicuously bad one. Let’s not stop with democracy. It isn’t, and can’t be, the best form of government, because it’s based on a scientifically invalid conception of man. It fails to take account of the fact that in the long run *man is determined by the state*” (Skinner 1976: 257).

In discarding democracy, Skinner makes two points: one about politics in general, and one about democracy as “based on a scientifically invalid conception of man.” Politics in general is poor since it is more of an art than a science. It relies on guesses rather than experimental knowledge. Frazier scorns:

> “Once in a while a new government initiates a program to put power to better use, but its success or failure never really proves anything. In science, experiments are designed, checked, altered, repeated - but not in politics. Hence our extraordinarily slow progress toward a science of government. We have no real *cumulative* knowledge. History tells us nothing. That’s the tragedy of the political reformer. He has nothing to work with but a spurious science of history. He has no real facts – no real laws. A pathetic figure!” (ibid. 181)

But if society is to be ruled according to scientific principles, there had better be some real facts and laws to experimentally ascertain: “You can’t have a science about a subject matter which hops capriciously about”, as Frazier declares at one point (ibid. 242). And this is where the denial of autonomous man comes in. Skinner explicitly rejects free will. Man is

324 The good life, Frazier asserts, consists of health, a minimum of unpleasant work, the chance to exercise talents, satisfying personal contacts, and relaxation and rest (ibid: 146f).
determined to act as he does, mainly by the environment. By using operant conditioning – by controlling the environment – the scientist can determine how a person acts. Walden Two is founded on the idea of such environmental control. But the inhabitants of Walden Two do not feel controlled. They always do what they want. But due to clever social engineering on behalf of the Planners, what they want to do is invariably to the benefit of the community. They have simply been conditioned to want what the Planners want them to want.325

The scientific understanding of human behaviour is thus what enables a better way to rule society, according to Skinner. There is no autonomous man. This means that there can be no special moral status of democracy, such as being the only form of government that allows the exercise of the uniquely human ability of autonomy. Man is governed by determinism, democracy or not. And while a case could be made for the merits of democracy as a way of aggregating preferences or achieving the good life, Skinner argues that democracy is inferior to the kind of political system adopted in Walden Two. It is inferior since it doesn’t build on scientific experimentation, but on power politics and guesswork.

As for the practices of praise and blame, Skinner does away with them for the same reasons. The notion of personal responsibility must too be revised in accordance with the scientific conception of man. In Beyond Freedom and Dignity, Skinner observes that, “It is the environment which is ‘responsible’ for the objectionable behaviour, and it is the environment, not some attribute of the individual, which must be changed” (1971: 74). In Walden Two, this means that practices of praise and blame are discarded entirely. Personal accomplishments are not praised just as personal failures are not blamed. Those who do a poor job are replaced. Those who disobey the rules are offered treatment rather than punishment. Fairness and progress requires, Skinner contends, that “the old notion of personal responsibility” be discarded (1976: 159-161).

325 Kane refers to the kind of control exercised by Frazier as CNC-control (“covert nonconstraining control”), the essence of which is to make sure that people only want to have what they can get and choose to do what they are allowed to do (Kane 1996: 67-68).
8.7 CONCLUSIONS

In Skinner’s writings we find a curious and largely alien take on moral and political matters. From the truth of determinism and the rejection of autonomous man he concludes things about democracy and punishment which are very interesting, but also disturbing. His denial of free will means that people are never responsible for their actions and never deserving of praise or blame. Wrongdoers are, just as Pereboom has said, similar to “earthquakes or epidemics.” Retributivism is of course unequivocally out of the picture if this is true. Furthermore, the causal factors that govern people’s behaviour make possible a kind of science of behaviour which can and should replace democratic rule. There is nothing special about human nature or action that renders democracy particularly fitting, Skinner argues – free will is a myth, so the exercise of free will cannot be used as a moral justification of individual autonomy or democracy.

I want to be perfectly clear on one point: I do not here claim that acceptance of HD entails a rejection of democracy. There are plenty of ways in which one could defend democratic rule – direct or representative, deliberative or party-based – even if one believes that autonomous man and free will are myths. There is, however, one justification of individual or collective autonomy which the truth of HD seems to rule out: the notion that it allows the exercise of the uniquely human ability of autonomous choice.

I also want to emphasise that the idea of making the significance of autonomous choice reside in its undetermination is highly contentious. First of all, there might be examples of determinism-in-action that are perfectly benign. Assume for instance a person who is always determined by good reasons – she always chooses the alternative which the balance of reason supports. Suppose such a person always does the right thing; she is literally incapable of doing wrong. Is there any reason to think less of that person’s actions just because they are determined? Wouldn’t we on the contrary admire her for her unwavering moral compass?

326 It has sometimes been suggested that a person who has not been presented with the opportunity to do evil cannot be raised for doing good (Kant at one
choices, we might be dangerously close to supporting the so-called “liberty of indifference”, which compatibilists have rightfully scorned. Are we the most free when maximal uncertainty reigns? Is the choice between X and Y the most free when the likelihood of choosing either alternative approximates 0.5?

These difficulties notwithstanding, the thesis of hard determinism introduces ideas which seem to require substantial revision of a number of practices in place today, not least those who concern penal policy. Since HD challenges the notion of personal responsibility, it seems difficult to defend a retributive penal regime based on giving rule breakers their (nonexistent) desert. But as we have seen, in challenging the idea of autonomous man HD also challenges a much wider set of concerns than simply punishment. It invalidates much of our traditional ways of moral thinking, such as the general and fundamental notion that humans are morally free beings whose choices carry a special significance and warrant our respect. I have in this chapter defended HD against some objections. The conclusion is that the thesis of HD is stronger than what one initially might think. Does this mean that we have to accept its truth and revise society in accordance with its conclusions, many of which seem hard to stomach? I think not. In the next chapter I will try to provide a way of rejecting hard determinism as a basis for policy.

point wrote that many blameless individuals have simply been fortunate to escape temptation; see Chisholm 2003: 36). The idea is here that good deeds are not worth much unless there is a genuine opportunity to evil ones. Sometimes this seems correct – when Alex, the main character of A Clockwork Orange, is conditioned into being unable to do evil, he certainly does not warrant being called “good.” But the notion of praiseworthiness depending on the opportunity for blameworthiness sometimes seems questionable. Suppose that a parent, when witnessing his or her child drowning, cannot but attempt to save the child. It is difficult to see why the opportunity to refrain from saving one’s child would be valuable. On the contrary, that the parent had to attempt to rescue the child - was unable not to - seems more a praiseworthy trait than something else.
9. Betting Against Hard Determinism

“The question now becomes a different one: not about what we know, but about how it is rational to behave given that we do not know. This question is one of practical decision-making against a background of uncertainty”

(Richards 2000: 34)

If it is true that explanations are excuses, we seem propelled towards the position where acts of rule breaking, as well as actions in general, are such that one cannot be said to be personally responsible for them. In the previous chapter I defended hard determinism, the controversial view that determinism is true and that responsibility therefore does not exist, against three common objections: various versions of compatibilism, probabilism, and the notion of explanatory incompatibilism. These objections, I argued, do not suffice to reject the thesis of hard determinism.

The purpose of the present chapter, however, is to reject that thesis. There are at least three ways in which one could attempt to do so. First, one could deny that determinism is true. This I shall not do. I shall at least not conclusively say that determinism is false. (I shall, however, exploit the fact that we do not know whether it is true). Secondly, one could deny that determinism is incompatible with free will and personal responsibility. As indicated in the previous chapter, I do not find this strategy plausible. A third way, and the way I opt for here, is to treat advocacy of hard determinism as a bet of sorts, and argue that there are
moral and political reasons for betting against it. My argument relies on three steps:

1) Whether we believe that, and act as if, hard determinism is true is a different question than whether hard determinism is true. The two questions are logically distinct. One may believe that, and act as if, determinism is true even though it is false.

2) A society where belief in free will and personal responsibility is prevalent is, other things equal, normatively preferable to a society where it is not.

3) We do not know whether determinism is true or false.

These steps are all indispensable for the argument, so denying any of them means that the argument will not work. I shall argue that since we do not know whether determinism is true or false, and since it would normatively be better if determinism was false, we should bet that it is false.

I want to immediately clarify that I am not here advocating any kind of “useful illusion”-argument; I do not claim that we should act as if a theory is false even though we are convinced that it is true, because doing so would provide certain benefits. Rather, my claim is that since we cannot say for sure whether it is true or false, moral and political considerations suggest that we should bet that it is false. This is not acting against our better judgement. It is to make a justified choice under uncertainty. The logic of the argument works best, as we shall see, if we as-

327 In other words, since I shall not contest the premise about incompatibility between free will and determinism, nor assert that determinism is false, there is a sense in which I am not attempting to reject the argument of hard determinism (which is clearly valid by the way, so there is nothing to contest there). What I do instead is that I connect hard determinism with taking action and changing policy; I shall argue that since we do not know whether determinism is true or false, we should refrain from acting on the recommendations of hard determinism. This is an argument not so much against hard determinism as such, but rather against hard determinism as an idea to act on.
sume that the principle of insufficient reason applies to whether determinism is true or false.

9.1 STEP 1. TWO INDEPENDENT QUESTIONS OF THE THREAT FROM DETERMINISM

It is important to realize that determinism as such has no threatening consequences of the kind people sometimes worry about. Only belief in determinism does. If we fear that determinism will reshape our societies, and in particular invalidate the notion of personal responsibility and our practices of praise and blame, it is important that these changes will not, strictly speaking, happen because determinism is true, but because people in general and policymakers in particular believe that it is true, and act in ways they perceive as coherent with its truth.

Though this point is rather obvious, one sometimes gets the feeling that some tend to overlook it. Consider again Berlin’s classical statement, which was quoted in the previous chapter:

“There are some terms which, if we took determinism seriously, we would no longer use, or use in only in some peculiar sense, as we speak of witches or the Olympian Gods. Such notions as justice, equity, desert, fairness would certainly have to be re-examined if they were to be kept alive at all and not relegated to the role of discarded figments [...] If determinism is valid, this is a price that we must pay” (Berlin 2002: 15).

Berlin here begins by suggesting that “if we took determinism seriously”, we would no longer use certain terms. He then adds that “this is a price we must pay” if determinism is “valid.” It is important to note, however, that the question of validity is neither necessary nor sufficient for the changes Berlin fears. All that is needed for the changes to occur is that we believe that determinism is true and act as if it’s true. This we may do even if determinism turns out to be false.

We thus need to distinguish between two types of questions:

Q1: Is determinism true or false?
Q2: Is determinism believed to be true or false?
Regarding Q1 and Q2, a second thing should be noted: if determinism is true, whether it is believed that determinism is true or false is also determined. This is self-evident: if determinism is true, but people in general believe that it is false, then their belief that determinism is false is itself determined. This is why determinism as such poses no threats to our practices: we might be determined to believe that free will and personal responsibility exist just as we might be determined to believe the opposite. Sometimes theorists write as if the truth of determinism admits some choice about how we respond to determinism. But this is incorrect: if determinism is true, our reactions to determinism are determined to be as they are. Consider the following statement by Fischer, a well-known compatibilist:

“[I]f I were to wake up tomorrow and read in the Los Angeles Times that scientists have decisively proved that causal determinism is true, I would not have any inclination to stop thinking of myself, my family and friends, and human beings in general as morally responsible. [...] Our reactive attitudes should not be held hostage to an esoteric scientific discovery of the kind in question. That is, the reactive attitudes, and our view of ourselves as morally responsible agents, should be resilient in a certain sense” (Fischer 1999: 129).

Fischer here seems to suggest that we have some leeway when it comes to how we react to the belief in determinism. Our “reactive attitudes” – love, gratitude, respect, resentment, and so forth – “should be resilient.” But if determinism is true (Q1), the “reactive attitudes” either are or are not “resilient.” We cannot genuinely choose how to react to the truth of determinism if determinism de facto is true. That determinism is proved to be true by scientists would itself be determined, as would our reaction on the proof in question, the fate of the reactive attitudes - and the reasoning of Fischer. Again, we might very well believe that determinism is false as a result of deterministic processes. If determinism is true, all reactions in Q2 are to be deterministically explained.

Now, bearing in mind the distinction between Q1 and Q2, we get a 2x2-matrix. I here assume that Q1 is dichotomous – determinism is either

---

328 Cf. Boström 1996: 149
true or false (with respects to human action). More problematically perhaps, I also assume that Q2 is dichotomous: any given agent will either believe that determinism is true or that it is false. For simplicity, I also assume that people in general, in particular policymakers, either believe that it is true or false, though this is of course more contestable.  

9.1 The separateness of Q1 and Q2

<table>
<thead>
<tr>
<th>Q2 – Belief in the truth of determinism</th>
<th>Yes</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>False</td>
<td>(3)</td>
<td>(4)</td>
</tr>
</tbody>
</table>

Note that if determinism is true, whether we end up in box (1) or (3) is itself determined; it is not subject to choice. Box (1) is the outcome where determinism is true and we are determined to believe so. Box (2) is the outcome where determinism is false but we nevertheless believe that it is true. Box (3) is the outcome where determinism is true and we are determined to believe that it is not. Box (4), finally, is the outcome where we believe that determinism is false – and it is.

329 It is contestable for a number of reasons. First of all, we may expect many people to have no opinion whatsoever over the matter. Secondly, among those who do have an opinion, we may expect some people to be genuinely undecided about whether determinism is true or false. Thirdly, a society may exhibit varying proportions of determinists and indeterminists.
9.2 STEP 2. A FREE WILL SOCIETY IS PREFERABLE TO A DETERMINISTIC SOCIETY

Having distinguished between Q1 and Q2, the next step is to argue that a society in which people in general believe that determinism is false is preferable to a society in which they believe that it is true.

This step might sound strange. If determinism is true, how could we say that it would be preferable to live in a society where free will in general is believed to exist? If we realize that determinism is true, aren’t we obliged to change our moral concepts and political institutions? In particular, if we realize that people are never able to do otherwise, aren’t we obliged to reject the concept of desert? Since we are not personally responsible, surely we cannot justify distributing praise and blame on the basis that people deserve it. So to claim that a society is better when it is not designed to fit the truth of determinism seems to be nothing more than to advocate useful illusions. It would perhaps be prudent to hold people responsible even though we know that they are not responsible. But we would be acting against our better judgement, and in a straightforward sense we would be committing an injustice.

This is correct. But when I claim that a society where determinism is believed to be true, and acted upon as if it was true, (hereafter a “D-society”) is normatively worse than a society where determinism is believed to be false (hereafter a “F-society”, as in “Free will-society”), I do so without taking the possible truth of determinism into account. This might seem awkward, for surely it is correct to say that the truth or falsity of determinism has consequences for how a society ought to be fashioned: If determinism is true, a D-society would be preferable; if false, an F-society. But I am here considering the merits of a D-society and an F-society while remaining agnostic about the truth and falsity of determinism. I judge them, so to speak, on their own merits. One might state it this way: does it seem preferable to live in an F-society if determinism is false than in a D-society if determinism is true?

Thus conceived, I do think one could make the case for why an F-society would be preferable to a D-society for basically the same reasons as in previous chapters. In a D-society a retributive penal regime cannot be upheld, so it is questionable if the institutional benefits of living under a penal regime based on desert will be enjoyed. We saw that hard determinists are especially fond of rehabilitation and incapacitation as re-
sponses to rule breaking, and Rachels provides a compelling case for why those measures can have dangerous implications. Furthermore, in a D-society we are not subject to reactive attitudes. Our good actions are not admired, and our bad actions are not resented – which boils down to the same thing; our choices – and status as free and responsible persons – are not respected. Of course, if determinism is true, the basis for that respect is false and should be rejected. But other things equal it is valuable, I have argued, to have one’s choices respected. It is flattering to be held to normative standards, even if this means that we are sometimes blamed or punished. Moreover, I suggested in the last chapter that the respect for choices might be a basis for democracy. And although I would not want to go as far as to say that the D-society has to be undemocratic, it cannot justify democracy on the ground that it is the only political system which respects the unique moral significance of persons’ choices. In sum, in the D-society we cannot take credit for our successes or blame for our failures. This makes us seem, in a lofty phrase, lessened as human beings.

The F-society, on the other hand, is a society where it is often (though not always) assumed that we have a choice about what to do and what to think. This means that we are personally responsible for what we do or think, and may deserve praise for our successes and blame for our failures. This expresses a kind of respect for us as persons, which will sometimes work to the benefit of our immediate interests, and sometimes not, but invariably express an attractive and empowering message - “I am not a vector sum of forces acting on me, but an autonomous agent.”

The F-society is basically the kind of society we live in today: a society of reactive attitudes which gives prominence to the notions of responsibility, praise and blame. It is possible that scientific advances will prove the F-society to be based on myths. But it has moral and political features which seem not only familiar, but also attractive. A few, like B.F. Skinner, see nothing but improvements in discarding the “old notion of personal responsibility” and wants to purge society of the prescientific muddle associated with it. My claim here is that there are perfectly understandable reasons why one would want to resist such a development, however, and conclude that it would be better to live in an F-society.
So far I have made the distinction between Q1 and Q2 and argued that the F-society is preferable to the D-society. This doesn’t mean that we should prefer the F-society in face of overwhelming evidence that determinism is true. The final premise in my argument, however, is that no such overwhelming evidence exists.

As I have previously said, most contemporary philosophers are compatibilists about free will. This usually means that they are determinists about human action: they regard libertarian free will - the ability to do otherwise - as inherently suspect, and instead analyze freedom in terms of the ability to do what one wants. It is important to note that compatibilists as rule place the burden of proof with the libertarians. Determinism is the common sense view, they argue, and it is those who defend libertarian free will who must explain themselves. The principal reason for placing the burden of proof with the libertarians is found in compatibilists’ commitment to so-called inclusive naturalism, or the view that human beings and human behaviour are completely included in the natural world. According to inclusive naturalism, we cannot accept the existence of a non-material soul (or something similar) that resides over and above the natural world. Just like everything else, the human mind is a fully natural phenomenon, and it plays by rules no different than other parts of nature. This means that there can be no privileged place in humans for “free will” to reside. In Brian Barry’s words:

“If the denial of determinism requires the belief in a ‘mind’ possessed of a ‘free will’ which enables it to induce human actions contra-causally, it becomes increasingly hard to see how anti-determinism can continue to be sustained except by resorting to the belief in five billion daily miracles” (Barry 1990: lxii).

Now, it is perhaps surprising that determinism is taken to be the common sense view at a moment in time when scientists, via the consol-
dated standing of quantum physics, are more convinced than ever that the basic foundations of nature are indeterministic.\textsuperscript{331} There are quantum phenomena going on in all of us (we all ultimately consist of particles), so the notion that there might be indeterminacies in our agency should not be inconceivable. As a result of quantum physics, few today are universal determinists. Some are determinists about human action, however, arguing that indeterminacy on a subatomic level does not affect our behaviour, which takes place at a “global” or “macro” level. To them, some sort of psychological determinism might well be true even though physical determinism is false. The majority of writers, however, conclude that even if quantum indeterminacies affect our behaviour, this does not count as evidence of “free will.” It only shows that some aspects of our behaviour happen by chance and others by necessity. Free will, in the sense of being genuinely able to do otherwise, remains obscure; very much a thing in need of clarification.

Empirical social scientists, to the extent they think and write about the nature of agency, tend to be sceptical of determinism (although some argue that they are nevertheless committed to a worldview that is tacitly deterministic; see Wendt 2006). As we have seen, appeals to probabilism are frequent, and few social scientists are willing to advance deterministic explanations. The reason for this is straightforward. The old covering law model of scientific explanations, which stated that successful explanations take the form of sound deductive arguments which subsume particular occurrences under covering laws or lawlike generalizations, was too demanding.\textsuperscript{332} Social reality in general and human action in particular does not seem regular or fixed enough to justify talk of scientific laws. We only have observed patterns of varying uniformity: for every proposed explanation we seem doomed to find exceptions and anomalies. Thus the probabilistic turn in the social sciences.

\textsuperscript{331} Apparently there is some debate over whether quantum physics is indeterministic – in particular David Bohm is famous for defending a deterministic hidden-variable interpretation of quantum mechanics – but the majority of physicists seem to accept the indeterministic interpretation.

\textsuperscript{332} The covering law-model is of course Hempel and Oppenheim’s Deductive-Nomological Model of scientific explanations. See Hempel & Oppenheim 1948; Hempel 1965. For an account of the influence this model exerted on early political science, see Gunnell 2006.
That the science of human action, social scientific or otherwise, has conspicuously failed to identify non-trivial laws or lawlike generalizations is widely admitted.\(^{333}\) It is possible that this merely testifies to our inability to see the regularities in the vastly complex system which is the social world. But even so, the failure to discover laws of human action does make one wonder about the plausibility of treating determinism as a common sense position. There certainly seems to be little observational basis for inferring that human action is governed by determinism.\(^{334}\) Determinism is the thesis that every occurrence (event or action) has sufficient causes; that every occurrence is determined to occur given the conjunction of prior events and causal laws. But if this is true, it seems as if we should be able to identify lawlike regularities in human action. And since knowing the initial conditions and relevant laws would allow us to forecast future events, it furthermore seems as if we should be able to predict human behaviour with accuracy.\(^{335}\) But we seem unable to do either: the explanations remain marred by anomalies, and accurate predictions are conspicuously missing. Is it plausible, given this track re-

\(^{333}\) See Elster 1998; Flyvbjerg 2001: ch. 4; Goodin & Klingemann 2000: 9; Shapiro 2002: 190-1; Rosenberg 1995: 13f.

\(^{334}\) Humphreys notes: “As a psychological habit acquired from everyday experience, a belief in universal determinism is peculiar in the extreme, because events in the ordinary world are neither sufficiently predictable nor sufficiently under our control that an inference of inviolable regular succession would be warranted from our experiences” (Humphreys 1989: 17).

\(^{335}\) Explanation and prediction are symmetrical according to the covering law model: with knowledge comes the ability to explain past events and predict future ones. The classical thought-experiment here is of course LaPlace’s observer: if universal determinism is true, then, “Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situations of the beings who compose it – an intelligence vast enough to submit these data to analysis – it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present for its eyes” (Laplace, cited in Dennett 1984: 50). It has been objected, however, that even the Laplacean intelligence might be unable to predict the future even if it was able to fully explain the past – for, as any social scientist will quickly remark, one also have to predict the effects of making a prediction.
cord, to treat determinism as the common sense view about human behaviour?

Those who are sceptical about libertarian free will can argue that inconclusive empirical results prove nothing. To the extent that quantum indeterminacy affects social phenomena or human action, we should not be surprised if uniform laws and accurate predictions are lacking. And even if human action is fully governed by determinism, nothing says that proving this in a complicated social context, where no laboratory control exists, is going to be easy or straightforward. Above all, sceptics could argue that inconclusive results do nothing to prove libertarian free will. That is a separate point and the fact remains that libertarian free will may seem “essentially mysterious and terminally obscure” (Kane 1994: 28).

Libertarians could respond that the psychological feeling of being able to choose between alternatives is overwhelming, and that we could discount this as evidence only by assuming that people are systematically deceived about themselves. As for the charge that free will is obscure since it violates inclusive naturalism, libertarians have tried to reconcile free will with a naturalistic understanding of the world (Kane 1996; Searle 2001; Balauger 2004; Wendt 2006), typically by locating the relevant indeterminacy in chaotically amplified quantum events. These attempts are admittedly often strange, but they exist.

What I am suggesting here is that we do not know if determinism is true of false when it comes to human action. One can suspect that we shall never come to know this. The crucial point is that libertarians and determinists both have to add unsubstantiated assumptions when judging whether an action is determined or not.336 Consider the average bank robber, Rob. Rob has no particular job skills, lacks an education, and has to finance a costly substance abuse. Being inclined to crime to get what he wants, Rob decides to rob a bank in order to get money, and proceeds

---

336 It will not do to claim that the libertarian, unlike the determinist, resorts to metaphysics. As made clear by the critique of logical positivism back in the day, to say something allegedly anti-metaphysic such as “there is only matter” or “there is nothing beyond the natural” is of course blatantly metaphysical. We cannot observe whether or not such statements are true; we assert them. Thus conceived, the statements “there is a libertarian free will” and “there is no libertarian free will” are equally metaphysical, or, at the very least, equally unverifiable.
to do so. Now, was Rob determined to rob the bank or not? Libertarians and determinists will agree that Rob’s action is intelligible, and that causal factors such as his substance abuse and social background help to explain it. But the libertarian will hold that Rob, given exactly the same past, could have chosen not to rob the bank, while the determinist will hold the opposite. How could we possibly tell which position is correct? Observe what people like Rob do in similar situations? Alas, neither the situations, nor the people, will be identical. Observe what Rob himself have done before in similar situations? Alas, time will have changed Rob even if the situations are identical in other respects.

Intuitions clash about whether libertarian free will is conceptually absurd or not, and it seems inherently difficult to prove or disprove it observationally. A substantial improvement of science’s ability to explain and predict action would suggest that causal laws do govern what men do. But in the absence of such scientific advances, I think that the sensible conclusion about the truth or falsity of determinism is: we cannot at this point tell for sure.

9.4 CONCLUSIONS: BETTING AGAINST HARD DETERMINISM

If the three steps above are correct, we can make the following argument: The F-society is preferable to the D-society on normative grounds. We do not know whether determinism is true or false, but we know that we may act as if determinism is true even if it is false. So when we choose between adopting the D-society or the F-society, we are essentially making a bet about under uncertainty. We know that we may act as if determinism is true when it is in fact false, and we know that we may act as if determinism is false when it in fact is true. But given that we do not know whether determinism is false we should bet that it is false and adopt (or uphold) the F-society.

Consider another the matrix, where Q2 has been modified to represent whether or not we bet that determinism is true:
9.2 Four outcomes of the bet

Q1 – Determinism

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>No</td>
<td>(3)</td>
<td>(4)</td>
</tr>
</tbody>
</table>

What hard determinism presents is actually a kind of bet with four different outcomes (keeping in mind that if determinism is true, the bet is also determined). I assume here that we should believe that, and act as if, determinism (hereafter D) is true if D is true, and that we should believe that, and act as if, D is false if D is false. If D is true, it would be better to act as if it was true, and if D is false it would be better to act as if it was false. This is what I meant by not advocating useful illusions – I shall not argue that we should act as if D is false even if it is true. Though there might be reasons in favour of, e.g., holding people responsible even when they really aren’t, I assume that it is valuable to have society’s practices correspond with the truth.\(^{337}\) At any rate, to hold people responsible when they in fact are not is a gross injustice according to any retributive conception of punishment.

\(^{337}\) Some rehabilitation techniques, for instance, try to change behaviour by making the offender feel a strong sense of personal responsibility. Even though the offender was not responsible for her crime, it might be an effective strategy to make her think that she was, as this makes sure that the offender cannot trivialize future crimes with being a victim of circumstances (see Pereboom 2001: 184).
So how should one bet? It depends on how the outcomes rank. Granted that the F-society is preferable to the D-society, I suggest we get the following ranking:

(4) is better than (1), which is better than (3), which is better than (2)

Let us go through this ranking step by step: (4) is the best outcome, since we here live in the preferable F-society, and are justified in doing so (e.g., we do not hold people responsible when they in fact are not responsible). The worst outcome is (2) – this happens when we falsely believe that determinism is true, and adopt the D-society. This means that we change society for the worse for no good reasons. The “middle” outcomes (3) and (1) are more difficult to rank. Is it better to “rightly” live in a D-society than to “wrongly” live in an F-society? The question might seem strange, since if determinism is true, whether we live in a D-society or not is also determined. Nevertheless, I have chosen to rank (1) over (3), since, as we saw above, we are coherent in the former but not the latter.

On to the bet. Note that whether we are optimists or pessimists, we should bet that determinism is false. Optimists, who want to choose the alternative whose best outcome is highest ranked, should bet this since if we bet that D is false, chances are we get the best outcome (4). And if we are wrong, and D is true, (3) is still better than (2). Pessimists, who choose in order to exclude the alternative where the outcome of being wrong is the worst, should also bet that D is false, since the worst outcome of being wrong when betting this is better than the worst outcome of betting that D is true – (3) is better than (2).338

Consider it as a payoff-tree where you have to bet on D being true or false. If you bet that D is true you will get the D-society, and if you bet that D is false, you get the F society:

---

338 For a similar argument, see Boström 1996: 152
If you bet on D being false and are correct, you get (4). If you bet that D is false and you are wrong, you get (3). Betting on D being true and being wrong gives (2). Betting on D being true and being right gives (1). It seems as if you should bet that D is false in this situation. If you want to play it safe and assume that you will bet wrongly, you should bet that D is false since (3) seems better than (2). If you are optimistic and want to go after the best outcome of being right, you should also bet that D is false since (4) is better than (1). Put shortly: it is better to be right in an F-society than in a D-society, and it is better to be wrong in an F-society than in a D-society.339

339 Andrew Williams has suggested to me that this might be wrong. It might be that the worst outcome is to falsely live in an F-society, as this would mean that we hold people responsible when they in fact are not. Being wrong in an F-society, for instance, would mean that we punish people who do not deserve it, which is a gross injustice. Miller, then, suggests that we rank the alternatives as follows:
This is perhaps a bit technical. We could state the argument in plain English as follows. The practical relevance of hard determinism is that it requests that we change society, in particular practices of praise/reward and blame/punishment. If we knew whether D is true or false, we would know whether to make these changes. But we do not know whether D is true or false, so what we are essentially doing when we choose to change (or not change) society is placing a bet. We know that whether we adopt the D-society or keep the F-society, chances are we might be wrong. What I am here suggesting is, assuming that an F-society other things equal is preferable to a D-society, that we should bet that D is false. This doesn’t mean that the thesis of hard determinism as such is wrong. It might very well be right. But if I am right we are justified in not acting on hard determinism. Hard determinism is thus rejected in the sense that it is not a thesis that should be the basis for policy. Put shortly, we should uphold the F-society.

An objection concerning probability

There are plenty of possible objections to this argument. The notion of the F-society being preferable to the D-society could be questioned. Perhaps a D-society, other things being equal, is on the contrary preferable to the F-society, with its guilt, resentment and blame. One could also

(4) is better than (1), which is better than (2), which is better than (3).

If this is correct, it follows that the pessimist, as defined above, should bet that determinism is true. However, I would hold that it would be a greater wrong to treat people as if determinism is true when it fact is false, since this would mean a profound violation of the appropriate ways in which people, as persons, deserve to be treated.

340 Of course, we wouldn’t change society out of choice in the genuine sense. If D is true, whether we choose to adopt the D-society or not is also determined. But I assume that if we knew that D is true, we would naturally change society to cohere with that knowledge, the human mind striving to rid itself of incoherencies and inconsistencies.

341 As I said before I accept fully HD’s first premise that freedom and responsibility are incompatible with determinism. I am here exploiting the uncertainty about the truth of its second, factual premise.
question the claim that we do not know whether determinism is true or false. Finally, one could question the extent to which belief in determinism (Q2) and the possible truth of determinism (Q1) are separate questions. It could be argued that people tend to believe that which is supported by strong arguments, and that determinism is supported by strong argument (and is thus believed in) because it is true.

These objections have to do with the three premises of my argument, and I said at the start of this chapter that denying any of them topples the argument. If you accept the three premises, however, I think you can draw the conclusions I have drawn. But there is one important qualification to be made, which concerns probability.

The argument I am defending only works when it is uncertain whether D is true of false; there must be some (subjective) probability $p$ of D being true and a $(1-p)$ probability of D being false. But I have tacitly assumed that it is equally likely that D is true as it is that D is false. I have tacitly applied the so called principle of insufficient reason. The principle of insufficient reason applies when we have a number of exhaustive alternatives or outcomes, but no way of telling which alternative or outcome is more probable or correct. Since we have no idea, we treat each outcome as equally likely. If the principle of insufficient reason is applied here, there is a 0.5 probability of D being true and a 0.5 probability of D being false. The actual work in my argument is then solely done by the assumption that the F-society is preferable to the D-society.

If we question the application of the principle of insufficient reason here, new considerations emerge. For even granting that the F-society is preferable to the D-society, there must be some probability between 1 and 0.5 at which we should bet that determinism is true. To put the point starkly: if the probability of D being true was 0.99, we wouldn’t bet that it was false even if the F-society, other things equal, is greatly more

---

342 The Principle of Insufficient Reason is sometimes also referred to as the Principle of Indifference. The principle is a method for assigning epistemic probabilities. A prototypical illustration is flipping a coin. We know that the coin will land either heads or tails, but we have no idea on which side the coin will land on any given toss. Therefore we assign a $1/n$ probability to each outcome, where $n$ represents the number of possible outcomes. In the case of flipping a coin there are two outcomes, so we get $n = 2$. We then perform the operation, i.e. $1/2 = 0.5$. Similarly, rolling a random six-sided die gives a $1/6$ probability of each outcome.
agreeable than the D-society. How we should bet is a function of the probability of determinism being true and the values of being right and wrong. And only if the F-society is made out to be infinitely more agreeable than the D-society could we maintain that one should bet against D regardless of the probability of D being true.

These considerations are clearly very important. So the question is: are we justified in treating the truth or falsity of D as a question where the principle of insufficient reason applies? I lean towards saying yes. There are convincing arguments in both directions, and peculiar difficulties are involved in proving either the truth or falsity of determinism. And even if one assigns specific probabilities to D being true or D being false, so that e.g. it is 0.6 likely that D is true and 0.4 likely that it is false, one should perhaps still bet on the slightly smaller chance that D is false. All of this is contestable of course, for it is unclear how we assign the probability of D being true, let alone how we determine the values of the D-society and F-society. The argument I have sketched does, however, provides a logic for why we should bet that D is false and resist acting on hard determinism.

9.5 A FURTHER OBJECTION: THE FISHINESS OF ADVOCATING DETERMINISM

Having provided an argument for rejecting hard determinism as thesis to act on, I want to end this chapter by offering some brief reflections about problems associated with advocating determinism. There is something deeply strange about reading hard determinists, and the moral and political views they espouse. Put shortly, the strangeness has to do with that the fact that hard determinists deny free will while their texts ooze of the very freedom – of belief and desire – that they supposedly reject. I do not know to what extent this curious feature of the free will debate counts as detrimental to determinism. I do think it deserves more attention, however, and that some political views are weakened by it.

Determinists write articles and books about that determinism given such-and-such considerations must be true, and that such-and-such consequences follow from it. The problem with espousing determinism is, as we have already mentioned, that the books and articles are determined to
be written as they are written (as are the reactions they evoke) if D is true. There is something weird - although it is hard to locate exactly what - about the whole idea of writing books about D in order to prod intellectual, moral and political discussions in the direction the author prefers. If D is true, it encompasses all human action. This includes academic debate. So determinists are those who are determined to believe in the truth of D, while indeterminists are those who are determined to believe the opposite. This need not mean that the debate itself is futile, however. Throwing away the ladder after having claimed it, we can see that arguments may be the causal impetus that serves to change somebody’s opinion about free will. But determinists often say that we should react in such-and-such a way to the truth of D: Compatibilists say that we should realize that D changes little about our moral and political world; hard determinists say that while D changes some things, it need not result in too disturbing a society. This tacitly assumes, I think, that we have some choice about how to react to the truth of D. But this is ruled out by the all-encompassing scope of D. So if a determinist says that even if D is true, we have no reason to surrender the reactive attitudes we have towards each other, it is important to note that – at least in an empirical sense – whether we give up the reactive attitudes or not is determined by sufficient causes just as everything else. We cannot choose how to react to the truth of D, nor choose whether to believe that D is true for that matter. The idea of determinism being a threat (to some important moral and political values) is thus in some sense suspect. If D is true, and have been true all along, the moral and political values in question emerged by necessity, and will decline or persist by necessity.

It could be objected that determinists, just as everyone else, have to act under the illusion of freedom. If I advance an argument in favour of hard determinism, or struggle with choosing one word rather than another while writing this chapter for that matter, I have to assume that the results are not given in advance. I have to assume that I can choose to craft the argument in different ways (evoking different reactions), or that I can choose to use one word rather than another.\footnote{For a further treatment of this point, see Searle 2001: 142; Dennett 1984: ch. 5.} This assumption, determinists will admit, is false. But unless I hold the assumption true, I shall not be able to construct an argument or complete a sentence in the first place. (It could also be added that the truth of D is something that
can only be momentarily realized, in moments of reflection, but that it is psychologically impossible to live everyday life according to it.) So there is perhaps nothing unforgivably incoherent about arguing the truth of D while assuming that one can choose between different ways of making the argument.

I am willing to concede this point (although I find it curious that people often assume that determinism has specific moral and political consequences). But there is one position that may rightly be called fishy: to assume oneself to be free while others are not.

Consider again the case of B.F. Skinner. Skinner’s theory of the behaviour of organisms, human beings included, is quite simple: “behaviour”, Skinner proclaims, “is shaped and maintained by its consequences” (Skinner 1971: 18). This notion of operant behaviour – behaviour is shaped by the consequences the behaviour elicits – enabled Skinner to suggest that many problems could be solved by controlling the environment. Skinner called such environmental control a “technology of operant behaviour” (ibid. 19). Walden Two, Skinner’s political Utopia, is as we saw his outline of a society which employs this “technology” in order to solve many of mankind’s most pressing problems. It will be recalled Walden Two is not democratic. It is those who have the necessary knowledge about human operant behaviour – the Planners and, ultimately, Frazier – who should rule.

Now, reading Skinner’s works, especially about operant behaviour and the rejection of autonomous man, one never gets the feeling that Skinner himself is completely controlled by the environmental contingencies of his behaviour. Yet, if his theory of human behaviour is true, Skinner’s own behaviour is also entirely shaped and maintained by its consequences. No special status can be bestowed on the messenger here: any universal doctrine must also encompass those who launch or defend the doctrine. In Walden Two, however, Skinner conveys a picture of the community as a sort of anthill. There is no free will in Walden Two, Skinner maintains (though he adds that it’s nevertheless the freest place on earth, because people do what they are allowed to do and want what they can have). But Frazier, behaviourist-king and Skinner’s supposed alter ego, seems very much a free agent; the primus mobile of the commu-

---

344 This point has perhaps been most suggestively articulated by Dennett, in his remarks about the wasp Sphex and “sphexishness”; see Dennett 1984: 10f, 168-9.
nity. He is the one who controls the anthill, and seems in a sense separate from it; existing, as it were, under different conditions.

In an ominous passage towards the end of *Walden Two*, Frazier reveals to Burris his more fundamental views of himself and the members of the community:

“‘Our friend Castle [a visiting philosopher] is worried about the conflict between long-range dictatorship and freedom. Doesn’t he know he’s merely raising the old question of predestination and free will? All that happens is contained in an original plan, yet at every stage the individual seem to be making choices and determining the outcome. The same is true of Walden Two. Our members are practically always doing what they want to do – what they ‘choose’ to do – but we see to it that they will do precisely the things which are best for themselves and the community. Their behavior is determined, yet they’re free. Dictatorship and freedom – predestination and free will,’ Frazier continued.

‘What are these but pseudo-questions of linguistic origin? When we ask what Man can make of Man, we don’t mean the same thing by ‘Man’ in both instances. We mean to ask what a few men can make of mankind. And that’s the all-absorbing question of the twentieth century. What kind of society can we build – those of us who understand the science of behavior.’” (Skinner 1976: 279)

What is telling about Frazier’s elitist declaration is the use of “us” and “them.” Frazier here distances himself from the members of the community; he places himself among the select few who seem capable of a different sort of agency. Thus, Frazier states that “we” – the knowledgeable Planners, presumably – “see to it that they [the members of the community] will do precisely the things which are best for themselves and their community.” The question, he declares, is what those who understand the science of behaviour can make of mankind. This is not only elitist. It also appears inconsistent. It is as if Skinner envisions a world where some are puppets and others puppeteers. But unless one wants to assert the different breed of the select few, this cannot be upheld. If Skinner is right to see free will as an illusion, it is only coherent to drop the talk of “us” and “them” in this respect and admit, without further ado, that the Planners themselves are subject to the very same deterministic processes. To hold the contrary view would indeed be fishy.

Skinner is not alone in seeming to fail to generalize the argument to himself. To inhabit positions and adopt views that seem hard to defend
given the theses put forward is a common academic disease. We see the same phenomenon, e.g., when relativists claim the non-existence of universal truths (something which seems to assume that it is universally true that there are no universal truths), or when Foucault insists on the all-encompassing ordering power of discourse (a power, it seems, which has one glitch: Foucault himself). In the case of advocating hard determinism while tacitly assuming that oneself is not governed by determinism, the problem has not only to do with adopting a vantage point that seems strange given the truth of D. It also seems to violate a basic ethical obligation of equality: do not assume others to be lesser to yourself. Skinner’s position is thus not only ontologically fishy. It can also feed into elitist, anti-democratic, and paternalist sentiments – such as those characterized, as we saw in chapter 7, as “objective” by Strawson.

9.6 CONCLUDING REMARKS

Questioning the difference between scientific explanations and excuses has taken us to strange and alien places. If all human action, and not only criminal human action, can be scientifically explained, are any of us ever responsible for our actions? Hard determinism is the position that answers that question negatively: since we do not have free will, we are never responsible for what we do. We are neither to be praised for our successes, nor blamed for our failures. Each of us is the product of, say, our genetic endowment and environment.

Hard determinism is a formidable thesis, the truth of which seems to require substantial changes in the way we think and act. I have in this chapter tried to reject hard determinism, or at least reject hard determinism as an argument to act on. First of all, I devised an argument to the effect that, when it comes to social and political practices, we should bet that D is false. Granted that we do not know whether D is true or false, that we may act as if it’s true even though it is false, and that it would be preferable, other things being equal, to live in a society where belief in D is not prevalent, we should bet that D is false, I argued. The argument is intended to show that it is reasonable to gamble that D is false while admitting that it might be true. Furthermore, I suggested some difficulties involved in advocating HD. Hard determinists sometimes reserve a posi-
tion for themselves, which, if their thesis is correct, cannot logically exist. While this might be psychologically unavoidable, it does suggest a performative inconsistency which puts the thesis of HD in further doubt.\footnote{A performative inconsistency involves acting in ways that invalidate the truth of statements put forward. Imagine that I, a human being, utter the following words: “all human beings are mute.” If my statement is taken literally, I have actually disproved it by speaking.}

Where does this leave the theory of excuses? Am I now abandoning the idea, sketched in chapter 7, that causal explanations of actions rule out personal responsibility? No. The gist of saying that we have free will is instead that not all actions can be given a complete causal explanation. The possible existence of free will means that there are sometimes irreducible limits to our ability to explain why one choice was made rather than another (McIntyre 1981; Shapiro 2002; cf. Chisholm 2003).

The free will problem is ancient and has survived thousands of years of “solutions.” It would be presumptuous to pretend that my solution will be any different, although I do find my approach reasonable (and also somewhat new). The considerations in this chapter at any rate suffice for me to reject hard determinism – at least as a thesis to base political action on – and I shall not deal any further with it.

Dispensing of hard determinism does not mean that explanations cannot be excusing, however, or that all acts of rule breaking are free choices and all rule breakers fit to be held responsible. In light of the considerations presented in part III in this book, retributivism is still severely challenged. It is possible that desert applies much more seldom than what is often supposed, and we cannot appease those who are sceptical of desert-based conceptions of justice with the observation that since hard determinism can be reasonably rejected it is \textit{possible} that we sometimes deserve certain things (such as punishment). The next and concluding chapter tries to sketch the outlines of a just penal regime, and provides conclusions that integrate the different and often conflicting ideas treated thus far.
Part IV.
Conclusions
10. A Justified Penal Regime?

“To decide precisely when the individual is acting as an independent agent, morally responsible for his or her own actions, is, however, one of the most difficult questions not just in relation to punishment, but in political theory itself.”
(Heywood 2004: 172).

10.1 THE STORY THUS FAR

It is time to take stock of what we have concluded so far. I began by assuming that there are worthwhile rules of behaviour, the violation of which may legitimately be punished or otherwise interfered with by the state. Given this, the overarching question was (Q):

(Q) Which principle or theory (or principles or theories) should serve as the basis for a state’s penal regime?

The investigation up to this point has mainly concerned the question of justification: upon which principle or theory the penal regime should be built; what aims or values it is to express or promote in order to be justified. Throughout, I have argued that retributivism provides a more compelling answer than its deterrentist and rehabilitationalist rivals: a penal regime which aims to treat rule breakers as they deserve seems preferable to penal regimes that seek to deter or rehabilitate.

Why is the retributive answer to (Q) preferable to the ones based on deterrence or rehabilitation? I have argued that there are two different sets of reasons that make a retributive penal regime preferable. On the one hand, such a regime seems more consistent with weighty principles
of justice; we saw, in chapter 5, that a penal regime based on retributivism is the only type of regime (among those considered in this book) which excludes such things as punishment of innocents and pre-emptive incarceration of risk-individuals. Though deterentists and rehabilitationalists can argue that such violations of the rule of law are unacceptable, they can only do so contingently. This follows from their utilitarian underpinnings: the problem with utilitarian theories of punishment/treatment is not that they must advocate violations of the rule of law, but rather that the violations are “open moral possibilities” – not wrong in principle, but contingently wrong. Retributivism, by contrast, squarely rejects the possible justifiability of such violations. I have argued that this is a merit of a retributive penal regime, and I referred to it as the institutional reason for retribution.

On the other hand, retribution, the penal response meted out by a retributive regime, respects the rule breaker as a person in a way punishment-as-deterrence or rehabilitation does not. Retribution, I have argued, expresses blame and resentment for the choices the rule breaker has made. This in effect presupposes many empowering messages. If you are blamed, it means that more could have been expected of you. Furthermore, it means that you weren’t forced, either by bad genes, a malfunctioning brain, or a rotten social background, to do what you did. It was your choice – and it should be respected as such. I referred to these considerations as the symbolic reason for retribution.

In favour of the line of thought inherent in the symbolic reason is that we do adhere to it in other areas of social and political life. Take for instance democratic elections. When a voter votes for a certain party, we normally do not suppose that there is some underlying disorder that triggers the behaviour. We do not assume that the voter is ignorant, compulsive, unable to vote otherwise, or suffering from a disease. Instead, we acknowledge the vote as a choice that warrants our respect – it should be duly counted simply in virtue of being made by the voter. Retributivism seeks to treat criminal choices in the same way. Admittedly, there are few political parties to vote for of such nature that questions of normality become as pressing as in cases of heinous crimes. But as a general principle, I submit, it is appropriate to treat criminal choices as other kinds of choices, and rule breakers as other kinds of agents. This follows, I argued, from the idea of respecting persons. One key aspect of respecting persons is precisely to treat their choices as significant.
10.2 WHY HOLD RESPONSIBLE?

The symbolic reason might seem worthwhile only to the extent rule breakers by and large are personally responsible for their crimes (or, more generally, only to the extent people are personally responsible for their actions). I proceeded to investigate whether there are factual reasons to reject retribution as a foundational principle of a penal regime. If there are such reasons, I argued, the arguable normative attractiveness of retribution must stand back.

It was argued that there indeed are reasons to think that scientific explanations are incompatible with responsibility, as they seem to have tendencies towards regress and determinism. Therefore, it does seem as if retributivism, with its insistence on personal responsibility and desert, can be criticized for being unscientific.

In chapter 8, the implication of explanations as excusing was followed to its end: what if all actions are explainable in such a way that responsibility is undermined? The thesis associated with that view is hard determinism, the nature and consequences of which we considered in some detail. Hard determinism, it was argued, would not only rule out retribution as a possible justification of punishment, but would also, by undermining the idea of autonomous man, undermine the significance of choice and the practices associated with it.

Still, while the case for hard determinism is stronger than one initially might think, it is still the case that we at this point do not know whether determinism is true or false. By implication, nor do we know whether all action is scientifically explainable in causal terms. I argued, in the previous chapter, that this state of uncertainty should advise us to bet that determinism is false. Having escaped from the clutches of hard determinism in this way, the normative reasons in favour of retribution as a foundational principle of a penal regime are back in force. We should bet that people are such that they can deserve blame in virtue of their actions.

Hard determinism, however, is an all-or-nothing thesis. It states that none is personally responsible. If this is too much to stomach, there are still reasons to think that many of the rule breakers we presently hold responsible in fact are not responsible for their actions. So the question of why it is important to hold responsible once again surfaces, if only in lesser scale. Many rule breakers are presumably disordered and hence
not ordinarily responsible for their actions. What, if any, is the point of “respecting” their choices by punishing them?

One possibility is that we are justified in treating individuals as if they are responsible. Dennett has argued this:

“Instead of investigating, endlessly, in an attempt to discover whether or not a particular trait is of someone’s making – instead of trying to assay exactly to what degree a particular self is self-made – we simply hold people responsible for their conduct (within limits we take care not to examine too closely). And we are rewarded for adopting this strategy by the higher proportion of responsible behaviour we thereby inculcate” (Dennett 1984: 164).

Dennett here advocates that we should hold people responsible for their actions even though we will sometimes fail to excuse those who in fact have good excuses. We should do so because we would thereby get a “higher proportion of responsible behaviour.” If people know that they will be held to account for their actions, they will behave in a better way. They are not encouraged to behave recklessly out of a belief that they will not have to answer for their actions anyway.346

On a retributive view, however, Dennett’s view must be taken with great scepticism. A basic desideratum for retributivists is that people get what they deserve. This entails excusing those with good excuses. The possibility that a regime of strict liability, in fostering responsible behaviour, may have beneficial consequences does not overrule the fact that such a regime would be unjust. Dennett raises a fair point, however, in questioning the sense of investigating “endlessly” the borders between responsibility and excuse. This difficulty, which I have addressed only briefly in the preceding chapters, has certainly been picked up by the

346 Wilson has similarly argued that: “A goal of the legal system is to foster self-control by stigmatizing and punishing its absence. Most people possess enough self-control to refrain from any serious from of criminality. But some observers wish to reduce the punishment of people who seem to lack self-control. There are two things wrong with this. The first is that excusing failures of self-control will increase the frequency of such failures. The second is that once we leave the narrow excuses of duress or insanity, we will have, in the main, a vast list of criminal causes that, if adopted, will erase the distinction between law and science” (Wilson 1997: 43).
reader. Does it make any sense for the courts to inspect whether a particular criminal is “really” responsible or not? Why not simply hold people responsible, within a few crude principles of the kind usually employed by existing states? We would then certainly get some cases wrong, but we would not have to delve too deeply into impracticable matters.

Such a view tacitly depends, I think, on an assumption that most rule breakers are responsible for their actions, and that a rough-and-ready regime of responsibility therefore would get more cases right than wrong. A retributive penal regime would have to examine this assumption – and in the event it turns out false, punishment would be ruled out. It is undeniable, however, that we cannot hope for much precision when it comes to drawing the lines between responsibility and excuse. But this situation is structurally similar to the one we faced in the previous chapter: since we do not know whether $X$ is the case (e.g., whether someone is responsible or not), we may need to take other considerations into account in deciding what to do. And given that there are values in holding and being held responsible, these should caution us not to excuse too lightly. I propose, therefore, adherence to the principle (P):

(P) Given the values inherent in being held responsible, the burden of proof should lie with those who want to excuse.

(P) states, then, that holding people responsible for their actions should be a default reaction, which may only be overridden if a strong case can be made that a good excuse exists. The general idea here is of course that we ought to excuse those who have good excuses, but not accept bogus ones. Rather than presuming that people should be excused, and place the burden of proof on those who advocate holding responsible, it is

---

347 Including, say, workable criteria for self-defence, legal insanity, and non-culpable ignorance.

348 Since it is unlikely that rule breaking can be tolerated by society, however, it is plausible that the state would respond to rule breakers in the language of civil compulsory commitment instead.
those who want to excuse that must explain why, and do so in a compelling way. 349

(P) is of course the position held by actual states, and in that sense is nothing new or revolutionary. But it is not until we realize that holding and being held responsible contains symbolic values that we also see the normative sense behind the principle. Some could maintain that making excuses rather than blame the default reaction is the enlightened and humane thing to do. In fact, it is the other way around – excuses do not respect rule breakers as persons, and as such may belittle or demean them.

As I argued in chapter 9, this does not mean that we are engaging in wishful thinking. In particular, it does not mean that we advocate useful illusions that we know are untrue. 350 We have reason to accept (P) precisely because there are so many difficulties and uncertainties involved in drawing the line between responsibility and excuse.

10.3 THE INSTITUTIONAL ARCHITECTURE OF A RETRIBUTIVE PENAL REGIME

Barring a scientific development that strongly indicates that our belief in the personal responsibility of rule breakers (or personal responsibility in general) is mistaken, our answer to (Q) should be cast in terms of retributive principles of punishment. A (version of) retributivism should serve as the basis for the penal regime.

We have yet to say something about the institutional architecture that follows from these principles. In chapter 1, I said that I would offer not only an account of according to what theory of principle the penal

349 Note that this is different from treating prosecuted individuals as innocent until proven guilty, a plausible principle which places the burden of proof on those who want to prove guilt. We are now considering what to do with those who have de facto been proven guilty of breaking a rule.

350 If I truly believed that personal responsibility is a value so important that it must take precedence over truth and justice, I surely would not have written a book which to a large part deals with strong reasons to doubt the idea that we are personally responsible for our actions.
regime should be fashioned, but also some reflections on the institutional make-up that follows from that account. Indications about what a retributive penal regime would look like in practice have been scattered throughout this book. Here I will offer a brief description of what I take to be central components of such a regime. I suggest that a retributive answer to (Q) would include at least the following six points:

1. *Punishment as penal response.* Retribution involves giving rule breakers what they deserve. If the rule breaker has done something morally blameworthy, he or she deserves to have something unpleasant happen to him or her. This means that penal responses are to be designed to impose pain or deprivation. They should not be geared towards benefiting the rule breaker, as in some accounts of rehabilitation. In short, penal responses should take the form of punishment. This does not say much in practice, as standards of what counts as punishment and what forms punishment can take will be shifting. Suffice to say that punishment should not be conceived as a paternalist practice.

2. *Proportionality.* As was stated in chapter 3, the available punishments should minimally have an ordinal match with the crimes, ranked by their degree of wrongfulness. More satisfactorily, but possibly utopian, crime and punishment should have a cardinal match (see Elster 2006). It is an open question whether a proportional penal regime will be of a nature that can be characterized as “soft” or “hard.” My own preference is a by most standards fairly soft penal regime, which does not include the death pen-

---

351 It seems that this clause commits us to a subjective criterion of punishment: a penal reaction is only punishment as long as the punished person experiences pain or deprivation. Thus, for a person whose life would actually improve by being taken off the streets and put into prison, imprisonment would not be an appropriate penal response. Such a position seems unwieldy, however – given that most people would experience fines and imprisonment as unpleasant, one could make a case for some reasonable person-standard of appropriate punishment. Note though that vast differences in wealth seem to undercut the rationale for a unitary system of monetary fines (a millionaire will presumably not experience the same loss from paying a speeding ticket as a poor person).
alty and is based on short prison sentences. As we saw earlier, however, it is important to note that different people will have different conceptions of what proportionality requires. Even given that we are able to agree on a complete ranking of crimes (in terms of their seriousness), we are left with the question concerning how severely these crimes should be punished, and where the upper and lower limits of the set of punishment are to be drawn. One retributivist might advocate much more severe punishments than another, and they will thus be in profound disagreement about what the principle of proportionality requires. Note that such a debate, which concerns not the principle but the implications of it, does not seem possible to settle by rational argument. What proportionality requires seems to be something which is settled on a more intuitive level, perhaps wholly on the basis of emotion. In a democracy, it is likely that the task of establishing appropriate proportionality in punishment would or should befall the *demos*, or its representatives.

3. **No pre-emptive strikes.** A retributive penal regime will hold that the state is only justified in punitively interfering in the life of an individual when he or she deserves it. But one only deserves punishment when one *de facto* has done something blameworthy. Therefore, a retributive penal regime will not admit pre-emptive action, such as when an as-to-yet innocent is incarcerated on the ground that he or she is likely to engage in future rule breaking.\(^{352}\) It is of course a difficult terminological and substantive question to say what kinds of pre-emptive responses that falls under such a ban. For example, is it an illegitimate form of pre-emptive strike to commit a mentally disordered person to coercive care on the grounds that he or she might pose a risk? Such a case is partly based on the notion of disordered people being in need of treatment, and it is probably fully com-

---

\(^{352}\) Whether we call such incarceration “punishment” or not is beside the point here. It is true that action in such cases is not taken because of some act of rule breaking, but given that the motive is to stop future criminality (as opposed to, say, infectious diseases as in the quarantine analogy), it is appropriate to call the action “punitive.”
compatible with a retributive penal regime. But the notion of, say, using instruments and techniques of risk prediction to pre-emptively interfere with people simply because they pose a risk of future crime – something which has been espoused by among others Morse (1999) – must be seen as inadmissible.

4. **Determinate sentencing.** A retributive penal regime will publicly declare the rules for punishment in advance (e.g., “doing X under circumstances Y yields 5 months imprisonment”), and then stick to those declarations. It will thus support determinate sentencing – since a retributive penal regime is in the business of giving people what their criminal actions merits, good or bad behaviour in prison should not lead to shorter or longer sentences. It is also highly doubtful if a retributive penal regime can support a scheme of increasingly severe punishments for recidivists, such as the “three-strikes-and-you’re-out”-scheme. A person who has chosen to break the rules and has paid the publicly declared price for this, should, as it were, be back at zero. In the case of future rule breaking, each act should be given what it merits.\(^{353}\) It is also doubtful if a retributive penal regime is com-

\(^{353}\) There are two ways in which a retributivist could defend something like the three strikes-scheme. First, one could argue that a criminal act is more blameworthy if it is preceded by previous criminal acts. The second kidnapping is worse than the first kidnapping, on this account. Secondly, one could defend a character-based retributivism as opposed to an act-based one (see Murphy 2003, who prefers to call what I refer to as act-retributivism as grievance-retributivism). According to the former, punishment is deserved on the basis of one’s wicked character, not the wrongfulness of the act as such. A recidivist, continually flouting the values of society, could plausibly be said to have a worse character (be a worse person) than a first-time offender, other things being equal. Therefore, it makes sense to punish recidivists with increasing severity. The penal regime sketched here is based on act-retributivism, however. Much of the appeal of retributivism is exactly that it is preoccupied with acts rather than persons (noting that planning or intending to commit a crime also qualifies as an act). We have strong intuitions against punishing legally innocent but morally repugnant people, for instance, as well as against increasing the punishment of people who are simply unpleasant. We can only make sense of these intuitions if we abandon character-retributivism in favour of act-retributivism.
patible with the practice of letting first-time offenders off more easily, as this would too undermine the notion of acts, not characters, being punished. Finally, and most clearly, a retributive penal regime will unequivocally rule out various forms of allowing crime victims to influence the duration or severity of punishments. Since it is the wrongdoing which should be punished, such considerations are irrelevant – it is not a greater wrong to commit a crime against an unforgiving person than against a forgiving one, so the thought of letting the reactions of crime victims influence sentencing is clearly a bad idea if the objective is to give rule breakers what they deserve.

5. Sensitivity to justifications and excuses. I have repeatedly argued that a retributive penal regime will not be insensitive to pleas of excuses or justifications. In placing the blameworthiness and responsibility of the criminal act in doubt, these pleas may serve to mitigate the appropriate punishment, or make punishment inappropriate altogether. Courts, aided by scientific expertise, should have procedures for assessing the validity of defences based on excuse or justification. Disordered rule breakers who are excused, furthermore, should sometimes be civilly committed to various non-punitive treatment institutions. If this is an invitation to arbitrariness, it is a necessary form of arbitrariness for a penal regime that seeks to treat rule breakers as they deserve.354 Recalling the taxonomy in chapter 7, however, we also have requests for mercy, which is simply the idea of admitting fault but asking for pardon. Though the retributivism defended here is non-strict, there is probably a strong case for being very cautious with pardons. It is true that a pardon still assumes and respects the personal responsibility of the rule breaker, and a judge could express appropriate blame verbally without meting

354 There is clearly a practical risk of this requirement coming into conflict with the notion of determinate sentencing. Logically, however, there is no conflict: a court should have procedures (publicly declared) to judge various mitigating and exonerating (as well as aggravating) factors, which affect the determinate sentence meted out. This is what the “Y” in “doing X under circumstances Y” in point 4 above stands for.
out punishment *per se*. But as we saw in chapter 6, given our present understanding of what counts as being held responsible for a crime, such a pardon would presumably seem like cheap talk. Also, rule breakers who have committed similar crimes but have received punishment can reasonably protest that they are being inequitably treated.

6. **Rehabilitation and deterrence as welcome side-effects.** A retributive penal regime is compatible with seeing any deterrence or rehabilitation of rule breakers as welcome side-effects of punishment, as long as these do not conflict with the aims of retribution. I have already argued, in chapter 6, that to comply with the laws out of fear of punishment does not in itself violate the autonomy of persons, and any reduction in crime rates produced by a proportional penal scheme is to be welcomed. However, punishment should not be meted out with the *intention* of producing deterrence. This would not only in all likelihood upset the proportionality between crime and punishment, but would also represent an illegitimate way of treating people. When it comes to rehabilitation a similar case could be made: any bettering of offenders is to be welcomed as long as it does not conflict with retribution. It follows from point 1 above that the penal response to rule breaking should be punishment, properly understood. This means that any rehabilitation which takes place must happen within the scope or context of punishment. In plain text this means that the time in prison should not – duly noting the fact that loss of freedom is probably painful enough in its own right – be too pleasant for the imprisoned. A retributive penal regime will also respect an individual’s right not to become rehabilitated. The state should not be in the business of remoulding people who prefer not to conform to the values of society. People should be free to choose a life of crime, should they want to.\(^{355}\)

\(^{355}\) This is a general conclusion which of course is difficult to defend in particular cases. There are some types of crime where such a stance of non-interference seems particularly hard to stomach. I would not, for instance, hold that a sex offender should be set free with unchanged motivations; at least not that he
These are the outlines of a penal regime based on retributive principles. Many details remain to be worked out, and many objections could be levelled at it. The most natural worry is probably that the regime will fail to provide sufficient public protection – that the regime, in ruling out indeterminate sentencing, pre-emptive strikes and increasing punishment for recidivists, will fail to effectively reduce or control crime, and therefore be a poor choice. If public protection were the only standard by which we judge a penal regime, that objection would probably be correct. But there are other values which should also be in the picture. And then it becomes possible to trade some losses in public protection for some gains in other areas. As Hood has famously written:

“I believe that a system which arrives at the lengths of sentences based more on moral evaluation than on appeals to the utilitarian philosophy of deterrence and reductivism would be fairer, not necessarily less effective, possibly less, not more punitive and appeal to that sense of social justice on which any acceptable system of social control must be founded” (Hood, cited in Raynor & Robinson 2005. Emphasis added).

Like Hood, I grant that a retributive penal regime may be less effective in reducing crime than the rivals considered in this book, and certainly less effective than some of the more draconian expressions of the Nothing Works view. But the crime-reducing effectiveness of the regime should not be the sole criterion by which we judge it. The institutional and symbolic reasons are reasons why we should prefer a retributive penal regime that do not have to do with crime-reduction.

There are other reasons to doubt the desirability of a retributive answer to (Q). One objection, which I have dealt with in chapters 7-9, is basically that the regime is built around a concept – desert – which does not exist. Since there is reason to bet that free will (and hence desert) exists, as was argued in chapter 9, that objection can be countered.

There is another objection, however - more political than metaphysical in nature - that can be levelled at a retributive penal regime. This ob-

should be left with the option of not undergoing various forms of rehabilitation. On the other hand, I find it completely unproblematic to release, say, property thieves with an undiminished desire to steal. How to draw the line between these types of cases in a concise way, I do not know.
jection runs as follows: a retributive penal regime would be appropriate in an ideal society, where rule breakers and rule followers enjoy roughly the same opportunity sets. But real societies are marred by deep inequalities. To say that the worst off deserve punishment when they break the laws of such a society is basically to say that they ought to repay a debt they do not owe. Therefore, a retributive penal regime does not apply in real societies.

10.4 THE JUST SOCIETY AND PUNISHMENT

As I have indicated elsewhere, Karl Marx was a retributivist of sorts. “From the point of view of abstract right”, he argued, “there is only one theory of punishment which recognizes the human dignity in the abstract, and that is the theory of Kant, especially in the more rigid formula given to it by Hegel.” However, Marx also held that retributivism, instead of looking upon the criminal as an actual social being, “elevates him to the position of a free and self-determined being.” This means, he continued, that retributivism serves as a “transcendental sanction to the rules of the existing society.” Marx wrote:

> “Is it not a delusion to substitute for the individual with his real motives, with multifarious social circumstances pressing upon him, the abstraction of ‘free will’ [...] Is there not a necessity for deeply reflecting upon an alteration of the system that breeds these crimes, instead of glorifying the hangman who executes a lot of criminals to make room only for the supply of new ones?” (Marx, cited in Murphy 1995: 3-4)

Marx’s case is made against a backdrop which is all too familiar. When we look at any given society, we tend to find that it is among the worst off that most crimes are being committed. As a rule, it is not the well-educated and successful who breaks the rules. It is the poor and deprived (Heffernan & Kleinig 2000).\textsuperscript{356} Given this fact, it may seem as if retribu-

\textsuperscript{356} See however Vold 2002, who argues that the evidence here is more contradictory than one might think.
tivism justifies, in so many grand words, punishment of the least fortunate — people who do not choose crime from a sizeable list of alternatives, but instead are coerced into it by the restricted hand they are dealt. How else would you explain the statistical patterns of crime? That the poor have innately worse characters or weaker will?

Marx is not alone in thinking that retributivism may have problems handling social and political inequalities in society. Murphy has argued that since retributivism is the morally correct theory of punishment, and since crime is a consequence of profound inequalities, “modern societies largely lack the moral right to punish” (Murphy 1995: 7). Furthermore, Anderson has criticized the recent trend of coupling retributivism with contractualist norms of reciprocity, since reciprocity presupposes a natural equality between individuals whereas crime is typically committed by the worst off (Anderson 1997).

What these thoughts strongly indicate is that the justification of punishment, or the design of a just penal regime, cannot be treated separately from reflections on what Honderich has called “the decent society.” We cannot discuss the question of punishment without also reflecting on what kind of society we should have (Honderich 2006: 201f). Along similar lines, Brettschneider has argued that punishment needs to be discussed as problem of political theory rather than one of ethics (Brettschneider 2007a). Punishment needs to be analyzed not only in terms of what it would be morally appropriate to do to rule breakers, but also in terms of what agent it is that does the punishing. In short, to justify punishment requires the legitimacy of the state. And if the playing field is very uneven, a case can be made that there is no legitimacy for the state.

357 After having described the life of a severely deprived but not atypical rule breaker, Murphy asks whether we, after having learned the gravity of the deprivation, would “still want to talk — as many do — of his suffering punishment under the rubric of ‘paying a debt to society’? Surely not. Debt for what?” (ibid. 28). And while admitting that not all criminals are severely deprived, Murphy argues that the severely deprived criminal is “a closer picture of the typical criminal than the picture that is presupposed in the retributive theory — i.e., the picture of an evil person who, of his own free will, intentionally acts against those just rules of society which he knows, as a rational man, benefit everyone including himself” (ibid.).

358 For another argument sympathetic to the idea of regarding punishment as a political rather than moral problem, see Binder 2002.
to claim. A severely deprived citizen could argue that he or she is under no duty to respect the laws of a polity which systematically mistreats him or her. In sum, then, there is reason to ask whether retributivism is an ideal theory that cannot be applied to non-ideal settings (i.e. actual societies).

In response, note first that it is questionable if an illegitimate state wholly loses the right to punish rule breaking. Suppose first a state that is illegitimate because it preserves very large differences in undeserved wealth between people. Its illegitimacy seems to undermine the state’s right to punish the worst off. But is the right to punish equally lost in the cases when the well off commit crime in this society? Such a conclusion would be highly problematic. Or suppose, second, a dictatorial state. Surely we find claims by the state’s authoritarian rulers to a moral right to punish political dissidents absurd. But as was noted in chapter 2, this need not mean that the dictatorship is unjustified in punishing other kinds of criminals, such as murderers. If these murderers have done something wrong, then, on a retributivist note, they deserve punishment. It does not seem plainly wrong for an illegitimate state to punish them simply in virtue of being illegitimate. Perhaps, then, an illegitimate state retains some right to punish some rule breaking.

The charge raised by Marx, Honderich and others, however, is that crime is a business mainly for the poor and disenfranchised, and that the profound injustices of society must be taken into account when considering the justification of punishing the worst off. Retributivism, in their mind, serves to make us inattentive to this fact. In emphasizing the autonomy of the rule breaker, it upholds a myth that reinforces and reproduces the status quo. In Honderich’s words:

“Retribution theories of the justification of punishment have always been discussed independently of the question of the decent society. That is to say that it has been supposed that it is possible to say what, if anything, justifies punishment, without attending to the other larger question. It has been tacitly assumed that punishment is being considered in connection with a great, decent, justifiable, defensible or perhaps somehow necessary society. Something like this was Kant’s confident assumption when he said […] that punishment must in all cases be imposed only because an individual has committed a crime” (2006: 202).
But if society is not “great, decent, justifiable, defensible or perhaps somehow necessary”, then claims that the state has a right, indeed maybe even a duty, to punish lawbreakers seem problematic, for we could only justify the defence of a society that is worth defending. And a deeply unequal society, where poverty breeds crime, may seem precisely an indefensible society. To say that deprived criminals deserve punishment here appears stark – indeed a “transcendental sanction of the status quo”, as Marx put it.

I have two things I want to say about this important critique of retributivism. They are less counterarguments than reflections of mine. First, in the case that retributivism should apply to non-ideal settings, should retributivists advocate policies that reduce crime? We need to ask whether there is something in retributivism which precludes us from addressing the kind of social injustices that breed crime. Second, if the charge that retributivism provides a “transcendental sanction of the status quo” is to be any different from the objection from non-existence, it is unclear whether poverty is a justification or an excuse. If the latter, we still have reasons to hold responsible.

Should we reduce crime? On preventive policies and retributivism.

A curious question, which is a consequence of the symbolic properties of retribution argued in this book, is this: if punishment is so great, why should we have less crime?

The awkwardness of this question is not difficult to pinpoint. Intuitively, most of us would prefer less crime to more crime. In fact, many of us would prefer a society where most of what we know as crime does not happen at all.359 But since retributivists argue that punishment expresses and reinstates the valuable notion of personal responsibility they seem forced to ask whether we should in fact prefer less crime to more. If

---

359 Following what I said in chapter 2, I here assume that there is a list actions or omissions that are crimes, and that if few/no actions on the list were performed, there would be little/no crime and hence little/no punishment. I thus do not assume that society, given that it cuts rates of crime as we know it, would proceed to criminalize other acts.
fewer crimes are committed and hence fewer persons punished – what happens then to the notion of personal responsibility?

To answer this question, we need to realize that punishment is not the only way to express and reinforce personal responsibility. First of all, we may get the same effect from praising and rewarding people. The principle of personal responsibility is equally served by rewarding praiseworthy people as it is by punishing blameworthy people. Moreover, our everyday interpersonal relationships and exchanges are absolutely steeped with expressions of personal responsibility - if you forget your spouse’s birthday, you will still be resented even though crime, and hence legal punishment, has been eradicated. In fact, the whole idea behind the retributivist case is that crime and punishment should be conceived just as any other area of social and political life. It is a matter of making legal punishment resemble the kind of appropriate blame your spouse will subject you to when you forget his or her birthday, not to make punishment the bastion and sole expression of personal responsibility.

Once we realize this, nothing stops a retributivist from desiring that crime be reduced. After all, crime is (most of the time) wrongdoing, and it would be odd to say that a state of affairs with more wrongdoing is better than one with less – we do not prohibit theft in order to punish thieves, but because we want to express that theft is unacceptable. However, a retributivist could not coherently defend any means of reducing crime. He or she could not, as we have argued, defend pre-emptive incarceration of risk-individuals. The same goes for “curing” recidivists by subjecting them to treatment that violates their status as autonomous agents, or deterring crime through a wildly disproportional penal scheme. Nothing stops the retributivist from addressing the kind of

360 If it seems backwards to criminalize theft simply to be able to punish thieves, note that some have argued in structurally similar ways. Aristotle, for instance, criticized Plato’s scheme of collective property on the ground that it would rob people of the opportunity to exercise the virtue of generosity (Aristoteles 2003: 1263b). One could say, following this line of thought, that we enact behavioural rules simply to make wrongdoing possible and thereby allow for the exercise of self-restraint, or similar virtues.

361 A puzzle, however, is what to make of a system of perfect deterrence. Perfect deterrence is when punishments or sanctions are so deterring that no unwanted
social injustices that breed crime, however. If crime happens because of the restricted opportunity set of the worst off, striving to enhance that opportunity set by offering better jobs, housing or education, is perfectly compatible with retributivist thought. Battling the social causes of crime of the kind raised by Marx is a preventive strategy to reduce crime, which in a direct sense at least violates the autonomy of none.

Is rotten social background a justification or an excuse?

Given that we should act preventively so as to reduce crime (as long as it does not violate the autonomy of agents), but that a society without crime is utopian, what should we do with those who break the rules if it turns out that they are predominantly poor and disenfranchised? Legal theorists have entertained the possibility of introducing “Rotten Social Background”, or “severe environmental deprivation”, as a mitigating or exonerating factor in sentencing. Offenders that come from dismal circumstances of crime, poverty, unemployment, abuse, etc., should be able to raise this as part of their legal defence. Suppose for now that it makes sense to attribute much crime to the causal influence of “severe environmental deprivation.” Unless such deprivation was the case, the deprived individual would not have broken the rules. Is this an excuse (partial or full) or a justification (partial or full) of rule breaking?

Rotten Social Background would be a justification if it was claimed that the rule breaker, in virtue of his or her poor circumstances, owes nothing to society, respect for its rules included. In a society which mistreats one, the mistreatment can be taken as ground for saying that one has not done something wrong when one breaks the society’s rules. Rotten behaviour happens in the first place. While utopian enough, under such conditions no one experiences disproportional punishment, and no one seems ill treated as there is nothing about penal threats as such which is problematic. The only autonomy-based charge against a system of perfect deterrence is, it appears, that it in effect robs people of the opportunity to do wrong.

Social Background is an excuse, however, if the claim is that the deprivation reduces the capacity for autonomous agency.

Suppose Rotten Social Background is used as a justification. For a retributivist, justifications pose a lesser problem than excuses because they do not threaten personal responsibility. So if the claim is that an environmentally deprived offender does not deserve punishment because he or she owes no allegiance to society, a retributivist could comfortably accept it. For the claim is now simply an unusual way of saying that one should not be punished unless one has done wrong. Intuitively, there seem to be cases where a person is disadvantaged enough to be able to break the rules justifiably. I would not be willing to specify at what level of social injustice the worst off cease to owe anything to society. It is clear that a person who has to break the rules in order to survive is justified in doing so. It is also clear that most actual crime does not meet that criterion, at least not in contemporary welfare states. Suffice to say here that there is nothing in principle that rules out that a retributivist accepts Rotten Social Background as a justification among others.

When it comes to Rotten Social Background as an excuse - which seems to be what Marx had in mind - matters become more intricate. As a rule, a retributivist would accept extreme environmental deprivation as an excuse as long as it is a good one (i.e., one which truly undermines or precludes responsibility). But Rotten Social Background needs to be treated as any other kind of proposed excusing explanation (cultural, biological, psychological, etc.). This, again, is because of the symbolic reason for retribution. Consider for instance the following remarks by Delgado about why Rotten Social Background, RSB, can explain rule breaking:

“[1] [Writers] observe that daily existence in a ghetto environment creates a reservoir of rage, which, if tapped, can take control on the individual’s actions...In this sense, the defendant’s conduct resembles a seizure or automatic reflex. The actor’s conduct is not voluntarily de-

---

363 These cases are of course more difficult to find in cases of crimes against life or person. Moreover, matters are only made worse if it is the case that members of the worst off are most likely to commit crimes against other members of the worst off. However, one can easily imagine crime against government or corporations which could be justifiable in virtue of severe deprivation.
terminated, but rather directed by the dominating emotional force of rage. [2] Even where the defendant’s conduct appears outwardly voluntary, the power of the RSB defendant’s mental or emotional trauma might cause a different excusing condition, the loss of ability to control conduct” (Delgado, cited in J. L. Anderson 1999: 368-9)

Here we face the same problem of explanations we have considered in previous chapters. It makes perfect sense to assume that many, including those who come from a Rotten Social Background themselves, will be reluctant to explain and excuse people’s actions on the grounds that they resemble “seizures” and “automatic reflexes.” This kind of humanitarian concern may well be demeaning; not respectful of the choices of the environmentally deprived. As Morse has argued:

“The existence of a deprivation excuse would cast doubt on ordinary notions of responsibility that are part of our self-conception and contribute a great deal to our sense of dignity and self-worth. The excuse would tend pejoratively to label large number of citizens as less than full moral agents, contributing to the continued degradation and exclusion of the worst off” (Morse 2000: 154).364

The sympathy we feel for the worst off, Morse holds, may well turn into a patronizing objective attitude, where the deprived are seen as helpless victims of their circumstances. Even the worst off themselves would or should shun being regarded in that way.365 Like other kinds of proposed

364 Morse and I are, as will be recalled from chapter 6, in profound disagreement over where the “dignity” of people resides, however. He places it in our practical reason, and dismisses, contrary to what I hold, any further need to invoke libertarian free will.

365 Some studies show the psychological benefits of something like a fatalistic worldview, however. Patterson for example reports that there is a “close association” between lower-class Afro-Americans’ sense of low personal control and high self-esteem. In other words, lower-class Afro-Americans feel good about themselves while – contrary to what I argue here – at the same time feeling that they are not masters of their destinies; not in the driving seat, as it were, of their own lives. Patterson writes: “Their is a world in which all sense of responsibility for failings has been abolished and in which the highly developed sense of being determined by external forces acts as a security blanket, a source not of shame but of pride” (Patterson 1998: 91). These results are surprising, and run contrary
excuses, then, the Rotten Social Background defence should be ruled by principle (P): we should only accept if it passes high epistemological requirements. Whether Rotten Social Background would pass such muster in individual cases is of course something I do not know. 366

In sum, we can say the following about the charge that retributivism cannot apply to real (unjust) societies: it is certainly true, as Honderich has claimed, that the justification of punishment has tended to be discussed without attention being paid to in what kind of society punishment is to be practiced. Retributivism is perhaps more guilty of this lacuna than other theories. Logically, of course, retributivism is compatible with a world where each and every rule breaker is justified or excusable. But substantively, as I have argued, the theory seems committed to the view that most rule breakers are personally responsible and deserving of punishment. In that way, the observation that the worst off – who commit most crimes – do not deserve punishment could be seen as challenging to the theory. But to advocate a retributive answer to (Q) does not preclude us from battling social injustice, thus reducing crime. Furthermore, to the extent social injustice is being used to excuse wrongdoing, we must realize that this comes with the price of undermining the significance of the choices of the worst off. A reasonable conclusion is that we should battle the causes of crime while at the same time adhere to principle (P) when it comes to using Rotten Social Background as an excuse.

to what I have argued in this book. However, it could still be the case that the sort of well-being one derives from seeing oneself as without responsibility is an undignified form of well-being that should not be cultivated, either by the state or individuals.

366 Given my analysis of probabilism in chapter 8, it of course does not suffice to make the trivial but uninformative observation that most RSBs do not commit crime. (Some do, and the question is whether RSB undermines their responsibility.)
10.5 THE ULTIMATE JUSTIFICATION OF A RETRIBUTIVE PENAL REGIME

Suppose that we by now accept that (Q) should be answered along retributive lines, and that there is no decisive reason why retributivism could not be put into practice. A final question, which I have so far postponed, is this: what is the ultimate moral justification of a retributive penal regime as I have defended it?

To be sure, this is not the only question which remains to be answered. It should be apparent to the reader that there are many unsolved problems in connection to my argument, as well as to retributivism in a more general sense. We still know too little about proportionality: it is unclear whether anything more than ordinal proportionality is needed to defend retributivism, and how, if this is desired, we could arrive at cardinal proportionality. Furthermore, many problems remain with respect to criminalization and the relationship between illegality and moral wrongdoing. There is also a substantial debate on whether retributivism presupposes libertarian free will. Some are convinced it does not. Obviously, I have taken issue with this, but it would be foolish to rule out the possibility of some coherent desert-based philosophy, about punishment or otherwise, which does not rely in any way on what I referred to as the Autonomy Model. More analysis is also needed to address the relationship between punishment and democracy in general, as well as the more specific question, insufficiently laid out in this book, concerning the relationship between free will and democratic rule. These questions seem to me in need of further investigation, and surely there are more like them.

Every book must end somewhere, however, and I choose to end it by reflecting on the ultimate moral justification of a retributive penal

---

367 Apart from the reasons against retributivism mentioned so far (most notably the objection that it does not apply in non-ideal circumstances and, as we saw in chapter 3, that there is no non-arbitrary way of establishing proportionality between crime and punishment), a further objection can be raised, which has to do with the consequences of human fallibility. Since the justice system is operated by humans, and since all humans err, innocents will inevitably be punished. This is a great injustice according to retributivist thought, and it can therefore be wondered whether a retributivist can support actual penal regimes.
regime. Given that we all have reason to value a retributive regime, from where does its value ultimately derive?

A popular idea, stemming from Hart and Rawls, is that retributivism answers questions about whom to punish and how, but that utilitarian considerations, in emphasizing forward-looking protection of society, answer the more fundamental question of the “general justifying aim” of punishment. Retributivists have objected that it is an affront to say that they only justify acts of punishment but remain ignorant about the justification of the very institution of punishment. In their mind, they have answers to both questions: the institution of punishment is justified because it gives rule breakers what they deserve. But that answer only seems to push the question one step further. Why is it important to give people what they deserve? This, it will be recalled, was the question raised by Dolinko’s gap.

One answer, which I entertained in chapter 3, is that giving people what they deserve is intrinsically good. If this is true, it is a small task to justify the practice of punishment. Punishment is then simply the way in which we bring about the proper state of the world; a kind of maintenance of the moral balance where wrongdoers get what they deserve and are not allowed to gain unfair advantage through their wrongdoing. To use yet another distinction, intrinsic-good retributivists have no difficulties being bold retributivists, i.e. retributivists who hold that desert answers the question of the general justifying aim of punishment.

Whereas intrinsic-good retributivists deny Dolinko’s gap altogether, I have tried to traverse it by identifying reasons why being treated in accordance with one’s desert is valuable. But none of these reasons seem

---

368 Corlett, for instance, argues that the institutions of punishment is justified because it “logically coheres with other moral/social/political principles, ones which do not imply or entail anything about the primacy of deterrence or future consequences of the state’s inflicting of punishment of wrongdoers” (2001: 84). As he has subsequently clarified (Corlett 2003), the principles in question are tied to the notion of personal responsibility.

369 “No difficulties” is surely overstating it. As discussed in chapter 3, and as noted by Dolinko (1997), a bold retributivist is faced with questions concerning whether acting on desert is a perfect duty even outside the realm of criminal justice.
to present the ultimate reason *why* we punish – they only address, as it were, the properties of a just penal regime once we have one.

Think of it this way: I have argued that everyone, rule breakers included, have reason to prefer a retributive penal regime over regimes based on deterrentism or rehabilitationalism. The merits of a penal regime which aims to give rule breakers what they deserve are such that the regime is justified, at least when compared to rival answers. In this sense, a retributive regime lies in the interest of all (is the "right" of all). But this argument, which the reader by now is more than familiar with, leaves two questions unanswered. First, if the symbolic and institutional reason are what justifies a retributive regime, what kind of reasons are they? And secondly, are these reasons sufficient to justify the existence of a penal regime in the first place?

Consider the first question. A penal regime which treats people with fairness and respect is justified, I have argued. But why is fairness and respect important? Here we may note that answers can vary between what are traditionally called consequentialist and deontological answers. It is possible that fairness and respect is good because it leaves society better off, in terms of welfare or utility, than it would otherwise have been. So conceived, fairness and respect is valuable simply because it has *consequences* which work to the benefit of society, perhaps even to each and every individual. James Rachels, the retributivist whom we met in chapter 5, has argued this:

“To justify punishing someone, we may refer simply to what he or she has done – we may point out that they deserve it. But when we examine the arguments that support the general practice of treating people as they deserve, it turns out that those arguments all refer to ways in which people are *better off* under such a practice. So, at least as far as anything here is said, the ultimate justifications could all be utilitarian.” (Rachels 1997: 479)

But it is also possible that fairness and respect is good because it represents some deontological obligations we owe to each other, irrespective of their consequences for overall utility or welfare. Whereas the institutional reason for retribution is particularly suggestive of a consequentialist answer (unless the penal regime aims to give rule breakers what they deserve, we risk suffering misuses of penal power), it still seems that both the symbolic and institutional reason can be valuable for deonto-
logical or consequentialist reasons. They are compatible, so to speak, with both conceptions of morality.

Saying this takes us into deep waters. The distinction between deontology and consequentialism has become increasingly opaque. Retributivism has of course traditionally been understood as a deontological theory (Dolinko 1997). But this need not mean that a retributive penal regime, built upon the basis of retributive principles, is ultimately justified out of deontological reasons. One could argue that a retributive penal regime is a good idea and hold that it is good because it maximizes overall welfare, just as well as one could argue that it is good for non-consequentialist reasons. To admit that a retributive penal regime might be consequentialistically justified of course opens up for the difficult questions concerning the status of rules in consequentialist/utilitarian thought, which we encountered in chapter 5. But given that the retributive regime I have defended is non-strict, we seem able to escape the charge that rule utilitarianism bestows an incoherent status to rules. Perhaps genuinely desert-based rules are justified because of their very strong, but not exceptionless, tendency to maximize utility.

Suppose, however, that we settle for the more traditional, deontological interpretation: a penal regime should give us what we deserve because it represents a way of treating us which is right or appropriate in itself. We are then left with the second, more fundamental question of why there should be a penal regime at all. In Rawls’s words, this question concerns “why we have the institution of punishment: why do people punish one another rather than, say, always forgiving one another?” (Rawls 2001: 22). Here it seems that the deontological interpretation faces problems. It is one thing to say that a retributive penal regime represents a more appropriate way of treating people than a deterrentist or rehabilitationist one. It is another thing to say this is sufficient to show that we should have a penal regime in the first place.

370 For a slightly complicated but rewarding discussion of consequentialism and deontology, see Scanlon 1999. Part of the problem here is that even some traditional retributivists have defined themselves as consequentialists, whereas the position has classically been understood as a deontological one. Michael Moore has for example argued that since retributivism conceives deserved punishment as a good “state” or consequence, retributivism is consequentialistic. For a criticism of Moore’s view, see Dolinko 1997.
Part of the appeal of utilitarian theories of punishment is of course that they offer a highly plausible answer to why we punish; why we, say, opt for action rather than inaction when it comes to rule breaking. We do so, utilitarians argue, since it is instrumental in protecting society and the well-being of its members. This rational justification of the penal practice easily undermines any attempts to morally justify a retributive penal regime (Dolinko 1991). If protecting society from crime is really what the penal regime is for, then what status can retributivist principles have besides curbing misuses of power that either impact overall welfare negatively or are wrong in themselves? The merits of building the penal regime on desert are not enough, it can seem, to justify the existence of punishment as such.

Again, intrinsic-good retributivism have no problems answering this second question: a world where the guilty are punished is simply better than a world where they are not. But in our case, to be a bold retributivist seems more problematic. The institutional reason is clearly not an argument for punishment; it is an argument for a specific penal regime when compared to others. The symbolic reason could be analyzed in a similar way. Perhaps it is not until it is settled that there will be a penal regime at all that it is preferable to live under a regime which treats one with respect. This supports that the ultimate justification of penal regimes is consequentialist, or straightforwardly utilitarian. We have them because we want to curb crime, which in turn makes us better off.

Note, however, that there is something to the symbolic reason which may suffice to justify the existence of a penal regime. Perhaps the value of holding and being held responsible is enough to argue that we should have a practice of punishment. For this practice, just as other responsibility-based practices, could exist simply because it lets us express our reactive attitudes to the praise- and blameworthy actions of others. Thus conceived, we have a penal regime because it allows us to respond to wrongdoing in the appropriate way, i.e. with resentment.

It is curious that one can spend a great deal of time writing a book and yet come out undecided on what seems like the most fundamental question of all. I do think that retributivist principles should serve as the basis of a state’s penal regime. But more theorizing is needed in order for us to say more precisely why this is so. It is one thing to claim that people with different conceptions of morality could agree that a retributive regime is justified. It is quite another thing to say which of these conceptions, and the different justifications they represent, is most compelling.
REFERENCES


Dixon, Nicholas, 2001: "Boxing, Paternalism and Legal Moralism", *Social Theory and Practice* 27:2, pp. 323-44.


Hegel, Georg Wilhelm Friedrich, 2005: Philosophy of Right, Mineola, New York: Dover


Miller, David, 2004: »Holding Nations Responsible«, Ethics 114, pp. 240-68.


—, 1999: *What We Owe to Each Other*, Cambridge: Harvard University Press.


Scheid, Don E., 1993: »Kant’s Retributivism«, *Ethics* 93:2, pp. 262-82.


Shapiro, Ian, 2002: »Problems, Methods, and Theories in the Study of Politics, or What’s Wrong with Political Science and What to Do About it«, *Political Theory* 30:4, pp. 596-619.


Tamburrini, Claudio Marcello, 1996: Crime and Punishment?, Edsbruk: Akademityck AB.


