JUSTICE AND PUNISHMENT: THE RATIONALE OF COERCION.

Martin David Matravers

Thesis submitted for the degree of Doctor of Philosophy,
London School of Economics & Political Science.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis attempts to answer the question why, and by what right, do some people punish others? I begin by examining the retributivist theory, largely through an analysis of the work of Kant and Hegel. I conclude that no adequate justification can be given for the core retributivist claim.

I then go on to examine consequentialist theories of punishment and "mixed" accounts. I find that the former, like consequentialism generally, cannot accommodate the special value of persons and, thus, cannot give an adequate account of just punishment. "Mixed" accounts are also found to be flawed as they do nothing to resolve the tensions between their retributive and consequentialist elements.

I go on to examine the "fair play" theory. I conclude that such a theory cannot justify punishment nor can it capture the truth of our moral obligations. I argue that fair play theorists rely on a contractarian understanding of morality and it is this that should underpin the account of punishment. Turning to contractarianism, I look at three different approaches, justice as reciprocity, as mutual advantage, and as impartiality. I argue that none of these approaches can give either an adequate grounding to justice nor to punishment. Instead I argue for a combination of mutual advantage and impartiality. That is, I argue that each agent has a prudential reason to enter into an agreement with others to co-operate on moral terms, (that is terms which would be agreed between agents conceived as fundamentally equal), but that such a reason is not sufficient. In the absence of any decisive reason, I claim that each individual so agrees as an act of existential commitment. The community thus formed is a coercive one, for it is necessary that the condition of sufficient security be fulfilled through the coercion of free-riders. Such coercion is converted into moral punishment only through being addressed to the offender as a member of the moral community. The theory of punishment combines these two elements in "hard treatment" and retributive blaming.

Table of Contents

Acknowledgements

Writing a thesis is a lonely business, and it is certainly something that I could not have achieved without the friends I have been fortunate to have over the last few years. I have spent the last seven years in the Government Department at the LSE and I would like to thank everyone connected with it, especially those academics who helped me through. Notable amongst my early teachers were Janet Coleman, Kenneth Minogue, and George Schöpflin. Brendan O'Leary contributed as a teacher, advisor and friend.

Administratively, and in a myriad other ways, my passage was aided by Claire Wilkinson (who was the Departmental Secretary). I saved many hours because of her generosity in allowing me to print chapters on her computer, and spent most of them sat in her office, drinking coffee, and complaining about life. She became a good friend, whose favours are too many to recall.

Financially, I could not have undertaken this thesis without a grant from the Economic and Social Research Council, aided by a loan from my brother, Graeme Matravers, which cleared my undergraduate debts, and support from my mother, Anne, who constantly came to the rescue of a destitute son, usually when I needed someone to drive my possessions from one lodging to another.

Chunks of this thesis have been read to the LSE's Political Philosophy Seminar and the Political Theory Workshop. I would like to thank the participants in both of those, and to single out Duncan Ivison and Tim Stainton, not for any particular comments that I remember, but more for setting examples of how doing a thesis need not stop one reading widely and being interesting. My brother, Derek Matravers, has also read and commented on various bits of the thesis, always bringing a depth of insight that I have tried unsuccessfully to emulate.

Turning to personal acknowledgements: I would first like to thank Joy Drucker, even now I do not believe she realises how much she contributed to this thesis and to the conditions that made writing it possible. I would also like to thank Janet Murdoch for seven years of company through good and bad. The last eighteen months have been extremely difficult ones for me, and without the following people neither I nor this thesis would have come through them. In this regard I would like to thank, for numerous meals and invaluable support, Brian and Anni Barry, and John Charvet. It is impossible to find the words to express how much their friendship and support have meant to me. I would also like to thank Russell Bentley and Brenda Lewis. At the height of the difficult times I spent a couple of months in Cambridge with my brother, Derek, and his girlfriend (now wife), Amanda Hall. Without those months I would certainly not have made it; their quiet understanding, the way in which they let me get on with things without questions and yet were there to help when I seemed to be failing, and the number of hours they must have wasted joining me in vodkas with orange, were more than I deserved.

Since then - because of all these people - things have looked up, and for many happy moments I would like to thank Stephanie Norfolk. It

would have been even more difficult to find the necessary motivation to complete this if she had not been there at the end of each day. Things have also looked up because, in the absence of a fourth year grant, I took up the offer of becoming a sub-warden at an LSE hall of residence. It is certainly not the ideal place to work, but it has provided me with free accommodation, prepared food, and the company of the warden, Kurt Klappholz and the other sub-warden, Julika Siemer. I have known Mr. Klappholz since my first year at LSE and it gives me great pleasure to thank him now, not only for the job, but for his friendship and for what he has taught me about rigorous argument. My debt to Julika could be measured in the cups of coffee we have shared over the last few months, and it would be astounding, but it goes far deeper. She has become a close friend, a confidante, and a source of good advice; certainly the last few months would have been very different, and much less fun, without her.

To be in a department which boasts Brian Barry and John Charvet is lucky, to count both amongst your friends is miraculous. Over the past four years I have had the opportunity to discuss with Brian any and every aspect of political philosophy that I cared to; I should have learnt more than I have, but if I now display more critical bite and a more profound understanding of the subject it is down, in large part, to the exposure I have enjoyed to one of the discipline's greatest contemporary practitioners.

My first year at LSE included a class taught by John Charvet. I remember as if it were yesterday the first time I attended it; we discussed the first book of Plato's Republic and I can trace my addiction to the questions of political philosophy from that hour. Since then my debt to John has only increased. I have continued to learn from him and he has taught me more about political and moral philosophy than anyone else. He has also set an example both as a philosopher and a supervisor that I can only aspire to duplicate. Every argument in what follows has been meticulously examined by John, and usually we have met to discuss pieces within a week of my giving them to him, a record few supervisors could claim.

Over numerous supervisions and dinners I have been forced to improve my ideas in response to the sustained assaults and incisive criticism of these two great teachers. If what follows has any virtues the reader has reason to thank them, if it has any vices they are probably my own.

Introduction

> "He asked a very simple question: 'Why, and by what right, do some people lock up, torment, exile, flog, and kill others, while they are themselves just like those they torment, flog, and kill?' And in answer he got deliberations as to whether human beings had free will or not; whether or not signs of criminality could be detected by measuring the skull; what part heredity played in crime; whether immorality could be inherited; and what madness is, what degeneration is, and what temperament is; how climate, food, ignorance, imitativeness, hypnotism, or passion affect crime; what society is, what its duties are - and so on …, but there was no answer on the chief point: 'By what right do some people punish others?'
>
> L. Tolstoy Resurrection[1]

The aim of this thesis is to answer Tolstoy's question: Why, and by what right, do some people punish others? It might be thought that this is a dangerous topic for a graduate student who has to fulfil the requirements of making "a distinct contribution to the knowledge of the subject and afford evidence of originality",[2] for the problem of punishment has, if we date the beginning of the discipline of political philosophy to Plato, been around as long as the discipline itself.[3] Yet, the study of punishment has tended to be concerned less with the question of whether punishment as a social practice can be given a moral grounding, than it has with the question of

---

[1]Quoted in Duff 1986, 187. Duff cites Pincoffs 1967.

[2]University of London PhD regulations, in the LSE Calendar 1993, 836.

[3]Rather than cite some of the huge literature on punishment here, I will refer to works that I think are particularly helpful on certain problems as these arise in the text. I have applied this principle, as best I can, to the whole of this introduction; thus, e.g., in the (very introductory), discussion of utilitarianism below, the reader will find no references to the arguments I have used (unless I have quoted or found that I owed a special debt to a particular work for my presentation of the argument). When the arguments reappear in their full form in Chapter 3 they are, of course, fully annotated.

how one justifies the imposing of punishment on an offender; that is, punishment theory has tended to put the question "we have an offender, what are we morally justified in doing to her?" Rather than the question, "how can we justify having a set of rules the contravention of which renders the contravener liable to punishment?" This is why I think the quotation from Tolstoy so revealing, because the concerns of the punishment literature have, at times, resembled those that are given (in the quotation), in answer to Tolstoy's question. If we start by saying "we have an offender, what can we legitimately do", it seems natural to go on to ask whether we can blame her, or whether we can somehow account for her actions in a way that relieves her of blame. If we start by asking the second of my two questions, on the other hand, these things are no less important, but their place in the debate is different; the answers to these questions change the way we view the system itself, not the culpability or appropriateness of the offender.

I do not mean to suggest that the conceptual distinctions between these questions can, or ought to, be maintained rigidly. For the moment, however, I just want to signal two different approaches to the problem, it will be my contention that what we can do to an offender will be clear only if we solve the question of how we may legitimately think of him as an offender in the first place.

Traditionally, the main theories of punishment that have dominated the subject have been retributivism and utilitarianism.[4] Retributivists claim that the voluntary commission of an offence constitutes a necessary and sufficient condition of justified punishment, this because the act of offending makes the person morally *deserving* of punishment. The quantum of desert being a function of the gravity of the harm and the culpability of the offender. The retributivist, thus, uses desert to identify who should be punished and why; indeed, the quantity of desert is also meant to determine the correct response of the authorities, that is, how much the offender should be punished. Retributivism, in short, looks to the nature of the act.[5]

Retributivism seems to get some things right, our intuitions tell us that there is a relationship between the past act of offending and the punishment of the offender, indeed, if we endorse a definition of punishment as "of an offender for an offence",[6] this is part of the very meaning of what it is to punish. Similarly, we feel that the punishment should be proportional to the level of guilt, an offender who acted negligently is, under normal circumstances, thought of as less culpable than one who acted intentionally, even if the end result is the same harm.

---

[4]The following summaries owe much to Lacey's excellent short entry (Lacey 1987), "Punishment" in Miller 1987, 409-12.

[5]See §1.

[6]See §18.

Yet, retributivism is also a strange doctrine; it no doubt appeals to a strong intuition that "the guilty deserve to suffer", but it contradicts another, that "two wrongs do not make a right". Most seriously, retributivists do not seem able to explain *why* the guilty deserve to suffer, and when asked to do so they often seem to do no more than restate this basic proposition in different, and often "mystical" terms.[7] Finally, whilst retributivism is able, it seems, to muster most of the intuitive arguments on its side, utilitarians can ask whether, if we were sure that punishment yielded absolutely no benefit in either social or individual terms, it would be right to maintain the practice of punishment.

By contrast, utilitarian justifications of punishment, in common with utilitarianism generally, hold that punishment is to be justified by its consequences.[8] The most common form of such theory claims that punishment is justified because it deters the offender from future criminal acts (individual deterrence), and it deters others, (general deterrence). Other benefits claimed by some utilitarians include the reform or rehabilitation of the offender, her incapacitation (during periods of imprisonment), and the satisfaction of victims' grievances.

---

[7]Lacey 1987, 410.

[8]See Chapter 3.

Utilitarianism, thus, sees the infliction of suffering not as intrinsically required in response to the desert of the offender, but as a means to realising good consequences. This means that the onus is on utilitarian defenders of punishment not only to show that punishment has desirable net consequences, but that the system of punishment itself is superior to other systems of "social hygiene".[9] It seems likely that on some occasions better consequences could be achieved by inflicting punishment on an innocent scapegoat, or innocent members of an offender's family. In short, because utilitarianism locates value in states of affairs, rather than in individual agents, it can always consider, and may demand as morally obligatory, practices which are unjust.[10] Utilitarian justifications of punishment suffer from the same shortcoming as utilitarianism in general, that is, they cannot account for the special value of persons.

Given that the problems associated with retributivism seem to stem from it attempting to account for why we should punish at all, and those associated with utilitarianism from its inability to determine to the satisfaction of justice who we should punish and how much, the answer seems to lie in some combination of these two theories. This was

---

[9]See Duff 1986, 1-3; Mary Mackenzie describes utilitarian justifications of punishment that do not address this question as "institution begging", (Mackenzie 1981, 41).

[10]It may sometimes seem to us to be a moot point whether it is right to sacrifice one to the good of the whole, but when it is so, we view this as a sacrifice, as a tragic moral dilemma, not as a morally obligatory act over which we should, presumably, feel no pangs of conscience.

the insight of H. L. A. Hart's seminal 1959 article, "Prolegomenon to the Principles of Punishment", which launched what have become known as "mixed" accounts of punishment. Hart's idea was that the problem of punishment could be divided into three conceptually distinct questions: "What justifies the general practice of punishment? To whom may punishment be applied? How severely may we punish."[11] Hart argues that the general justifying aim of punishment is utilitarian, but the distribution of punishment, who should be punished and how much, must be determined by retributive considerations.

Mixed theories have appeared in various forms since Hart's article first appeared, usually following Hart in giving a utilitarian answer to the question of why we punish and using retributivist principles as side-constraints. Others have attempted to account for punishment on the basis of desert but subject to limiting utilitarianism. The apparent attraction of mixed accounts, however, is illusory because they do nothing to overcome the fundamental conflict between retributivism and utilitarianism. As they stand retributivism and utilitarianism pull apart, and merely separating punishment out into separate questions cannot overcome this tension.[12]

In fact, in many ways, separating out the elements of punishment is unhelpful, for it takes away from the

---

[11]Hart 1959, in Hart 1968, 3.

[12]See §8; §§19-20.

approach that is most likely to yield success. That approach is to examine punishment as a whole, and in the context of much broader questions of political theory. The reason why such an approach is most likely to succeed is that, far from separating utilitarianism and retributivism, what is needed "is to reconcile the consequentialist and retributive principles at a far deeper level",[13] and to do this it is necessary to place punishment back amongst the questions from which it first sprang; questions such as, how do we understand the nature of morality and moral judgements? Why ought we to be moral? Why should some aspects of immorality be the subject of law? And why should persons obey the law?

Surprisingly, this approach to punishment is not common, although it is becoming increasingly so, and if there is a rallying cry contained within this thesis, and a claim to meet the stringent criteria laid out above for what it should do, it is in the claim that what I have attempted to do is recapture the philosophical study of punishment from the specialist sub-field of punishment theory. Having said that, I cannot claim to have answered all the questions above, nor, indeed, have I met the criteria laid out by Braithwaite and Pettit for a "comprehensive normative theory of the criminal justice system",[14] (which include not only giving a moral justification for punishment but also answering numerous other questions, including such

---

[13]Lacey 1987, 411.

[14]See Braithwaite and Pettit 1990, Chapter 2.

things as how resources should be allocated in the criminal justice system). So what have I done?

The answer is that I have tried to do two things; first, to look at the traditional approaches to punishment to see whether they do indeed have the problems I have ascribed to them above, and whether their responses have been sufficient to meet these objections. Finding that they are, and haven't, I have turned my attention to the sorts of questions identified above, and here I have restricted myself to a certain kind of approach, contractarianism. In examining contractarian accounts of justice I have asked what rationale they give for coercion, and how this informs their accounts of punishment. At the same time, I have used the analysis of coercion to shed light on, and examine, the cogency of the theories of justice themselves.[15]

---

[15]I would like to make it clear as early as possible, that I have presupposed an adequate theory of autonomy. This was an extremely difficult decision for me, but three considerations eventually convinced me that it was better to assume autonomous persons, than give a full account of how such a position can be justified. The first is that such a pattern is followed in the contractarian literature with which I am concerned. The second is that, insofar as some liberal contractarians have begun to approach the problem, they have done so very inadequately and without realising the significant tensions a full theory of autonomy would raise within, especially impartialist, contractarianism, (references to this work can be found in Chapter 5, note 57). The final, and decisive, consideration was, therefore, that, whilst I am convinced that a full theory of autonomy can be given, it would have required an at least Chapter length disquisition. Given that such a Chapter would raise deep and general philosophical issues, I decided that this would have the great disadvantage of breaking up the argument unnecessarily, and that in a work of political philosophy it was better, all things considered, to leave such an enquiry to another time.

This is not an attempt merely to try to apply the techniques of distributive justice to the problem of punishment, viewing social protection (or the distribution of harm), as resources to be dealt with like any other.[16] Rather, it is an attempt to meet Lacey's challenge and approach the subject at "a far deeper level". My conviction was, and is, that punishment can only be morally justified if we can correctly ground necessarily coercive moral norms. That is, if we can give a satisfactory account of morality that does not contain a satisfactory rationale for coercion then coercion (and its legal analogue, punishment), will have to be justified by an appeal to a second argument, and this can only take the form of a fundamentally retributive, fundamentally consequentialist, or mixed account, all of which are unsatisfactory.

The task, then, is to ground morality in such a way as to make sense of the notion of moral judgements, of the idea that the individual ought to be moral, and that in some cases failure to be moral necessarily invokes coercion. This is what I have tried to do in this thesis, however, I do not want to say too much here to anticipate the argument that follows.

There remain two tasks for this Introduction; to give an account of the structure of the thesis and to justify the choices I have made in what I have examined. I begin with

[16]For an analysis of this kind see Harel 1994.

an analysis of retributivism, trying to make sense of the most basic claim that there is some relationship between the past act of the offender and his deserving punishment. I have chosen to look for a justification of this claim in the work of Kant[17] and Hegel.[18] The reason for choosing two theorists from the history of political thought, in what is a work of analytic political philosophy, rather than attempting to examine this claim in abstract terms, is that my conviction that punishment theory can only really be understood if it is located within a larger context meant that, whilst attempting an analytic examination of retributivism, I wanted to give it the best chance I could. It seemed to me then, and still does now, that retributivism (both contemporary and classic), often relies on background Kantian claims about respecting persons, and it thus seemed to make sense to examine it in the work of Kant himself.

Finding no satisfactory account of the retributive principle in Kant, I turn to Hegel. Again, this is because I believe that Hegel offers a profound and influential theory of morality and, since he enjoys the reputation of being a seminal retributive theorist, it seemed to me that if a justification of retributivism could be found, it was likely to be somewhere in Hegel's *oeuvre*. In fact, my conclusion is that Hegel is a great deal less retributive than is commonly thought.

---

[17]Chapter 1.

[18]Chapter 2.

There are some additional reasons for my choice of these two theorists; in the case of Kant, it is clear that a great deal of contemporary liberal theory, (most explicitly, John Rawls's A Theory of Justice), claims his work as its most important philosophical antecedent.[19] I, thus, wanted to get Kant clear before I moved on to such theorising. Hegel, likewise, is often claimed by communitarian critics of liberalism, and he has informed a philosophical tradition that is as illustrious as his forbear. Having generally found myself sympathetic to Hegel, especially in comparison with Kant, it seemed natural to move to him in search of a justification for retributivism, and in search of a deeper understanding of the problems I knew would appear in the second part of the thesis. Although I would not want to be associated with the current communitarian school, the theory I propose in this thesis owes, or so it seems to me, a great deal to the influence of Hegel, although seldom is this obvious.[20]

Having found no coherent way of grounding the retributivist claim that the guilty deserve to suffer, and, indeed, having found that this principle is often "fudged" in Kant, and never so simple in Hegel, I turn to consequentialism.[21]

---

[19]Rawls 1971. See Appendix A, §C for a discussion of how influential the "Kantian Interpretation" is in Rawls's theory.

[20]In addition, I found the existing literature on Hegel's theory of punishment very unsatisfactory and, thus, I hoped the Chapter would have something to add to Hegel *exegesis*.

[21]Chapter 3.

Here the reason is obvious, consequentialists have provided one of the most influential theories of punishment, and, as I point out above, one that seems to match at least one intuition, that the system of punishment must have something to do with the future behaviour of the population subjected to it.

Rather than choose to examine consequentialism in the work of, say, Bentham, I have chosen a more independent analytic approach. This is partially because it is easier to grasp the basic thesis of consequentialism in the abstract, but also because the background moral theory on which it relies is the same as the theory of punishment, the latter is more clearly just an application of the former, (in a way in which it is not in retributive theory). An additional reason is that there have been a number of developments in consequentialist theorising about punishment which could not be understood on a basic Benthamite model. The conclusion of Chapter 3 is that consequentialism cannot adequately include a respect for justice and that this is fatally undermining for consequentialist accounts of punishment.

Having found retributivism, consequentialism and mixed accounts[22] unsatisfactory, and, indeed, having found the very approach of such theories unhelpful, I turn to contemporary accounts of justice. I have already explained above why I believe that the route to morally justified

[22]Mixed accounts are dealt with in §§19-20; see also §8.

punishment lies in such a direction, however, I have not defended my limited choice of contractualist theories of justice. One possible justification would be the limits of time and space on a thesis such as this, however, I think something more substantial can be said.

First, I have rejected, for the purposes of this thesis, realist accounts of morality. I have not defended this choice but it seems to me that we have no good reasons to believe in, say, God, Platonic forms, or, for that matter, in Kantian noumenal selves. Starting from this position, it seems to me that we can only ground morality in human reason and will, and if we are true to our conviction that morality is not "out there", then this must be our starting point. The challenge is to avoid the slide into relativism on the one side, and the appeal to metaphysics, or God, or some teleological theory, on the other. Of course, such a position does not necessarily have to be unfolded in contractualist terms, but here I can say two things. First, I believe such terms offer the best hope. Second, insofar as this thesis is part of contemporary Anglo-American political philosophy, there is no doubt that this has been the predominant approach of the last twenty-five years by theorists keen to keep to the path between relativism and God, and it seemed to me to be necessary, in the first instance, to attend to their attempts.

With this in mind, I have divided[23] current accounts of justice into three; Chapter 4 examines justice as mutual advantage and justice as reciprocity, the latter encompassing what has become one of the cores of the retributivist revival; theories of punishment as "fair play". In Chapter 5 I turn my attention to theories of justice as impartiality.[24] Throughout, the intention is to use the rationale of coercion as a prism through which to examine the theories themselves, whilst at the same time trying to see whether the rationale, itself, is satisfactory.

I do not want to attempt a simplified version of the argument here, but, there is a sense in which the conflict between justice as mutual advantage and justice as impartiality mirrors the conflict between consequentialism and retributivism; justice as mutual advantage relying on the benefits of co-operation and justice as impartiality on a fundamental claim to respect others. Overcoming the objections to these theories and giving an account of my own, involves reconciling their separate demands at a deep philosophical level, and this is the argument proposed in Chapter 6, which claims to offer an anti-realist account of justice that successfully navigates the path identified above. Chapter 7 then extends this reconciliation to punishment, and concludes that a rationale can be found for

---

[23]Following Barry 1989; Kymlicka 1991.

[24]Because of the peculiarities of Rawls's account, but keeping in mind his importance in the discipline, I have examined his account of punishment in a separate Appendix.

coercion at the abstract level of the theory of justice, and that this has a legal analogue in morally justified punishment, I then go on to examine the account of punishment in greater detail.

One final comment; this is a thesis in political philosophy and, although I have enormous sympathy for those who complain that philosophers of punishment are criminologically and sociologically naive and enormous respect for those philosophers who are not, it, therefore, falls far short of offering a complete theory of punishment. As a matter of fact, I do not believe that providing such a theory is possible philosophically, because, as will be clear below, I believe that the abstract philosophical demands that can be justified may take many different concrete forms in the real world. Nevertheless, the best I can maintain is that I have given an abstract justification for coercion, and, in the right empirical circumstances for punishment.

Chapter 1: Retributivism

> "The law of punishment is a categorical imperative, and woe to him who crawls through the serpentine windings of the doctrine of happiness looking for some advantage to be gained by releasing the criminal from punishment or by reducing the amount of it!" (Kant 1797, 141)[1]

1. Retributivisms

Retributivism has traditionally been regarded as one of the "great schools" of punishment theorising, opposed by the other, consequentialism. Yet, retributivism has had a decidedly mixed history. In 1969 one standard text on punishment recorded that "there no longer are defenders of the traditional retributive theory…. At any rate, there are no defenders writing in the usual places".[2] By 1984, however, the new edition of this book had a postscript dedicated to the "new college industry [which] turns out theories of retribution".[3] What is, perhaps, more interesting is that retributivism became discredited not only because it is, as we shall see, difficult to find philosophical arguments in its support, but because the task of finding such arguments itself became the subject of disdain. Just at the time when contemporary political philosophy found its greatest defender of the importance of our "considered moral judgements", in John Rawls's A Theory

---

[1]I have followed A. C. Baier's translation (see Baier 1993, 441 and 17n), rather than that of Mary Gregor.

[2]Honderich 1969, 148.

[3]Honderich 1984, 10.

of Justice,[4] it became intellectually fashionable to deny that the sincerely held conviction that punishing offenders was morally good could be anything other than an expression of barbarism or a throwback to a more primitive age.[5]

One standard problem with considering the retributivist justification of punishment is that there is no consensus on what retributivism actually is. Retributivism has no clear definition, it resembles an "essentially contested concept" rather than a rigorous term which can be used to distinguish clearly one non-consequentialist theorist from another. Indeed, given the bewildering number of writers who lay claim to the ascription "retributivist", it is difficult to imagine giving anything but an affirmative answer to Antony Duff's question, "has the label 'retributivist' been applied to such a diversity of views and principles that it now lacks any unambiguous or unitary meaning?"[6] Nonetheless it is possible to make certain distinctions which, if they will not solve the question of what retributivism is, might go some way to telling us what it concerns and what it is not.

The first thing that requires clarification is that in this chapter I am primarily concerned with the question of what

---

[4]See Rawls 1971, 21; Kymlicka 1990, 7-8.

[5]For an excellent discussion of this see Moore 1987.

[6]Duff 1986, 4. Cottingham likewise remarks that, "the fact is that the term 'retributive' as used in philosophy has become so imprecise and multi-vocal that it is doubtful whether it any longer serves a useful purpose." (Cottingham 1979, 238). See also Ten 1987, Chapter 3.

justifies the general practice of punishment, and it is thus not sufficient to identify retributivism with the *jus talionis*, or any other answer to Hart's third question;[7] "how much?" Having established this, I do not believe the second of Hart's questions, "whom should we punish?" can be separated off as easily; retributivism is closely linked with culpability and desert, and thus who might qualify as an object of punishment is entailed by any retributive theory, although the question of whether the offender should then be punished might depend upon the precise formulation of the theory in question.

A quarter of a century ago, H. L. A. Hart suggested a model of retributivism as follows:

> "Such a theory will assert three things: first that a person may be punished if, and only if, he has voluntarily done something morally wrong; secondly, that his punishment must in some way match, or be the equivalent of, the wickedness of his offence; and thirdly, that the justification for punishing men under such conditions is that the return of suffering for moral evil voluntarily done, is itself just or morally good."[8]

Although it has been largely ignored in the literature, (and is only introduced by Hart in order to be "modified"), this simple model repays serious consideration. The first claim has it that it is permissible to punish someone if,

---

[7]Hart 1959, in Hart 1968, 3.

[8]Hart 1968, 231. For a useful discussion of this model see Bedau 1978.

and only if, they are morally culpable. The second that punishment must be proportional, and the third that if these conditions are met, punishment is just or "morally good". I shall, for the moment, ignore the second of these claims[9] and examine the first and third.

The first claim matches up to two of the most controversial questions within retributivist writing: Is punishment permissible (given that the other conditions for justified punishment are fulfilled), or obligatory? Second, is punishment a response to legal or moral culpability? Both of these questions will be considered below, however, what is important to note here is that there is nothing in this principle that makes it particularly *retributivist*. It could for example be accepted by a rule-utilitarian theory, an expressivist theory, a victims' grievance satisfaction theory, a moral education theory, as well as by other more marginal approaches such as restitution theory. What, then, must be at the core of retributivism as a general justification for punishment is the third of Hart's claims. Yet Hugo Adam Bedau, in considering this third claim, echoes Honderich when he says, "I cannot think of any retributivist today who defends R_3 [Hart's third claim]".[10] And Hart, himself, says of it that it,

> "appears to be a mysterious piece of moral alchemy in which the combination of the two evils of moral

[9]On the grounds identified above that I am not concerned with the question "how much?" but rather with retributivism as a justification for the practice of punishment.

[10]Bedau 1978, 615.

> wickedness and suffering are transmuted into good…[or]…it seems to be the abandonment of any serious attempt to provide a moral justification for punishment."[11]

I shall, for the greater part of this and the next chapter, be concerned with this "moral alchemy", largely through considering two of the most seminal retributivist thinkers, Kant and Hegel.

There is, however, still the issue of precisely what Hart's third claim amounts to. In at least one version - normally proposed by anti-retributivists - this claim is absurd. That version has it that punishment in itself is a moral good. This is absurd because it would mean that the behaviour of the offender in offending was a means to the production of a moral good and thus of moral value, and this is obviously false. Rather, what this claim, at its strongest, must mean is that if and only if an offence has occurred, the punishing of the offender is morally good; the world would, of course, be a better place without either the offence or the punishment. In other words, what the third claim must amount to is that punishment is morally justified by the past act of the offender, for it is the past act which creates the conditions by which (a certain regulated and correctly imposed), harming of an agent (a *prima facie* wrong), is converted into a punishment which is morally right. Let us try to make sense of this claim.

---

[11]Hart 1968, 234-5.

Michael Moore states that the central core of retributivism is,

> "the view that punishment is justified by the moral culpability of those who receive it. A retributivist punishes because, and only because, the offender deserves it".[12]

He goes on to distinguish this from other claims which in his opinion are labelled "quite misleadingly, 'retributivism'".[13] These other claims concern such things as the amount of punishment (the *jus talionis* for example), victims' grievance satisfaction, society's grievance satisfaction, the avoiding of private vendettas, expressivist theories, and formal theories of justice.[14] Further, Moore argues that Hart merely establishes guilt as a necessary rather than a necessary and sufficient condition for justified punishment. Interestingly Moore appears to take his own proposal to be that guilt is a necessary and sufficient condition, and this to mean that his proposal "gives society more than merely a right to punish culpable offenders… but also gives society the *duty* to punish."[15] However, it is not at all clear how this follows; that guilt is a necessary and sufficient condition for justified punishment means just that if, and only if, someone is guilty then their punishment would be justified. In other words society has a good reason to punish and, in

---

[12]Moore 1987, 179.

[13]Moore 1987, 179.

[14]Moore 1987, 179-181.

[15]Moore 1987, 182.

the absence of reasons to the contrary, the offender ought to be punished. That is not necessarily the same as saying that the offender *must* be punished all things considered. This would only be so if one assumed a moral theory of a particular kind. Crude utilitarianism, for example, admits of no distinction between actions which are morally justified and those which are morally obligatory, but this is not often thought to be a characteristic in its favour, not least because of the problem of supererogatory acts.[16]

Nonetheless, Moore's distinction is crucial; as noted above with respect to Hart's first claim, a theory of punishment which held that guilt was merely a necessary condition for punishment could be consequentialist - rule-utilitarianism, for example - or the guilt condition can be used to answer the distributive question, "whom should we punish?", rather than the question of justification, "why punish?".

The problem is to find a set of claims that, whilst not tightly defined, provide the framework for understanding certain theories as being retributive. The most obvious similarity between theories competing for the term is in what they are not, they are all not consequentialist. This is far from being as glib a remark as it may seem, for what follows is that if a theory claiming to be retributivist can be revealed to be relying in some manner on a version

---

[16]In short, necessary and sufficient means that all and only those who are culpable offenders *may* be punished, and other things being equal ought to be. Moore seems to think that it means that they "must" be in all cases, i.e., all things considered.

of consequentialism, then it may be dismissed *as* a retributivist theory.[17] Put positively the retributivist claim is that punishment is to be justified through its relationship to what has already happened, and, further, retributivist theories explain that relationship through a notion of desert. Thus, in essence, Moore has the heart of retributivism right.[18]

The problem with this, however, is that this description of the central core of retributivism does not get us very far because the nature of each of the key terms is, itself, a subject of controversy. Central to Moore's way of characterising retributivism is the notion of desert, yet one can surely ask what it means to say that an offender deserves punishment, just as one can ask what it means to say that the punishment of an offender is morally good. This is, perhaps, why Hart avoids recourse to the notion of desert in characterising retributivism.[19] In the end, however, one surely has to take Hart's model as expressing the conditions for deserved punishment; the notion of desert is simply unavoidable if retributivism is to sustain its traditional form.

---

[17]I see no reason to limit non-retributivist theories to those which are concerned with "crime control" (as does Scheid 1983), or conversely to define retributivism as a theory in which crime control is not "morally relevant".

[18]And, indeed, he is right in distinguishing retributivism from other non-utilitarian theories of punishment.

[19]Cf. Walker 1969, 1-22; and Walker 1991, 72-82.

The multiple examples of retributivist theories derive from the differences in understanding this central core, that is, the idea that the offender's past action makes that offender morally deserving of punishment. Two points of the debate are crucial: First, the question of whether guilt is not only necessary and sufficient for justified punishment, but also creates an obligation on society to punish. Second, the question of whether guilt is to be construed in legal or moral terms.

Within retributivist writing one must, then, be careful to distinguish what I shall term hard moral retributivists, i.e., those who believe that punishment is an obligation once moral guilt has been established; hard legal retributivists, who likewise would claim punishment is an obligation once legal guilt has been established; and soft moral (or legal) retributivists, who would argue that moral (or legal) guilt makes punishment morally permissible.[20]

In the remainder of this chapter, and in the next, I intend to examine the views of two seminal retributivist

[20]It might be argued that a pure moral retributivist position on *punishment* is unsustainable because legal guilt is a necessary condition of punishing, the definition of punishment being that it is imposed by an authority for the contravention of a law. A moral retributivist could, of course, agree and argue that all moral offences ought to be legal offences, but, more likely he might argue that moral guilt is a necessary and sufficient condition for the imposition of suffering and that, when it is accompanied by legal guilt, such suffering is imposed as punishment. This could take the form of arguing that *some* law should be found to punish the morally guilty because they deserve to be punished, even if their moral offence is not the same as the legal charge. Thus, The Guardian (25 January 1993), reported that a Thai monk had been charged with damaging a coffin because no offence barring sexual intercourse with a corpse (the monk's "moral" offence), could be found under which he could be charged.

theorisers of punishment, Immanuel Kant and G. W. F. Hegel, and in so doing I will examine interpretations of their work which have often been adapted as free standing theories in the modern literature.

2. Kant

There is little doubt that punishment was for Kant, and is for Kantian theorists, an especially difficult problem. Kant recognises the necessity of punishment yet faces the fact that it is, *prima facie*, an infringement on the individual's autonomy. What is more it is an infringement that is made necessary by a contingent fact - that of the individual having committed a criminal act. Indeed, in The Critique of Pure Reason, Kant argues that punishment would be eliminated from the ideal society.[21] Nonetheless, Kant believes that punishment is obligatory once guilt is established, so, if he turns out to be a retributivist, he will, in my terms, be a hard one.[22] At this stage it would be premature to give an answer as to whether his concern is with moral or legal guilt.

It may seem to some extraordinary to write in this manner about Kant; that is, as if there were any ambiguity in his arguments concerning punishment. The traditional view has

---

[21]Kant 1787, 312: "The more legislation and government are brought into harmony with the [ideal], the rarer would punishments become, and it is therefore quite rational to maintain, as Plato does, that in a perfect state no punishments whatsoever would be required".

[22]See, for example, Kant 1797, 142.

it that if anyone holds to unreconstructed (and in the opinion of some unthinking[23]), retributivism then that person is Kant,[24] and quotes such as the one which prefaces this Chapter abound to show just how retributivist Kant's philosophy is. Yet, as several recent commentators have pointed out,[25] such a reading of Kant has to account for other passages which are as clear, but as clearly not retributivist. Consider the following two passages from the Lectures on Ethics:

> "Punishments are deterrent if their sole purpose is to prevent an evil from arising; they are retributive when they are imposed because an evil has been done. Punishments are, therefore, a means of preventing an evil or of punishing it. Those imposed by governments are always deterrent. They are meant to deter the sinner himself or to deter others by making an example of him."
>
> "Ruling authorities do not punish because a crime has been committed, but in order that crimes should not be committed."[26]

---

[23]Cooper, for example, says that Kant has "no theory at all beyond the denial of utilitarianism". (Cooper 1971, 160). For similar (and other) criticisms of Kant's views on punishment, see Arendt 1982, 7-8; Brown 1962; Cohen 1939.

[24]E.g., "The most thoroughgoing retributivists, exemplified by Kant, maintain that the punishment of crime is right in itself…." (Benn 1967, 30).

[25]See, for example, Byrd 1989; Tunick 1992b, 95-101; Riley 1983, 107-110; Scheid 1983.

[26]Kant 1930, 55 and 56.

It will be the claim of this Chapter that Kant's theory encompasses deterrent, retributive and (to a lesser extent) fair play elements. In order to understand how this is possible, however, it is first necessary to look more generally at Kant's philosophy, although it will only be possible to do this in an extremely restricted manner.[27]

3. The Noumenal and Phenomenal Worlds

At the heart of Kant's philosophy is a distinction between the sensory, phenomenal, world and the intellectual, noumenal, world. This bifurcation is repeated in various forms; in the distinction between things as they appear to us and things in themselves, between man as a determined and man as a free being, and between the moral and juridical realms. The most important form of this distinction, as it concerns punishment, is between the moral and the juridical, each of which is characterised by very different features. Taking Kant's ethics - the moral world - first; Kant argues that the only thing that is good in itself is the good will, which, as will be shown below, means that our actions to be moral have to be free, in the sense that an individual is free when his will is not determined by anything other than itself. Pure practical reason - i.e., practical reason which is not determined by contingently given ends - must, then, be sufficient to command action, and the individual in according his

---

[27]For good brief introductions to Kant's moral philosophy see Allison 1990; Hill, 1992; O'Neill 1991.

behaviour to such reason achieves autonomy;[28] it must also be formal and unconditional because it must be sufficient to command independently of contingent factors about the world (including, of course, factors about the agent). The form given by Kant to pure practical reason is the categorical imperative, which in its most general form is the injunction to: "Act only on that maxim through which you can at the same time will that it should become a universal law".[29]

The categorical imperative embodies the "pure idea of law" that is, it commands "unconditionally, …universally or without qualification."[30] The content of the law is derived from testing subjective maxims against the requirement of the categorical imperative; those which pass are, then, substantive moral laws. I shall consider one example, that of promise keeping below, but first, I wish to press ahead with the characterisation of Kant's ethics.

The question which must arise concerns the relation between actual agents, that is between members of the phenomenal world, and the moral law, for the moral law must, as noted above, be sufficient to motivate action. The agent must not act simply in accordance with the external demands of the moral law, but it must be the case that his actions could have been motivated purely by duty to that law, in

[28]For an alternative view of Kantian autonomy see Hill 1989.

[29]Kant 1948, 84.

[30]Charvet 1981, 70.

the absence of any other contingent reason.[31] But how is it possible for the moral law to determine the phenomenal agent? Kant, himself, admits that for this to be possible man must exist not only as a phenomenal being, but also as a rational being whose end is to live in accordance with its own nature which is itself rational. If morality is to be possible this being must be entirely free, i.e., self-determining, because it must not be the case that adherence to the moral law is a matter of contingency or circumstance.

Kant's argument for our being free depends again on this bifurcation between the phenomenal and noumenal worlds. Kant argues that whilst we can only have access to things in the world through our sensory faculties, there must be some way these things are, independently of our relations to them; they must exist also as "things in themselves". Applied to ourselves, that is to human beings, we know ourselves as phenomenal beings but we must also suppose that each of us exists as a "thing in itself", as what Kant calls a "noumenal" being. We can, of course, have no empirical knowledge of things in themselves, of the noumenal world, but through our special capacity of reason we can have rational belief about it. Reason attempts to go beyond experience and empirical knowledge and seeks an originating cause which is itself not an effect of some other cause; reason demands that at some point there is an "unmoved mover".

---

[31]See *infra*, note 36.

Kant claims that freedom is such an idea of reason; man is to be thought of as causally determined when conceived as a phenomenal being but as also belonging to another realm, a realm of reason, in which conceived as a member of this realm - as a noumenal being - man is independent of causality and is in possession of a free will, an originating cause. The moral world is thus coextensive with the noumenal. Moral action is that which is in accordance with the categorical imperative and in the absence of other contingent reasons it is performed out of a sense of duty to the moral.[32] In short, the agent in living in accordance with the moral law is living in accordance with his own nature as a free rational being.

Turning now to the juridical realm. Unlike the moral, the juridical realm is the realm of externality. In other words, its concern is not with the internal motivations of the agent but only with the agent's external actions, man conceived as a phenomenal rather than as a noumenal being.[33] Similarly, juridical legislation is addressed to the contingent, subjective motivations of the individual, as against the moral law which addresses the objective

---

[32]Where other motivations are present it must be the case that in their absence the motivation of duty to the moral law would have been sufficient to ensure the agent's performance of the action. See *infra*, note 36.

[33]"Kant is clear that what matters in morality is the good will, or the incentive of one's actions, while all that counts in politics and law is that one's external behavior (however motivated) be consistent with everyone's freedom under a universal law", (Riley 1983, 108). See also Byrd 1989; Scheid 1983, 266-71.

motivation of duty to the categorical imperative. This, of course, means that the fulfilment of the agent's juridical obligations is empirically verifiable; all one needs to do is check that the agent did indeed do that which he was obliged to do.[34]

Compare this with the moral realm. What is at issue in the juridical world is, as noted above, the freedom to form and act upon subjectively chosen maxims. The motive, or end, which guides the agent is irrelevant in juridical terms. The morality of the agent's actions, on the other hand, is determined only by the motivation; if, and only if, there is an *internal* correspondence between the subjectively chosen maxim and the demands of the categorical imperative can the action be said to be truly moral. Because the correspondence has to be internal it is obviously not the case that the subjective maxim must simply result in *actions* which the categorical imperative would endorse as right, for that would only be an external correspondence. Consider Kant's most famous example, that of promise keeping.

Kant argues that promise keeping is a moral duty[35], in other words it is commanded by the categorical imperative.

---

[34]Byrd 1989, 156-162; Fleischacker 1992, 202; Scheid 1983, 262-68. Further characterisation of the juridical realm follows below, §4.

[35]For the sake of simplicity I am, for the moment, ignoring Kant's classification of duties (perfect and imperfect; to self and to others). Good discussions of these can be found in Byrd 1989, 168-9 and Scheid 1983, 266-7. Kant's argument that promise keeping is a requirement of the categorical imperative can be found in Kant 1948,

Let us say that person P makes a promise to person A to do X at some specified time Y. The categorical imperative tells us that P ought to do X at time Y. At time Y, P does X, and thus his external actions are in accordance with the demands of the moral law, nonetheless the performance of X at time Y by P was only a moral action if P would have been sufficiently motivated by duty to the moral law in the absence of any contingent factors, such as that it was in his interest to do X, or even, that he wanted to be the sort of person who performs his duty.[36] The morality of an action, thus, is only discernible to one who knows the internal motivations of the agent, and Kant claims that the only being with such a capacity is God. Not even the agent, himself, may be sure of his own reasons for action.[37]

The significant differences between the moral and the juridical realms should now be clear. Moral and juridical legislation may both command the performance of an action,

---

67-68. It is not my purpose here to evaluate this claim, a brief discussion of its problems can be found in Charvet 1981, 72-74.

[36]Byrd seems to argue that Kant demands that for an action to be moral it must be the case that the performance of the action was motivated only by duty to the moral law. This is too strong; if the agent is motivated to perform the action by some contingent motivation (say a desire to be the sort of person who performs his external duties) then the action is still moral *if* in the absence of this motivation the agent would still have performed the action out of pure respect for the moral law. It is a common error to suppose that because Kant demands the presence of the motivation of duty, he demands that it is in any given case the actual motivation for the action. Of course, we cannot know whether, in any given case, the motivation of duty would have been (or was) sufficient.

[37]"The real morality of actions, their merit or guilt, even that of our own conduct, thus remain entirely hidden from us." (Kant 1787, 475n); See also Kant 1948, 74-5; Byrd 1989, 161; Fleischacker 1992, 202.

but from the perspective of morality, that is conceived independently of any contingent factors, it is also demanded that the action would be performed for the right reason, and thus morality addresses itself only to the internal motivations of the agent. Finally the performance of a moral action is unverifiable except by God. Juridical legislation addresses itself only to the external motivations and the performance of the agent. It asks merely that the agent do the action, not that she do it for a particular reason. It is thus easily verifiable, we know whether P kept her promise to A simply by asking whether P did X at time Y. We cannot know whether P thus deserves the accolade moral, that is only for God to judge.[38]

4. Legal and Moral Punishment

Kant's bifurcation of the moral and juridical realms means that it is neither wise nor helpful to talk of punishment *per se* in Kant's philosophy; one can only speak of legal or moral punishment. For the moment I shall put aside the moral realm and the question of the relation between the juridical and moral perspectives; both of these concerns will be dealt with below, but first to the question of what

---

[38]This interpretation, in which the motivations of the agent are opaque to others, seems to me to be the most coherent if Kant's distinction between the moral and juridical realms is to stand up. It has to be admitted, however, that Kant, when considering the *jus talionis*, occasionally talks of legal punishment in proportion to "inner wickedness", (see Kant 1797, 142). The balance of his remarks, however, support the reading given above; see, for example, Kant 1960, 87 and 91; this is also the conclusion reached by Hill 1978 in Hill 1992, 185-87.

justifies legal punishment, that is, punishment in the juridical realm.

The best way to approach this question is to ask what role the juridical realm plays in Kant's philosophy. In order to answer this two things must be held constantly in mind: First, the differences between the juridical and the moral perspectives discussed above, and second that Kant was profoundly concerned with the quest for stability and peace. One can best examine the role of the juridical realm through the device of the social contract, envisaged as a device for discovering the rational principles for regulating the inter-personal co-operative actions of a group of people, that is, of a society. This is a device which was familiar to Kant:

> "The problem of organizing a state, however hard it may seem, can be solved even for a race of devils, if only they are intelligent. The problem is: 'Given a multiple of rational beings requiring universal laws for their preservation, but each of whom is secretly inclined to exempt himself from them, to establish a constitution in such a way that, although their private intentions conflict, they check each other, with the result that their public conduct is the same as if they had no such intentions.'"[39]

Kant argues, on this interpretation, that the rational course of action for a group of people, when that group

---

[39]Kant 1795 in Kant 1963, 112.

contains likely criminals,[40] is to institute a system of rules and a system of punishment. The latter is needed because without it the enterprise would fail and the society would decay into anarchy. This outcome is not only because of the sum of individual misdeeds, but because without punishment what Hobbes called the condition of "sufficient security"[41] could not be attained, and thus the trust necessary to the success of any co-operative venture would be absent. The condition of sufficient security holds when each individual is sufficiently convinced that others will co-operate (by for example keeping their word), so that he is prepared to enter co-operative ventures, (for example, a transaction based on a promise).[42] But this only tells us that the juridical realm applies to civil society and the maintenance of peace. What it does not tell us is why civil society is important.

Clearly the answer to this question lies in the fact that civil society creates and maintains the conditions necessary for external freedom, i.e., for the freedom to develop and act upon subjectively chosen maxims, without interference from others, with the important *caveat* that such freedom must be equal. In more modern terms, civil

---

[40]This condition is what distinguishes the actual from the ideal world, in which punishment would be unnecessary; laws "will naturally also be penal laws if there are any criminals among the people". (Kant 1797, 143)

[41]Hobbes 1651, Part I, Chapter 15, 215.

[42]The problem of the necessity for conditions of trust to prevail is now referred to as the "assurance problem". It is the subject of an extended discussion below, §36.

society functions to provide the greatest possible external freedom compatible with equal freedom for others.[43] Consider the case of possession, which is central to Kant's justification of civil society. Kant argues that possession is necessary to freedom, possession being required for use, and use for freedom of choice. Justice as such, or the "mere concepts of Right",[44] cannot yield the right to possession and therefore it must be a postulate of reason. This claim,

> "put[s] all others under an obligation, which they would not otherwise have, to refrain from using certain objects of our choice because we have been the first to take them into our possession."[45]

But if this is so, and Kant succeeds in establishing the necessity of possession, then it establishes a universal law placing an obligation on all to refrain from using the possessions of others. If this were non universal, it would merely represent a one-sided attack on the freedom of some, for in this instance one would be obliging another to respect one's possessions without taking on a reciprocal obligation. The security that each will undertake his obligations is given, in Kant's view, by civil society.[46]

---

[43]"A constitution allowing *the greatest possible human freedom* in accordance with laws by which *the freedom of each is made to be consistent with that of all others…*" (Kant 1787, 312 emphasis in original). Cf. John Rawls's first principle (the "liberty principle"). Rawls 1971, 250.

[44]Kant 1797, 69.

[45]Kant 1797, 69.

[46]See Byrd 1989, §4; Murphy 1970 presents an interpretation of Kant's philosophy based very largely on this idea of reciprocity.

> "When I declare … something external to be mine, I thereby declare that everyone else is under obligation to refrain from using that object of my choice, an obligation no one would have were it not for this act of mine to establish a right. This claim involves, however, acknowledging that I in turn am under obligation to every other to refrain from using what is externally his…. I am therefore not under obligation to leave external objects belonging to others untouched unless everyone else provides me assurance that he will behave in accordance with the same principle with regard to what is mine. … It is only a will putting everyone under obligation, hence only a collective general (common) and powerful will, that can provide this assurance. But the condition of being under a general external (i.e., public) lawgiving accompanied with power is the civil condition."[47]

We are now in a position to consider some of the accounts Kant gives of the role and justification of legal punishment in civil society.

5. Deterrence and Respecting Others

Civil society, then, is formed by the agreement of the members to live under a common authority which organises itself through law. This agreement must be understood as juridical, that is between phenomenal beings who are self-interested and concerned to safeguard their external

---

[47]Kant 1797, 77.

freedom.[48] For this reason, the laws "will naturally also be penal laws if there are any criminals among the people".[49] Punishment, thus, appears consequentialist in nature; if the role of civil society is seen as providing a condition of sufficient security then it seems to follow that punishment is to be understood as a mechanism for maintaining civil society through deterring potential offenders, thus, creating the condition of sufficient security.[50] In short, the justifying purpose of the system of punishment is to protect the social order; to maintain civil society so as to ensure continued external freedom.[51]

If such a agreement were purely to be understood in juridical terms, then the deterrence aspect of legal punishment would not contradict Kant's conviction that morally good action is done with a pure will, because such punishment is addressed to the external motivations and performances of the citizens as phenomenal beings. This analysis would also quite satisfactorily explain the deterrence based comments of Kant quoted above, and it can also claim textual support from elsewhere in Kant's work. For example, in the discussion of the drowning man which

---

[48]By far the best exponent of this idea is Jeffrie G Murphy, 1970, 1972 and 1973; see also Williams 1983, 98-102; Aune 1979, 166; Shell 1980, 161. Arguments against this interpretation can be found in Byrd 1989, 181; Scheid 1983, 265 n7.

[49]Kant 1797, 143.

[50]See Byrd 1989; Tunick 1992b, 95-102; Scheid 1983.

[51]Byrd would prefer to say that the "threat of punishment" is deterrence based, rather than that the justification of the system of punishment is so. (See Byrd 1989, 183 105n) However, this distinction does not seem to me to be particularly useful, see §8.

appears twice in his writings, Kant argues that a man in a shipwreck, who, faced with certain death by drowning, "shoves another, whose life is equally in danger, off a plank on which he had saved himself", could not be *legally* culpable because "the punishment threatened by the law could not be greater than the loss of his own life."[52] In other words, the use of the law in such cases could not be justified because the law could not have a deterrent effect, the certainty of death from drowning being greater than that of judicial execution, for the latter depends upon one's being apprehended and convicted.

The problem with this analysis is that it ignores the noumenal world altogether. What it has ignored is the important *caveat* that civil society is for the protection of *equal* freedom. This must be the case, for although the agreement to organise in civil society is between phenomenal beings, it must, if it is to be compatible with the requirements of the universal law, respect the categorical imperative's demand that one should:

> "act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end."[53]

---

[52]Kant 1797, 60. See also Kant 1974, 68. Nonetheless the man is not *inculpable*. See below, §9.

[53]Kant 1948, 91. Hence the argument for the universal demand to respect the right of possession.

The deterrence based system of punishment, then, is not an unfettered consequentialism, subject to all the objections that plague that theory, most of which revolve around the fundamental claim that consequentialism treats people as means to better states of affairs.[54] Rather, the agreement is to co-operate under equal laws with the condition that whoever breaks those laws will be coerced.

The rationale of coercion, however, is unclear. At first, it appears as if Kant's argument is that those who do not obey the law should be coerced in order to satisfy the demand of providing the condition of sufficient security, but this would still be to treat people as means for a social good, it would not treat them as "ends in themselves". The problem is in seeing how the demand that the law guarantee equal freedom is translated into punishment practices. If Kant is to constrain the consequentialist rationale by the demand that the agreement be such that it is compatible with what could be agreed between noumenal beings, he needs to show that retributive punishment is what is required to treat people equally, in which case the "added bonus", so to speak, would be the provision of the condition of sufficient security.

Certainly there is plenty in Kant to support his fearful reputation as a seminal retributivist theorist. Consider one of Kant's more (in)famous passages from The Metaphysics of Morals:

---

[54]See *supra*, Introduction; §18.

> "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; for otherwise the people can be regarded as collaborators in this public violation of justice."[55]

The key lies in the language of doing to the offender what "his deeds deserve", this suggests a far more retributivist theory than the deterrence based analysis would allow, as does Kant's unambiguous support for the *jus talionis*.[56] As Patrick Riley has pointed out, Kant "often wants to be able to say that punishment must be deserved or merited", and in fact he often does say just that.[57] Kant must find an argument, then, that will allow him to constrain the agreement between phenomenal beings by the demand that treating people equally requires that the guilty suffer a punishment that is imposed on the offender "only *because he has committed a crime*."[58] The central question, then, is

---

[55]Kant 1797, 142.

[56]See Kant 1797, 141: "But what kind and what amount of punishment is it that public justice makes its principle and measure? None other than the principle of equality. …only the *law of retribution (ius talionis)* … can specify definitely the quality and the quantity of punishment".

[57]Riley 1983, 109. The discussion of punishment found in Kant 1797, 100-102, contains numerous retributivist sounding passages.

[58]Kant 1797, 141.

given the necessity of a system of punishment, how is it compatible with the requirement to treat people as ends? This is, in a sense, no more than to ask how punishment can, in Kant's view, be deserved, and we have thus reached the core intuition of retributivism. There are two arguments marshalled by Kant to defend the idea that the criminal deserves punishing, both derived from the idea that the criminal wills his own punishment. The first attempts to fill this out in the nature of the agreement, and the second in the nature of the offender's act.

6. Willing One's own Punishment: Contractarian Retributivism

Jeffrie G Murphy[59] has argued that the social contract picture presented above is a valid representation of Kant's views, what is more, he thinks it contains a way of combining the autonomy of the individual and retributive punishment. Murphy suggests that the social contract presentation of Kant's understanding of the legal realm means that people can be said to have consented to the system of punishment, and thus to their own punishment should they commit an illegal act:

> "What is needed, in order to reconcile my undesired suffering of punishment at the hands of the state with my autonomy, …is a political theory which makes the state's decision to punish me in some sense my own decision. If I have willed my own punishment (consented to it, agreed

[59]See Murphy 1970, and, more especially, 1973.

> to it) then - even if at the time I happen not to desire it - it can be said that my autonomy and dignity remain intact."[60]

This interpretation can claim textual support from Kant:

> "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 along with everyone else to the laws, which will naturally also be penal laws if there are any criminals among the people."[61]

It is important to realise that what is at issue here is not the justification of punishment as a practice, it is not that punishment is just because the offender would have consented to the institution of punishment in a rational contract, it is that punishment does not impinge upon the autonomy of the offender given that, conceived as a rational being, he consents to the practice of punishment.

The problem with such an argument is that as phenomenal beings agreeing to be bound by the rules of civil society, we cannot agree to *deserved* punishment, we can agree to be subjects of punishment when we break the rules, but as such subjects we are merely means to the maintenance of the agreement. We cannot agree to anything more - to deserved

---

[60]Murphy 1973, 287.

[61]Kant 1797, 143-4.

punishment - because as phenomenal beings we cannot choose to be criminal, we are simply caused to be. As a heteronomous, phenomenal being, I break the law because it is in my prudential interest to do so, and this is what motivates me. In order to be deserving of punishment, I would have to be free to choose between good and evil, and I am not.

What is even more undermining for this defence of Kant is that even the noumenal will cannot choose to do evil; the good will is the only thing that is good in itself, that is, autonomy is defined as living in accordance with one's rational nature, which is the same as saying in accordance with the demands of pure practical reason, or the categorical imperative. That is why evil is heteronomous, a consequence of the contingent desires of the phenomenal being motivating the agent to act contrary to the demands of his rational nature as a moral being. Evil, in short, is a consequence of natural causation. However, *even if* we granted Kant the idea from common sense morality, that as free beings we could choose to do good or evil, this would still not rescue the position, for the agreement to form civil society cannot be between noumenal beings, for reasons I have highlighted above. The noumenal realm is concerned with the internal motivations of the agent and, thus, an agreement between noumenal wills could not be one based on *coercive* terms.

In summary, it seems that desert can only be attributed to the agent conceived as a noumenal being (and even here it is doubtful), but, as a noumenal being the agent could not agree to a mutually binding system of coercive laws to regulate behaviour, for the noumenal self is free only insofar as it is self-determining.[62] However, the self that is capable, and needs, the agreement - the phenomenal self - is heteronomous, and thus cannot be deserving.

The contractarian approach to the idea that the agent wills her own punishment, then, does not succeed, however, Kant has a second argument, derived from the nature of the offender's act.

7. The Rebounding Maxim

Consider the following the remark by Kant:

> "For the only time a criminal cannot complain that a wrong is done to him is when he brings his evil deed back upon himself, and what is done to him in accordance with penal law is what he has perpetrated on others…."[63]

Samuel Fleischacker has argued that it is only by taking passages such as this seriously that we can understand Kant's theory of legal punishment.[64] The claim is that when an individual performs a criminal act he acts *as if*

---

[62]See §3.

[63]Kant 1797, 169.

[64]Fleischacker 1992. A similar argument is presented in Pincoffs 1966, 8-9.

the maxim upon which he acts were a universal one, and thus the punishment is merely that maxim being reapplied to, or *rebounding* upon, him; in Pincoffs' words, the criminal's

> "crime of violence against another's freedom implies a maxim sanctioning such actions; as a rational being - and only as such does his autonomy have value - he must be prepared to will that maxim as a universal law, and thus to will that others should violate his freedom as he violates theirs. In punishing him, the essence of punishment being the hindrance of freedom, we therefore do to him just what he has as a rational being willed that we should do - we respect his autonomy and his will."[65]

The argument is that it is only by taking the agent's maxim *as if* it were a universal one that the punisher treats the offender as an end, and thus punishing the offender is a manifestation of the state's respect for his autonomy, it takes his heteronomous will as if it were autonomous and thus treats him as a rational being, and as an equal under the law.

Fleischacker claims that it is not disingenuous to treat the offender as if he had willed the maxim to be universal because, as noted above, the judge can only judge the *action* of the individual not the individual's motivational state; the action is one which can be interpreted as in accordance with a freely chosen maxim, even if that was

---

[65]Pincoffs 1966, 8-9.

actually not the case and the individual was, say, acting out of passion.

Whilst on the surface this account seems attractive, not least for the fact that it can accommodate Kant's rigorous implementation of the *jus talionis*, it is open to three serious objections. The first concerns the ascription of a particular maxim to a particular action. Kant, for example, discusses the case of bestiality and Fleischacker follows him in arguing that, "one who commits bestiality acts as though he had willed that all men deny their human (social) nature".[66] This is not, it seems to me, an uncontroversial description of bestiality, and it is not at all clear how the judge is to capture the best description of the action in the form of a maxim upon which the offender can be said to have acted. This is significant because it is *that* maxim (and no other), which is to rebound upon the individual. The case of bestiality is a hard but good one; Kant proposes banishment - "permanent expulsion from civil society"[67] - for the offender on the grounds that the crime is against one's nature as a human. Of course one can understand Kant's desire to avoid a more forthright description given the possible consequences under the *jus talionis*, nonetheless it is difficult to see why one should accept his, and Fleischacker's, particular version of this crime.

---

[66]Fleischacker 1992, p200. See Kant 1797, p169.

[67]Kant 1797, 169.

It is important to distinguish this problem - of which maxim the judge should *ascribe* to the action - from the sort of criticism that is found, for example, in Duff's Trials and Punishments.[68] Duff argues that one cannot ascribe a maxim to the individual because we do not know on what precise maxim that individual was actually acting. Such a problem is avoided by Kant; the judge is not required to evaluate the actual maxim upon which the agent acted - because, as noted above, motivations are opaque in Kant's philosophy - rather, he is required to ascribe a description in the form of a maxim to the action.[69] The problem highlighted above is that Fleischacker and Kant assume there to be a (single) maxim which uniquely describes every act, and this is, frankly, implausible.

Duff's other criticism is also not appropriate here. Duff argues that in reapplying the maxim of the offender, we are doing to the offender what he has willed and, if this is the case, this cannot be punishment - a view with respect to the offender's will with which Kant entirely concurs.[70] But the criminal does not will his punishment, he wills the offence. Society in reapplying the maxim under which his actions can be described does not act in accordance with

---

[68]Duff 1986, 201-2. It should be noted that Duff does not take the position being described to be Kant's.

[69]In Fleischacker's telling phrase "not dishonesty, but an essential blindness, characterizes the workings of justice." (Fleischacker 1992, 204).

[70]"…for it is no punishment if what is done to someone is what he wills, and it is impossible *to will* to be punished." (Kant 1797, p143); Duff 1986, pp201-2;

his phenomenal will, but in accordance with this will as if it were his rational, autonomous, will.

The second objection is even more serious; what is difficult to understand is why the rebounding of the maxim is to be understood as (justified) *punishment*. The criminal wills, say, to kill another, and on this account we respect his autonomy by interpreting this as if he had willed it to be universal, and thus we are entitled to kill him; he cannot, so to speak, rationally exempt himself from the consequences of his own maxim. This seems to be an attractive solution, since there is general agreement that reasons do possess this feature of impersonality, but why is our rebounding the maxim onto the offender, punishment?

In the first instance the criminal might agree that the community is entitled to kill him; "yes", she might say, "I accept that I (in some sense), willed that killing is permissible and you can kill me", but she may legitimately object if, in killing her, the community claims that it is doing anything different from what she has done. On this interpretation the criminal is simply claiming that it is a "dog eat dog" world in which killing to advance one's interests is permissable; of course, she must accept that this means the community can kill her in its interests, for in this amoral world she has lost.

In order to convert the execution of the criminal into punishment, it is necessary for the community to claim that

the criminal's real will is to live in accordance with morality and that her empirical will has put her in conflict with herself. The claim must then be that the infliction of suffering negates the empirical will and restores the offender to her real will, and in this sense it is not merely doing to the offender what she has done, but is punishing the offender on her behalf.[71] The problem with this is that if what is of concern is that the offender should be brought back to morality, it is not at all clear why this should be done through the infliction of punishment, or, indeed, whether it *can* be done through punishment.[72]

Even if we accept that the impersonality of reasons can give us an account of punishment (rather than something else), there is a third objection. What this account fails to provide is any explanation for why we have a *duty* to punish in this manner; why should we treat the heteronomous actions of a phenomenal being as if they were the autonomous actions of a noumenal one? Combining it with the general justifying aim, we could say that given that punishment is necessary for the maintenance of civil society and that we must never treat another as a means, the only way to reconcile these demands is to punish in this manner - i.e., all and only the guilty in accordance with the rebounding of the maxim ascribed to the criminal

---

[71]This is Hegel's interpretation of the Kantian idea that the maxim rebounds on the offender, it is pursued in much greater length in §§14-15.

[72]See §14.

action. But this seems to establish only that punishment when required by deterrence is *permissible* if it is in this form. It does not seem to make punishment *obligatory*; that would only follow if punishment itself was a good in that it respected persons, that is independently of its deterrence value.

Fleischacker suggests something very similar to this; he argues that in the moral realm punishment is good, and the legal realm is, in some sense, a bad copy of the moral. This seems to challenge the interpretation of the legal given above, however, when considering why the legal ought to be this bad copy of the moral, Fleischacker falls back on a consequentialist concern, that of preserving "perpetual peace", and thus his reading fails in the end to provide the retributivist core Kant needs.

8. Threat and Execution

I would like to make one final point; throughout this Chapter I have used the distinction between the general justifying aim of punishment and the distribution of punishment. The general justifying aim of legal punishment is, I have claimed, deterrence based in the absence of a free standing account of desert with which to constrain it. But a number of the theorists I have considered, most notably B. Sharon Byrd and Don Scheid, argue that it is precisely this distinction which solves the apparent inconsistency in Kant's theory. Punishment, they argue, is

deterrence based in its general justifying aim but retributive in distribution.[73] And this is, of course, also at the centre of H. L. A. Hart's "Prolegomenon to the Principles of Punishment".[74]

In some senses I have accepted this distinction in my search for an adequate defence of retributivism in Kant's work. Byrd's argument relies on the idea that the threat of punishment is addressed to phenomenal beings involved in the agreement to co-operate in civil society. Although it does threaten them, it is not autonomy infringing because the subjects it addresses are not being addressed as autonomous beings. When it comes to implementing the punishment, however, this is done on the basis of desert, and is, thus, retributive.

What strikes me about this argument is that it presupposes that Kant has an adequate theory of desert, and that is something I have found conspicuous only by its absence, but also, that if one granted that Kant had a free standing argument for desert that would undermine the structure of Byrd's argument. Given that Kant regards the imposition of punishment as obligatory, an account of desert would give a *prima facie* general justifying aim for punishment; one would have a system of punishment in order to give people what they deserve. No doubt having the system of punishment would also deter individuals seen as phenomenal

---

[73]See Byrd 1989; Scheid 1983.

[74]Hart 1959.

beings, but this could no longer be regarded as the reason for having punishment in the first place, it would, as I said above, just be an added bonus.

9. Conclusion

In conclusion, it seems as if there is no satisfactory way to theorise the relation between the noumenal (moral) and phenomenal (juridical) in Kant's philosophy. Within the latter Kant argues for a largely deterrence based system of punishment, nonetheless he thinks that this is compatible with a retributive element based on desert, which is the theory that dominates his discussions of the noumenal realm. The various specific attempts to show the compatibility and relationship between these realms and theories are all unsatisfactory.

In the end, the reason for Kant's belief that punishment is fundamentally retributive which leads him to characterise the moral realm in the way he does, is, I believe, revealed in the connection he sees between the idea of transgression and punishment. As of now, even if the gap between the idea of responsibility and the heteronomous actions of the offender could be bridged, no argument has been given, beyond deterrence, to actually inflict punishment. The retributive claims in much of Kant's writing rely on a simple statement that there is a relationship between transgression and punishment; for Kant, there seems to be something innate in the notion of transgression that

implies retributive punishment. Kant's argument for this is largely intuitive, and so we come back to the position from which respect for persons and contractualism were supposed to have rescued us; that is that Kant simply assumes that punishment has a retributive element, and takes the form of the *jus talionis*, because it is self-evidently so as part of justice. There is a "celestial mechanics",[75] in which "every criminal action metaphysically deserves 'an equal an opposite reaction', in the shape of punishment."[76]

Thus, if we return to the example of the drowning man cited earlier. Kant says that although the sailor who shoves the other off the plank on which he had saved himself is not legally punishable, because the cause of deterrence cannot be served in such a circumstance, the sailor is "not to be judged *inculpable*".[77] And, again, in the diaspora example, it is clear that no deterrence benefit arises from executing all the murderers before civil society disperses. Likewise, if one considers the rebounding maxim argument as a free standing account of deserved punishment, Kant provides no answers to the questions I posed there, why should the maxim rebound and why is this punishment? In the end, one can only conclude that Kant has no argument; in the retributive frame of mind for which he is famous, Kant's position is genuinely without foundation and thus of

---

[75]Cohen 1939, 279.

[76]Fleischacker 1992, 204.

[77]Kant 1797, 60.

little use to a defence of retributivism. He does adhere to the two fundamental tenets of this doctrine: that punishment is obligatory and the suffering of the offender is both morally good and justified by his past act, but he takes these as self evident truths derived from the ideas of justice and transgression.

Nevertheless, although Kant is probably the most famous retributive thinker he is by no means the only one, and before dismissing retributivism as invariably based on the mere mystical restating of its fundamental intuition, I will turn to another such theorist, G. W. F. Hegel, in an attempt to find the grounding for the conviction that the guilty deserve to suffer.

Chapter 2: Hegel and the Annulment of the Wrong

> "Those who have gone beyond Hegel are like country people who must always give their address as *via* a larger city; thus the address in this case read - John Doe *via* Hegel."(Søren Kierkegaard, 17 January 1838)[1]

"Hegel's philosophy", writes Bertrand Russell "is very difficult - he is, I should say, the hardest to understand of all the great philosophers".[2] Even Alan Wood, in his excellent study of Hegel's ethical thought, admits that whilst "the retributivist intent of Hegel's theory [of punishment] is clear enough … its central claims are shrouded in obscure metaphors."[3] Nevertheless, Wood believes that Hegel offers an account of the core of retributivism. Hegel, he believes, argues that, "the justification of punishment is that it is inherently just to inflict some evil on those who have done wrong."[4]

The "obscure metaphors" of Hegel's theory are notorious; punishment is an "annulment",[5] or a "cancellation", which

---

[1]Quoted in Stern 1993, epigraph.

[2]Russell 1946, 701. Russell's Chapter shows strong evidence that he found Hegel impossible rather than difficult to understand. Good introductions to Hegel's social and political thought are Hampsher-Monk 1992, Chapter 9; Plant 1983; Taylor 1979; Wood 1990, although in an otherwise excellent study Wood's discussion of Hegel's theory of punishment is disappointing.

[3]Wood 1990, 109.

[4]Wood 1990, 108.

[5]Hegel 1967, §99R.

is "an infringement of an infringement".(§101)[6] In this Chapter I shall try to make sense of these metaphors and, in so doing, look to Hegel for a coherent account of retributivism. First, however, it is necessary to explain what it is that Hegel means by crime; in the second part of the Chapter I shall then go on to show how Hegel develops two accounts of coercion, which, despite their obscurity, shed light on contemporary approaches to punishment theorising, although unfortunately not on the core of retributivism.

10. Wrong

Before one engages with Hegel on the idea or justification for punishment one has first to examine his analysis of wrong, and thus of crime.[7] The discussion of wrong, crime and punishment in The Philosophy of Right,[8] occurs at the end of Hegel's analysis of Abstract Right, before the transition to Morality.[9] Hegel is still concerned with the

---

[6]All such parenthetical references are to Hegel 1991. R denotes a Remark, A an Addition. All italics appear in the original.

[7]Two authors who explicitly admit this to be the necessary starting point for an analysis of Hegel on punishment are Nicholson 1982; and Tunick 1992a, 25-29.

[8]In this chapter I shall be primarily concerned with The Philosophy of Right, this is because it contains Hegel's best discussion of punishment and because it is part of his mature work. For a discussion of early versus late Hegel and the problem of punishment, see Tunick 1992a, Chapter 4.

[9]That is, it lies between §82 and §104. The term "punishment" does not appear in this part of the text of The Philosophy of Right because the sphere of Abstract Right is prior to society. Instead Hegel favours the term *Aufhebung*, translated as "cancellation" by Nisbet (Hegel 1991) and "annulment" by Knox (Hegel 1967). Hegel is less careful in the Remarks and Additions where he makes it clear

abstract will; what has been shown so far is that if the abstract will is to have determinate existence then it must be embodied (§§34-40), and manifest itself in property (§§41-65), which can be alienated (§§65-70). Further, one's will must be given objective standing through the recognition of one's right in property by an other and this recognition is manifested through contract (§§71-81).

Hegel divides wrong into three categories, "unintentional wrong" (§§84-86); "deception" (§§87-89); and "coercion and crime" (§§90-103). Unintentional wrong occurs when two wills' claim rights over a piece of property to which only one is in fact entitled. This is "the sphere of *civil actions*" (§85). Neither party denies the claims of rights, but one is mistaken in making his claim. For example, if I mistakenly drink your pint of beer in a bar, believing that it is mine, I do not assert that you have no right to *your* pint of beer, or even - should I be disabused of my incorrect belief - to the pint of beer that I have drunk. I am simply mistaken, and by drinking a pint of beer that is in fact yours I have denied your legitimate right to that pint. In unintentional wrong,

> "Right is … recognised in this case. Each person wills what is right, and each is supposed to receive only what is right; their wrong consists solely in considering that what they will is right." (§86A)

---

that punishment is what he has in mind [e.g. "If crime and its annulment (which later will acquire the specific character of punishment…)" Hegel 1967, §99A]

Such wrongs are non-malicious. The offending party does not deliberately challenge the right of the other party, he is simply in error over the content of that right. In Hegel's view the offender is wrong because he relies on his particular, subjective judgement about what is right. Importantly, neither party denies the legitimacy of rights, that is, once society is established, of the foundation of law. Thus, "this first kind of wrong negates only the particular will, while universal right is respected" (§86A).[10] Such unintended wrong is not a subject for punishment. Rather, both parties submit themselves to the authority of the law - that is to a disinterested judge - who settles the dispute. Compensation (or restitution), not punishment is required.[11]

Deception, or fraud, is a more difficult category.[12] In his discussion of fraud Hegel seems most concerned with contract, and it is this which allows him to distinguish the deceiver from the criminal. In fraud, the deceiver appears to act in accordance with right, that is, his fraud depends upon his appearing to his victim to recognise the claims and binding nature of law. Nicholson suggests that

---

[10]For further elucidation of this wrong see Nicholson 1982, 109-111; Tunick 1992a, 26.

[11]The argument for compensation should be obvious: If I have drunk your pint of beer, thinking that it is mine, and then discover my error, I am obliged to replace the pint (and also compensate you for any further damages caused by my action).

[12]Nicholson remarks that "it suits the triadic pattern common in Hegel's presentation that fraud should be a category on its own" although Nicholson goes on to argue that Hegel has a line of argument open to him that would "justify making fraud distinct from crime". (Nicholson 1982, 114-5)

there are parallels between Hegel's distinction between fraud and crime and the contemporary distinction between white-collar crime and crimes of violence against persons or property.[13] White-collar crime is parasitic on the existence of good, stable business relationships and thus needs a relatively established and healthy "host". Similarly, for Hegel, fraud requires the existence, and ostensible recognition, of right in contract. But Hegel also appears to go further; he describes fraud as a wrong against only the universal will - that is only against right - not against the particular will. In other words when one is defrauded one believes (in order for the fraud to be successful), that one is not being defrauded (in crime the case is different - one does not need to believe one is not being mugged for the mugger to be successful). Thus, Hegel claims, one's particular will is not harmed although, I take it that it has to be added that this claim, even if true, only applies for as long as the victim remains ignorant of the deception.

This idea - that one's particular will remains unharmed as long as it remains ignorant - is surely a bizarre one; if I desire to buy Tower Bridge, and somebody passing on the street (with no legal right to Tower Bridge), sells it to me, it seems clear that my particular will is frustrated from the moment that I pass over my money in return for a worthless "contract". I did not desire a piece of paper with words to the effect that I owned Tower Bridge on it; I

---

[13]Nicholson 1982, 114-5.

desired ownership of Tower Bridge. The general idea behind Hegel's distinction is that fraud has all the appearance of right whereas crime, as we shall see, appears as a direct negation of right, but I have grave doubts that it can be sustained as a separate category of wrong in the way that Hegel desires.[14]

Crime, Hegel's third category of wrong, is an offence against both the particular and the universal will. It is important, at this point, to remember that for Hegel contract and property are not simply useful devices for regulating co-operative enterprise, but are mechanisms through which abstract personality becomes actual in the world. Contract extends the concretisation of the abstract will because it involves the recognition of the person's rights in things by another. The criminal by, for example, stealing property, not only appropriates something that is already owned - that is treats the property as if it were still a part of the natural world of things - but treats the owner as if he, too, were a thing rather than a being with rights.

---

[14]Tunick makes heavy weather of Hegel's invocation of the categories of "simple negative", "positive infinite" and "negative infinite" judgements. Despite his self-aggrandising claims to being the only commentator he knows to take seriously and make clear these categories, his analysis of the differences between fraud and crime in the end amounts to the claim that "I think Hegel is saying that the criminal acts *as if* he lacks this understanding [the understanding of right]". (Tunick 1992a, 27-9) The meaning and purpose of the correspondence between the categories and the forms of wrong is certainly no more clear by the end of Tunick's discussion and is, it seems to me, certainly treated no more "seriously" than it is by Nicholson 1982.

> "The initial use of coercion, as force employed by a free agent in such a way as to infringe the existence of freedom in its *concrete* sense - i.e. to infringe right as right - is *crime*. This constitutes a *negatively infinite judgement* in its complete sense … whereby not only the particular - i.e. the subsumption of a thing under my will - is negated, but also the universal and infinite element in the predicate 'mine' - i.e. my *capacity for rights*." (§95)

Hegel claims that such action is "null" (§97), because it denies that the owner of the thing has rights, and thus contravenes the universal will. This raises two important questions that are essential to an interpretation of Hegel's theory of punishment: Why does an attack on a particular right to a particular thing constitute a denial of all rights, that is on the idea of right? And second, in what sense is such a denial invalid? The first of these questions will lead to one element of Hegel's theory of punishment and the second, through a less direct path, to the other.

Crudely, Hegel argues for the necessity of coercing the coercer on two grounds. First, it annuls the crime and restores right. Second, it is required if society is to treat the criminal as a rational being - it is the criminal's right to be punished. These two reasons, which, following Mark Tunick, we can call the "objective" and "subjective"[15] reasons to punish, lie at the heart of

---

[15]Tunick 1992a, 35.

Hegel's theory, although different commentators on Hegel have given varying interpretations of these claims. I shall consider each of the reasons below, analysing the possible readings in an attempt both to clarify what Hegel was saying and to see whether it has any merit. I am concerned, here, with the justification for *coercion*, and punishment justified as a harm to the offender, and it is important that this is clear, not least because the conclusion of this chapter will turn on the claim that Hegel may have either or both an expressivist or moral educative account of punishment, but he lacks a coercive (and in that sense, retributivist) account.

Before beginning, however, it is necessary to enter a methodological warning. As noted above,[16] Hegel's most important discussion of punishment occurs in the section on Abstract Right, prior, in other words to organised society and thus to punishment. In Abstract Right, Hegel's concern is with "annulment" or "cancellation". This does not mean, however, that punishment is to be justified in ethical society, or in the state, and not in the sphere of Abstract Right. The "truth" of punishment, like the "truth" of rights, will be finally revealed in the ethical life of the modern state, and for that reason the discussion of rights and annulment in Abstract Right is necessarily incomplete. However, the argument for the annulment of the crime in Abstract Right must remain at the centre of the argument

---

[16]See *supra*, note 9.

for punishment in ethical society, it is incorporated not annulled.

Furthermore, wrong and its annulment plays a crucial role in the transition from the sphere of Abstract Right to the sphere of Morality, and this is important; the structure of The Philosophy of Right cannot be ignored in the analysis of the individual arguments contained therein. When considering Hegel's justification for the use of coercion, however, it is difficult to show each argument separately, so in what follows I shall jump between the spheres of Abstract Right and the State in an attempt to show how each argument is to be understood. I shall, at the end, however, consider Hegel's full account of punishment in the ethical state.

11. The "Objective Reason": Punishment as the Annulment of Wrong

Three of the most acute commentators on Hegel's theory of punishment have all argued from Hegel's charge that crime is null to a particular interpretation of Hegel's justification for a retributive form of punishment.[17] In

---

[17]Cooper 1971; Stillman 1976; Nicholson 1982. There are differences between the accounts, noticeably in that Cooper places more emphasis on the idea of "right" than on its infringement - that is, on "crime". Nicholson is quite explicit in denying that his is an analysis of Hegel's theory of punishment, instead focusing on crime. Nonetheless, he suggests that his (excellent) analysis of crime supports Stillman's account of punishment, and, I take it, endorses that account. Despite some important disagreements, I think that they can be treated as variations on a single theme, and certainly Stillman pays tribute to Cooper as being "especially valuable" (Stillman 1976, 169), and Nicholson cites both Stillman and Cooper as

brief, the argument is that crime is wrong and null, and must itself be annulled so that it is not held as right. This annulment becomes, in society, punishment. Trying to make sense of this requires, as noted above, that one first makes sense of the claim that crime is null.

The clearest passage in which Hegel makes the claim that crime is null is in the addition to §97:

> "Through a crime, something is altered, and the thing exists in this alteration; but this existence is the opposite of the thing itself, and is to that extent within itself null and void. The nullity is [the presumption] that right as right has been cancelled. For right, as an absolute, cannot be cancelled, so that the expression of crime is within itself null and void, and this nullity is the essence of the effect of crime." (§97A)

Hegel's claim is clearly that crime denies the validity of right and this is incoherent, or inconsistent, and thus "null". But why, to return to the question asked at the end of the last section, should a specific crime be seen to attack the very foundation of right itself?

Hegel makes two claims, one is that when a criminal, say, steals property, he denies the property owner's status as a rights bearing person; that is, he treats his victim as a thing and denies the victim's "capacity for rights" (quoted above). The second claim is more general and is that the

---

amongst "the best accounts" (Nicholson 1982, 103) Particular details of their accounts and the differences between them follows below.

criminal denies the existence of rights *per se*, in other words, not simply of the victim, but of everybody else (including himself).[18]

The first claim makes more sense once one recalls that, for Hegel, rights are not simply useful devices for regulating social conduct, but are claims that abstract will makes to embodiment in the world; that is, they are essential to human freedom. The argument is thus that should one challenge the right of an individual to property one challenges not only that claim to that property, but also the victim's claim to anything other than abstract freedom.[19] One is, in Kantian terms, treating the other as a means and not as an end. This claim, furthermore, is inconsistent and this inconsistency marks the passage to Hegel's second argument, that the criminal denies the validity of rights as such. The inconsistency stems from the fact that the criminal as a rational being requires, as a necessary condition of his embodied freedom, the sphere of right. In Abstract Right all persons are identical,[20] and thus to deny the rights of one person is to deny the rights of all (including, of course, the person doing the denying), and thus, given the derivation of right from the

---

[18]See Knox's note 86 to §99 of his translation of The Philosophy of Right. (Hegel 1967, 331).

[19]See Nicholson 1982, 112.

[20]"In terms of personality", that is, in Abstract Right, "persons are equal". (§49R)

necessary embodiment of the abstract, free, will, it is to deny the existence of right.[21]

Hegel's claim that crime is a "negatively infinite judgement" (§95, quoted above), makes this point clear. A negative infinite judgement is an assertion of total incongruity between subject and predicate, crime asserts incongruity of person and right and this is absurd; it is, in Tunick's memorable phrase, like saying "the rose is no elephant". To say,

> "'your right to your property is no right' [indicates that] the criminal is so utterly ignorant of what a right is, flouts it to such an extreme, that he could not understand what it would mean to say something is not a right; he might just as well say 'your right is no elephant'".[22]

Of course, Tunick is forced to retreat slightly because the criminal cannot lack the understanding of what right is, because if he did so he would not be responsible,[23] he merely "acts *as if* he lacked this understanding".[24]

---

[21]See Stillman 1976, 171.

[22]Tunick 1992a, 29.

[23]That Hegel thought this is clear in his wonderful (because one suspects entirely serious) note concerning "Christ's intercession on the Cross for his enemies: 'Father, forgive them, for they know not what they do'". This is, Hegel says, "a superfluous request if the fact that they did not know what they were doing removed the quality of evil from their action so that it did not require forgiveness." (Note to §140) Michael Mitias has argued that because to commit a crime requires that one understand right, but crime is differentiated from fraud precisely by the denial of this understanding on the part of the criminal, Hegel has committed himself to the view that there can be no crime. (Mitias 1978).

[24]Tunick 1992a, 29.

Hegel's argument thus far is a variation of a fairly standard universalist theme, which retains support through to contemporary political philosophy. That is, the argument that it is inconsistent to claim rights on the basis of something shared in common with others, and then to deny others rights claimed on the same shared feature. The Hegelian twist is that the rights are derived from the necessity for embodiment of abstract wills, rather than directly from claims of value inherent in such wills, and this means that the existence of such rights as actual is dependent upon their recognition by other independent individuals. Because the criminal's will, to be actual, depends upon the existence of rights, his apparent assertion that rights do not count (as it were), is, Hegel claims, inconsistent and incoherent. It is "null".

Hegel's next step is, in brief, to say that the criminal's denial of right must be countered; it, itself, must be annulled. From the "objective" perspective, the primacy of right must be restored. Before we ask why such annulment should take the form of punishment, however, there is a prior question not sufficiently identified in the theorists who follow Hegel's own steps; this is, why need the crime be annulled at all?

One interpretation that has been offered takes seriously Hegel's claim that without coercion right would be invalidated and wrong held to be right. Above, it was

claimed that the criminal will was inconsistent because it denied the existence of rights which it, itself, must claim in order to be realised as actual. In less opaque terms, the criminal will is, by denying that its victim had rights, denying the rights of all, including itself, whilst at the same time it makes a claim to those rights. One interpretation of the claim that the criminal will must be annulled takes the idea that, if it is not, wrong will be held to be right to mean that this will be destructive of the system of rights.

The problem with such an account is that it is *not* the case that the criminal's mere assertion that right is invalid could possibly have such an effect, unless that assertion has some efficacy in the world. If I assert that the world is flat, I may be wrong and it may be irrational for me to do so if I am confronted with evidence, but that assertion in itself can have little impact on the world, or on my ability to get around it, unless I refuse to travel past a certain point, despite my needing to, because I think that I will fall into a great void. In other words, if the criminal will must be annulled because it negates the sphere of rights which it requires to be itself actualised, then Hegel must defend this claim that it does indeed negate the sphere of rights to the extent that the criminal (and all other), wills no longer have the capacity to be actualised as concrete free wills in the world. On this interpretation, Hegel is saying that the criminal will is self-destructive because it destroys the conditions

necessary for its own freedom. This is simply false, one criminal act is extremely unlikely to destroy a system of rights.[25] It may, as I have said, be inconsistent for the criminal to claim rights which he has just denied to an identical other, but it cannot be the case that the criminal will must be annulled for the weak, and doubtful, contingent reason that it has the destructive capacity to invalidate the sphere of rights.

The second interpretation, which is closely related to this argument, is a great deal more cogent. This holds that it is not simply the criminal's will that is destructive of the system of rights, but the criminal will if combined with the acceptance of that will by others. In other words, whilst the criminal will on its own could not destroy the system of rights, if that will is not renounced by others the combined effect is to destroy rights.[26] This latter claim is presented as an empirical one, as when Stillman writes that:

> "In general, even outside the Hegelian context, it holds that, when rights are not enforced and denials of them not refuted, the rights fall into disuse and disappear; rights thus must be enforced…"[27]

---

[25]See Wood 1990, 112.

[26]I do not see how this claim can be taken to amount to the idea that "if nobody declares [a criminal act to be] wrong it's not." (Tunick 1992a, 80) If that were the case the necessity for public renunciations of the criminal will would be obvious, however, it clearly is not the case. What makes an act wrong is that it *is* contrary to right, not that it is declared to be contrary to right.

[27]Stillman 1976, 172. This claim is repeated by Tunick: "if the state fails persistently to punish crimes its citizens are likely no

According to Hegel crime is null, the denial of right is an absurdity, however, this does not mean that crime has no external manifestation. As Hegel puts it,

> "when an infringement of right as right occurs, it does have a *positive* external *existence*, but this existence *within itself* is null and void." (§97)

Its positive external existence is, of course, the crime itself; in the case of car theft, the criminal does in fact appropriate your car. The criminal acts as if[28] the car owner has no rights, and by extension as if no-one has rights to anything. The essence of the theory of just coercion is, then, that punishment is the reiteration or restoring of right; it "annuls" the crime and in so doing restores right. That is why Hegel calls it the "negation of the negation" (§97A). The crux of Hegel's argument is presented in the following passage:

> "The *positive existence of the injury* consists solely in the *particular will of the criminal*. Thus, an injury to the latter as an existent will is the cancellation of the crime, *which would otherwise be regarded as valid*, and the restoration of right." (§99)

---

longer to regard crimes as wrongs". (Tunick 1992a, 35) Cooper (1971) also makes a similar claim, see below §12.

[28]The "*as if*" is important because it is extremely unusual for criminals to be "intellectual anarchists" who in fact deny the existence of rights. (Cooper 1971, 166) Most criminals are perfectly happy to admit that their victims have rights of ownership over their property. Tunick makes much the same point. (Tunick 1992a, 29) See *supra*, note 23.

This argument alters the emphasis from the criminal will itself to the reaction to that will by others. Above, it was claimed that an individual criminal will, acting as if it denied the existence of rights, was unlikely to have any important consequences, other things being equal, to the system of rights; more explicitly the claim that it would destroy the system of rights was denied. But the argument for the necessity of cancelling the crime is slightly different; what is at its heart, is the claim, not that the criminal will itself can be destructive of the system of rights, but that the primacy of right must be restored. Two questions immediately need to be addressed: Why? And how?

12. Cooper and the Logical Relationship of Rights and Punishment

Before proceeding to offer a suggestion as to how Hegel addressed these questions, I want to pause to consider one given by David Cooper.[29] Cooper's account of Hegel's theory of punishment represents the clearest attempt to show the necessity of punishment in order to protect rights. His much respected article is a classic example of what Nicholson calls "the terminology of modern Oxford philosophy".[30] Cooper explicitly denies that the theory of punishment can follow from the actions of the criminal being "inconsistent", and he also claims to find Hegel's

[29]Cooper 1971.

[30]Nicholson 1982, 114.

charge that the criminal will is "self-destructive" unfathomable.[31] It might thus appear odd to categorise him, as I have done, with Nicholson and Stillman who base their arguments on Hegel's idea of crime. However, Cooper's argument is essentially similar, in that he argues from the idea of rights to punishment.

Cooper moves from rights to punishment without an intervening consideration of wrong because he believes the relationship between rights and their protection to be a conceptual one. As my concern is to show both that Cooper's analysis is not Hegel's, and that it is flawed as an independent attempt to justify punishment, I shall examine his argument in detail beginning by quoting the argument at some length.

"Legal rights", writes Cooper, "are performatees", by which he means that "whether or not a person has the right to do *x* is logically dependent upon some rule or convention by reference to which such a right may be ascribed".[32] Furthermore:

> "There is one very important question to ask if we are trying to decide whether persons have rights…. The question is: what happens to those who try to prevent these persons from doing what they have a supposed right to do? If nothing happens to them - if no attempt is made to apprehend and punish them - there is very strong

---

[31]Cooper 1971, 160-1.

[32]Cooper 1971, 162.

> reason to suppose that the persons had no such right at all."[33]

Cooper both strengthens and weakens this claim during the course of the article; strengthens, in that the "very strong reason" becomes (without further argument), a "logical" connection between rights and punishment;[34] but he then weakens the claim, applying it not to rights *per se*, but "felicitous" rights.[35] The closest he comes to defining "felicitous" is by arguing that "one way in which a performance may be infelicitous…is if certain subsequent behaviour does not take place".[36] The example he gives is of the giving and then taking back a gift, but it is clear that punishment is going to justified through a similar mechanism. Crudely, the "subsequent behaviour" which makes one's rights felicitous, is punishment for anyone attempting to frustrate an individual in the *using* of their rights. Thus Cooper claims:

> "Unless other people are generally apprehended and punished for preventing others doing *x*, there is reason to suppose that the latter do not have the right to do *x* - certainly not a 'felicitous' right… [thus]… the justification of punishment is very simple. It is the same as the justification of the rights which crimes violate. If it is important that men have legal rights,

---

[33]Cooper 1971, 162.

[34]Cooper 1971, 162.

[35]The two moves are obviously related. As I point out below, Cooper makes the relation between "felicitous" rights and punishment tautologous.

[36]Cooper 1971, 161-2.

> it is important that there be punishment - for without the latter, there could not, logically, be the former".[37]

In brief, "punishment must logically follow crime if we are to speak of there being rights and crimes at all".[38]

Cooper's argument can be addressed from two perspectives, that of whether it succeeds as exegesis, and of whether it is plausible as an independent argument. Beginning with the latter criterion, a number of points are immediately apparent. The first is that Cooper's argument, once the element of "felicity" is introduced, is circular. A felicitous right is one which exists both formally and commands certain subsequent action. Cooper does no more than assert that to possess felicitous legal rights is to possess rights which command "the attempt to apprehend and punish" anyone who acts so as to frustrate the rights holder in the exercise of his rights. If we follow Wood and replace the slightly ridiculous term "felicitous" with "socially guaranteed", the tautologous nature of Cooper's claim becomes obvious; the existence of socially guaranteed rights depends upon their being socially guaranteed.[39]

The second point, is that even if Cooper's argument were to work, it does not advance us very far. In essence we would be left with the claim that to possess a *legal* right, is to possess a right which has as a component part the

---

[37]Cooper 1971, 162-3.

[38]Cooper 1971, 163.

[39]See Wood 1990, 111.

stipulation that the guarantor of that right will act (or attempt to act), so as to guarantee the legitimate use of it. This is quite obviously the case[40] (whatever the relation between such rights and their enforcement), all it does is beg the question of why such rights exist and how do we justify enshrining them in law. The question, "what is the justification for punishment?" may very well be (and I believe it is), best answered by asking the question "what is the justification for having rules the contravention of which render the contravenor liable to harm imposed by the state?",[41] but just stating that this happens does not get us very far. Of course, Cooper might claim that he has reduced the question to "what is the justification for rights", and this furthers the argument because Hegel (and possibly contemporary political philosophy), has a (more) plausible account of rights, but Cooper is not entitled to this on two counts. First, he has not reduced the question to one of rights, but to one of "felicitous" rights which are not, despite his occasional equivocation, simply rights, but are, rather, rights which carry a legal and social guarantor. Second, he has not shown that punishment is the only mechanism through which such rights could be guaranteed.

---

[40]Maurice Cranston makes Cooper's point in far fewer and more elegant words when he writes "to say that I have a [legal] right [to various things] … is to say that I live under a government which allows me to do these things, and will come to my aid if anyone tries to stop me. … A positive right is necessarily enforceable; if it is not enforced, it cannot be a positive right." (Cranston, 1973, 4-5)

[41]See *supra*, Introduction.

But, if Cooper has done nothing to address the question of why the law guarantees certain rights through punishment, he has, it seems to me, done even less to further our understanding of Hegel. As Alan Wood has pointed out, there is simply an absence of textual evidence; Cooper's conceptual link between right and punishment is not to be found anywhere in Hegel.[42] Cooper constantly makes reference to Hegel's text and claims that his account is the only plausible framework within which Hegel's darker passages can be understood, but nowhere does he actually find a defence of his primary claim.

This is, however, not as serious as the second charge which can be brought against Cooper's interpretation of Hegel. Cooper, in support of his argument that legal rights require enforcement if they are, in some sense, to be real, cites the example of Jews in Nazi Germany prior to their rights being formally abolished. Presuming that the Nazi authorities did nothing to enforce the rights of Jews, Cooper makes the claim that therefore the Jews did not possess felicitous rights.[43] As we have seen this is tautologically true, the question is, is this all that can be said about rights? Of course, it is entirely plausible that rights (understood as performatees - that is as part of a rule or convention governed practice), depend upon enforcement. Consider the case of a game of cricket; if the umpire consistently refused to allow the batsman to run

---

[42]Wood 1990, 111.

[43]Cooper 1971, 163.

when he struck the ball, the batsman would begin to question whether he had the right to run under the laws of cricket, or whether he was still playing cricket. But there is certainly another sense of rights that is commonly invoked in both philosophical and ordinary discourses. In this sense the Jews could have complained after the Nuremberg edicts abolished even their formal rights that such edicts infringed their "human rights". For Cooper, once the formal rights of the Jews had been abolished there was no question - they certainly no longer possessed felicitous or any legal rights, and this is again tautologously true. But if this is all there is to rights discourse then a great number of claims, including those of the declaration of the rights of man, are going to have to be amended.

Of course, Cooper's defence here is that I am committing a simple philosophical confusion, that is, between legal and some idea of universal rights (an idea which is possibly nonsensical). Cooper, remember, began his defence of Hegel by claiming that legal rights were performatees and everything follows from this. But this is especially bizarre when defending a Hegelian interpretation of rights, and furthermore, one located in the sphere of Abstract Right. Legal rights just are not the issue. The annulment of wrong - indeed wrong itself - is an infringement of right, and the last thing that right is for Hegel,

especially at this stage of the argument, is dependent on "some rule or convention".[44]

Given that Cooper's contention that the idea of rights conceptually requires punishment for violators of rights is unsustainable, what other account can be given of the initial Hegelian claim that wrong calls forth its own annulment so that right might be restored?

13. The Objective Reason Reconsidered

The key objective reason for the annulling of the criminal will is that right must be restored and the nullity of the criminal will made manifest. It is important to note that punishment is addressed to the criminal will, and not to society or the positive act of the criminal.[45] I have argued above that it is difficult to know quite how to conceive of the claim that the criminal will threatens the system of right, unless we endorse an empirical assertion that if the system is not validated then it will be endangered. In the sphere of Abstract Right the negative command to respect the freedom of others creates a system of (purely negative and incomplete), rights which the

---

[44]Cooper 1971, 162. This criticism is aimed merely at the claim that rights, for Hegel, are performatees. It could, therefore, be accepted by commentators on Hegel who take a different view of the status of rights in Hegel's philosophy (itself a controversial issue), from me; just so long, of course, as they do not endorse Cooper's claim. For a good introduction to the debate on the status of rights in Hegel see Smith 1989.

[45]Although obviously it is necessary that the positive act be undone, and compensation provided.

criminal will violates. This system, so the argument goes, must be held to be valid, and thus the criminal will negated. Thus, for example, Stillman argues that,

> "with the injury of the particular will of the criminal, the positive existence of the crime no longer exists, the infringement of right is annulled, and right is restored. Conceptually, crime is nothing at all; with the annulment of crime in the objective world, crimes nullity is manifest. Criminals are punished in order to ratify in the objective world the truth within the conceptual world: that crime is null."[46]

But, even if we accept the necessity of publicly annulling the criminal will, that is of publicly denying that such a will is valid, does it follow that coercion is the best way to achieve such an end? Why could society not simply condemn the offender and his actions (combined with demanding compensation, as in civil wrong)?[47] Such an expressivist response would satisfy the demand that the criminal will be publicly denied as valid, and the compensation element would negate the positive existence of the crime, that is, it would restore the situation to that which existed before the criminal act. Of course the compensation would have a coercive element - the offender *must* compensate the victim - but this coercion is not at the heart of the justification for punishment. Rather, the justification for punishment is the expression of the

---

[46]Stillman 1971, 172.

[47]For an interesting account of expressivist punishment, which has become standard, see Feinberg 1965. For an Hegelian account of punishment based solely on restitution see Day 1978.

invalid nature of the criminal will, the compensation is merely an additional requirement. Yet, Hegel argues for a retributive theory of punishment and one that goes well beyond mere condemnation.

An expressivist interpretation of Stillman's argument is reinforced if one shifts from coercion to law and punishment. Consider a society in which petty burglary is largely unreported and, even when it is reported, the clear up rate for such crimes is extremely low. In short, punishment for petty burglary is proportionally very rare. Does this challenge a Stillman Hegelian? The answer surely is no, I take it that a defender of Stillman's thesis would argue that the fact that petty burglary is not often accompanied by punishment is less significant than the fact that the society has a law against burglary which condemns convicted burglars. The fact that there are few convicted burglars is neither here nor there, what matters is the expressivist content of the law.[48]

---

[48]It is often taken to be the case that legislators should not allow the law to enshrine unenforceable rules, on the grounds that it brings the law into disrepute. That is why, for example, one might argue that the British law which makes it a criminal offence for a man to sodomise a woman in private (even if the act is consensual) ought to be repealed. It would be bizarre, however, to argue that one has the right consensually to sodomise one's female partner because of the absence of any convictions and punishments for those who engage in this act with the consent of their partners. If one did want to argue that this law was now "dead" the absence of punishments would surely be merely epiphenomenal - the important point would be that nobody had invoked this law (at least not in consensual circumstances) in the moral and legal discourse for many years. In other words, as I shall argue below, it would be the expressivist element of the law that would be more likely to ground one's argument, not the absence of criminal proceedings and punishments.

14. The Subjective Reason: Punishment as the Right of the Offender

Hegel, as we have seen also appeals to a second, subjective, reason for the necessity of annulling the criminal will. This is that it is the right of the criminal:

> "The injury which is inflicted on the criminal is not only just *in itself* (and since it is just, it is at the same time his will as it is *in itself*, an existence of his freedom, *his* right); it is also *a right for the criminal himself…*". (§100)

What sense can be made of this rather extraordinary "right"? As Ted Honderich has argued it is a very odd right when it is not in the power of the right holder to renounce the entitlement conferred by the right.[49]

The first, and most important, element of the subjective account of punishment is not, in fact, to be found in the claim that the criminal has a right to his punishment, but, rather, in the secondary claim that the criminal wills (or consents to), his own punishment. Clearly the consent of the criminal is not explicit, but Hegel argues, the criminal's implicit will is in accordance with the justice of his punishment. Thus the passage quoted immediately above continues,

---

[49]"A right that cannot be escaped is an odd right." (Honderich 1984, 47).

> "…, that is, a right *posited* in his *existent* will, in his action. For it is implicit in his action, as that of a *rational* being, that it is universal in character, and that, by performing it, he has set up a law which he has recognised for himself in his action, and under which he may therefore be subsumed as under *his* right." (§100)[50]

Mark Tunick has attempted to make sense of these claims by invoking the idea of split-level self.[51] Translating this into the language of contemporary theories of autonomy, the essential claim is that the criminal will is wanton in the technical sense of that term adopted by Harry Frankfurt.[52] In short, the criminal's real will is in accordance with the universal - it is a rational will - the criminal's actions are thus not his (in the sense that they are not what we might now call autonomous), although, other things being equal, the criminal is still responsible for his wanton actions. The criminal's actions are not in accordance with his second order desires, and his second order desires are a truer reflection of him, because he is a rational being participating in the universal will.[53] In this sense the criminal is, according to Tunick,

---

[50]Stillman sums up the nature of Hegel's argument in the claim that the criminal "implicitly consents to the punishment and sees it as just" (Stillman 1976, 174).

[51]Tunick 1992a, §2.2.

[52]See Frankfurt 1971.

[53]Tunick is right in saying that this is very similar to Rousseau's idea that a particular will can be coerced into acting in accordance with the universal (general) will because to do so is not to infringe upon the freedom of the particular will but rather to "force it to be free".

> "like the smoker who knows that smoking is bad, and who wants very badly to quit, but who nevertheless continues to smoke and enjoys doing so. This smoker's de facto will is to smoke, but his real will is not to smoke".[54]

In the contemporary literature an account such as this would suffer from allegations that it relies on an intuition about second order desires being closer to the individual's "true self", or on unsustainable claims about rationality, or that it fails because it opens an infinite regress, or, finally, because it is incomplete.[55]

In Abstract Right, Hegel faces no such problems as the universal will is present; and right, which is the form of the universal will, commands a negative sphere of respect for the property and personhood of others. The question at issue is, rather, why annul the irrational will - or, why does the irrational will demand to be cancelled? Tunick sums it up as follows, "if the criminal's real will is not to commit the crime, then it follows, Hegel argues, that the criminal wills his own punishment".[56] Since it is not at all self-evident that any such thing *follows* this is the claim that must be explicated. The first answer is the objective one; because otherwise right would not be restored. That has been considered above, what we are concerned with here is the subjective reason for

---

[54]Tunick 1992a, 30.

[55]Anyone wishing to follow up on the autonomy literature referred to here can do no better than by beginning with the Introduction to Christman 1989, and the articles contained therein.

[56]Tunick 1992a, 30.

punishment. There is little doubt that the form of Hegel's subjective argument is Kantian - punishment supposedly treats the criminal's will as if it were rational by universalising it, by holding it as valid, and allowing it to return, or rebound, on the offender.[57] It thus treats him as a free being.[58]

15. Hegel and Kant

On such an interpretation Hegel's view is very similar in form to the Kantian argument I examined above. Consider the following two passages:

> "Accordingly, whatever undeserved evil you inflict upon another within the people, that you inflict upon yourself. If you insult him, you insult yourself; if you steal from him, you steal from yourself; if you strike him, you strike yourself; if you kill him, you kill yourself."[59]

> "If you rob someone, you rob yourself; if you kill someone, you kill yourself; the perpetrator may be subsumed under the manner of treatment he established."[60]

I have argued that on a Kantian interpretation the criminal wills a criminal act and is supposed thereby to make the

---

[57]See §7.

[58]See §99A.

[59]Kant 1797, 141.

[60]Hegel 1970, 244. Quoted in Tunick 1992a, 31; and Wood 1990, 113.

same act permissible when it is employed against him. In Kantian terms, the criminal acts according to a maxim which rebounds upon him.[61] This is the interpretation given to Hegel's theory by Alan Wood,

> "when I commit a crime, I set up a law making it permissible for others to violate my right to the same extent that my crime violates the right of its victim".[62]

But as Wood admits, this theory is incomplete, it merely makes it permissible for the criminal to be punished, not obligatory.[63] Indeed, it is not at all clear why the existence of the irrational will should be regarded as establishing a right for others to punish the possessor of that will. The standard Kantian move here is to say that we are obliged to respect the criminal will *as if* it were a rational will and thus reapply that will to the criminal in the form of punishment, but we have seen that this is a very difficult claim to understand.

Fortunately, although it is very similar, this is not Hegel's argument, although at times his writing about respecting the criminal (especially when he is concerned with condemning utilitarian justifications for punishment), suggests otherwise.[64] To understand the relationship

---

[61]See §7.

[62]Wood 1990, 114. Cf. Hegel: "For it is implicit in his action, as that of a *rational* being, that it is universal in character, and that, by performing it, he has set up a law which he has recognised for himself in his action, and under which he may therefore be subsumed as under *his* right." (§100)

[63]Again, see §7.

[64]See, e.g., §99A.

between the crime committed by the particular will and the subjective reason for annulling that will, it is vital that one appreciate the role of that annulment in the transition of the particular will from the sphere of Abstract Right to Morality. The particular will, through its annulment, is brought back to the universal will - like Tunick's smoker who in being denied cigarettes is forced to live in accordance with his "real" desires. But, unlike Tunick's smoker, this deepens the moral nature of the particular will; in being "forced to be free" the particular will makes the transition from the sphere of Abstract Right to a deeper understanding of itself in Morality.

In the transition from Right to Morality, Hegel says that the will

> "…first posits itself in the opposition between the universal will which has being *in itself* and the individual will which has being *for itself*; then, by superseding this opposition - the negation of the negation - it determines itself as will *in its existence*, so that it is not only a free will in itself, but also *for itself*, as self-related negativity. Thus it now has its *personality* - and in abstract right the will is no more than personality - as its *object*; the infinite subjectivity of freedom, which now has being *for itself*, constitutes the principle of the *moral point of view*." (§104)

Hegel claims, in other words, that in Abstract Right the will has an immediate relation to its content. That is the absolutely free will is contentless, it is unconstrained by any particular embodiment; in Abstract Right the will arbitrarily embodies itself in a content, (that is in desires, personhood and property), that is unmediated by Right. The will simply claims a content, which it identifies as itself. Given this, the will is capable of wrong - of claiming a content that rightfully belongs to another - and, in the annulment of such a wrong will, the abstract will must come to understand its actions *as wrong*, and thus come to realise the necessity of the mediating role of Right, and, thus, of the universal in its fulfilment as free. In other words, the will must come to understand itself as a self-determining will, that is, it must realise that it is free only insofar as its content - what it wills - is adequate to the idea of its freedom. It is coming to its understanding as self-determining (in accordance with Right), rather than merely as determined by its capricious desires:

> "The immediacy which is superseded in crime thus leads, through punishment - that is, through the nullity of this nullity - to affirmation, i.e. to *morality*." (§104A)

The crucial claim is that it is coercion that makes this realisation possible, the particular will must be brought back to the universal and Hegel's claim is that this is achieved through coercion; a coercion made permissible because it is implicit in the action of the criminal will

as an inconsistent and irrational will. It is difficult to see what underlies this claim, as David Cooper puts it,

> "no doubt inconsistent behaviour should be brought to the attention of the agent - this might be the job of a psychoanalyst - but I do not see how inconsistency *per se* merits punishment."[65]

Cooper's remark comes under attack in Peter Stillman's account where he asserts that

> "Cooper ignores the universality of the person and of the rights of persons, the 'equality (identity) of men in terms of personality.' (Hoffmeister, 333.[66]); identity means that what is true for one person is true for all."

Stillman adds, gleefully, that "since there are no intentions nor morality nor society, there are no psychoanalysts in 'Abstract Right.'".[67] Of course there are no psychoanalysts in Abstract Right, however, it seems unlikely that Cooper was suggesting that there were; as with Hegel we can, I think, treat Cooper's aside as a remark which does not have to be strictly taken to apply only to Abstract Right. In fact, Cooper's case seems a little more damaging than Stillman allows, for if there is no justification for coercion in Abstract Right, then, once

---

[65]Cooper 1971, 160-1.

[66]This reference is to Hegel's marginal notes, published in Hoffmeister's edition of The Philosophy of Right, the number being that of the page, not the section. (Hegel 1955)

[67]Stillman 1976, 171.

society is established, the annulment of the crime may take the form of moral education or psychiatric help.[68]

Stillman's substantive point that coercion is justified because all are identical in Abstract Right is perhaps less funny than his comment on psychoanalysts, but is also considerably less perspicuous. The identity of persons has been taken into account above and yet no relation between the criminal will and coercion has so far been forthcoming, either for objective or subjective reasons. If the purpose of punishment viewed subjectively is to deepen the understanding of the particular will then the intuition lying behind Cooper's claim that this is likely to be achieved by something other than coercion holds. But more than this, even if it were true that coercion is the best method of correcting the will this would be a purely contingent truth, it just may be that certain particular wills respond differently to moral education, denunciation or coercion.

In short, the subjective reason for the annulment of the particular will is that the will must be corrected - it must be forced back to the universal; but neither this correction nor this forcing is identical to, nor perhaps

---

[68]Indeed, Jean Hampton in her excellent "The Moral Education Theory of Punishment" suggests that Hegel is best read as a believer in punishment as moral education. (Hampton 1984) J. McTaggart also, or so it seems to me, suggests that this is the best way to read Hegel see McTaggart 1896. Tunick writes that "…we might see punishment as a sign of affection, the community's expression of concern for the criminal's own well being." (Tunick 1992a, 88) He fails to see, however, that there may be an alternative to coercion implicit in such a view.

best served by, coercion. To think that it necessarily is, is to confuse *correction* (in which education may be coercive), and *coercion*. (in which coercion is used to educate).

So far my discussion has centred on the annulment of the criminal will in Abstract Right, however, before concluding that Hegel fails to justify the coercive form of this annulment on either objective or subjective grounds, I want to move the argument forward to punishment and the State. This is clearly vital if we are to consider punishment in its final, true, form.

16. Punishment, Sittlichkeit and the State

The arguments from Abstract Right transfer to, and are completed in, the ethical realm of the State, and they thus suffer from the same flaws as in Abstract Right. The annulment of the criminal will becomes, in the modern state, punishment. The ethical community is to be understood as the form taken by a self-conscious and self-determining ethical whole. Further, as the ethical community shapes and moulds our ethical nature, so it is that in breaking the rules of the community we break our own rules. In other words the ethical community is that through which our own rational natures become fully actual, but at the same time as our own development as rational beings is dependent upon the ethical community, so the

continued existence of the community is dependent upon our participation within it.

The self-actualisation of the ethical whole - that is the pursuit of its good - is expressed through its positive law; the legal system of rights which is denied by the criminal will. As in the sphere of Abstract Right, the primacy of the right - now expressed in law - must be asserted and the crime must be annulled. Likewise the criminal must be brought to see that his act is self-destructive, in that in acting against the rules of the community he frustrates his own real will, as it is expressed in those rules. As Hegel puts it:

> "We must accept that the absolute will of the criminal also is that he be punished. Insofar as he is to be punished, the demand is present that he understands that it is just that he be punished, and though he understands he can of course wish that he be liberated from punishment as from external suffering; but insofar as he admits that it is just that he be punished, his universal will is in agreement with the punishment."[69]

The argument is, briefly, as follows. In Hegel's philosophy we come to freedom, to full actualisation, when we comprehend the institutions and practices of the modern state as necessary, and as such, as in accordance with our own will. As Alan Wood puts it,

---

[69]Hegel 1970, 225. Quoted in Tunick 1992a, 30.

> "freedom is actual, therefore, only in a rational society whose institutions can be felt and known as rational by individuals who are 'with themselves' in those institutions".[70]

It is in this sense that Hegel can be summed up in the aphorism, "I am at home in the world when I know it, still more so when I have understood it". Punishment is a necessary practice of the rational state and thus is in accordance with our own, universal will. The agent, as a member of the ethical community implicitly wills the practices of that community and thus wills his own punishment, a punishment that, furthermore, when inflicted restores him to that community and to his real will, from both of which his criminal action had alienated him. He is thus bought back to freedom, understood as living in accordance with his real will.

This is why Hegel applauds the ancient Athenian practice of demanding that the accused, once found guilty, should then participate in his own sentencing by proposing a just punishment,[71] and why he argues in The Philosophy of Right for trial by jury.[72] The argument is that whilst it is most desirable that the criminal confess, because "only when the criminal confesses does the judgement no longer contain anything alien to him" (§227A), the criminal may

---

[70]Wood 1991, xii.

[71]See Hegel 1974, Volume I, 440-441.

[72]§227A. It should be noted that trial by jury was not, by any means, universal at the time, and this makes Hegel something of a reformer in the modern sense.

refuse to do so. In this case it is not satisfactory for the judge to be the final arbiter because this introduces too great an element of subjectivity - in this sense it would resemble revenge (carried out by a third person)[73] - rather, the judgement must emanate from the criminal herself, and if she refuses then from a representation of the criminal conceived as part of the ethical community, that is, from a jury made up of members of that community:

> "If...the subjective conviction of the judge is to prevail, an element of harshness is again introduced, for the person in question is no longer treated as a free individual. The mediation [between confession and the judge] is the requirement that the verdict of guilt or innocence should emanate from the soul of the criminal - as in *trial by jury*." (§227A)

Again the parallel with Abstract Right is apparent, as is the worry that nothing in this justifies coercion. The idea is that the criminal in being tried and punished realises that her deed was wrong, and that as an action against the ethical community it was also against herself, and her freedom. Punishment is supposed somehow to reinforce this realisation, to bring forth a sincere confession and a correction of the criminal will.[74] However, this is a realisation that might better be

---

[73]See §§101-102.

[74]It is interesting to compare this with Michel Foucault's claim that the judge, when asking whether the criminal has anything to say before he passes sentence, is asking for the criminal's confession so that he (the judge) may feel legitimised; in other words, it is a favour done by the criminal for the judge so that the latter may be reconciled to his use of coercion against another without, as Nietzsche put it, "even emotion to excuse [him]". Nietzsche 1967, 67.

accomplished through the use of moral education not coercion, and again, even if coercion were the best method in some cases, this would be merely an empirical truth.

Of course, in the final argument - that is in the modern State - it is necessary to unite the two positions - that is subjective and objective. However, such an argument is in danger of collapsing the subjective into the objective. The argument would be this; not only does the punishment of the criminal reveal to the criminal the nullity of his will, but it is also in accordance with his universal will. That is the criminal as part of the ethical community wills the structures and institutions of that community, and one of those structures is coercive punishment. In other words, if the objective reason for punishing was such that the institution of punishing was shown successfully to be part of the necessary - that is rational - structures of the State, then the criminal as part of the State would, conceived as a rational being, consent to his punishment. This, however, depends upon coercive punishment being a necessary feature of the State, and I have shown above that Hegel cannot achieve this; rather, what is left is a powerful argument for the expressivist view of law and punishment, combined with a claim to the moral education of the offender.

In neither the work of Kant nor Hegel, then, have we been able to find a justification for the core of retributivism, the idea that the guilty deserve to suffer, and that this

is what justifies the practice of punishment. This despite their calling upon three of the most powerful retributivist arguments, that the act of the offender rebounds upon her, that punishment is somehow entailed by the idea of rights, and that the offender wills her own punishment. However, I have not completed my analysis of retributivism. In the next Chapter I shall turn to consequentialism, first as a complete account of punishment and then as a general justifying aim which can be constrained by retributivist principles. In so doing, I shall consider what many take to be retributivism's real contribution to the punishment debate, its role as a side-constraint on consequentialism. In Chapter 4, I will consider one further retributive argument, derived from the principle of fair play, as part of an analysis of the role of theories of justice in justifying punishment.

Chapter 3: Consequentialist Justifications of Punishment

> "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 to be admitted in as far as it promises to exclude some greater evil."
>
> Jeremy Bentham 1789, Chapter 13, §1, 2.

17. Consequentialisms

Consequentialism has had a hard time of it recently. Once philosophically dominant, it was still taken to be the theory to beat in Rawls's A Theory of Justice, yet now the philosophical opposition to liberalism is more likely to be made up of post-modernists of various descriptions, feminists, and communitarians.[1] In this change, the sub-field of the study of punishment acted as something of a harbinger; whilst in the post war years consequentialist justifications of punishment of one sort or another dominated, there is little doubt that, as Antony Duff puts it, "the 1970's saw a growing revolt against this consequentialist orthodoxy".[2] This revolt was not simply a philosophical one, indeed, it was most notable not in academic journals but in political and judicial debates over sentencing policy. Put simply, there was a move away from indeterminate to relatively fixed sentences.

---

[1] "In Anglo-American philosophy, the prevailing way of looking at morality used to be consequentialist, with some qualms about lying, breaking promises or killing innocent people in a good cause. It has now shifted towards absolutism, with residual doubts about catastrophic consequences." (Barry 1979, in Barry 1991b, 67-8).

[2] Duff 1993, xii.

In this Chapter I intend to examine various consequentialist justifications of punishment. I shall argue, first, that in its traditional form, consequentialism cannot include a respect for rights, or more generally, for the special claims of individuals which underlie our intuitions about justice (and inform our idea of rights). I shall then go on to examine a number of arguments which purport to show how such claims can be accommodated in a consequentialist theory.

At the heart of consequentialism is the argument that punishment is to be *justified* only by reference to its good consequences, if it did not yield a net gain in whatever measure the particular consequentialist endorses, punishment would not be justified. The retributivist punishes, let us say, because to do so is morally required by the offending act, he may check after punishing that he has correctly performed the task, but he does not justify his punishing the offender by anything other than the relationship between the past act and morally deserved punishment. In contrast, a consequentialist punishes because it has good consequences, and, there is a net benefit (in terms of the measure the particular consequentialist endorses), compared with not punishing. Punishment is justified by reference to the benefits it yields, usually in terms of the benefits of preventing or reducing the incidence of crime.[3]

[3]For a standard account of this sort see Benn 1958. For a modern, 'unreconstructed' consequentialist account of punishment see Walker 1991.

The usual, philosophical, approach to consequentialism (and, specifically, utilitarianism), is to distinguish between "act" and "rule". J. J. C. Smart contrasts them thus:

> "Act-utilitarianism is the view that the rightness or wrongness of an action is to be judged by the consequences, good or bad, of the action itself. Rule-utilitarianism is the view that the rightness or wrongness of an action is to be judged by the goodness or badness of the consequences of a rule that everyone should perform the action in like circumstances."[4]

Although this distinction has become standard in discussions of utilitarianism it seems to me to be of only limited use when applied to society rather than individual actions. This is because no society could possibly operate on an act-utilitarian level, that is without rules. The maintenance of a society needs more predictability than could be achieved if it were founded on an act-utilitarian theory. Where this distinction is of use is when considering the nature of the rules and the morality of rule following. This is because, even if a system of rules is necessary, there is always the question of whether, in a particular instance, it would be better to break the rule; that is, the question of whether act-utilitarian considerations can take precedence even given the existence of a system of rules that are supposed to be action guiding. I shall assume that an act-utilitarian theory

---

[4]Smart and Williams 1973, 9.

applied to a society answers this question affirmatively, i.e., although the society has action guiding rules, it is morally obligatory to break a rule if doing so in that instance will yield a net increase in utility compared with keeping to the rules, taking into account the effect on rule following of breaking the rule in that instance. Let us, then, begin with this sort of theory as our subject.

18. Consequentialism, The Locus of Value and Punishment

Even given the plausible assumption made above that all societies need to have some systems - some rules - even if they can be broken when consequentially justified, (in the most consequentialist society perhaps they would be more like routines than rules), we can still ask whether a thorough going consequentialist society would have anything that looks enough like a system of criminal law and criminal sanctions to make asking the question of whether a consequentialist justification of punishment can be found, meaningful. I shall deal below with whether inflictions of "special treatment" justified by consequentialism count as punishments, here I want to ask whether a consequentialist society would generate anything remotely resembling the whole "penality"[5] system.

---

[5]The term is borrowed from David Garland; "Penality … refer[s] to the complex of laws, processes, discourses, and institutions which are involved in … the whole process of criminalizing and penalizing". Garland 1990, 10.

The first point of divergence between our ideas and practices and the system that might emerge in a consequentialist society, would be that if the consequences that count are the reduction of certain anti-social acts, the past act of the offender would play little role in determining which individuals should receive special treatment and what that special treatment should be. If the point is to deter the offender from committing a similar act again, the important thing will be the particular psychological makeup of the person; if the purpose is to rehabilitate the offender, the same thing applies; and if the purpose is to deter others, the special treatment will be determined by certain empirical conditions in the society. Offenders convicted of similar offences, then, might - and probably would - receive very different treatment.

However, there is no reason to assume that special treatment will be confined to offenders, it might be the case that a criminal act alerts the society to the existence of a potential new offender[6] but there is no reason to assume that this is the only, or even the best, way of identifying such people. Duff has suggested that a consequentialist society may instead opt for "Adult Panels", tribunals that "would deal with *anyone* alleged to be in need of … special treatment, whatever the grounds for

---

[6]By "new offenders" I do not mean people who have not offended but people who might offend again. They are, of course, by virtue of having committed a past criminal act also "old offenders" but this would not be relevant to a consequentialist society except, as I say, as one indicator of potential behaviour.

that allegation".[7] Such "diagnostic tribunals" may resemble criminal trials in presuming innocence and granting a right to be heard, not because these things are grounded in an account of justice, but because they are likely to reduce disutility through subjecting the wrong people to special treatment. On the other hand, in different - less stable - conditions, both of these procedures might be foregone and, indeed, a presumption of guilt might be required.[8]

This idea, that determining who should be the subject of what special treatment is a matter of diagnosing those who are dangerous or useful (in terms of being made an example of and deterring others), rather than discovering the guilt or otherwise of the individual for an offence, is part of a family of problems that is attached to any act-utilitarian theory. Other examples are well known and documented.[9] The core of the problem, in essence, is that in such a theory value is identity independent, (what matters is states of affairs), rather than identity specific, (that is attached to individuals), and this is the basis for Rawls's charge that "utilitarianism does not take seriously the distinction between persons".[10]

---

[7]Duff 1986, 104. See also 102-106; 164-72.

[8]See Duff 1986, 102-106.

[9]See, amongst many other things, Williams' contribution to Smart and Williams 1973.

[10]Rawls 1971, 27.

Punishment is one of the standard practices from which examples are drawn to demonstrate this, (and it is, therefore, one of the clearest precedents for what I shall try to do in the rest of this thesis, that is, use punishment as a concrete practice through which to examine accounts of justice). The examples are not so much well worn as exhausted and I shall spend little time elucidating them. In addition to the problems noted above concerning the distribution and nature of special treatment, a fully consequentialist society of the type I have been discussing could, for example, sanction the punishment of an innocent person, that is, an innocent could be tried, convicted, and punished with the full complicity of the authorities because there is greater benefit realised through his punishment than the loss of benefit felt by him and those who care about him.[11] Second, it might be morally obligatory for a judge to sentence a convicted criminal to

---

[11]This example is used in a great many discussions of both utilitarianism and punishment. If its philosophical fame can be attributed to anyone, it is, perhaps, H. McCloskey 1968. It has been objected that such "fantastic examples" do not aid moral argument, (see Anscombe 1957; Hare 1981, especially Chapters 1-3, 8-9; Sprigge 1968; and the reply to these by Ten 1987, 18-32), however, it is not clear to me that such analysis is all that fantastic. The arguments are not quite analogous, but the following example, nevertheless, seems to me instructive. Lord Denning, who at one time was England's third most senior judge, was asked (in a BBC Television News interview), whether the quashing of the conviction of six Irishmen jailed for the bombing of a Birmingham pub (in which there were fatalities) because the police evidence was found to be unsafe, had altered his belief in the rightness of capital punishment. His answer included the argument that at least if the "Birmingham Six" had been hanged there would have been no campaign to re-examine the evidence, the conspiracy and failures of the judicial system would not have come to light, and the integrity of that system in the eyes of the public would have been maintained.

A variation on this example, which is less often cited, but which I have found students think more plausible, is the punishing of an offender's family. This was, apparently, a fairly strong disincentive to potential defectors from the Soviet Union, and remains so in contemporary China.

an extremely harsh penalty in order to make an example of her and deter others.

The point is that in all these cases - the diagnostic tribunal, the punishment of an innocent, the exemplary punishment - the claims of the individual to equal respect are lost in the concern for the whole. The individual is simply a value carrier, and if greater value can be achieved by sacrificing this particular vessel, then that is what is morally required.

Consequentialists have attempted to rise to this challenge - and, in punishment theorising, to the resurgence of the "new retributivism".[12] However, before I consider some of these developments in consequentialist theory, I want to dismiss two defences sometimes invoked by defenders of consequentialism.

Taking the example of punishing an innocent, it is sometimes claimed that, as a matter of fact, the consequences of punishing an innocent will always be worse than not doing so, because of the impossibility of keeping it quiet, because of the immense unhappiness caused to the innocent and her family, because of the uncertainty it might lead to in the society, etc., and that therefore a consequentialist society will never, in fact, punish an innocent. This defence, however, just misses the point. The problem is that the consequentialist could consider

[12]Honderich 1989, 208.

punishing an innocent as a morally permissible - or, worse, obligatory - action. The fact that the empirical conditions in which this was required might never occur does nothing to alleviate the worry that such thinking is deeply offensive.[13]

The second defence offered by some consequentialists is the "definitional stop".[14] In a well known paper,[15] Anthony Quinton argued that there is a logical connection between punishment and guilt; punishment *means* the infliction of suffering by a duly constituted authority of an offender for an offence.[16] Utilitarians, therefore, cannot be committed to *punishing* the innocent, although they may be committed to subjecting "innocents" to special treatment.[17]

There are two responses to this argument, the first is to deny that punishment means any such thing. This is sometimes implicit in revised definitions of punishment that build on the Flew definition; thus, Hart simply

---

[13]"The crucial charge is not that a consequentialist will in fact punish the innocent, but that she is ready to contemplate it as an open moral possibility." Duff 1986, 160.

[14]Hart 1959, 5, all page references are to the reprint in Hart 1968.

[15]Quinton 1954.

[16]The classic formulations from which these characteristics are drawn are Baier 1955; Benn 1958; Flew 1954; Hart 1959. See also Davis 1983, 728.

[17]"Even if the world gathered all its strength, there is one thing it is not able to do, it can no more punish an innocent one than it can put a dead person to death." (Kierkegaard 1961, 85; quoted in Duff 1986, 152.) Of course, relative to the "special treatment" the individual might not be "innocent" (which is why I have put that term in quotation marks, although they may be innocent of any past behaviour that requires punishment or special treatment).

includes, as a "substandard or secondary case", the "punishment of persons … who neither are in fact nor are supposed to be offenders".[18] Alternatively it can be the explicit claim that punishment is not a simple practice that can be captured in any definition on the Flew model, and certainly it is not so simple as to allow the definitional stop. Something like this seems to be behind David Garland's, (and, for that matter, Michel Foucault's), sociological approach.[19] At one point Garland cites, approvingly, the following passage from Nietzsche's The Genealogy of Morals:

> "I would say that in a very late culture such as our present-day European culture the notion 'punishment' has not one but a great many meanings. The whole history of punishment and of its adaptation to the most various uses has finally crystallized into a kind of complex which it is difficult to break down and quite impossible to define."[20]

Garland's general point (like Nietzsche's and Foucault's) is that punishment is a deeply embedded social practice that functions at a number of levels and performs a variety of tasks (some, perhaps, unknown or unstated). Any attempt to capture it on something like the Flew model is, therefore, likely to yield only confusion, and likewise, any attempt to invoke a definitional stop is specious.

---

[18]Hart 1959, 5.

[19]Garland 1990; Foucault 1977.

[20]Nietzsche 1956, 212; cited in Garland 1990, 17

Neither of these criticisms of the definitional stop seem to me to be appropriate, instead what must be asserted is that the definitional stop is true but trivial. The reason I think them inappropriate is that they give up too much, it is important that we maintain that punishment is "of an offender for an offence" because that is what it is, and it is only by remembering this that we can identify what is to count as punishment and what is a perversion. In other words we are not arguing just about the meaning of punishment but also about the way in which we must go about justifying punishment practices:

> "Questions about the meaning of punishment are not merely terminological: they concern the justificatory criteria internal to the concept of punishment, which enable us to criticise, as well as to identify, punishments and systems of punishments; and an account of punishment must explain and justify these criteria."[21]

Hart's addition of punishing the innocent as "a substandard or secondary case" must, therefore, be rejected, for otherwise we lose the central core of what it is that we are trying to justify and in so doing our idea of how to go about justifying it.

David Garland's argument is less simple, and there is no doubt that the work of sociologists, and sociological approaches to punishment, have a great deal to teach philosophers and criminologists about punishment. Nevertheless, no matter what role the practices that are

---

[21]Duff 1986, 151.

called "punishment" play in any given society, we must, if we are interested in analysing and justifying them, identify what it is that is to be analysed and justified. A society may call some "special treatment" doled out on the basis of a witch-doctor's prediction of future actions, "punishment", but that does not mean it is justifiable as such; if we want to argue that it is not justifiable as punishment, or indeed, undertake any analysis of it, we must have an idea of what we are analysing and how it is to be justified. To repeat Duff's point, the meaning and justificatory criteria are intertwined. Garland, himself, unintentionally highlights the importance of this, within eleven lines of criticising "formulaic … philosophers of punishment", he declares, "punishment is taken here to be the legal process whereby violators of the criminal law are condemned and sanctioned in accordance with specific legal categories and procedures".[22]

Nevertheless, although my criticism of Hart and Garland commits me to accept the premise that punishment means the infliction of suffering on an offender for an offence, I do not see that I am thereby committed to a definitional *stop*. There are two reasons for this, the first is that one question in which we are interested is whether a system of punishment, as against some alternative system, is to be preferred. Hart, for example, in dismissing the definitional stop, argues that it is hardly sufficient when asking whether a consequentialist society could inflict

---

[22]Garland 1990, 17.

suffering on an innocent, or on the relatives of an offender, to be told that "*that*, by definition, would not be 'punishment'".[23] The question we want to address to the consequentialist is would a consequentialist society have a system of punishment (which only - or attempts to only - punish offenders for their offences), or a some other system which did not include these constraints.

When asking other questions it is important what the system of "social hygiene"[24] is called. Thus, if someone claims to be punishing an offender when he really knows the person to be innocent, or a society is claiming to punish when it is deliberately inflicting suffering on the family of an offender, then it is important both that what is being claimed is that the procedure is part of a system of "punishment" and that we have the idea of what punishment is; here the definition of punishment does not stop us, in fact, it allows us to judge what is going on and denounce it.[25]

Having considered and dismissed the empirical and the definitional consequentialist ripostes, I now want to turn my attention to more serious attempts to get out of the problems that arise because of consequentialism's commitment to an identity independent locus of value. I shall begin with the most well known move, the adoption of

---

[23]Hart 1959, 6.

[24]Taken from Hart 1959, 6.

[25]See Duff 1986, 152.

rule-utilitarianism, before considering some, perhaps more esoteric, developments in punishment theory.

19. Rule-Utilitarianism

Fortunately, the best defence of rule-utilitarianism, John Rawls's "Two Concepts of Rules" includes a detailed argument about the punishment - or "telishment" - of an innocent scapegoat, this makes it particularly useful and appropriate for my purposes.[26] Rule utilitarianism, according to Smart, "is the view that the rightness or wrongness of an action is to be judged by the goodness or badness of the consequences of a rule that everyone should perform the action in like circumstances."[27] Rawls gives this a concrete form by distinguishing between "justifying a practice and justifying a particular action falling under it".[28] His argument is, essentially, that once a practice is up and running the participants cannot appeal outside of the rules that govern and define the practice, even to considerations that inform the structure of the practice itself.

---

[26]Rawls 1955, reprinted in Foot 1967, 144-70, page references are to this reprint. "Telishment" is the term Rawls adopts in order to avoid definitional problems. There are numerous discussions of rule- (or indirect-), utilitarianism, (e.g., Smart and Williams 1973; Lyons 1965); for a discussion which is historically informative, see Gray 1989, Chapter 8.

[27]Smart 1973, 9.

[28]Rawls 1955, 144.

Rawls's simplest example is of a game of baseball, he notes that if a batter asked whether he could have four strikes, this would normally be taken to be a request to have the rules clarified for him. If, when he was told that the rules only entitled him to three, he said that he knew this, but he thought that on this occasion it would be better for all concerned if he had four, "this would be most kindly taken as a joke".[29] Of course, this does not stop the player trying to change the rules of baseball in the out season, if he believes that it would be a better game with four rather than three strikes allowed.

With respect to punishment, Rawls argues that instituting a practice of "telishment" - in which officials "have authority to arrange a trial for the condemnation of an innocent man whenever they are of the opinion that doing so would be in the best interests of society"[30] - would be so hazardous that a "utilitarian justification for this institution [would be] most unlikely".[31] Instead, Rawls believes, the practice most likely to receive utilitarian backing would be one similar to the practice of punishment, with all the checks of the right to audience, the presumption of innocence, etc., in place. "It happens in general", Rawls remarks, "that as one drops off the defining features of punishment one ends up with an

---

[29]Rawls 1955, 164.

[30]Rawls 1955, 151.

[31]Rawls 1955, 152.

institution whose utilitarian justification is highly doubtful."[32]

Let us assume that Rawls is right, and that empirical conditions would mean that a consequentially justified system of rules for regulating society and dealing with offenders would include something which resembles a system of just punishment, although this is by no means clear.[33] Rawls's argument, then, is that having established the practice of punishment with rules governing who may be punished and how, those rules must take precedence over any direct appeal to the consequences in any particular instance. Here we have something that looks very much like Hart's "mixed account" in "Prolegomenon to the Principles of Punishment", in which the general justifying aim of punishment is consequentialist but the principle governing the distribution is retributive.[34]

At its simplest, the idea is that, e.g., given the circumstance in which an innocent individual comes up charged with an offence in a magistrates court,[35] the magistrate is not at liberty to convict and punish that individual because, in this instance, there would be a net gain in utility, because she can only operate on the rules

---

[32]Rawls 1955, 152.

[33]See McCloskey 1972.

[34]See Hart 1959.

[35]The advantage of using a magistrate is that no jury is involved and the example is thus more simple.

that define, and are thus internal to, the practice; she cannot appeal to the utilitarian grounding of the practice itself.

Rawls does not mean this to be taken in the practical sense, often ascribed to Kant's view on promise breaking, that if one allowed the breaking of the rules in any given practice, that practice would wither away because people would lose confidence in it.[36] Rather, Rawls is making a logical claim, that if one does not follow the rules of a practice, one is simply not doing the same kind of thing, "the rules of practices are logically prior to particular cases". Thus, although one can swing a piece of wood at a ball and run if one hits it, one cannot play baseball unless one is engaged in the practice of baseball, and following the rules that define it. Similarly, then, the practice of punishment is defined in such a way as to make the statement "you are innocent but, for the benefit of society, I am going to punish you anyway", *nonsensical*.[37]

This looks right, but would anyone say such a thing. Consider the practice of promise keeping, Rawls's argument must be that if anyone said "I have no intention of doing X, but I promise to do X", they would be speaking nonsense, they would not understand what it is to make a promise. But, if one thought that all things considered it would be

---

[36]Rawls explicitly rejects this argument with reference to promise keeping, Rawls 1955, 154-55.

[37]Cf. Duff 1986, 162-63.

better to make a false promise would one say such a thing, or even describe one's actions in such terms? Surely not, what one would say is, "all things considered, it would be better if I *pretended* to promise to do X, even though I know that I will be unable to do X when the time comes." This is not logically incoherent, it is simply deceitful, and, of course, if one is trying to realise consequentialist ends it is likely that such deceit will be necessary. Similarly, the magistrate does not admit in open court, "I know you to be innocent, but for the good of society I am going to punish you", rather she says, "you are guilty and I am going to punish you", although she knows the person to be innocent.[38] As Mabbott puts it, "indirect utilitarian arguments can always be met by 'Break the rule and keep it dark'".[39]

There seems to be no logical conflict, then, in the agent pretending to be engaged in a defined practice whilst breaking the rules for consequentialist purposes. To deny that the agent should do such a thing introduces a moral schizophrenia, the agent knows that the right thing to do is to punish this individual given the meta-theory that defines what is right, but knows that the rule requires that she release him; if it is not logically incoherent,

---

[38]Cf. Duff 1986, 163. As Duff makes clear, the importance of this argument is that it reminds us again that punishing the innocent violates criteria which are internal to the concept of punishment; deliberately punishing an innocent is a perversion of punishment, and this is why it must be accompanied by dishonesty and deceit on the part of the punishing authorities.

[39]Mabbott 1955, 128, see also Mabbott 1939.

why should she give precedence to the rule instead of breaking it and "keeping it dark"?[40]

Logic, in fact, seems to work against the rule-utilitarian. This is because any practice justified on utilitarian grounds is clearly going to include acceptable exceptions to the rules, and these will be justified by the meta-consequentialist theory. Thus Rawls says, of promise keeping, that there are "various excuses, exceptions and defences, which are understood by, and which constitute an important part of, the practice".[41] Simplifying the situation, let us take the rule, 'in circumstances A, do B'. If this rule is to be justified on consequentialist grounds, we must admit consequentially justified exceptions, so in cases C, let us say, it will always be better, all things considered, not to do B when A. Thus, the rule will now read, 'in circumstances A, do B except when conditions C apply'. But, one can conceive of the possibility (although one could not hope to achieve it), of capturing all possible exceptions to the rule, that is all conditions when it would be better, all things considered, not to do B when A. But that means that the rule is now, 'in circumstances A, do B except when conditions C - N

---

[40]See Ten 1987, 71. This argument applies, *mutatis mutandis*, to Hart's, 1959. The point is, that "mixed theories" of the Hart type, offer no account of how the different demands of yielding good consequences and matching moral desert can be reconciled. Instead they try to confine them to different spheres of punishment. However, since each appeals to a comprehensive moral theory, each makes claims that conflict with the other, see Goldman 1979; 1982, 62.

[41]Rawls 1955, 156.

apply', and this is identical to the rule, 'in circumstances A, do B except when better consequences would be achieved by not doing B', in short, rule-utilitarianism logically collapses into act.[42]

The final objection I wish to levy against this kind of theory is that it is inappropriate in a way similar to the argument that a consequentialist society will be unlikely to punish an innocent because of empirical factors. As Rawls admits "a utilitarian justification for [the systematic punishing of the innocent] is most unlikely",[43] but "most unlikely" just isn't good enough; punishing the innocent is wrong for some reason other than that it is unlikely to yield good consequences.

20. Consent, Threats and Self-Defence

The failure of rule-utilitarianism and the problem of making consequentialism compatible with respect for agents' autonomy or rights has led consequentialist punishment theorists away from the traditional tripartite goals of deterrence, rehabilitation, and incapacitation, and towards a more singular focus on deterrence. Deterrence has the immediate advantage of seeming to appeal to the agent as a rational being, in the sense that punishment, on such a justification, provides the agent with a prudential reason

---

[42]See, amongst many others, the discussions in Lyons 1965, Smart 1968.

[43]Rawls 1955, 152.

not to commit the act he might be considering. The immediate disadvantage is that it seems that deterrence based punishment, if it aims to deter others as well as the offender from further infringements of the law, use the offender as a means of affecting the behaviour of others. The challenge, then, is to show that using the offender thus, nevertheless treats her as an autonomous agent.[44]

C. S. Nino has suggested that in committing an offence, given certain conditions, the offender consents to his liability to punishment.[45] The argument is most easily understood through an example.[46] When an individual gets into a cab and gives an address then, under normal conditions, that agent has consensually entered into a contract; "when [a] particular legal consequence of the voluntary act is known by the agent, we may say that he has consented to it".[47] He is legally obliged to pay the driver on arrival, even if when the time comes he does not want to, or believes that it would be better all things

---

[44]See Duff 1993, xiii-xiv.

[45]Nino 1983.

[46]I am much indebted to Ted Honderich's discussion of Nino's article (in Honderich 1984, 219-224). Honderich treats Nino as one of the "new retributivists", however, I think that he is wrong to do so. Nino clearly believes that the authorities should issue threats (in the form of punishments attached to certain actions) as a mechanism of social protection. Furthermore, his model of consent eschews moral desert, which is one reason not to treat him as a retributivist - even a "new" one. In his own words, he is seeking "a line of argument which shows that the practice of punishment can be patterned after a commonly accepted principle of distribution that does not rely on the moral blameworthiness of people and does not require us to relinquish the conception of punishment as a measure of social protection". (Nino 1983, 293).

[47]Nino 1983, 296.

considered if he did not do so, etc. This is subject to the important *caveat* that "the relevant laws be in some sense just",[48] this is to avoid the claim of a contracting party "that the law coerces him into accepting the terms of an offer should he want something over which the offerer has legal power".[49] Accepting the "legal normative" consequences of an act (e.g., getting into a cab), has moral consequences in that "the individual who … consents to undertake some legal obligation is, in principle, morally obliged to do the act which is the object of that obligation",[50] and, thus, others have a *prima facie* right to force the individual to meet his obligations. None of this is affected by the fairness or otherwise of the distribution of benefits and burdens that result from what the individual has consented to do, that is, it doesn't matter whether the cab driver is an eccentric millionaire and his passenger relatively poor, this is what Honderich calls "the fairness owed to consent".[51]

The liability to punishment, Nino thinks, is to be understood on this model. The offender, when he voluntarily commits an offence (in the knowledge that it is an offence), knows that to do so has certain legal consequences, specifically, he loses his right to immunity from punishment. As with the cab passenger, this is

---

[48]Honderich 1984, 220.

[49]Nino 1983, 302.

[50]Nino 1983, 296.

[51]Honderich 1984, 220.

independent of his desire, when the time comes, not to be punished or his belief that, all things considered, it would be better if he were not punished. Notice that the offender need not consent to the law, or accept the justice of the law, he need merely know that it is the law and that the action he has performed carries certain consequences. Nevertheless, it is necessary that the law be just, that it not be "discriminatory [or] proscribe actions that people have the moral right to do".[52] The next step is pretty clear, as Honderich puts it, "given the fact of the offender's consent, the authorities have at least a *prima facie* moral justification for exercising their legal power to punish him",[53] and this is unaffected by the justice or otherwise of the distribution of burdens that result from the offender's punishment. In summary,

> "The justification of punishment defended [by Nino] relies on the consent to assume the liability to suffer punishment involved in the voluntary commission of an offence with the knowledge that the liability is a necessary consequence of it."[54]

It is important to note that the offender does not consent to her punishment, but to her "liability to suffer punishment", this must be the case because the offender only consents to the necessary consequences of her actions, and punishment is not a necessary consequence.

---

[52]Nino 1983, 302-3.

[53]Honderich 1984, 221.

[54]Nino 1983, 305.

There are a number of problems with Nino's account; Ted Honderich, for example, argues that the problem with the analogy Nino draws is that the offender might be said to deny the legal normative consequences *at the time* of committing the act with those consequences. Imagine, Honderich says, if an individual gets into a cab and gives an address and then says, 'I know that the normal practice is to pay cab drivers, but I have no intention of paying you for this ride.' As Honderich says, it is certainly a moot question whether a contract has issued from the individual's actions, and it can only be more controversial still to claim that her actions issue "in a moral justification for [her] paying a fare".[55] Honderich's general claim "that if we begin with an offence, and find a close analogy of it that might turn up in civil law, we do not find anything remotely like a clear case of consent and contract",[56] seems to me to be right, and to cast doubt on the foundation of Nino's account.

The reader might feel, however, that this objection does not strike at the general idea that a contractual type consent might be the route to justifiable punishment, so it is necessary to address a second, more general, criticism of Nino's account. What is most ambiguous in Nino's account is the role of justice. Clearly justice is important; it is not enough to say to someone 'your money or your (claim to) life', and then to argue that they

[55]Honderich 1984, 223.

[56]Honderich 1984, 223.

consented to losing their claim against being murdered because they knew that this was the necessary consequence of not handing over the money. Nino commits himself to a Kantian ideal of treating each person as having an equal claim to well-being and he sees as the burden of his theory the task of reconciling the equal claims of individuals and the requirement that punishment diminishes future crimes by means of "general and special deterrence".[57] Nino says so little about the requirements of justice that his theory can most sympathetically be regarded as incomplete. However, his insistence that punishment is only to reduce the incidence of crime, and that it performs this function through general and special deterrence, does seem to open him up to Duff's attack that he does nothing to show that the use of threats is compatible with "proper respect for the potential criminal".[58]

What is surprising is that Nino does not regard the requirements of justice as affecting the fundamental consequentialist purpose of punishment. It may be that some degree of threat is compatible with respecting the offender as an autonomous being,[59] but it is difficult to know without giving a fuller account of the requirements of justice. Likewise, Nino argues that no matter how unjust the consequences of punishing, if the laws are just and the

---

[57]Nino 1983, 292.

[58]Duff 1986, 180.

[59]See §43 for a further discussion of this, and of my differences with Duff.

penalties known, then the fairness we owe to consent means that we should punish. The conflict between the requirement that the law distribute benefits and burdens in a just manner and the argument that a kind of consent can justify an unfair distribution of these things, is not addressed,[60] and, although, as Honderich says, they are not "clear enough to be inconsistent",[61] they do seem to pull apart in important ways. What is needed is a much better account of how we are justified in issuing threats, and how these threats can be compatible with the account of justice, for, as Daniel Farrell argues,

> "we are supposing that despite the fact that a murderer had no right to murder, it is problematic as to whether or not we have a right to kill him in order to prevent other potential murderers from murdering. And if all we can say, by way of showing that we have this right, is that we have told him we would kill him if he killed, it seems we have no said nearly enough. For our question is exactly what makes us think we have the right to tell him this, meaning to do what we say we will do if *he* does what we have warned him not to do."[62]

Farrell believes that this right - to threaten and carry out our threats - can be generated from the more familiar

---

[60]Nino does see the problem, "following out these suggestions [about the justification of punishment] would lead to a discussion of the extent to which the consent of the person affected can justify measures and political arrangements which may imply inequitable burdens upon him" (Nino 1983, 305), but he doesn't pursue it.

[61]Honderich 1984, 224.

[62]Farrell 1985, 379.

right to self-defence, and this is a position which has also been advocated by Warren Quinn.[63] In marked contrast to Nino, the 'self-defence thesis'[64] construes the "direction" of justification to be "from the justifiability of the threats to the justifiability of imposing the threatened penalties".[65] Both Quinn and Farrell begin with thought experiments; because Quinn's is the simpler theory, and the more enjoyable *Gedankenexperiment*, let us take his. Quinn imagines a system of mechanical (or m-) punishment. M-punishments are carried out by devices that can "detect wrongdoing,… identify and apprehend those who are responsible, establish their guilt, and subject them to incarcerations (and perhaps other evils)".[66] He assumes that we have "lost whatever taste we once had for retribution",[67] and thus, the system of m-punishment is grounded in trying to deter certain conduct by constructing the "automated retaliation devices" (ARDs).[68] Quinn sees

---

[63]See Farrell 1985; 1988; 1989; 1990; Quinn 1985. There are important differences between them, see Farrell 1989.

[64]I shall use this phrase to denote the basic approach which is shared by Farrell and Quinn.

[65]Farrell 1989, 126, emphasis suppressed. As Quinn notes, this is structurally similar to the rule-utilitarian accounts of Rawls (1955) and Hart (1959). The crucial difference is that Rawls and Hart rely on the distinction between the practice and the act, the self-defence thesis relies on a distinction between a justified earlier threat and a later punishment.

[66]Quinn 1985, 337. The ability of these devices to administer a wide range of m-punishments is what distinguishes them from Alexander's "Doomsday Machines" which merely imposed death for every transgression, (Alexander 1980). Cf. also Buchanan's "automatic enforcing agents" in Buchanan 1975, 95.

[67]Quinn 1985, 337.

[68]Farrell 1989, 127. See also 1989, 305.

the creation of ARDs as analogous to the creation of "threats" whose function is to deter individuals from certain conduct, if we are to justify creating these devices then, we must give an account of our right to threaten in this manner.[69] This account is based on drawing parallels with "other, more familiar, self-protective rights", such as the right to protect ourselves and our property, including with violence if necessary; to "erect barriers"; to arrange an "automatic cost" to "precede or accompany the violation of some right" and to confine those who show themselves to be both dangerous and uncontrollable by other methods.[70]

There are two questions that must concern us: First, can we show that from some general commitment to self-defence we are morally justified in creating ARDs? Second, if the creation of ARDs were morally permissible, can we understand the infliction of punishment as justifiable on the same grounds? Because I do not believe that the first question can be answered in the affirmative I shall not discuss the second in any detail.

The problem with the first question is that it is not clear what sense of "justifiability" is being appealed to. Quinn's thought experiment posits "a new community … at

---

[69]As Farrell points out, this is a rather idiosyncratic use of the term "threat". In creating ARDs, we create "a situation where potential wrongdoers are at a higher risk of being harmed, if they wrong us, than they would otherwise have been." (Farrell 1989, 151). Cf. Quinn 1985, 335.

[70]All quotations are from Quinn 1985, 341, emphasis suppressed.

some time in the future … our social structures having been destroyed by earlier upheavals", and Farrell's "something like a Lockean state of nature".[71] Since Farrell's is marginally the more informative let us follow the self-defence claim through as he offers it. In this 'almost Lockean' state of nature he asserts, "one right most of us would claim … is the right to resist, directly, others' attempts to violate our rights".[72] What strikes one here is the fact that there are a lot of rights kicking around, the grounding for which is entirely unaccounted for. Let us assume, and I take it that it is plausible to do so, that Farrell's state of nature is only something like Locke's because there is no reliance on a system of natural law, and accompanying natural rights, guaranteed by God. In such a case it seems to me that it would be much more honest for Farrell to say that his state of nature was "something like" Hobbes's. Differences might be that it involves a greater degree of social interaction, but this is a good deal less significant than the non-existence of what is possibly the most important feature of Locke's account.

Of course, the problem for Farrell in using Hobbes, is that the next claim, to the right of self-defence, would have a much more prudential, and a correspondingly decreased moral, ring to it. Indeed, once one claims a prudential grounding for the right of self-defence all sorts of things

---

[71]Quinn 1985, 337.

[72]Farrell 1985, 371.

might follow, Hobbes' "warre … of every man, against every man",[73] is not a consequence (as is often thought by students), of man's brutal nature, but of prudential *reason*. Reason tells each person that, in the absence of a common authority, it makes sense to pre-emptively strike at others.[74]

Farrell describes the right of self-defence in terms of a principle of distributive justice, some version of which appears in all his papers. The principle states:

> "When someone knowingly brings it about, through his own wrongful conduct, that someone else must choose either to harm him or to be harmed herself, justice allows the latter to choose that the former shall be harmed, rather than that she shall be harmed, at least if the harm inflicted on the former is roughly proportional to the harm that would otherwise be inflicted on the latter."[75]

I have to say that I find it remarkable, that after four papers on the topic, in a paper entitled "The *Justification* of Deterrent Violence",[76] Farrell adds to this principle a note, stating that,

---

[73]Hobbes 1651, Chapter 13, 185.

[74]See Barry 1968 for the best account of this interpretation.

[75]Farrell 1990, 303. Farrell is following Phillip Montague 1983, 31-36. Cf. Scanlon 1988, 172: "Some have held that from the fact that a person is morally blameworthy it follows that it would be a good thing if he or she were to suffer some harm (or, at least, that this would be less bad than if some innocent person were to suffer the same harm)" Scanlon attributes this insight to Derek Parfit.

[76]Farrell 1990, emphasis added.

> "it will be obvious … that I do not intend anything I say in the present paper as a defense of [this principle]…. The point is simply to show how far we can go, in the justification of deterrent violence, if [this] principle is granted".[77]

I suppose this is, in a sense, a similarity between Locke and Farrell, the similarity being that if you start by assuming P, proving that P follows is made a great deal easier.

Farrell admits that Quinn does not appeal to such a principle but says that he assumes "Quinn would welcome the line of argument [and that] some such principle … is implicit in Quinn's remarks".[78] I have no idea whether the former is true, but the latter seems to me quite accurate; Quinn, like Farrell, has to get a moral right to self-defence from somewhere.

Perhaps I am being a little unfair, Quinn and Farrell between them do a great deal of interesting work in showing how an individual right to self-defence can, once granted, be used to explain certain aspects of punishment and Farrell does much to distinguish how this relationship changes between special and general deterrence, but the fundamental point remains; the answer to the question of whether one would be morally justified in constructing ARDs

---

[77]Farrell 1990, 303, 2n. Because of this, it is possible to read Farrell's argument from this principle rather than from any more fundamental right to self-defence, see, e.g., Tamburrini 1992, 71-2.

[78]Farrell 1989, 131.

depends upon the moral theory that one endorses, and it is not enough just to be given one principle, unexplained and out of any theoretical context. For instance, how might one address the sort of challenge that Duff's work exemplifies, that is, that any system that relies only on threatening the agent so as to alter her behaviour, fails to treat the agent with due respect as an autonomous member of the moral community? Farrell and Quinn, like Nino, are careful to include strong voluntaristic elements in their accounts, the offender must perform his actions knowingly, both in the sense of knowing what he is doing and knowing the possible consequences, and Farrell, especially, is well aware that simply announcing in advance that one is going to do a thing, is not sufficient to morally justify doing that thing,[79] yet when the time comes all he offers is the argument that "if we suppose … that threats of harm are likely to reduce such violations [of innocent persons' rights] and, moreover, are necessary if we are to reduce them to a tolerable level, it would seem that such threats are perfectly justifiable".[80]

The second question, as to whether punishment can be understood as analogous to a morally permissible system of m-punishments, or on the basis of some other account of the right to self-defence, cannot be answered, because the conditions under which the right to self-defence is morally justifiable are not known. It may be, and almost certainly

---

[79]See Farrell 1985, 379.

[80]Farrell 1985, 380.

is, true that under certain conditions one is morally permitted to harm another to defend oneself, but saying any more about that, and about what else may be understood on this model, requires an account of the conditions and the moral theory which informs them.

It may seem as if I have come a long way from consequentialism, and it, perhaps, is not too useful to try to make the self-defence thesis into some white knight riding to the rescue of consequentialist accounts of punishment. Yet, they do offer to take a right that many of us assume to have some grounding, and explain a deterrence based system of punishment. I have, as will be seen, a good deal of sympathy with this approach; as I said above, punishment theorising needs to start not with an offender but with the question of what justifies the system that includes such a category of persons.[81] I also believe that a good deal of the purpose of the penality system can only be understood as there to deter actual and potential offenders. Farrell and Quinn try to show that to do so does not involve the problems associated with consequentialism highlighted at the beginning of this chapter, but their accounts are unpersuasive because they assume a non-consequentialist moral theory which they neither ground nor really examine. Of course, if we are obliged to treat people as "ends in themselves", or if we assume that each person is a rights bearing agent entitled to equal respect, then a consequentialist theory of

[81]See *supra*, Introduction.

punishment will have to be constrained or replaced by such considerations.

Below, I aim to show both that we can give an account of the grounding of the equal value of agents, and that this is compatible with a punishment system partially justified by its deterring actual and potential offenders. To do this, I must first turn my attention to other attempts at grounding the claims of agents, and to the rationale of coercion that can be located in such theories.

Chapter 4: Playing Fair and Playing Rough

21. Fair Play Theory

Despite the fact that political theory has been dominated in the last twenty years by the subject of distributive justice punishment theory has been slow in catching up, perhaps because the relationship between distributive and retributive justice is not at all a clear one.[1] Yet a number of writers have responded to Ted Honderich's claim that the completion of punishment theorising is dependent on finding "a clear principle of *distributive* justice, an answer to the question of how all the benefits and burdens in society are to be distributed."[2] In what follows below I shall begin by considering one justification for punishment based on the idea of reciprocity, or fair play. Fair play theory claims to give a justification for punishment which is independent of any substantive theory of justice or account of morality. I shall show, first, that there are insurmountable difficulties for fair play theory in attempting this, and second, that it in order to

---

[1]For a debate on the relationship of punishment and political theory see Philips 1986 and the reply by Davis 1989.

[2]Honderich 1984, 239. Honderich's appeal is, in his case, motivated by a particular view of the free will/determinism debate, a view which is wholeheartedly determinist. (See Honderich 1988, 1993) A number of writers have recently attempted to use the language and techniques of distributive justice theorising in discussions of just punishment, notably Michael Davis (1983, 1986a, 1986b); David Hoekema 1980; J. Narveson 1974; W. Sadurski 1989; Sterba 1977.

overcome these difficulties fair play theorists rely on an implicit (and usually contractarian) account of morality. Fair play theory is, in short, incomplete and the theoretical resources needed for its completion cannot be found within the theory itself. In the final part, I shall examine the contractarian account of justice as mutual advantage, illustrate its relations with the theory of fair play, and examine its account of punishment.

In its modern form, fair play theory developed from H. L. A. Hart's seminal paper "Are There Any Natural Rights?" in which he argued that the obligation to obey the law was derived from a "mutuality of restrictions"; that is,

> "When a number of persons conduct any joint enterprise according to rules and thus restrict their liberty, those who have submitted to these restrictions when required have a right to a similar submission from those who have benefited by their submission".[3]

The idea is that fair play theory - the mutuality of restrictions - should be a new and independent mechanism for creating rights and duties; "independent" in the sense that it "differs from other right-creating transactions (consent, promising)."[4] In fact, I shall argue below that punishment theorists have had to rely on ideas such as "consent" to bolster the claims of fair play theory. Even Hart, in the formulation given above, seems to indicate the

---

[3]Hart 1955, page references to reprint in Waldron 1984, 85. An earlier statement of much the same view can be found in Broad 1915-16; a later, more influential, statement in Rawls 1964.

[4]Hart 1955, 85.

incompleteness of the idea of reciprocity with his talk of "joint enterprises" and "rules". What is more, although Hart is concerned to distinguish obligations derived from fair play from those derived from consent, he nonetheless, in the sentence immediately before introducing the idea, states that he believes that "it is true of all special rights that they arise from previous voluntary actions".[5]

Hart's basic principle has been developed and debated at some length in the forty years that have followed its publication,[6] but as Richard Dagger has recently pointed out,[7] most of that debate has remained centred on the question of whether one has on obligation to obey the law; the application of the principle to the justification of punishment has had, by comparison, a relatively short popular life.[8]

The starting point for a fair play justification of punishment is with a conception of society as a co-operative endeavour. Co-operation brings advantages to the

---

[5]Hart 1955, 85. Whether he means that the agent needs to have voluntarily joined the "joint enterprise" or voluntarily accepted its benefits is unclear. On this point see Simmons 1979, 108; this criticism is pursued in Barry MS.

[6]See, broadly in support, Arneson 1982; Becker 1986; Dagger 1985, esp. 443-46; Fishkin 1992, §3.3; Gibbard 1991; Klosko 1992; Sadurski 1989; Sher 1987. Broadly against, Nozick 1974, 90-95; Simmons 1979, Chapter 5; Smith 1973. Other references appear below.

[7]Dagger 1993, 474.

[8]The two most influential versions applied to punishment were Morris 1968, and Murphy 1973. Other contributions have been made by Finnis 1972; von Hirsch 1976; Sadurski 1989; and Sher 1987, Chapter 5. Both von Hirsch and Murphy have since expressed dissatisfaction with the theory; von Hirsch 1985, 57-60, 1990, 264-5; Murphy 1985, 1990.

members who compose the society, however, these advantages can only be realised if the members of the society interact on the basis of shared rules that govern the co-operation. The basic idea, then, is that insofar as each enjoys the benefit of the co-operative enterprise, each is obliged to co-operate on the basis of the shared rules.[9] This of course means that each member is obliged to undertake activities which he finds burdensome - paying taxes for example - if such activities are commanded by the rules of the co-operative enterprise.

The law - which is to be thought of as the form given to the norms governing co-operation - provides enormous benefits to the participants, most notably in attempting to provide conditions of "peace, security and freedom".[10] It protects the individual from unwanted interference by others, and thus allows her to formulate and attempt to carry through her chosen desires or plan of life in conditions in which her legitimate expectations are likely to be fulfilled and in which she will be free from unwanted and unwarranted interference from others. The law secures these benefits for the individual, however, only by imposing on her a burden; the burden of self-restraint. That is, the individual must restrain herself should she

---

[9]At this stage I am trying to keep the characterisation of fair play theory to a minimum because, as I said above, I think there are various versions each relying on different background assumptions about the nature of justice. As it stands here (without an element of consent) this claim has been challenged by Nozick. (Nozick 1974, 90-95)

[10]Duff 1986, 206.

desire to interfere with others in ways which are illegal; she must respect the prohibitions of the law and moderate her behaviour accordingly.[11]

Characteristically the next move in fair play theory is to say that in a just society the distribution of the benefits of the law and the burdens of having to keep the law are distributed fairly; the law protects everyone and provides the benefits of peace, security and freedom for everyone. Likewise, the burden of having to obey the law falls equally on everyone; that is, the rule of law applies. There is, then, a pattern of distribution of benefits and burdens that is just, and when this pattern is disturbed it is a requirement of justice that it be restored. This, it is claimed, is the role of punishment: to restore the *status quo ante*.

If we are to find a moral justification for punishment in the idea of restoring the distribution of benefits and burdens it must be the case that for punishment to be justified the pattern of distribution must be just in the first place. What a just distribution is, however, is not a question that fair play theory need address. Indeed if it is to retain its character as a justification of punishment based on Hart's simple idea it must not do so. To see this it is merely necessary to examine the claim that is needed by fair play theory: "in a just society the distribution of benefits and burdens is just"; such a claim

---

[11]See Duff 1986, 206; 1993, xii; Sadurski 1989, 355-6.

tells us nothing about the nature of that distribution, it is, in fact, tautological. All that fair play theory needs is to posit a just starting point and advantage from participation for all members of the society. This is important because fair play theory essentially holds that given a just society it can justify punishment as a second, independent, stage of moral theorising. That is, given just and non-coercive norms, punishment is justified through the ideas of fair play. The derivation of the norms (and their legal analogue, laws), and the question of whether they are just is a different matter from the derivation of just punishment. Below, I shall argue that this position is both untenable and uninteresting (and that it undermines any attempt to found justice on ideas of reciprocity), what is of interest and what must be considered if we are to justify punishment[12] is the derivation of just *and* necessarily coercive laws.

Before addressing this concern - of whether the fair play project is a sensible one - I want to consider whether fair play theory works on its own terms. Does it, in fact, justify coercive punishment? To do this we need to examine the argument in more detail.

The argument is that punishment is a requirement of justice, for (given a just starting point), it is just that each carry out his share of the burdens, and should someone

---

[12]Without reference to God or some other variant of moral realism and whilst avoiding the pitfalls of utilitarian justifications.

fail to do so, a situation of injustice results that has to be rectified. The criminal disturbs the just distribution by refusing to mediate the pursuit of his self interest by the norms of the society; that is, the criminal does not bear his burden of self restraint.[13] The benefit to the criminal is *not* the property he steals or the pleasure he gets from raping someone, the benefit is, in Duff's word, "*intrinsic*",[14] it is the evasion of self restraint. The justification for punishment is that it imposes on the criminal an extra burden, in some way equivalent to the burden which she evaded, and thus it restores the balance of benefits and burdens in the society. Punishment is this restoring of the balance, just as the crime was the disruption of that balance. On this account, then, punishment is a demand of justice: it falls on those who fail to shoulder their burdens or those who claim a greater share of benefits than that to which they are entitled. The correct distribution, the *status quo ante*, is restored through punishment. As Richard Dagger has put it in a recent article:

> "Criminals act unfairly when they take advantage of the opportunities the legal order affords them without contributing to the preservation of that order. In doing

---

[13]In Murphy's rather hard hitting prose, "The criminal is a *parasite* or *freerider* on a mutually beneficial scheme of social cooperation". (Murphy 1985, 7)

[14]Duff 1986, 207. Of course, the criminal does benefit from the gain in property or from the pleasure of raping someone. What is being claimed here is that the benefit which the fair play theorist - as against, for example, a restorationist - is concerned with is that of the offender not mediating his behaviour in accordance with the norms the obeying of which by others leads to benefits for the offender. This argument is set out in greater detail below.

> so, they upset the balance between benefits and burdens at the heart of the notion of justice. *Justice requires that this balance be restored*, and this can only be achieved through punishment or pardon."[15]

It is important to get clear that this is a retributive conception of punishment. Justice demands that the criminal be punished, that the just distribution of benefits and burdens be restored. This might have deterrent effects - presumably co-operators will know that the society operates a system of punishment and this will thus make it less plausible that they will find it in their interests to break the law - but the reason for punishing is not that it serves to increase or decrease the likelihood of disobedience, the reason for punishing is that it is *just* to do so.[16]

---

[15]Dagger 1993, 476, emphasis added. See also Morris 1968, 478.

[16]Andrew von Hirsch has argued that fair play theory generates a *prima facie* reason for punishment based on the desert of the offender. Against this *prima facie* reason he argues is a "countervailing moral obligation of not deliberately adding to the amount of human suffering". To get to justified punishment he thinks that one needs the assumption that punishment has deterrence effects, and thus in fact punishment reduces rather than increases the total of human suffering. Although these extra steps are interesting they do not, it seems to me, change the basic justification for punishment offered by von Hirsch as being based on desert generated by fair play. "Step 3", the deterrence argument, merely makes "Step 2" (subtract from rather than add to human suffering) irrelevant; "the prima-facie case for punishment … based on desert … stands again." (von Hirsch 1976, 54) This is the theory that he has since retracted, see *supra*, note 8. Dagger thinks that the general justifying aim of fair play theory is deterrence (see Dagger 1993), but this is because he fails to distinguish fair play theory from contractualist accounts of justice.

22. Two Objections

There are a number of problems with this account considered strictly from the perspective of whether it succeeds in justifying coercive punishment,[17] however, I want to consider only two.[18] The first is that fair play theory cannot provide a sufficient reason for coercive punishment. The second goes far beyond the problem of punishment and to the heart of theories based on reciprocity or fair play; it claims that the fair play thesis fundamentally misdescribes the nature of law and the character of moral actions.

The first problem is that it is unclear why coercive punishment flows from the requirement of fairness to erase the debt owed by the offender. Punishment is supposed to restore the balance of benefits and burdens, but it is difficult to see how it does this. In the ordinary language of morals "two wrongs do not make a right" and it is not clear that in this case punishment (a *prima facie* harm), can achieve the desired end. If the offender has claimed greater liberty than is compatible with the demands of the co-operative enterprise, how does removing his liberty restore the balance? The intuition being traded on here is clearly that the criminal has taken more liberty than that to which he is entitled, and in punishing him -

---

[17]That is, even assuming that fair play theory correctly characterises society and the nature of obligation (which, I shall argue below it doesn't), can it justify punishment?

[18]For extended discussions see Duff 1986, Chapter 8; Dolinko 1991; Fingarette 1977; Wasserstrom 1980. Replies to some of the criticisms can be found in Burgh 1982; Dagger 1993.

in taking away some of his liberty - the situation is restored. But the liberty the criminal has taken is the liberty of not regulating her behaviour in accordance with the law, in punishing her we do not make her sacrifice her self-interest to the interest of the co-operative venture, nor in punishing her is there any obvious repayment of liberty to the other members of the co-operative venture. If punishment were conceived on a model that attempts to capture this simple intuition it would surely have to take the form of the criminal contributing to the co-operative venture whilst being deprived of the benefits. It might be possible to conceive of punishment as contributing in such a way by thinking of it as a contribution by the offender to the system of deterrence, however, this seems a little odd, and in any event the criminal certainly benefits from the deterrence effects of the system of punishment.

If it is not obvious quite how punishment restores the *status quo ante* neither is it obvious that it is the only mechanism for approaching such a goal. One alternative, admitted by fair play theorists such as Herbert Morris, may be pardoning:

> "Forgiveness - with its legal analogue of a pardon - while not the righting of an unfair distribution by making one pay his debt is, nevertheless, a restoring of the equilibrium by forgiving the debt. Forgiveness may be viewed, at least in some types of case, as a gift after the fact, erasing a debt, which had the gift been given before the fact, would not have created a debt.

> But the practice of pardoning has to proceed sensitively, for it may endanger, in a way the practice of justice does not, the maintenance of an equilibrium of benefits and burdens. If all are indiscriminately pardoned less incentive is provided individuals to restrain their inclinations, thus increasing the incidence of persons taking what they do not deserve."[19]

This is not pardoning as traditionally understood. Traditionally pardoning is not an alternative to punishment, but is something done once it has been established that the offender is deserving of punishment; it is, in the strict sense, an act of grace.[20] Morris, however, is accepting that pardoning restores the balance of benefits and burdens rather than is (as it would be on the traditional picture) an acceptance of the disequilibrium as the new *status quo*. But on Morris's account, pardoning is an alternative to punishment (both being justified only by appeal to their efficacy at restoring the *status quo ante*) and this makes Morris's distinction between the righting and restoring of an unfair distribution (and, by extension the difference between "the

---

[19]Morris 1968, in Gross and von Hirsch 1981, 95. Quoted in Duff 1986, 216.

[20]It is no doubt true that sometimes there are calls for pardoning when it is thought that there are mitigating circumstances, such as that the offender is young or naïve. In such cases the offender is said to "deserve pardoning". Such calls, it seems to me, are in fact motivated by something like the thought that the criminal justice system has failed to take such things into account (perhaps because of a mandatory sentencing policy) but rather than amend the system to meet a particular need it is easier to circumvent it by means of an extrinsic intervention such as a pardon. In such cases, the offender may be legally pardoned but in fact the offender deserved less punishment, and pardoning is being used to achieve this, it is still not the case that the offender deserved pardoning.

practice of pardoning" and "the practice of justice") difficult to understand. Morris's argument for coercive punishment as the unique solution to the disequilibrium caused by the criminal act is, in fact, explicitly consequentialist and thus is unacceptable if fair play theory is to retain its retributive character. For Morris we must punish because the indiscriminate use of pardon would not deter individuals from "restrain[ing] their inclinations…"[21] In short, the theory provides a necessary but not sufficient reason for coercive punishment.[22]

The second objection - that fair play theory characterises law and moral actions in entirely the wrong way - is more difficult to get to the heart of, and will develop into the criticism that fair play theory is radically incomplete. To start, consider three criticisms of fair play theory. First, given by Fingarette, that fair play theory fails to capture the nature of legal prohibitions:

> "On [Morris's] view … I would seem to have two equally legitimate options - paying my debts earlier in cash, or paying later in punishment. But surely that is not the intent of the law *prohibiting* stealing. The intent is

---

[21]Morris, quoted above. See Duff 1986, 216.

[22]Dagger appears to accept this: "Justice requires that this balance be restored, and this can only be achieved through punishment *or pardon*" (Dagger 1993, 476 emphasis added). Dagger has open to him the extra claim - which is the subject of part III of his paper - that punishment is justified because pardon would not prove sufficient a guarantee to solve the assurance problem, although nowhere does he make this explicit. In fact, pardon doesn't reappear in Dagger's article, although he spends some time considering whether punishment is the best response to crime.

> precisely to *deny* us a legitimate alternative to paying the storekeeper for what we take…"[23]

Second, that fair play gives rise to a theory of compensation or restitution but not punishment. Again Fingarette provides such a criticism when he says, going on directly from the quote above:

> "And even if I restore the balance by returning the stolen goods, and by paying back any incidental losses incurred by the storekeeper, it still remains intelligible and important … to ask whether I should *also* be punished."[24]

If it were the case that the balance could be restored solely by this kind of restitution, the critic claims, the theory would be severely undermined. Apart from the obvious difficulties with crimes such as rape, it would not even be satisfactory in cases such as car theft. If a criminal steals your car and is caught, it would not be a punishment for him to be forced to return your car, paying you for any wear and tear, petrol used and inconvenience caused; that is not punishing the criminal it is making the victim into an (unwilling) car hire firm.[25]

---

[23]See also Fingarette 1977, 502.

[24]Fingarette 1977, 502.

[25]This example, which convinced me as an undergraduate of the unsatisfactory nature of justifications of punishment based solely on compensation, I owe to a political philosophy lecture given by John Charvet.

Third, in what sense does the rapist owe a debt to the law abiding members of the society? Duff has claimed that the fair play theorist has it that the rapist has done something that we (law abiding members of society), all would want to do, but which we don't, because we mediate our acts in accordance with the demands of the norms of co-operation. Duff argues that it is simply not true. We - or most of us - do not wish to rape or murder; obeying those laws just isn't a burden:

> "talk of the criminal's unfair advantage implies that obedience to the law is a burden for us all: but is this true of such *mala in se*? Surely many of us do not find it a *burden* to obey the laws against murder and rape, or need to *restrain* ourselves from such crimes: how then does the murderer or rapist gain an unfair advantage over the rest of us, by evading a burden of self-restraint which we accept?"[26]

The response to all three of these objections (and it should be obvious that without a response the theory would be fatally undermined) can be encapsulated in the claim that all three rest on a fundamental misunderstanding of the ideas of benefits and burdens in fair play theory. Fair play theory understands the benefits to be intrinsic to not obeying the law - what one gains is not having to obey the law. The burdens, likewise, are to be understood as those of having to mediate the pursuit of one's self

---

[26]Duff 1986, 213. A similar criticism appears in Wasserstrom 1980, 143-146.

interest by the demands of the norms governing the co-operative enterprise.

Thus, the fair play theorist can respond to the first criticism that it is not that the criminal may either pay for his goods now in cash or later in punishment (if he steals the goods), because the debt he owes if he steals the goods is not for the goods. Likewise, to address the second, the benefit (to be considered in justifying punishment), is not the illicit gains of material possessions, or the pleasures of rape; it is the benefit of not having to obey the law, that is, of not regulating one's actions by the norms of co-operation established in law.[27] Again, the criminal would not "restore the balance by returning the stolen goods" because those are not the goods that punishment addresses.[28] Finally, Duff's argument that most of us do not find the prohibition on rape burdensome can likewise be rejected; it is not the case that what the criminal has done that the law abiding members would want to do is rape or murder, but that they have acted as free-riders, and this is something that law abiding members would also want to do, but don't because they recognise their obligations. That is, what the rapist

---

[27]See Sadurski 1989, 360-2. Clearly from the criminal's point of view the benefits of crime are primarily the material or physical gains achieved in robbing or raping someone. A fair play system of punishment would presumably advocate the restoring of possessions and compensating of the victim (insofar as it is possible), this would be in addition to punishment, however, the punishment itself must be conceived in the terms given above if fair play theory is to avoid the objections being met here.

[28]See Dagger 1993, 478 & 484-87.

has done that the law abiding members of the society want to do, is not rape but free-ride.[29]

This defence seems to me the only one available to the fair play theorist to meet three otherwise devastating critiques. It can, however, only do a certain amount of work for the fair play theorist, because although it is clear that the balance of benefits and burdens could not be restored by simply restoring the distributive pattern of *goods* that held before the crime, it is not clear that coercive punishment can magically restore the distribution of benefits and burdens either, and it seems even less likely that punishment will be the only way to do so.

Far more important, however, is the fact that the characterisation of benefits and burdens that is at the heart of this defence is such that it reinforces the view that fair play theory fails to capture what it is that is wrong in many criminal activities. It may be true that some crimes can be understood as taking unfair advantage of others, (*mala prohibita*), but others are surely wrong in some more fundamental sense. To return to the examples of rape and murder, these acts are wrong in some other way than because they violate a principle of fair play.[30] It is not the case that what is wrong with rape is that it is

---

[29]This defence is given by Dagger 1993, 479-480; Sadurski 1989 offers something similar caged in the language of choice options, 357-360.

[30]The precise nature of such wrongness depends, of course, on the grounding that one gives to morality. I shall argue for a particular moral theory below in Chapter 6.

an example of someone not playing fair, what is wrong with it is that it *is* rape.[31] Fair play theory reduces all crimes to the non-mediation of one's self interest by the duty of fair play,[32] and this is problematic for a theory of punishment because, in the absence of any additional argument, it seems that if there is only one crime then there should be only one punishment.

Richard Dagger has attempted to give an answer to this problem by arguing that whilst, of course, crimes such as rape and murder are *mala in se* they are also crimes of unfairness.[33] He goes on to argue that it is for their character as crimes of unfairness that they must be punished. The additional wrongness of an act might be the subject of "punishment of the gods" or "revenge"[34] but the law has no business replicating the former and once society has come into being the right to private revenge is replaced by the rule of law. But does this mean, as it seems to, that all crimes are equivalent in terms of the

---

[31]See Duff 1986, 212.

[32]This reduction of all crimes to a single offence mirrors Klosko's argument that one has a duty to obey all laws (including those whose function is to provide discretionary goods) because (if the principle of fair play applies) one has an obligation to play one's part in the provision of the presumptive good of the rule of law. Klosko 1992, 101-3

[33]"All crimes, I have said, are in some sense crimes of unfairness. They may be *more than* crimes of unfairness, as rape, robbery and murder surely are, but they must be *at least* crimes of this sort." (Dagger 1993, 479)

[34]Dagger 1993, 479.

punishment they are to receive?[35] Dagger answers no, and his answer appears to unravel the case for fair play theory:

> "This is not to say that the murderer and the tax cheater should receive the same punishment *tout court*. For the murderer has committed two crimes, in a sense, but the tax cheater only one. The murderer has simultaneously committed a crime of unfairness (a *malum prohibitum*) *and* a crime against her particular victim (a *malum in se*). For these two offenses, as it were, she must suffer two punishments. The first serves to discharge her debt to society by restoring the balance of benefits and burdens under the rule of law. The second punishment must be justified and established on other grounds."[36]

This passage seems to deprive fair play theory of much of its force as a justification of the practice of punishment in any form with which we are familiar because it seems a "considered conviction" (to employ a phrase from Rawls[37]), that different crimes deserve or merit (for whatever reason), different punishments, (and that more morally serious crimes merit more severe punishments). Further, such an account separates the offence (or, at least one

---

[35]Or, perhaps even more damaging, that crimes are to be punished in proportion to how great the burden is on the average person not to commit that act. This problem is discussed in Burgh 1982, 209. The reason such an outcome would be unsatisfactory is that it seems likely that most people find it harder not to fiddle their tax returns than they do not to commit murder; tax fraud, on this scheme, would then be more serious and punished more seriously, than murder.

[36]Dagger 1993, 484.

[37]Rawls 1971, 19.

description of the offence) and the moral character of the act.[38]

Even ignoring questions about whether fair play theory can account for our obligation to obey the law, then, it cannot convert its account of obligation into an account of justified coercive punishment. The problems are twofold; on its own terms, it cannot explain why the debt owed to society needs to be repaid in the form of coercive punishment. Viewed from the outside, it characterises the debt owed in a manner that is unsatisfactory. This second problem is a variation on a wider theme; as I noted at the beginning of my remarks, fair play theory is silent on the content of justice, it begins "in a just society…". The problem with this is that it separates the account of justice from the account of punishment. Such a theory is formal; any and all violations of the positive law are conceived of as benefits and all actions in accordance with the positive law as part of a scheme of burdens. The content of the positive law is open.[39] This is unsatisfactory both because it has not been shown that fair play theory can account for punishment, and because it separates the account of moral wrongness from the account of punishment.[40]

---

[38]See Duff 1986, 212.

[39]This mirrors the objection given by Brian Barry (amongst others), to theories of justice based on the idea of reciprocity, such as that offered by Gibbard. See Barry 1995a, Chapter 2; Gibbard 1990.

[40]See Sadurski 1989, 368.

Fair play theory must, then, provide something more substantive in its characterisation of "benefits and burdens" and we have seen that it cannot do this without appealing to arguments which go well beyond its initial claims and which, I believe, threaten to negate its own role. To return to the question I left hanging some time ago: most fair play theorists begin much as I did with the statement "assuming that the benefits and burdens are distributed in a society in a just manner…", they then go on to consider punishment as a response to a disruption in the pattern of distribution, and then perhaps say something at the end about how this makes just punishment only possible in a just society.[41] What has become apparent is that this is not sufficient, because fair play theory cannot even complete the intermediate stage of showing that punishment is necessarily to be understood in this way. It needs completing, and in what follows I aim to show that the theory that informs the first claim (as to the justice or otherwise of the distribution of benefits and burdens) generates the understanding of moral wrong and punishment; fair play theory, if it has any function at all, has a mere fragmentary role at a much later theoretical stage than has been recognised by its supporters.

---

[41]See, e.g., Murphy 1973.

23. Contractualism and Consent

The contention of this section is that fair play theorists have relied on an implicit contractualist account of moral wrong. Indeed, often this commitment is made explicit.[42] The advantages of contractualism for the fair play theorist are numerous: First, contractualism offers an account of morality that is different from the traditional options of intuitionism, Kantian deontology or some version of consequentialism. Second, contractualism is often thought to involve consent, and if consent can be worked into the theory then the worry of people such as Robert Nozick that the foundational claim of fair play theory[43] is itself unsustainable, can be met. One is obliged to contribute in accordance with the demands of the joint enterprise because one has agreed so to do. Third, a further advantage of invoking consent is that it bolsters the claim that punishment is not the *prima facie* harm, or invasion of the offender's autonomy, that it is thought to be. This is because the offender can be said to have consented to his punishment. In considering contractualist accounts of morality I want to address two questions: Is it true that contractualism embodies consent in such a way as to allow the punishing authorities to claim that the offender has

---

[42]See Murphy 1973; Dagger 1993 represents his position as a summary of fair play theory, but in fact relies on two not terribly well integrated arguments, a contractualist account of the general justifying aim of punishment and a fair play account of the distribution of punishment.

[43]That is, that advantageous membership of a joint enterprise obliges one to do one's bit.

consented to his own punishment? And, second, what type of punishing system would emerge from a contractualist account of morality?

One of the most important statements of fair play theory - Jeffrie Murphy's 1973 article "Marxism and Retribution"[44] - is explicit in claiming precisely these advantages. The fundamental problem, Murphy argues, is to reconcile individual autonomy - individual rights - and legitimate punishment: "Even if punishment has wonderful social consequences, what gives anyone the right to inflict it on me?"[45] Murphy's answer to this question is to invoke a mixture of contractarianism, consent and fair play theory. His first move, and the one that is of concern here, is to say that "one fairly typical way in which others acquire rights over us is by our own consent." In what follows I want to examine the nature of consent in Murphy's article before turning to different contractarian accounts of justice and analysing whether consent has a role in any of them. I do not follow Murphy's account of the contractarian theories of Kant and Rawls because I believe he is fundamentally mistaken in his account of both of them.[46] In short, what I want to consider first is the question of whether it follows from a contractarian account of morality that an offender can be said to have "willed

---

[44]Murphy 1973.

[45]Murphy 1973, 223.

[46]For Kant on willing one's own punishment see §6; for Rawls on justice and punishment §27; Appendix A.

his own punishment". Murphy begins with the following example:

> "If a neighbor locks up my liquor cabinet to protect me against my tendencies to drink too heavily, I might well regard this as a presumptuous interference with my own freedom, no matter how good the result intended or accomplished. He had no right to do it and indeed violated my rights in doing it. If, on the other hand, I had asked him to do this or had given my free consent to his suggestion that he do it, the same sort of objection on my part would be quite out of order…even if, at the time of his doing it, I did not desire or want the action to be performed."[47]

Murphy believes that it "is obvious [how] … this applies to our problem",[48] but as it stands this example is too open to draw any conclusions. Let us say that on Monday I ask my neighbour to lock my drinks cabinet on Tuesday. Between Monday and Tuesday I do not get drunk, nor in any other way impair my mental or physical faculties. On Tuesday my neighbour comes around and I tell him that I have changed my mind and that I no longer want him to lock the cabinet. If he does so now, the intuition on which Murphy is trading seems undermined; one is entitled to change one's mind, and there is no reason why precedence should be given to the desires an individual expressed at one time over the desires the individual expresses at any relevant later time. A second story could have it, however, that on

---

[47]Murphy 1973, 223. Cf. Duff's example of Jane. (Duff 1986, p220)

[48]Murphy 1973, 224.

Tuesday I am going to an important evening meeting about which I am very nervous. I know that when I am nervous I tend to drink to excess to fortify myself with "Dutch courage" but that this is unlikely to impress the business partners with whom I am meeting. I therefore ask my neighbour to come around on Tuesday and lock up the cabinet because I am aware that if he does not I will get drunk. My neighbour comes around and I tell him that I have changed my mind. I have, however, forewarned my neighbour that I would do this and told him to ignore my expressed wishes and go ahead and lock the cabinet.

In some sense I have given my consent to my neighbour's actions - but to what have I consented? I have surely agreed to his subverting my autonomy; that is, given that I am self-reflective being capable of understanding and acting upon reasons, but that I know that I am likely to suffer weakness of will in this particular context, I abrogate responsibility for my actions to my neighbour - I ask him to be responsible for me. If I wished to retain my autonomy and yet seek help, I could ask my neighbour to come around on Tuesday and reason with me, pointing out all the arguments I have given him for why I should not drink, why the meeting is important to me, etc. If I simply tell him to come around and lock the cabinet irrespective of whether I concur with these arguments or not, then I consent to the sacrifice of my autonomy because of a concern for my business future, that may well be sensible but it is still a sacrifice of autonomy.

There is one further development of the example that may aid us in getting to a position in which acting against the express wishes of an agent is compatible with respecting his autonomy. That is by removing the stipulation that the agent is still capable of self-critical reflection and rational action. It is possible to amend the given example to illustrate this, by say assuming that "I" am a chronic alcoholic. However, an intuitively stronger example is provided by Murphy, and so as to take on this argument at its best let us turn to this case - that of the psychotic depressive.[49] Let us say that I am a friend of a psychotic depressive, call him Richard, who, when not in the throes of depression shows no suicidal tendencies. He otherwise lives a normal, full human life with a job he values and a wife and children for whom he cares very deeply.[50] When he is in the depths of his depression, Richard is incapable of rational self-reflection and rational action and he is prone to attempt suicide. He asks me forcibly to stop him from committing suicide when he is depressed and to accompany him to a hospital. Now, it is not clear to me that in so doing I would be respecting his rational will;[51] however, it is the case that in interfering with Richard's

---

[49]Murphy 1973, 230.

[50]I am not adding these details simply to make the example more interesting, rather it is necessary that one has at least *prima facie* evidence that Richard does not, when capable of rational action, desire to commit suicide. I take it that there is nothing that makes suicide necessarily irrational. (For a fascinating, if idiosyncratic study of suicide see the case of Ellen West. Binswanger 1958; M. Foucault 1976, 54-55; 1984-5, 62.)

[51]This is a claim made by Duff. (Duff 1986, 221)

suicide attempt I am not acting so as to remove his autonomy. Given that at that time Richard is incapable of autonomous action, autonomy simply does not arise as an issue; it is, rather, analogous to stopping a goldfish from eating so much food that it kills itself.

Having established three models of consent I want to turn to contractarian accounts of morality. What Murphy argues is that if we understand moral injunctions and the nature of society in contractarian terms then the criminal when punished will retain his autonomy because he will have willed his own punishment. Murphy believes that this combines with fair play theory because the contractarian story he tells is one based on reciprocity, but that does not have to be the case. Contractarian accounts of morality, however, do aim to tell us the nature of wrong acts. Whether they need a fair play account of punishment in addition to this account of wrongness is something that will be considered below. I now want to turn to contractarian theories and ask what kind of consent they embody, specifically what kind of consent can be imputed to an agent who appears to do anything but consent to his own punishment; in fact, who does everything he can to avoid capture and punishment. In addition, I want to consider the question of what type of punishing systems would emerge from differently constituted contractarian accounts. First, then, it is necessary to describe the purpose and main features of contemporary contractarian accounts of justice.

24. Contractualism Considered

Contractarian accounts of justice have dominated the landscape of political theory since the 1971 publication of Rawls's A Theory of Justice. Contemporary contractarianism can be distinguished from its more traditional Enlightenment versions (of Hobbes, Locke and Rousseau) because the latter were embedded in non-contractarian moral theories; they explained how political authority could be made legitimate - why we should obey the sovereign - in terms of promise keeping. The terms of the contract and the moral force given to it, however, both stemmed from background theories of natural law. The traditional contract theorists, in short, claimed that one ought to obey government because one had promised so to do - one had given one's word - however, as Hume put it, such theorists 'find [themselves] embarrassed when it is asked, *Why we are bound to keep our word?*'"[52] Contemporary contractarianism asks a different question and does not rely on the analogy with promise keeping. The question upon which it has focused has not been that of political obligation but of justice. In short, it addresses the question "What is justice?" Beyond this, it is difficult to say anything of contemporary contractarianism as a whole because of the numerous varieties that have been suggested over the past twenty years. What the approach embodies, however, is a

---

[52]Hume "Of the Original Contract", quoted in Kymlicka 1991, 188 citing, incorrectly, Barker 1960, 229. The correct page number is 161.

commitment to the idea that moral realism in any of its traditional forms is unsustainable and that we are to understand morality as the norms which govern the co-operative practice of some group (from small societies to the human species depending upon the theorist), and in accordance with which individuals should regulate their actions. Its starting point is not with an historical account of the genesis of society but with the question of how we are to understand and (de)legitimate the norms which currently govern society. It is thus a reflective practice, requiring that the theorist distance himself from the practices and norms current in his society and question how they are to be understood.

Such a characterisation of contemporary contractarianism is what leads thinkers such as Murphy and Sadurski to think that punishment must be understood on some fair play model. If society is a co-operative endeavour and morality the voluntarily agreed rules that govern the co-operative endeavour, they ask, how else is punishment to be understood except as restoring the balance of benefits and burdens which are distributed in accordance with the agreed rules? Further, if the agreed rules are to be understood as voluntarily accepted by the members of the society, how can punishment be contrary to the offender's will? The first of these questions has been partially addressed above, and in what follows I shall examine whether it is necessarily true that contemporary contractarian thought yields fair play type justifications of punishment. The

second question immediately brings back into focus the role of consent in contemporary contractarianism, or more plausibly, it questions the motivation contemporary contractarianism gives *to the agent* for theorising his relations with others in this manner. To consider this a better, more detailed, account of contemporary contractarianism is needed.

Contemporary contractarianism is sometimes referred to as a species of constructivism, perhaps because John Rawls adopted that term when describing his own approach.[53] The clearest statement of what is meant by constructivism can be found in Brian Barry's Theories of Justice, and because of its clarity and importance it is worth quoting in full. Barry sets two conditions in defining constructivism:

> "First … there must be a theory to the effect that what comes out of a certain kind of situation is to count as just. "What comes out" might be a principle, a rule, or a particular outcome. Justice can be predicated of any of these, and the point is that we can derive its justice from its having emerged from the situation. A "situation" is specified by a description of the actors in it (including their knowledge and objectives) and the norms governing their pursuit of their objectives: what moves are to be legitimate. And the "emergence" is to be a particular kind of emergence, namely the result of the actors in the situation pursuing their given objectives within the given constraints.

---

[53]See Rawls 1980; 1993a, III:1 89-99, 110-116, III:3, III:5.

> That is a necessary condition of a constructivist conception of justice but not a sufficient one. The second requirement is that the constructing is to be done by a theorist and not by the people in the situation themselves.[54]

I shall argue in Chapter 5 that impartialist theorists build in to the situation sufficiently substantive moral claims to make the agreement component of constructivism redundant, and I shall call such a position, *faux* constructivism. In other words, if the theorist constructs the situation in such a way as to determine the outcome, then the construction is left to be nothing other than a *heuristic* device. The importance of that here (apart from introducing the idea of constructivism), is that in order to bolster the claim that the offender has willed his own punishment, the Murphyan contractualist theorist needs to show that the offender is in some way committed to the contractualist scheme that justifies his punishment. This brings up the important question of the motivation posited by contractualists, and it also reveals a problem *faux* constructivism poses for Murphy's argument.

If, as I have claimed, *faux* constructivism embodies a commitment to a particular version of morality, then the offender might be thought to embody a different account, and we are returned to the seemingly arbitrary imposition of punishment, or, if justified in some other manner,

---

[54]Barry 1989, 266.

certainly not a punishment imposed with the offender's consent. This is why Murphy attempts to portray Rawls (and Kant in his political theory), as deriving his conclusions from rationality alone.[55] Murphy converts the Rawlsian contract into a meeting of noumenal selves in "the kingdom of ends":

> "On this theory, [the social contract theory of "Kant and Rawls"] 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. This avoids arbitrariness because, according to Kant and Rawls at any rate, the question of whether such a rule would be picked in such a position is objectively determinable given certain (in their view) noncontroversial assumptions about human nature and rational calculation. 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 might break the other generally beneficial social rules that had been adopted."[56]

---

[55]This is why I said earlier that I would not be particularly concerned with Murphy's own characterisation of Rawls on the grounds that it was wrong. The portrayal of Rawls as attempting to ground his principles of justice on rationality alone was one which had some currency but which is based, as Brian Barry has put it, "on a reading of Rawls that is pure fiction". (Barry 1995a, Chapter 1).

[56]Murphy 1973, 230.

Two issues, then, still concern us: first, whether contractarian accounts of justice do, in fact, embody Murphyan type consent, and second, on a more coherent account of the contractualist scheme, the question of what kind of moral injunctions and punishing practices would emerge from differing characterisations of the initial situation.

Murphy is clearly not appealing to the type of explicit consent found in the liquor cabinet example, rather what he is saying is that his version of contractarianism includes the claim that all rational beings (including the offender) would will the system of punishment (that was agreed, ignoring for the moment what its particular form would be) and thus the offender can be said to have willed his own punishment. Let us for the moment grant Murphy's characterisation of the contractarian scheme, is it true that the offender has willed the system of punishment? Clearly not. Murphy's strongest claim cannot be that the offender has willed his own punishment but that punishment is not an infringement upon the offender's autonomy because autonomy is to be understood as living in accordance with one's rational self, and one's rational self would commit one to the principles agreed in the initial situation, including a principle of punishment. To substantiate such a claim, however, it needs to be shown that the initial situation and the moves permitted within it are generated from rationality alone. This is not plausible, but to see why we need to embark on a closer examination of

substantive contractarian theories of justice and in so doing approach, as well, the second question: what type of punishing system would emerge from a contractarian starting point?

The answer to this second question, unfortunately but unsurprisingly, is going to depend upon the theorist who performs the task of constructing the initial situation, for just as there is an initial situation for every theory of distributive justice so there is one for every theory of punishment. Thus, Jan Narveson has argued that if one removes (from a Rawlsian starting point) the Rawlsian stipulation that the people in the original position suffer from extreme risk aversion then the social contract will yield utilitarian principles for punishment.[57] On the other side, so to speak, James Sterba has argued that from the contractual situation one can generate "a morally adequate form of retributivism",[58] the particular form of which has been challenged (in my view successfully) by T. M. Reed.[59]

---

[57]Narveson 1974. This argument is employed by Harsanyi to show that average utility, rather than the difference principle would be chosen as the principle of distributive justice, once risk aversion has been removed. (Harsanyi 1982)

[58]Sterba 1977, 352.

[59]Reed 1978. For a discussion of both the problem of willing one's own punishment and of deriving any determinate theory of punishment from a Rawlsian scenario see Duff 1986, 217-228. On the general problem of the derivation of profoundly different results from small changes in the construction of the original position see Fishkin 1993, 79-81. What Fishkin overlooks is the other side of the reflective equilibrium equation - our considered convictions - which would eliminate most of the alternatives, a possibility which as I point out here is not available to punishment theorists. (See M. Matravers 1995) Another notable attempt to apply Rawlsian theorising to punishment can be found in Hoekema 1980.

At this stage, however, my primary concern is with the nature of contractarianism and especially with the motivations imputed to the parties in the initial situation and thus it is not necessary to examine each and every version of the contractualist argument. Following Barry and Kymlicka[60] I shall rather look at two general forms of contractarian theorising, one embodying justice as mutual advantage in this Chapter and, in the next, two versions of justice as impartiality.

25. Justice as Mutual Advantage

Contract theories embodying justice as mutual advantage assume that society is to be conceived as a co-operative endeavour; each individual theorises his relations to the norms of the society from the point of view of furthering his own interest and nothing else. Each realises that the unconstrained, independent, pursuit by all of their own goods is likely to result in conflict and that, therefore, agreeing to a common set of constraints is likely to result in greater success for all. The motivation for participation is that it is in one's interest to do so, and that is all. In Barryan terms, the actors in "the situation" are concrete individuals, aware of the details of their lives - such as their talents, abilities,

---

[60]Barry 1989; Kymlicka 1991.

conceptions of the good etc. - and motivated by furthering their interests.[61]

In such a situation what "emerges" is likely to be in conflict with our moral intuitions because the strong are in a position to exploit their strength to the point where those with whom they want to co-operate are only just better off than they would be outside of co-operation but far less well off than the strong.[62] Those in a position to contribute little or nothing, with whom the strong will have no reason to desire co-operation, will be excluded altogether; "beyond the pale of morality", in the words of David Gauthier, a recent exponent of justice as mutual advantage.[63] Any set of principles which mirrored relative bargaining strength, then, would hardly be recognisable as a theory of justice; perhaps, as Kymlicka suggests, it is not really a theory of our moral experiences and practices but an entirely alternative morality.[64] Of course, amongst groups with relatively equal bargaining power the outcome

---

[61]The *locus classicus* is, of course, Thomas Hobbes's Leviathan (Hobbes 1651). For extended discussions of the theory of justice as mutual advantage see Barry 1995a, Chapter 2; Hampton 1986; Kymlicka 1991, 189-91.

[62]I am here allowing that justice as mutual advantage would settle on a particular determinate answer to the distribution of goods (liberties etc.); that is, one which reflects relative bargaining power. As Brian Barry has pointed out, however, there is no reason to think that in a world in which information is anything but perfect this will be the case. Barry 1995a, Chapter 2.

[63]For an indication of just how such a theory might effect the weak see David Gauthier's discussion of the "congenitally handicapped and defective." (Gauthier 1986, 286, 18 note 30) For a discussion of Gauthier on this point see Barry 1995a, Chapter 2; Buchanan 1990, 230-2.

[64]Kymlicka 1991, 190-1.

would better approximate to our convictions about justice, but this is similar to Socrates' claim that thieves must recognise justice amongst themselves, it does nothing to address the fact that the relations between the thieves and the powerless traveller are not likely to be just. I can think of no more eloquent statement of the problem than that made by Ouray, the chief of the Ute indigenous American tribe which is cited by Brian Barry:

> "The agreement an Indian makes to a United States treaty … is like the agreement a buffalo makes with his hunters when pierced with arrows. All he can do is lie down and give in."[65]

But, if justice as mutual advantage doesn't match our considered convictions about justice, nor is it clear that it can succeed on its own terms; that is, as a method of ensuring stability. In the first instance, any agreement arising from mutual advantage is inherently unstable given that relative bargaining powers are liable to change. The moment one group increased its relative power it would be in its interests to re-negotiate the terms of agreement.[66] In addition, there is in such a theory a free-rider problem: given that each member co-operates on the basis of

---

[65]Barry 1995a, Chapter 2, quoting from Dee Brown 1971, 368.

[66]This, of course, is reflected in the experience of indigenous Americans; as soon as the settling whites found themselves in a position to further increase their territory at the expense of the American "Indians" they did so, first tearing up the relevant treaty - "treaties of perpetual peace" - then renegotiating another agreement after securing their territorial goals. See Barry 1995a, Chapter 2. David Hume, it seems, admitted this element of justice as mutual advantage. (See Barry 1989, 162 for a discussion of this point).

the 'moral' norms because (and only because) it furthers his interest to do so, he will be motivated when he believes that he can get away with it to break the terms of co-operation and free-ride.[67] This, in addition, gives rise to an assurance problem; it is only in the interests of each member to regulate his actions in accordance with the norms that emerge from the initial situation if every other contracting party does so, and if he is sure that they will do so.

This, of course, is why the co-operative practice to which each agrees much contain a coercive element. The coercive element is necessary to deter the individual who, in this instance, would be better off by breaking the norms of co-operation, from so doing. If it does this effectively, it provides the condition of "sufficient security",[68] (it answers the assurance problem), by convincing each party that the others will not free-ride. But what are the chances of a coercive element succeeding? Justice as mutual advantage yields prohibitions that can in a sense be best characterised as *legal* - they are prohibitions against which one's self interest strains - it does not yield

---

[67]As Brian Barry puts it: "Suppose you accept that some set of rules would advance everybody's conception of the good (including yours) if generally complied with, in comparison with a situation in which each person pursued his conception of the good independently. The question is: why does that give you a reason for complying with the rules on an occasion when you believe that you could advance your conception of the good more effectively by breaking the rules?" (Barry MS, 10)

[68]Hobbes 1651, Part I, Chapter 15, 215. It should be noted that I am assuming that a system of coercion (and, where appropriate, punishment), does have deterrent effects. I defend this in §39, see Chapter 7, note 17.

authoritatively binding *moral* norms through which one comes to understand one's self interest.[69] This is unsatisfactory because, in addition to yielding counter-intuitive results, it does not even yield the stability which is the point of the contract: whereas some simple societies may have succeeded in remaining relatively stable without separating moral from legal injunctions, no society could (or has) relied exclusively on legal sanctions. The reason is clear: coercion can be effective only if the majority of the people comply with the law because they believe it to be just, or legitimate, or just because it is the law. If the majority decided whether to obey the law on the basis of furthering their interests, albeit that the cost of being caught would be factored into their thinking, coercion would have little hope of succeeding in ensuring stability. Even if the sanctions for crimes with low rates of conviction (such as those against property) were extremely stringent it would be unlikely that the society could hold together. The problem is that what is missing is a recognisably moral (and not necessarily legal) element; the fabric of any advanced society is largely maintained by non-legal, moral, approbation and opprobrium.[70] This is just to restate what I said above: justice as mutual advantage fails to yield authoritatively binding moral norms.

---

[69]This understanding of morality is given fuller consideration in §32.

[70]Hobbes recognised this, men, he says "need to be diligently, and truly taught; because [civil society] cannot be maintained by … terrour of legal punishment". (Hobbes 1651, Part II, Chapter 30, 337).

On such a theory, then, the coercive element - which for the moment let us assume takes the form of a system of punishment - has as its justification the deterring of potential offenders and, if successful, the solution to the assurance game. The parties to the initial situation believe that it is generally in their interests to co-operate in accordance with the laws if every co-operating party does so; if punishment is to guarantee compliance what would the system of punishment have to look like? Given the failure of justice as mutual advantage I shall not examine this in too much detail here, confining myself to a few general remarks[71].

Punishment, then, is to be conceived as nothing more than an extra cost, a payment that one has to make if one commits an act contrary to the terms on which co-operation is to be undertaken.[72] Given this, the system of sanctions would have to be extremely harsh; each individual is motivated to break the law when it is to his advantage to do so and the *only* countervailing motivation is provided by the system of legal sanctions. Second - given the plausible assumption that people are more likely to be motivated to break tax laws, commit motoring offences etc., than commit murder or serious bodily harm, and given that conviction rates for the former type of offences is far

---

[71]The outcome of what I take to be a correctly characterised contractualist account of justice for punishment is examined below in Chapters 6 & 7.

[72]Cf. Rawls 1971, 215.

lower than for the latter - the punishments for the former offences may have to be relatively more severe compared with the latter.[73] It may still be that people contracting on the basis of mutual advantage may wish the penalties for crimes against the person to be more severe than for crimes against property (given that they fear the greater harm of crimes against the person) but it is still likely that the penalty for crimes against property (and minor offences such as motoring offences) would have to be much more severe than they are even in present day Britain, and thus relatively more severe when compared with the penalties for crimes such as murder. Finally, it is worth noting, that in the implausible world of a society based on justice as mutual advantage, the scope of the law would have to be at least as wide as that employed by the more "totalising" of totalitarian regimes. This is because, without effective moral sanctions - that is without sanctions that appeal to the agent being condemned in such a way as to get him to reconsider his understanding of his self-interest - the society would have to rely on legal sanctions.

The only appeal available to the those doing the condemning, then, is to say that the society taken generally would be better off if people were not successful

---

[73]As J. F. Stephen put it, "surely it is at the moment when temptation to crime is strongest that the law should speak most clearly and emphatically to the contrary." (Stephen 1883, Vol. 2, 107). Cf. Bentham: "The strength of the temptation (*ceteris paribus)* is as the profit of the offence: the quantum of the punishment must rise with the profit of the offence: *ceteris paribus*, it must therefore rise with the strength of the temptation". (Bentham 1789, Chapter 16, §9).

when performing whatever action is being condemned. This casts some illumination on the question of whether it can be said that the offender willed his punishment. Whilst it is clear that the starting point would be to say that to the offender that he benefits from the co-operative scheme and that the co-operative scheme is only possible if non-compliance with the laws is kept to a manageable level, the offender can concur with this but claim that in this instance it was in his interests to ignore the law, and given that he is a single individual it is not the case that he endangered the co-operative practice itself. He calculated the risks, both to himself and to the practice, and the benefits and chose to break the law. In some sense his argument suggests that he accepts his punishment; it is just a factor to be included in the calculation of whether the criminal action was worthwhile, but it surely extends the meaning of the term too much to say that he willed it, rationally or otherwise. I certainly accept that if I drink too much in an evening I will suffer a hang over, and I consider this cost when I decide whether to have a(nother) drink, but I would balk at the suggestion that I willed my hang over.

The idiocy of attempts to ground morality solely in terms of mutual advantage, then, explains why theorists of justice as mutual advantage are keen to import the motivation I identified above in theories of fair play, or reciprocity. That is, that one should be motivated to behave fairly; not to free-ride. This is not available in

a theory of mutual advantage, but the combination of an agreement reached by justice as mutual advantage and a motivation of acting justly has recently had something of a revival in the work of Alan Gibbard.[74] As I noted above, however, fair play or reciprocity leaves the content of the principles of justice open, and if these are given by an agreement based on mutual advantage then the theory is susceptible to all the objections formulated against the content of such theories above. But, further, if the content of justice is given by mutual advantage the proposition that the motivation the agent has for compliance - to behave fairly - will be sufficient to meet the objections formulated above is fantastical.

Consider the argument: justice as reciprocity is more stable because the motivation for complying with the norms of justice is not that penal sanctions apply to free-riders, but that the participants are motivated by the desire to act fairly. Is this plausible? The stability problem arose from two sources: the more powerful press for re-negotiation if their relative power increases; and the weak (like everybody else) are constantly motivated to break the norms when it is in their interests to do so (and, of course, given that the norms are constructed to benefit the strong, the weak, presumably, will often be in such a position). The motivation to act fairly does nothing to address the first problem, and, it seems to me, can only plausibly address part of the second. Perhaps the

---

[74]See Alan Gibbard 1990, 1991.

motivation to act fairly may convert the norms of mutual advantage into authoritatively binding moral norms for those who feel that the system benefits them in relative terms, but it seems odd (to say the least) to say that the motivation to act fairly will act in such an efficacious way for the less advantaged;[75] we are being asked to believe, say, that after the initial agreement has deprived the American Indians of much of their territory and condemned many of them to poverty and starvation, they have a good reason for keeping the agreement given the motivation to act fairly. Rather, I would have thought, tell them that they should keep the agreement to avoid complete destruction (which after all is the motive for their signing the treaty) as would be the case in a pure theory of mutual advantage.

Of course, as I noted above, theories of punishment which depend upon this motivation to play fair, begin with a commitment to a fair starting point. But, as I pointed out there, they cannot do this without importing a moral theory to give them this fair starting point. We have seen that justice as mutual advantage cannot be that moral theory. I now wish to turn to the remaining category of contractualist theories, theories of justice as impartiality, and see whether they can be any more successful, and whether fair play has any role in justifying punishment within such theories.

---

[75]This is the point made by Jeffrie Murphy in "Marxism and Retribution" (1973).

Chapter 5: Impartial Justice and Punishment

"If the original position is to yield agreements that are just, the parties must be fairly situated and treated equally as moral persons." Rawls 1971, 141.

26. Justice as Impartiality

In Chapter 4 I claimed that recent contractarian accounts of justice could be understood best in the light of Brian Barry's definition of constructivism. I then identified one set of contractarian theories, theories of justice as mutual advantage, and discarded them both because they led to outcomes fundamentally at odds with our considered convictions about justice and because they failed to ensure stability by providing authoritatively binding moral norms. In this chapter I want to consider theories of justice as impartiality, again from the perspective of punishment and again trying to use this perspective to examine the cogency of the theories themselves. I shall begin with some general reflections on the nature of such theories and then turn to a closer examination of the function of punishment.

To anticipate the conclusion of the chapter: justice as impartiality is an example of what I shall call *faux* constructivism, that is, it builds the answer into the construction in such a way as to leave the construction to do little work in the theory. As a consequence, the coercive element must be justified by a second, independent, argument of the form of one of those examined

and rejected in the preceding chapters; a metaphysical (and pre-institutional) appeal to desert; a two level consequentialism; or a fair play theory.[1]

Justice as impartiality differs fundamentally from theories of justice based on mutual advantage. This is not only a matter of coming to different conclusions about the content of justice, it also uses the contractualist procedure to a completely different end. As Kymlicka puts it in his review of contractarian approaches, justice as impartiality:

> "uses the device of a social contract in order to develop, rather than replace, traditional notions of moral obligation; it uses the idea of the contract to express the inherent moral standing of persons, rather than to generate an artificial moral standing; and it uses the device of the contract to negate, rather than reflect, unequal bargaining power."[2]

Kymlicka's characterisation of theories of justice as impartiality makes clear that although what emerges from the contract is, indeed, justice, what goes in to the initial choosing situation is a commitment to the moral equality of human beings.[3] We know that people must be treated as (in Kantian terms) ends in themselves or (in

---

[1]Indeed, in one case - John Rawls's A Theory of Justice - all three of the above appear at various points in the argument to justify punishment, see Appendix A.

[2]Kymlicka 1991, 191.

[3]The original position "represents equality between human beings as moral persons." (Rawls 1971, 190)

more modern terms) with "equal consideration". What we are unsure of is what, precisely, this means. The idea of the contract is that it is a procedure for determining the content of our moral duty to treat others with equal consideration. This is, in my view, a *faux* constructivism; the construction does not ground morality but merely explicates currently shared moral convictions. The "situation" is determined by the outcome - by what "emerges" - rather than the other way around. Thus, as Rawls admits, "for each traditional conception of justice there exists an interpretation of the initial situation in which its principles are the preferred solution".[4] Such *faux* constructivism is then some mixture of coherentism grounded (in the cases in which I am interested) in liberal post-enlightenment intuitions. The construction is left with a considerably less ambitious function than the grounding of justice; as Brian Barry has put it:

> "... once we admit that substantive intuitions have to go into the specification of the original position if we are to derive definite implications, the case for saying constructivism is something different from intuitionism becomes weaker. But it seems to me that there is still a good case for saying the construction is doing some real

---

[4]Rawls 1971, 121. This understanding of the role of the contract is the context in which Rawls's often quoted and equally often misunderstood statements about the principles of justice being the outcome of rational deliberation (or, the theory being an example of "moral geometry") need to be read. For example, Rawls says that to "say that a certain conception of justice would be chosen in the original position is equivalent to saying that rational deliberation satisfying certain conditions and restrictions would reach a certain conclusion". (Rawls 1971, 138) It is vital to remember when reading such quotations that the "certain conditions" and "restrictions" guarantee the "conclusion".

> work provided what is put in is more general than what comes out."[5]

The easiest way to further unpack the idea of justice as impartiality is to consider examples of such theorising; I shall begin with the most famous recent attempt, John Rawls's A Theory of Justice[6] I shall try to be brief, however, as I take it that the basic tenets of Rawls's work are familiar and I am trying to make quite specific points and not to provide a general overview of the main arguments of A Theory of Justice. I shall then go on to discuss Scanlon's development of Rawls.

27. Rawls

If what I have said above is right then the function of the Rawlsian contract is to capture and make more specific some basic "considered convictions" we share about morality.[7]

---

[5]Barry 1989, 275. The belief that this is all there is to constructivism is what allows Barry to derive the difference principle without reference to the contract argument, see Barry 1989, Chapter 6.

[6]Much work has gone into seeing whether Rawls's more recent statements on justice, and indeed on A Theory of Justice (see Rawls 1980, 1982a, 1982b, 1985, 1987, 1988, 1989, 1993a, 1993b) are compatible with his statements made earlier in that book. I am concerned here just with one - although I think the right - reading of A Theory of Justice. This is because I am not particularly interested here in Rawlsian scholarship, or in Rawls's answer as such, I am interested in a particular account of justice, one I think is embodied in A Theory of Justice. An analysis of recent developments can be found in Mulhall and Swift 1992; Barry 1995a, 1995b.

[7]Rawls 1971, 19. I take it that the question of who "we" are is not controversial, we are modern liberals (using that term in its broadest, political, sense). It does not follow from that, however, that what emerges from is necessarily relativistic. It is true that impartialist justice is ours in the sense that it is according to us,

Most importantly, it is to embody and reflect the fundamental equality[8] of moral agents. The contract, then, must be negotiated from such a position, and this is ensured by placing the contracting parties behind a "veil of ignorance".[9] The people in the original position - in the contracting situation - are to choose principles of justice from a position of ignorance about their talents, abilities and infirmities, their position in society and the position of their society. They are to be motivated by a concern to further their conception of the good through maximising their access to social primary goods (liberty and opportunity, income and wealth, and the bases of self-respect[10]). The veil of ignorance, here, is clearly the key; it commits each individual to choose principles as if he were anyone, indeed, in removing all the differences between people in the original position it removes the need for there to be people rather than a person,[11] all that

---

it does not follow that it applies only to us. As Scanlon has put it, "the fact that in the method of Reflective Equilibrium we begin with 'our' considered judgements does not make that relativistic, because the fact that others, beginning with different judgements, would arrive at different conclusions through the same method does not mean that we must regard their conclusions as just as valid as ours. 'Our' considered judgements and the principles they lead to can include judgements about what is right and wrong 'for them', including judgements about the correctness or incorrectness of their starting points."(Scanlon 1992, 22) It must be said that Rawls has since become considerably more relativistic in his interpretation of the claims made by his own theory, see especially Rawls 1993a, Rawls 1993b.

[8]I shall use "fundamental" as a term of art denoting that individuals are to be treated as abstract selves, and thus as identical for the purpose of initial distribution of resources.

[9]See Rawls 1971, §24.

[10]See Rawls 1971, 62.

[11]"To begin with, it is clear that since the differences among the parties are unknown to them, and everyone is equally rational and

matters is that it ensures that the principles are chosen from an impartial standpoint; an impartial standpoint which is built in to the theory.[12]

Rawls in fact goes even further than this, in "defin[ing] the original position so [as to] get the desired solution",[13] "we work from both ends",[14] that is, we characterise an initial choosing situation and then see what emerges from it. Should the principles that emerge conflict with our "considered convictions of justice … we can either modify the account of the initial situation or we can revise our existing judgements." "By going back and forth", argues Rawls, "eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgements duly pruned and adjusted." This is the now famous condition of "reflective equilibrium".[15]

The final point that should interest us is the congruence of Rawls's two arguments for the principles of justice presented in Chapters II and III of A Theory of Justice. That is between the "intuitive argument" from equality of

---

similarly situated, each is convinced by the same arguments. Therefore, we can view the choice in the original position from the standpoint of one person selected at random." (Rawls 1971, 139).

[12]Consider the entry for "Rawls" in The Philosophers' Lexicon; "Rawls: A fishing line baited with a few apparently innocent intuitions about fairness but capable of bringing in such big fish as Pareto Optimality and God knows what." (Dennett 1987).

[13]Rawls 1971, 141.

[14]Rawls 1971, 20.

[15]All quotations from Rawls 1971, 20.

opportunity and the argument from the veil of ignorance.[16] The "intuitive argument" develops the conviction that whilst persons are deserving of equal consideration they are not deserving of much else;

> "it seems to be one of the fixed points of our considered judgements that no one deserves his place in the distribution of native endowments, any more than one deserves one's initial starting place in society."[17]

This, of course, finds its place in the argument from the original position in the thickness of the veil of ignorance.

Before turning to impartiality and punishment I want to look at one other impartialist theorist, T. M. Scanlon. The reason for this is that it is important to remember that, although Rawls is an exemplar of the approach, justice as impartiality is a broader position than Rawls's theory. In addition, Rawls's account of punishment is ill worked out and often confused. Scanlon, on the other hand, attempts to formulate a coherent account of distributive and retributive justice, one which I shall examine below. I shall discuss Rawls's account of punishment in an Appendix (for reasons elaborated there).

---

[16]For more on the idea that Rawls is best understood if one separates out these two arguments see Barry 1989, Chapter 6, especially 213-14; Kymlicka 1990, 50-76. For a summary of the argument from equality of opportunity, see *infra* 43n.

[17]Rawls 1971, 104; the whole of §17 is relevant to this argument.

28. Scanlon

Tim Scanlon's version of the original position is presented in his 1982 article, "Contractualism and Utilitarianism".[18] According to Scanlon:

> "An act is wrong if its performance under the circumstances would be disallowed by any system of rules for the regulation of behaviour which no one could reasonably reject as the basis for informed, unforced general agreement."[19]

The motivation given to the agent for theorising his relations with others is given as a psychological fact:

> "According to contractualism, the source of motivation that is directly triggered by the belief that an action is wrong is the desire to be able to justify one's actions to others on grounds that they could not reasonably reject."[20]

This psychological fact is not "natural", it is a consequence of "moral education" and thus it resembles Rawls's concept of the "sense of justice".[21]

In Scanlon's alternative formulation the parties to the contractual arrangement are aware of their identities and interests, so equality is not guaranteed, as it is in

---

[18]Scanlon 1982.

[19]Scanlon 1982, 110.

[20]Scanlon 1982, 116.

[21]See Scanlon 1982, 117; Rawls 1971, §86; 1993a, 19; Appendix A, §C.

Rawls, by a "veil of ignorance", and this is why the motivation of "the desire for reasonable agreement"[22] is so important. In other respects, the Scanlonian formulation is not, or so it seems to me, a significant departure from the Rawlsian project. This is not to say that there are no substantive differences between Rawls and Scanlon - I have identified two important ones above - but merely that Rawls and Scanlon are engaged on doing much the same type of thing. Rawls's original position and the veil of ignorance collapse the plurality of individuals into the view of a single person, and thus lose the idea of different points of view being represented in the (thus, nominal) contract. Scanlon retains these points of view but still wants to guarantee impartiality. This is achieved through the conditions which surround the agreement; it is to be "informed", "unforced" and the content is to be subject to individual veto (so long as that veto is "reasonable"). Examining these conditions makes it clear how impartiality is built in.

By "informed", Scanlon means to "exclude agreement based on superstition or false belief about the consequences of actions, even if these beliefs are ones which it would be reasonable for the person in question to have."[23] This is a strong requirement, it is clearly intended to rule out agreements based on beliefs about such things as the catastrophes which will be visited on any society which

---

[22]Scanlon 1982, 115n.

[23]Scanlon 1982, 111.

does not give proper place to the worship of God. In the form that Scanlon has it, however, it requires that we give some account of truth, a task that might be at least as difficult as providing an account of morality.[24]

The condition of being "unforced" is, Scanlon asserts, not simply meant to rule out coercion, but also the exploitation of the weak by the strong; that is, it is meant to rule out the agreement mirroring relative bargaining strength, (the outcome of justice as mutual advantage). This, however, can only be the case if one assumes an initial base line of equality. If the powerful present the mutual advantage argument to the weak - that the agreement should benefit both groups, although given relative bargaining strengths, it should favour the powerful - and the weak assent to this argument, then it can hardly be objected that this agreement is illegitimate because it is "forced". The normative strength of the "unforced" condition, therefore, comes only with the assumption that there is a base line of equality which the weak would only leave (to the advantage of the strong), if they were forced to do so. Equality, in short, is presupposed as the initial position if Scanlon wants the "unforced" condition to do what he clearly does want it to do, which is rule out agreements based on relative bargaining strengths.

---

[24]Scanlon's requirements are reminiscent of Susan Wolf's that autonomy is acting rationally, where the latter is defined by being action based on true beliefs about the nature of the physical and the moral worlds. A requirement justifiably criticised in the literature for being unreasonably strong. (See Wolf 1989; Christman 1989, 12)

Nonetheless, despite these requirements, one is still left with the feeling that the real work is being done by the condition of reasonableness. The idea, "is to exclude rejections that would be unreasonable *given* the aim of finding principles which could be the basis of informed, unforced general agreement". "It would be unreasonable", Scanlon tells us, "to reject a principle because it imposed a burden on you when every alternative principle would impose much greater burdens on others".[25] Scanlon is saying, or so it seems to me, that one's rejection of such a principle would not be absolutely unreasonable but contingently so; it is unreasonable because one could not expect others to shoulder the extra burdens consequent to one's rejection of the principle. It seems clear how this might work for something like political rights; to take an example, let us say Rawls's first principle[26] is suggested, could it be reasonably vetoed?

There seem to be two grounds for saying no; first, the grounds on which an agent might veto such a principle might be subject to the requirement that the agreement be "informed". That is, if I attempt to veto the principle on the grounds that apostates ought not to have freedom of expression because they are unworthy in the eyes of God, then my veto could be declared invalid because it is based

---

[25]Quotations from Scanlon 1982, 111.

[26]"Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all." (Rawls 1971, 250)

upon a "superstition or false belief". This objection might sit uncomfortably with a number of liberals, notably those with a sceptical turn of mind.[27] The second ground upon which my veto might be overturned is that it is not reasonable; I could not expect apostates to accept the characterisation of themselves as unworthy in the eyes of God. This is clearly similar to the moral argument of Rawls that those who do worst have to be able to accept the reasons for their position as valid, otherwise they will find the "strains of commitment" too great to bear.[28] But, if reasonable is to be separated from truth then what is the content of this condition? As Allan Gibbard points out in a review of Barry's Theories of Justice, a great deal of emphasis is being put on the idea of "reasonableness".[29] Clearly if the rejection of a system of rules has to be reasonable it is going to be the characterisation of reasonable that does the work in the argument, rather than the idea of agreement.

Before considering quite how the condition of reasonableness might be filled out let us look at a further example, both because in more controversial areas it becomes clearer how much is being rested on the notion of reasonableness, and because the idea of desert is destined

---

[27]See for example Brian Barry's commentary on Scanlon's "informed" requirement, Barry 1995a, Chapter 3 MS.

[28]See Appendix A, §B.

[29]Gibbard 1991. Russell Hardin has put it another way, describing "reasonable agreement" as "the 'open sesame' of much of contemporary moral theorising". (Hardin 1991, 667).

to reappear later in this chapter. Is it reasonable for those who are more advantageously naturally endowed (with talents and abilities, advantageous character traits etc.), to veto principles which ignore these features as relevant to the distribution of, say, income? Here the truth claim does not seem as easy to invoke. The question of whether it is false that genetically determined features of an individual merit greater reward is amongst the questions that the contractual situation is supposed to answer; to claim that it is uncontroversially false in advance is, therefore, unsatisfactory. It is, then, the condition of reasonableness that is going to have to do all the work. The claim, of course, must be that it is unreasonable to expect others not genetically advantaged to accept that there is some pre-institutional sense of desert relevant to genetic endowments.

The connection with truth, then, cannot be completely broken in giving some content to the notion of reasonable, instead it must find a role in the idea of a good reason.[30] What must be at the heart of reasonableness is the idea that the participants to the contractual situation have good reasons to accept the veto. The thought is that I have no reason to accept that I am less worthy of "equal

---

[30]Brian Barry has attempted to give a fully worked out Scanlonian theory in his Justice as Impartiality, (1995a) in terms of a limited scepticism. This amounts to the idea that it is unreasonable to veto any rule - given the fact of pluralism - on the grounds of one's personal conception of the good because one has to admit the possibility that one might be wrong in one's beliefs about the good life. One cannot, therefore, legitimately expect anyone who disagrees with you to live in accordance with your views on the nature of the good life.

consideration" because I am a woman, or black, or disabled. The others in the contractual situation accept that this constrains their veto on equality because they are motivated by "the desire to be able to justify [their] actions to others on grounds [that the others] could not reasonably reject".[31]

Although it is clear that this is how the reasonable condition is supposed to advance the argument, it is not clear why these particular constraints should apply. If I hold that it is reasonable that those who contribute more to the co-operative enterprise should receive greater rewards, even if they are only able to contribute more because of natural or social advantages, then I can still claim that everyone receives equal consideration if the basis on which rewards are distributed reflects these advantages. As with the "unforced" condition, the reasonable only does the work that it is supposed to if a base line of equality is built in to the theory. It is unreasonable to demand that justice rewards natural or social advantages only if it is unreasonable not to treat everyone as fundamentally equal.

The psychological motivation likewise fails to perform the necessary conjuring trick to get equality out rather than having to put it in. The "psychological fact" that people "desire to be able to justify [their] actions to others on grounds they could not reasonably reject" and that this

---

[31]Scanlon 1982, 116.

motivation "is directly triggered by the belief that an action is wrong" is no doubt "strong in most people",[32] at least if we count those to whom we have the desire to justify ourselves, to be those in "our" community (or, perhaps, those who are recognisably "like us"[33]), but, returning to the example above, if I hold that desert is a relevant and reasonable basis for reward, then I will feel that demanding more for those who contribute more is perfectly compatible with my desire to justify myself to others on the terms specified by Scanlon.

Scanlon recognises this fact in his Tanner Lectures.[34] Here he says that the disagreement between Rawls and Nozick is not over free will, and its compatibility or otherwise with the truth of causality, but is a moral argument about "doing enough" for people. "Of course", he says,

> "Rawls and Nozick disagree over what constitutes 'doing enough' for a person. For Nozick, one has 'done enough' as long as the person's Lockean rights have not been violated; for Rawls, the standard is set by the principles which would be accepted behind the Veil of Ignorance."[35]

---

[32]Scanlon 1982, 116; see also Scanlon 1988, §5.

[33]That is, I think it would be difficult to maintain that we have this desire with respect to those who are beyond our immediate experience; for example, do we feel that we have to justify the developed West's greater share of the Earth's resources to people in the third world?

[34]Scanlon 1988.

[35]Scanlon 1988, 187.

Scanlon is explicit in accepting that this is a moral, not a factual, argument; "it locates the disagreement in what seems, intuitively, to be the right place - in a question of justice rather than in a separate (and I believe spurious) question of causal determination".[36] The problem is, however, that Scanlon still does not explain why we *ought* to "do enough" for people or, indeed, the grounding for his Rawlsian sympathies as to what that entails. That is, "doing enough", rather than giving content to the idea of reasonable, simply becomes its synonym. It is unreasonable to reject principles that do enough for you precisely because they do enough, but what it is to do enough is to do that which couldn't be reasonably rejected. Nozick can, at least, say why he believes what he believes about the content of doing enough - we have inalienable rights and it is doing enough when we respect those rights and do not invade them - albeit that he cannot explain where the rights come from. Scanlon can only exclude justice as mutual advantage if he builds in an equally substantive moral commitment such that mutual advantage is not "doing enough" for the individual.

The basic problem, in summary, is that "reasonable" and "doing enough" are not sufficiently defined to exclude relative bargaining strengths. When they are defined so as to exclude relative bargaining strength or, for that matter, utilitarianism, it becomes clear that they come to mean "from a position of fundamental equality". However,

---

[36]Scanlon 1988, 187.

if this impartialist moral stance is built in then it seems to eradicate the role of any deep moral reflection; in this moralised form this is precisely the motivation for which the individual is trying to find a grounding.

In Scanlon's account, the contractualist procedure remains a reflective practice, a method of distancing oneself as a self-reflective agent from the norms and practices of one's community. However, the agent is not asking why she should live in accordance with such norms as much as she is enquiring into the precise (or more precise) content of her obligations. The moralised contract means that the agent is never in a position to question the most fundamental moral commitment, to treat others as fundamentally equal. This is surely unsatisfactory even for a *faux* constructivism, for in such a position the agent can still ask whether this motivation is one which she wants to endorse. Even if the contract is simply a heuristic or coherentist device, then, this moralised motivation cannot be accepted unconditionally by the agent when reflecting on justice or else the answer seems to flow too directly. If we know that we cannot justify truth claims beyond limited scepticism, and we know that to treat others with equal consideration means to treat them as fundamentally equal then most of the work of moral theory has been done.

I shall return to the question of "doing enough" for an agent, when I consider just how far impartialist theory can take the idea of responsibility and how this informs the

account of punishment. Before I do so, however, it is necessary to make some preliminary remarks about the rationale of coercion in any theory which builds in the impartialist element prior to the construction.

29. *Faux* Constructivism and Punishment

The consequence of the *faux* nature of Scanlon's and Rawls's constructivism is that the individual has a reason to be just which is independent of the construction. This follows from the independent grounding of justice; if justice is treating people as fundamentally equal then (presuming that one takes an internalist position on the relationship between morality and individual motivation), one has a reason to treat people in this way independently of any other consideration. The construction functions merely to tell one the precise nature of this obligation.[37] This means that *faux* constructivists are in the position of having to find a justification for punishing an offender which is not entailed by the theory which defines him as an offender, that is unless they endorse a metaphysical retributive theory in which the mere fact of transgression justifies the imposition of punishment.

---

[37]It might be objected that it is unfair to use "merely" in this sentence insofar as telling us the content of our moral obligations to others is hardly to be thought of as an easy task. I am prepared to accept that this is the case; the "merely" is only justified given that I believe that we can ground morality through the constructivist procedure. Faux constructivism, thus, "merely" does less than this. If it were the case that this is all we could ask of moral theory, the "merely" would become ridiculous.

John Rawls, who I have treated as an exemplar of such theorising, gestures in the direction of such a retributive theory, however, he does not present any detailed argument for it. He also makes use of a second option, a fair-play theory, although again the argument is not fully worked out, and insofar as it is, it relies on elements of Rawls's account that bring out his proximity to theories of justice as reciprocity rather than impartiality. I have relegated the examination of punishment in Rawls to an Appendix both because his account is so confused and because it is not helpful to my task which is considering the role of punishment in justice as impartiality considered as a broad position rather than as a particular individual's theory. The only remark I will make here is that Rawls seems to me to canvass all the possible options available to the *faux* constructivist. In committing themselves to a justification of punishment which is independent of the derivation of the moral norms, impartialist theorists have to rely on the three options of metaphysical retributivism, fair-play theory or a complicated two level theory (two level, presuming that justice as impartiality does not yield consequentialist principles of justice).

I have considered and dismissed as inadequate the first two of these options - metaphysical retributivism and fair-play theory - however, I have not adequately considered the two level theory that underlies, I think, both the impartialist account of distributive justice and offers the best hope for the impartialist of providing a justification of

punishment. In so doing I will also re-examine Scanlon's idea of "doing enough" for someone. The easiest way to approach this will be to look at the economic realm in which the theory is most clear; the question, then, will be whether the distribution of economic rewards and punitive sanctions are analogous.[38]

30. Impartiality, Legitimate Expectations and Desert

The liberal commitment to fundamental equality has sometimes been taken as a response to liberalism having embraced a naturalist understanding of the universe, including social life.[39] It is assumed that liberals have given up on individual responsibility because of the perceived truth of what Scanlon calls "the Causal Thesis".

> "This is the thesis that the events which are human actions, thoughts, and decisions are linked to antecedent events by causal laws as deterministic as those governing other goings-on in the universe. According to this thesis, given antecedent conditions and the laws of nature, the occurrence of an act of a specific kind follows, either with certainty or with a certain degree of probability, the indeterminacy being due to chance

---

[38]Rawls specifically denies that distributive and what he calls "retributive" justice are analogous, at other times, however, he appears to endorse the view presented below. See Appendix A.

[39]See for a subtle and informative account of this type Scheffler 1992.

> factors of the sort involved in other natural processes."[40]

On one such reading, then, one could claim that to "do enough" for someone requires that one ignore all factors that distinguish one person from another because such factors are not under the control of the agent, thus generating a Rawlsian type position of fundamental equality. Rawls, at times, seems to suggest just such a reading; crucial to his whole theory is the claim that,

> "it seems to be one of the fixed points of our considered judgments that no one deserves his place in the distribution of native endowments, any more than one deserves one's initial starting place in society."[41]

From this claim - that natural endowments are "arbitrary from a moral point of view"[42] - the argument for the difference principle from equality of opportunity follows.[43] The important question, though, is to ask what it means to say that something is "morally arbitrary". Scheffler and those who believe that liberalism endorses naturalism at some deep level, interpret this claim as

---

[40]Scanlon 1988, 152.

[41]Rawls 1971, 104.

[42]Rawls 1971, 72.

[43]Barry summarises the argument thus, "(1) the (liberal) ideal of equality of opportunity is that all environmental differences that affect occupational achievement should be eliminated; (2) this will entail that all remaining differences are of genetic origin; but (3) if (as is assumed) the case for eliminating environmental differences is that they are morally arbitrary, all we should be doing is making occupational achievement rest on genetic factors which are (in exactly the same sense) morally arbitrary; therefore (4), since what is morally arbitrary should not affect what people get, differences in occupational achievement should not affect incomes." (Barry 1989, 225).

following from the truth of the Causal Thesis, but this is fundamentally mistaken. If it were the case that moral arbitrariness followed directly from acknowledging the truth of the Causal Thesis then one could make no sense of Rawls's use of primary social goods as the "currency of justice",[44] or the answer he gives to the problem of expensive tastes (in which the agent is held responsible for his preferences and, thus, not deserving of a greater share of resources because she has preferences for goods which are more expensive[45]), or, indeed, the lexical priority of the first principle.

It is certainly not exactly clear how Rawls combines his basically anti-choicist starting point with his choicest conclusions,[46] in the end it seems to depend, rather implausibly, on the idea that "people's tastes, aspirations and beliefs are always open to modification, so people can properly be held responsible for them",[47] but quite how this is done is, as I say, never clear. Nevertheless, the best interpretation of Rawls, and certainly the most coherent, is to claim that he relies neither on the

---

[44]The term is taken from Cohen 1989.

[45]See Rawls 1982b, 168-9; cf. 1975, 553; 1980, 545; 1978, 63; 1985, 243-4.

[46]The language of "choicism" and "anti choicism" is taken from Barry 1991a. A choicist is "someone who wants to show that it is consistent with justice to give the principle of personal responsibility a lot of scope at the expense of the principle of compensation", an anti-choicist "holds that people cannot be held personally responsible for their tastes, aspirations, or beliefs and this entails that they cannot be held responsible for the outcomes of actions that flow from them." (Barry 1991a, 144, 150).

[47]Barry 1991a, 155.

subversion of responsibility nor on a crude idea of pre-institutional desert (and in that claim Scheffler is right), but on a model of legitimate expectations. If this is right it brings Rawls into line with Scanlon; in any event, as I am not concerned with Rawls *exegesis* but justice as impartiality, and since I believe legitimate expectations offers the most coherent model for both distributive and retributive justice within that approach, I shall now turn to what, precisely, this means.

The legitimate expectations model takes the form of a two level theory; the principles of justice are derived from an independent account of justice, once that theory has established the rules of justice then the provisions of those rules apply even if they seem to conflict with the fundamental convictions that inform the derivation of the rules themselves.[48] This is clearly the argument that underlies Rawls's comments on the differential economic rewards sanctioned by the difference principle:

> "It is perfectly true that given a just system of cooperation as a scheme of public rules and the expectations set up by it, those who, with the prospect of improving their condition, have done what the system announces that it will reward are entitled to their advantages. In this sense the more fortunate have a claim to their better situation; their claims are legitimate expectations established by social

---

[48]This characterisation should make it clear how close this view is to Rawls's 1958 article, "Two Concepts of Rules", see §19.

> institutions, and the community is obligated to meet them. But this sense of desert presupposes the existence of the cooperative scheme; it is irrelevant to the question … [of how] in the first place the system is to be designed."[49]

This is a very different conception of desert from that which informs our brute intuitions. If we tried to capture our intuitions we would design institutions so as to reward desert; on the account given above such an appeal to pre-institutional desert is avoided, "the only notions of desert which [are] recognized are internal to institutions … the notion of desert is replaced … by the idea of legitimate (institutional) expectations."[50]

The basic idea, here, is that liberal institutions are justified through an independent notion of justice, in the impartialist theories of Rawls and Scanlon, a notion of justice that gives no role to desert. Once these institutions have got off the ground then individuals are held responsible for how well or badly they do *within* the institutional system. Within the scheme, agents are held responsible in that where they end up (economically) is a consequence of their choices; they are "entitled to their advantages". But, the system has been set up to reward

---

[49]Rawls 1971, 103.

[50]Scanlon 1988, 188. As Scheffler says of this understanding of desert "the idea that social institutions should be designed in such a way as to ensure that people get what they deserve makes about as much sense as the idea that universities were created so that professors would have somewhere to turn in their grades…" (Scheffler 1992, 306).

these talents not because they deserve rewards, but because differential rewards are to the benefit (in the Rawlsian case) of the least advantaged. Those who do least well cannot complain that the others do not deserve their benefits because the system does not reward desert, so to speak, but rewards in accordance with a principle of justice that is in accordance with reasons that no-one can reasonably reject or which would emerge from the original position.

It is important to understand how this theory differs from that which would follow if one read the claim of "moral arbitrariness" as derived from a commitment to a world without responsibility; on the legitimate expectations view if I have natural talents and abilities, (which, if we admit the truth of the Causal Thesis, are at some level arbitrary), then not only can I take pride in my talents and abilities - they are, after all, mine - but it *may* be the case that I am entitled to greater income than someone with fewer natural assets. Whether I am or not will depend entirely on whether a just system would reward me, and a just system is one that "does enough" for everyone.[51]

The basic structure of the legitimate expectations view, then, is relatively straightforward. By building in

---

[51]The link with the theories of punishment examined at the end of Chapter 3 should be clear. Nino's consensual view comes very close to this kind of argument, and again, what is crucial is not whether the offender has "free will", but whether the system is just and did enough to make the offender's choice one of which we may posit responsibility, see §20.

impartiality one gets the requirement that everyone should be treated as fundamentally equal in the initial distribution of resources, (including political rights). Beyond this distribution, the system announces that certain activities will gain greater rewards and those that undertake those activities are thus entitled to the greater rewards they receive. The fact that those that reap the greater rewards are only capable of doing so because of talents that are theirs by chance does not matter, for those that do not gain the better rewards have still had enough done for them in design of the system.

It should be clear, then, that the introduction of legitimate expectations has not cleared justice as impartiality of the charge of being a *faux* constructivism, for no account is offered of the building in of impartiality. There is, however, one outstanding question; why does the system include a role for choice and differential rewards?

Differential entitlements might be justified simply because they improve the position of the worst off, and it is possible that the role of choice could be justified in a similar manner. Rawls certainly, on occasions, seems to endorse this view, as when he writes that "the function of unequal distributive shares is … to attract individuals to places and associations where they are most needed from a social point of view",[52] and this dismissal of the

[52]Rawls 1971, 315.

importance of choice has led to one of the most important criticisms of Rawls's theory; that it is not "ambition sensitive".[53] But this is not the whole answer as endorsed by Rawls and Scanlon; choice is important not just because the position of the least well off (or of everyone) will be better if choices are allowed, but also because we value choice, and we value having the world such that a proportion of our lives is subject to our own decisions.[54] It is simply implausible to claim of a theorist who writes that a person's "highest order interest" is their capacity "to frame, revise and rationally to pursue"[55] their conception of the good, that he is only interested in choice as it affects the society as a whole.[56]

The fact is that Rawls, Scanlon, and Barry build in the impartialist element because of a commitment to the fundamental equality of persons; this commitment does not stem from a view of persons as equally capable of feeling pain or any such thing, but from a Kantian type endorsement of a view of each human as a potentially rational, self-directing chooser.[56] In order to hold this view, however,

---

[53]See, e.g., Kymlicka 1990, 73-76.

[54]See Scanlon 1988, Lecture Two. Scanlon differentiates three grounds upon which we value choice; instrumental (because if we choose for ourselves this, in most cases, will ensure that our preferences are met); demonstrative (because in choosing I may demonstrate some feature of oneself); and symbolic (because the act of choosing symbolises that one is competent, just as to be deprived of choice, even on paternalistic grounds, might represent a judgement that one is not capable of choosing, or choosing well).

[55]Rawls 1982, 16.

[56]Percy Lehning has argued that liberal neutrality is grounded in the idea that other people are shown equal respect by virtue of their

impartialist theory needs an account of responsibility which does not commit them to implausible Kantian metaphysics, and which is compatible with the truth of the Causal Thesis.[57]

If we grant that such an account can be given, and I think it can, the question is, can the legitimate expectations view be applied to the distribution of retributive justice, thus generating an account of morally justified punishment?

31. Legitimate Expectations and Punishment

Scanlon clearly believes that the answer to this question is yes, he argues for a direct analogy between the cases of distributive and retributive justice:

> "In approaching the problems of justifying both penal and economic institutions we begin with strong pretheoretical intuitions about the significance of choice: voluntary and intentional commission of a criminal act is a necessary condition of just punishment, and voluntary economic contribution can make an economic reward just

---

capacity to work out their own conception of the good life, see Lehning 1991.

[57]As I noted in the Introduction, (note 15), I am presupposing such an account. In fact, liberal theorists have been slow to develop such a theory, although the literature has recently shown signs that they are coming to realise its importance (see Barry 1991a; Ripstein 1994, but cf. Scheffler 1992; Scanlon 1988). So far, liberal autonomy theory has followed a fairly standard Dworkin/Frankfurt line (see Frankfurt 1971; G. Dworkin 1981), but without taking into account the problems associated with that approach, notably its tendency towards infinite regress or incompleteness (see Friedman 1986, but cf. Christman 1987; Thalberg 1979; for a general review see Christman 1989, Introduction). As I noted in the Introduction, liberal hesitancy could be related to the likely tensions an adequate theory of autonomy would cause in the general approach impartialist liberals take towards distributive justice.

> and its denial unjust. One way to account for these intuitions is by an appeal to a preinstitutional notion of desert: certain acts deserve punishment, certain contributions are just if they distribute benefits and burdens in accord with these forms of desert.
>
> The strategy I am describing makes a point of avoiding any such appeal. The only notions of desert which it recognizes are internal to institutions and dependent upon a prior notion of justice…".[58]

There seems to be something right about this argument. If we have a system of sanctions established and justified by a theory of justice, then if someone voluntarily and intentionally performs a criminal action they become liable to the sanctions which are specified; or to some sanction from a range depending upon certain other considerations.[59] The argument, however, does not get us far, the questions remain: What justifies the system of sanctions? And, how do we "do enough" for the individual in the distribution of punishment?

Nowhere does either Scanlon or Rawls say a great deal about the purpose of punishment, and what Rawls does say conflicts with the rest of his theory;[60] Scanlon (in a lecture on punishment), offers only half a sentence:

---

[58]Scanlon 1988, 188.

[59]Again the link between Scanlon's view and those of the consequentialist theorists examined at the end of Chapter 3 should be clear, the problem there was that no account of justice was given, the problem here is that the account of justice is ungrounded and the account of punishment underdeveloped.

[60]Rawls's account is examined in Appendix A.

"protecting ourselves and our possessions".[61] Scanlon's answer is a consequentialist one, the institution of punishment is a necessary means to achieving this "public goal",[62] in other words, having established a society based upon the rules of justice, punishment is needed to keep the society going. This raises a question about the analogy between economic justice and the system of sanctions. In the economic sphere impartialist theorists "do enough" for an individual by organising the system of rewards so that differential rewards are to the advantage of the worst off, or are such that no-one could reasonably reject them given a starting point of fundamental equality. In other words, distributive justice starts from a conviction that everyone is to be treated as fundamentally equal, and the principles which allow deviations from that equality must be at least sanctioned by the worst off. The less talented do less well than the more talented, but not less well than they would under an initially equal distribution; one can do better or worse but one cannot do so badly that one has grounds for not accepting the reasons for one's position.[63] In punishment, however, the situation seems to be different. In short, the punished may have reason to suspect that they are being sacrificed to the benefit of others. Why?

---

[61]Scanlon 1988, 201.

[62]Scanlon 1988, 201.

[63]This is the "strains of commitment" test in a nutshell. See Appendix A, §B.

The reason is, in a sense, encapsulated in the language we would have to use to describe what had happened to the offender. In addressing the offender, the punishing authorities would say, "You [the offender] voluntarily and intentionally chose to do X, although it was made clear that if you did X you would suffer Y". Therefore, "you are now *entitled* to Y". This might strike the offender as pretty bizarre, as indeed would the next step; when asked by the offender why Y's are attached to X's, the answer would have to be "because without the system that includes the rule that Y's are attached to X's the consequences for everybody would be bad".

This raises two problems; the first is one commonly attached to two level theories;[64] the point is that there is a fundamental division between the theory that grounds the norms of co-operation and the theory that grounds punishment. Up to the point where the individual "chooses badly" and commits an offence, the system "does enough" for her. Having chosen badly, the system sacrifices her to the good of the whole, and can do so without constraint; it could, for example, use her as an example, imposing a disproportionate punishment for her offence.[65]

The impartialist theorist would like to claim that constraints of justice apply to the system of punishment,

---

[64]See *supra*, Introduction; §8; §19.

[65]I shall claim below that this is not really moral punishment, at best, if correctly justified, it is justified coercion, see §§37-38.

but there is no reason why they should protect against such abuse. Justice requires that we treat everyone as fundamentally equal in the distribution of, say, primary social goods, but the theory of legitimate expectations then allows deviation from this equal distribution where it is to the advantage of the least well off (or could not be reasonably vetoed), and in recognition of agency. In the economic case, constraints on how badly the agent can do are built in to the system. In the case of punishment, however, the constraints which define "doing enough" do not seem to restrict how badly the offender may be treated, that is determined by the needs of the society.

To see this, consider Scanlon's account of "doing enough" with respect to punishment: To do enough we must have "requirements of due process" to "protect those who choose to stay out of the affected area";[66] and to "make it less likely that people will choose to enter [this area]", we must provide,

> "education, including moral education, the dissemination of basic information about the law, and the maintenance of social and economic conditions which reduce the incentive to commit crime, [finally, we must provide] … restrictions on entrapment by law enforcement officers".[67]

---

[66]By "affected area", Scanlon means those "activities which have been declared illegal". (Scanlon 1988, 202).

[67]Scanlon 1988, 202.

From here the argument is clear, having done enough for the agent, if she "voluntarily and intentionally" commits a criminal act, we have "done enough", she may be coerced so as to maintain public order. Of course, given the deep arbitrariness that underlies the legitimate expectations view, we should not feel "'You asked for this' but 'There but for the grace of God go I.'"[68]

The offender seems to be sacrificed to the whole on the basis that it is in everybody's interests (except her own) and, if she complains, she is told that a just system of laws with the normal checks and protections, which is announced in advance, "does enough" for her; her choice was a bad one, but having made it she cannot legitimately complain about the consequences, no matter how harsh. "There but for the grace of God", indeed.[69]

The problem is, as I have said, that by building in the requirements of impartiality, the theory separates the account of justice from the account of punishment. Justice is not only groundless on this theory but toothless. Punishment, then, has to be given some other rationale, and the most obvious is consequentialist. However, the

---

[68]Scanlon 1988, 216.

[69]In a sense the problem is the reverse of that encountered by rule-utilitarians. Rule-utilitarians know what the right answers are in defining the practice of punishment, but they are let down by the normative theory which is supposed to ground these answers. Scanlon has an ungrounded normative theory of justice, but he cannot bridge the gap between it and the account of punishment. Thus, whilst he gets pretty much the right answers in distributive justice, he cannot get the right ones on retributive questions.

requirement to maintain public order may demand exemplary punishments, or even (in a well ordered society), only the punishing of a proportion of the convicted offenders. The demands of justice to treat everyone equally do not apply because we have already done so in "doing enough" for the person before she chose to offend.

If we are to overcome this problem, the rationale of coercion must follow from the account of justice itself, it must not be an extra argument tacked on at the end to explain the necessary but disconcerting social practice of punishment. If it is linked to the account of justice, then the offender can be put in a position where, as a moral being, she accepts her own punishment, and thus coercion is converted from a, perhaps justified, act of societal self-defence into a moral practice. To show this, it is first necessary to give an account of justice, and in the next chapter I will argue for a true constructivism, which does not build in the requirement of impartiality prior to the construction. Such a theory, I believe, can yield the necessarily coercive terms of agreement that can, in turn, yield a justified system of moral punishment.

Chapter 6: A Constructivist Theory of Justice

32. The Aspirations of Constructivism

In the two chapters which preceded this one, I argued first, that if the pursuit of one's own advantage (understood from an individualist perspective) was the sole motivation for the agent in choosing principles to govern his relations with others, then those relations would neither be recognisably moral nor particularly stable. I insisted that if moral principles were to emerge from the constructivist procedure then some way must be found to reconcile individual interest and an impartial standpoint. I then turned to impartialist theories and argued that, in their drive to turn out the right answers, such theories build in the impartialist element. In so doing they open up an unbridgeable gap between the individual and moral standpoints. From the individual perspective the demand for impartiality appears as alien and serving only to frustrate the fulfilment of the agent's good.[1] The arguments that are offered as to how the two standpoints can be satisfactorily combined are specious, taking the form of doubtful empirical claims about moral motivation or Kantian claims that in acting justly we fulfil our natures as free beings. This gap is replicated in the account impartialist theorists give of coercion, for the theory

---

[1]For a good introduction to these perspectives and to the problems they give rise to, see Nagel 1986, and Nagel 1991, Introduction.

commits them to the view that morality is individually binding on the agent independent of any agreement by others to likewise adhere to moral norms. A separate justification, then, has to be found for coercion, either in bridging the motivation gap for everyone's advantage (i.e., in consequentialism) or in a metaphysical retributivism.

In this Chapter I want to present a constructivist account of morality, one which is neither grounded solely on mutual advantage nor in a *prior* commitment to fundamental equality. To do so I first need to address the question of what I mean by "moral". Morality is normally taken to be a system of rules, norms or standards which distinguish what is acceptable from what is not in the pursuit of one's aims, or in a society's pursuit of its aims, or, indeed, that of a species pursuing its.[2] Morality is thus often thought of as either an independent constraint upon the pursuit of self-interest or as justified because in the long run "crime does not pay" (i.e., over time moral behaviour serves self-interest). In the latter case the theory is not really a moral one as it subordinates morality to self-interest; if I behave in accordance with accepted moral standards *only* because not to do so will lead to punishment, my actions are not moral, and if the theory that grounds those standards is of this *modus vivendi* type then it is not moral either. For example, the injunction to do unto others only that which you would have

---

[2]Cf. the entry under "ethics" in Speake 1979.

them do unto you, could be understood in purely prudential terms, that is, it might be read as saying do unto others only that which you would have them do unto you only because there is a good chance that they will get to do back to you what you have done to them. This, however, would not be a theory of morality but of prudential rationality and, insofar as it informed the structure of relations between a group of individuals, of co-operation.[3]

In the former case, where morality is thought of as an external constraint on the pursuit of self-interest, some reason must be given to the agent as to why he should accept such a constraint. The problem here, of course, is that such a reason can either be a moral one, in which case the theory seems to be going around in circles, or one which appeals to the agent's self-interest, in which case the theory's credentials as moral seem to be undermined. Thus, for example, if we posited a set of norms established by a God which claimed to bind the pursuit of self-interest - that is which made the claim of being moral norms - then the agent, reflecting on his relationship with that God and those norms, could ask for reasons as to why he should limit the pursuit of his self-interest in the manner prescribed. If we were to say to the agent that he ought to behave morally because that is the right thing to do we have offered a moral reason and the agent would be justified in returning us to square one by asking why he

---

[3]It is, of course, the underlying motivation in accounts of justice as mutual advantage, see §25.

should "do the right thing". However, if we were to argue that he should obey the commands of God only because otherwise he would burn in hell for eternity (or, for that matter, only because he would otherwise not get to live in perpetual peace and tranquillity in heaven) then we seem to have given a prudential not a moral account.

The alternative route is to say that if there is an objective - or realist - account of the good (which we can assume there is in a theistic theory), then we have reason to be good because in so doing we flourish as moral beings, that is in accordance with our (externally given) good. For the purposes of this thesis I am largely ignoring such a possibility, however, because there are no good reasons to accept a realist account of the good and several for rejecting it. What I aim to show below is that an anti-realist account of morality can be given and the dichotomy between the moral and personal perspectives (what Nagel calls the "personal" and "impersonal standpoints"[4]), overcome. Nagel, himself, regards giving a rational basis for combining these perspectives as impossible given the "disenchanted" world we now confront as inheritors of the Enlightenment. I aim to show in this Chapter that such moral pessimism is largely (but not completely), unwarranted.

---

[4]"The impersonal standpoint in each of us produces … a powerful demand for universal impartiality and equality, while the personal standpoint gives rise to individualistic motives and requirements which produce obstacles to the pursuit and realization of such ideals." (Nagel 1991, 4); see also Nagel 1991, Chapter 2; 1986.

If the personal and moral perspectives are to be combined, the moral must not be grounded in a manner independent of human self-interest but rather must take the form of norms through which the agent comes to understand and pursue his own flourishing. This is the argument I shall offer below. The aspirations of such a constructivist scheme are not modest; I intend to give an account of justice grounded in human reason and will, an account which will allow us to evaluate the ethical standing of a community's rules and practices. In so doing, I shall argue that political obligation and the rationale of coercion are not to be understood and accounted for as separate moral arguments once justice is established, but as intimate elements of the account of justice itself.[5] In one sense, however, the claims made in this Chapter are modest, for in arguing that morality is possible only as the form given to the norms which govern the co-operation of a community of rational self-reflective agents, I shall argue that the precise content of the moral norms is a subject for the agreement of the co-operating members of that community. Thus, the analysis of the content of morality that can be undertaken in the abstract is necessarily limited to identifying the general claims that must find embodiment in the practices of such a community.

---

[5]I am following Rawls in regarding justice as "the first virtue of social institutions" (Rawls 1971, 3). Thus, although I shall sometimes talk of morality, justice should be taken as the highest order concept in the application of morality in political philosophy.

The starting point for a constructivist moral theory, then, must be with the personal perspective. A moral theory must address the question of what morality is from the perspective of a self-reflective agent; it must address the question, "given that I have value for myself what is the nature of my relations with others?".[6] It must not only give an account of the content of morality, however, but - if it is to answer the question, "why should I do the right thing?" - it must also provide the agent with good reasons to find moral injunctions compelling.[7] We begin with a self-conscious social agent possessed of language who is already a member of a community. The agent stands back from his social ties and questions the norms and practices of the community of which he is a member. Thus, although there is a sense in which "the constructing is to be done by a theorist",[8] it is done from the perspective of an

---

[6]That is, each person must consider her relations with others in the reflective stance not from a position of objective value, but from a subjective one, in which although the individual claims value for herself he does not claim value *in* herself and is thus not committed to recognise the value of others, (see Charvet 1981, 157-61; Charvet 1995). Assuming that the agent in the reflective stance claims objective value is the mistake made by Gewirth (1978). For a good, brief, account of Gewirth's argument see Moore 1993, Chapter 2. Moore makes a similar criticism of Gewirth: "It does not follow from the fact that something is good from a subjective standpoint, i.e., good from the standpoint of subjective desire-fulfilment, that it is good from an objective (or inter-subjective) standpoint; for example that it has the characteristics or qualities which make it objectively valuable or worthwhile. And it is the latter kind of claim which Gewirth needs in order to claim that their are duties incumbent on each person to enable or ensure that others secure the goods of freedom and well-being". (Moore 1993, 24). Nagel likewise commits this error: "some of the most important [things in your life] have to be regarded as mattering … so that others beside yourself have to take them into account". (Nagel 1991, 11).

[7]A point emphasised in the title of Nelson 1990: Morality: What is in it for Me?

[8]Barry 1989, 266.

agent reflecting on his commitments, interests, etc. Any answers must, then, take the form of answers for such an agent.[9] It is important to note that the agent is already engaged in social interaction, he is asking what the content of his moral commitments is, and whether it is rational to acquire commitments that he does not already recognise, and maintain and cultivate those that he does.

Of course, such a question is only available to a self-conscious individual able to distance himself from the community which formed him. An individual formed without social interaction (should one exist) could not take up this question not only because, *ex hypothesi*, questions of his relations with others would not occur, but also because he would be incapable of determining his conception of the good as a self-conscious rational being without either language or standards of rationality both of which are socially engendered; he would not in fact be a self-conscious rational being. Likewise the individual, even if brought up in a society, has to be capable of formulating the question, of achieving the distance between self and community. The absence of an external authority distinguishes the theory presented here from the classical contractarianism of Locke, and less explicitly, Hobbes and Rousseau,[10] and the achieving of distance between the agent and the community distinguishes it from some recent

---

[9]Cf. Williams 1985, Chapter 1.

[10]Kant could also be included in this list although he replaces natural law and the reliance on God with the claims of transcendental reason.

communitarian thought.[11] In the past (and perhaps in certain very simple societies), either the gap between the individual and moral perspectives was insufficient for the individual to question the norms and practices of the community - his identification with the community was, in Hegel's term, "naïve"[12] - or the source of authority for the community's norms and practices was posited in an external authority, God. In the former case the motivation problem does not arise[13] and in the latter it can be overcome by arguing that the agent, as possessor of an immortal soul, has good reason to be moral because of the chance of posthumous rewards and punishments.[14] As

---

[11]Something like the claim that the individual cannot escape his community so as to reflect upon it is often, and perhaps rightly, the interpretation given to the "embedded self" arguments of some communitarians (most notably Sandel 1982, see, e.g., 150). If the claim is this strong we surely have good reason to reject it, as Kymlicka puts it, Sandel "violat[es] our deepest self-understandings … . We do not consider ourselves trapped by our present attachments, incapable of judging the worth of the goals we inherited or ourselves chose earlier." (Kymlicka 1990, 213). It is a matter of reasonable dispute, however, whether Sandel or other leading communitarians (Taylor, MacIntyre and Walzer), are really making so strong a claim, see *infra*, note 13, although if they are not it is difficult to see what their criticism adds up to.

[12]See esp. Hegel 1977, §§464-476; also Hegel 1974, Vol. 1. As Alan Wood puts it: "until the rise of the subversive idea of subjective freedom in fifth-century Athens, the distinction between different interests, as something that might be mutually opposed, was not a natural one to draw for people living in the naïve harmony of Greek culture." (Wood 1991, 57)

[13]It seems to me fantastic to suggest that any society (past or present) could be so monistic as to eradicate the individual perspective. Rather, the motivation problem does not arise because when conflicts between individual interest and the interests of the community occur the agent gives precedence to the community. If this is a goal of communitarianism - and my suspicion is that it is more often imputed to communitarians by critics than expressed by communitarians themselves - we have good reason to reject it, for the outcome would be a society of unreflective moral cripples.

[14]In order to avoid the claim that one's reason to be moral is fear of punishment or desire for reward, an additional claim might be that

inheritors of the Enlightenment we have to recognise that the source of moral authority does not lie with God, and that morality and the motivation to be moral both have to be grounded in human reason and will.

33. The Personal Perspective

I have suggested above that if the separation of the moral and personal perspectives is to be overcome, the agent must come to understand and pursue his self-interest through, rather than in opposition to, the moral norms. I intend now to show how this is possible by presenting a "true" constructivist theory,[15] that is, a theory that does not rely on a prior commitment to a substantive moral position. In accordance with the nature of constructivist theory and the characterisation of morality given here, I shall present this in the form of a contractualist argument. This is not to say, of course, that we are to imagine society and morality as having come about through a contract, or that anything like a state of nature from which a moral community emerged has ever existed. The original position is not an historical event but a thought experiment; specifically, as I have argued above, the agent needs to distance himself from the community that formed him and question the nature of his attachments. It is this

---

the agent better flourishes, or fulfils his own nature, by behaving in accordance with his creator.

[15]To be compared with the *faux* constructivism of Rawls and cognate approaches.

self-reflective stance that can be characterised as putting oneself into an original position.[16]

Let us, then, begin with the personal perspective. If the agent is to overcome the separation of the moral and personal perspectives, and endorse and maintain himself as a moral being, it must be the case that the identification of his self-interest with that of the community is, in the first instance, in his interest as this can be conceived independently of the community. That is, when the agent takes up the reflective stance he must first ask himself whether it is in his interests to enter co-operation at all. Of course, if he is to do this, he must conceive of his interests as if he were not engaged in co-operation, that is independently of his social formation. In such terms the only sense that can be given to the individual's interests is to consider the acquisition of goods which are themselves independent of social formation; that is, goods which are naturally required by the individual *qua* human being. These are broadly the goods of food, sex, shelter, etc., and the freedom necessary to gain access to these, which are required by humans to flourish as separate individuals. A necessary condition for the agent's

---

[16]The question the agent asks is whether she should endorse, maintain and cultivate her moral dispositions, and this can be read in a manner that introduces the idea of consent; a notion that is often confused in the contractarian literature. In this sense the agent's consent is a running endorsement or rejection of certain dispositions she has as the person she is and given the type of person she wants to be, see §40. Rawls's strains of commitment test can, likewise, be read in this manner, that is, as a running demand on the agent that she be able to say "yes", (see Appendix A, §B) Ivison MS makes a similar point and I am indebted to his discussion in formulating my own ideas.

identifying her self-interest with the communal interest must, then, be that the community provides secure access to a greater share of these goods than the individual could obtain on her own. This is clearly the case for most individuals in many societies and for others the co-operative surplus, in terms of both quantity and security of access, is of enormous proportions.

The agent's self-interest, however, cannot be adequately understood in these narrowly individualistic terms. The agent, taking the reflective stance, realises that her flourishing depends upon the co-operation of others in that it is only insofar as there is mutual recognition of terms of co-operation to govern interaction between persons that co-operation and the co-operative surplus can be realised. The agent has reason to will the existence of terms of co-operation through which the co-operative surplus and her self-interest is served, and in doing so her self-interest becomes tied to the interests of her co-contractors. This is because in willing the existence of terms of co-operation she wills the existence of terms which are to the benefit of all the co-operating members (if organising the pursuit of their good through the terms of co-operation were not in their interests they would have no reason to co-operate). This, then, provides the idea of a common good through which the self-interest of each is aligned with that of every other. The terms of co-operation are not moral terms, however, the agent's self-interest merely coincides with the self-interest of other agents; each

views the terms as a necessary means to the realisation of her ends.

The idea of agents coming from the personal perspective to will terms of co-operation can be contrasted with the Lockean (and Nozickean) idea of individuals coming to the contract as possessors of absolute rights[17] and the ensuing problem of finding a mechanism through which to mediate the claims of each individual with respect to those rights. The Lockean scenario cannot overcome the problem that when conflicts arise between rights holders in the use of their rights each has a well grounded claim in his possession of absolute and natural rights. Any mediation of those rights claims through a collective determination of the content of the rights undermines the individualist core of Locke's account. If we reject the external account of rights as guaranteed by God, or derived from transcendental reason, then the terms which govern co-operation must be grounded solely in the wills of the contractors; each has reason to will the existence of terms of co-operation and to understand their self-interest as tied to the maintenance of these terms, and this is the theory I am proposing here.

This is only a necessary condition, however, because whilst it provides a good reason for the individual to co-operate, it provides good reason to enter co-operation on the best

---

[17]"Individuals have rights, and there are things no person or group may do to them (without violating their rights)." Nozick 1974, ix. There are, of course, the Lockean provisos, however, I do not take these to alter the fundamental character of the natural rights to life and liberty. See Locke 1960, §§25-51.

terms available, and if this is the case then the community could only organise around principles of justice as mutual advantage, an outcome I have shown to be unsatisfactory in Chapter 4. The idea that the agent's self-interest is tied to the interests of others through the existence of terms of co-operation cannot exclude those terms being based on mutual advantage because it does not tell us about the content of the terms beyond that they should be to the benefit of everyone.

One reason for the contractors to reject principles based on mutual advantage might be the instability of justice as mutual advantage discussed above;[18] if adherence to the terms of co-operation is contingent upon its being to the advantage of the agent compared with free-riding, then the system of co-operation would be unstable as each considered in every case whether her self-interest would be better served by acting in accordance with the terms of co-operation in that case. If these terms are to be stable it must be the case that they appear to the agent not as injunctions against which her self-interest is constantly pressing, but in the form of authoritatively binding norms. That is, they must be such that the agent comes to understand her self-interest as best pursued in accordance with the terms of co-operation. Stability cannot be achieved if the individual conceives of these terms as only contingently binding, the decision to obey being determined

---

[18]See §25.

by self-interest in each case, and stable co-operation is in the self-interest of each co-operator.

In order to realise the self-interest of each through the founding and cultivation of stable terms of co-operation such terms must be accepted by all co-operators. This means that they must exclude the use of force and fraud to improve relative bargaining strengths. To avoid the instability of mutual advantage and to arrive at an agreement to co-operate on moral terms, therefore, each must agree to conceive of his self-interest in abstraction from his relative bargaining power, and, thus, to agree to terms from a position of impartiality; the powerful must not exploit the weak and the weak must not band together in an attempt to become a powerful and exploitative class.[19] Under such terms the contractors will base their agreement on a principle of fundamental equality.[20]

---

[19]I am assuming that no fundamentally hierarchical society could remain stable over time because the oppressed class would eventually rebel. To an extent this relies on one of the other conditions of constructivism - that the agent is able to distance himself from his community - as I take it that the most successful hierarchical communities are those in which the oppressed class believe that the hierarchy is legitimate because determined by God or nature or some such external authority. Once the oppressed group realise that the grounds of the norms governing co-operation lie only in human will they will reject their oppression.

[20]John Harsanyi (1982) has suggested that the outcome of the impartiality requirement (with respect to Rawls's theory), will be maximising average utility (Rawls considers this in A Theory of Justice (1971), §27). Average utility cannot be accepted here, however, for two reasons. First, average utility is unstable because it cannot meet what Rawls calls the "strains of commitment test" (see Appendix A, §C for a discussion of this test); that is, it asks some "to accept the greater advantages of others as a sufficient reason to for lower expectations" and this is "an extreme demand". (Rawls 1971, 178) Second, we will see that the requirements of moral agency include the endorsing of a deeper sense of equality than is compatible with treating some less well, in order to benefit others.

Precisely what this commitment entails is discussed below, however, for the moment it should be clear that the appeal to stability is going to be insufficient motivation for the agent to ignore his relative bargaining strength and endorse a principle of equality. In the appeal to stability each agent is still in a position where he understands his self-interest from an individual perspective, best served when free-riding subject to the limit that such free-riding should not threaten stability. In short, we have not yet managed to give the agent a decisive reason to integrate the moral perspective into his personal conception of his good. It cannot be a sufficient reason for the agent to maintain himself as a moral being that the society is more stable if he, and everyone else, does so. Indeed, it could not be, because the agent would have to accept that he is maintaining himself as a moral being - that is as a being whose his self-interest is formulated through, and mediated by, the principles of justice - for purely self-interested reasons. The agent would be acting in "bad faith".[21]

It is, of course, a *necessary* condition of the agent's contracting that his self-interest is better served, as it

---

[21]This is the issue that dogs Hobbes. Hobbes has the basic structure of the problem right, or so it seems to me, however, he cannot explain that motivation of the agent to commit himself to the agreement without recourse to an implausible *summum malum* - the fear of violent death - and to a God who works through natural laws and sanctions. A contemporary Hobbesian, David Gauthier, facing the same problem is similarly reduced to an implausible empirical claim; that human motivations are sufficiently transparent to others to provide a sufficient reason to act morally, see Gauthier 1986, Chapter 4.

is that the society is stable. I shall argue below that it is also necessary that the society take the form of a coercive community, because the agent's participation is conditional on the agreement of others to commit themselves in a like manner. Taken together, however, these do not provide a sufficient reason for the agent to maintain himself as a moral being through theorising his relations with others in accordance with a principle of fundamental equality. What else can we offer?

34. Be Moral!

The motivational question we are facing is what decisive reason can be offered to the agent to theorise his relations with others on a basis of fundamental equality? I have argued that self-interest and the stability problem provide necessary but insufficient reasons. Is there an additional reason that can be offered to the agent which will be decisive for that agent? The negative answer we have to give to this question is no doubt disconcerting, however, it must be faced if we are to avoid importing false metaphysical claims into the account of moral motivation.

Perhaps this negative answer should not surprise us. Reasons for being moral are notoriously difficult to find, not least because it is not clear what form such reasons could take.[22] I have argued above that self-interest,

[22]See §32.

conceived narrowly in terms of getting more of certain goods, provides *a* reason to be moral. This is insufficient, which is in a sense reassuring, because it does not seem to be the right kind of reason anyway; if one raised one's child to be good only by rewarding it with a £1 every time it was so, one should be legitimately worried that the only reason the child has for being good is not a moral one. Likewise, I will argue below that the function of punishment is not merely to encourage the thought that one should not do X *only* because doing X will result (if caught) in certain bad consequences.[23]

Beyond self-interest, then, is there some other universal standard of practical reason to which we could appeal to offer the agent a decisive reason to be moral? Rawls, considering the same question, invokes the two most well respected options.[24] First, a Kantian argument that in acting in accordance with the demands of justice we realise our natures as free beings and, thus, express our independence from the natural world of causation:

> "The desire to express our nature as a free and equal rational being can be fulfilled only by acting on the principles of right and justice as having first priority.

---

[23]The structure of this paragraph owes much to D. Matravers 1995, 4MS.

[24]See Rawls 1971, 571-72; Appendix A, §C. There is nothing particularly valuable or interesting about Rawls's discussion of this point, I refer to him here and below because he neatly captures both positions. What is interesting is that a liberal theorist, concerned with the thinnest possible theory, should use two arguments derived from very different traditions: orthodox, transcendental Kantianism and neo-Aristotelianism.

> … it is acting from this precedence that expresses our freedom from contingency and happenstance."[25]

The agent, however, has good reason to reject this motive for it rests on an implausible metaphysical argument.[26]

Rawls also offers a second argument which he associates with Aristotelian virtue ethics: "The Aristotelian Principle runs as follows: other things equal, human beings enjoy the exercise of the realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity",[27] and, claims Rawls, "it follows from [this] Principle (and its companion effect), that participating in the life of a well-ordered society is a great good".[28] On the face of it this argument is open to two, very different, interpretations. On the one hand, Rawls could be endorsing a neo-Aristotelian virtue perfectionism. Such an argument would be that the agent has a decisive reason to be moral because in so doing he realises his nature - his *telos*. This, however, is unsatisfactory for we have no reason to think that humans share a *telos*, let alone one which is uniquely fulfillable through living a moral life. What, however, if we drop the teleological element and claim that the agent has a decisive reason to be moral

---

[25]Rawls 1971, 574.

[26]The role of the Kantian Interpretation within Rawls's theory is examined in Appendix A, §C.

[27]Rawls 1971, 426.

[28]Rawls 1971, 571.

because in doing so he shares in something which he will find of value?

Such a position would be a long way from Aristotelianism proper, and could be developed in quite complicated ways. We could say, with MacIntyre, that an agent needs to feel "at home" in the world - as if his life has purpose and meaning - and this is only achievable if the agent is able to make sense of her life in terms of "a narrative order". Such an ordering requires that the agent is situated in a moral tradition, engaged in certain practices in accordance with practice specific standards of excellence.[29] This position, however, does not seem to be compatible with the personal reflective stance taken by the agent. It might be the case that the agent would feel emotionally happier and more secure (in the sense that they would feel that their life had a point, or that they had, after all, been created *for* something), in a moral tradition, but that is possibly true of all moral traditions (irrespective of their plausibility). However, once one has asked the question, "do I endorse the tenets of the moral tradition which informs my way of life?" then the problem of modernity has already struck; a problem, that is, if one cannot find grounds for the values that define one's tradition.[30] It

---

[29]See MacIntyre 1984, especially Chapter 15.

[30]I, for example, used to complain to my girlfriend, when this thesis was proving particularly gruelling, or when my ability to "get on with life" was being impaired by my habit of analysing it, that things would have been much easier if I were a Klingon or an orthodox Catholic. I take it, however, that the accompanying thought that one can't *decide* to believe, or decide to endorse the values of a warrior culture, unless one considers that there are good grounds for

is no good *pretending* to be a Klingon, or a Catholic in the hope that one will feel reassured as part of some greater whole, for the attempt is bound to be unsuccessful.[31] It may be that the agent has reason to maintain herself as a moral being because given the type of person she is, she will be happier if she does so. This might be true (although it may well not be; sin, after all, is sometimes jolly good fun), but it will not do; the question the agent is asking, after all, is "do I have reason to be the sort of person I am?"

What, then, is left? I have shown that the agent has some reasons for theorising his relations with others in accordance with a principle of fundamental equality but I have admitted that these reasons are not decisive. Further, I have rejected the two most commonly invoked "turbochargers"[32] to get the theory over the motivational hurdle. Is it, then, the case that as Williams and Anscombe have argued,[33] the sorts of questions I am asking are the "ethical progeny"[34] of distant times when we had

---

thinking that the ranking of goods in these traditions is the right one, was right, albeit occasionally depressing.

[31]One cannot fool oneself if one knows that that is what one is trying to do.

[32]I have borrowed this term from Brian Barry; a "turbocharger … is an optional extra that provides additional power but comes at a stiff price." Barry 1995b, 13MS.

[33]Anscombe 1958, 1: "the concepts of obligation, and duty - *moral* obligation and *moral* duty, that is to say … ought to be jettisoned if this is psychologically possible; because they are survivals, or derivatives from survivals, from an earlier conception of ethics which no longer generally survives, and are only harmful without it." Williams 1985, especially 160-63.

[34]The term is borrowed from Pence 1991, 252.

good reason - or thought that we did - to believe in a punishing God or an enchanted world? Yes and no. Yes, in that there is no decisive *reason* that we can give to the agent to be moral; no, in that this does not mean that we have to remain silent. Given that the agent is already a social being engaged in co-operation we can press upon him the injunction, "be moral", not for any decisive reason but because in evaluating, endorsing and acting on the principles of justice he expresses himself as a being capable of taking responsibility for his choices and acts. In ordering his life in accordance with the demands of justice he gives his life a "narrative unity" of the sort MacIntyre so desires, and he expresses his independence of his contingent desires and preferences. It is, so to speak, a Nietzschean act of self-creation. Of course, it would be just such an act - as Nietzsche recognised (and pressed upon us) - for the individual to reject morality outright and reject the constructivist scheme being advanced here. Nietzsche calls for the individual to recognise that the death of God has freed the individual to reject morality, what we must respond is, first, that such a course of action would be prudentially irrational - rejecting co-operation would be to the detriment of the agent's self-interest - however, we have to recognise that this cannot be a decisive reason for the agent. We must, therefore, add that the death of God has merely made endorsing morality the responsibility of the agent. Here we have reached "the limits of philosophy" with the injunction "be moral", however, we have not exhausted our

resources. Although we cannot offer a decisive reason to the agent, we can attempt to bolster our injunction by showing the agent that the moral life is an attractive one, and for this we have much of the canon of great literature at our disposal.[35]

Before going on to consider the theory presented here any further, I want to say something to meet the objection that what is being proposed is not a theory of morality at all, but simply one of co-operation. To an extent the intuitive drive behind this objection lies in a realist conception of morality and I have rejected such an account for this thesis. Instead I have defended a particular idea of morality that is more than simply any system of terms of co-operation.[36] However, there is one particular facet of the theory outlined above - in addition to the absence of a decisive reason to be moral - that might cause considerable unease. This is the fact that one reason the agent has to be moral is that it is in his self-interest. As I noted above, this contradicts an intuitive feeling we have about morality which is that one ought to obey moral commands simply because they are such commands, independent of one's interests. However, to a degree this objection misses the point; it would of course be an odd theory of morality that said that when considering which of two acts to perform from a moral point of view, the agent ought to do the act

---

[35]Rorty makes a similar point in Contingency, irony, and solidarity (1989), see especially Part III; See also D. Matravers 1995.

[36]See §32.

that best serves his self-interest subject to the stability of the whole, (this sort of problem is what underlies our intuitive negative reaction to utilitarianism); but that is not the theory I am advocating. What I am considering is an agent who is asking whether he should endorse, maintain and cultivate being a moral person,[37] that is as someone who is not considering whether to do this or that act but rather whether to be a moral character. As a moral character the agent is predisposed to perform moral acts just because they are moral, and thus the objection is met. In this sense the theory I am proposing can be identified with those of "character" rather than "act" morality.[38]

35. The Terms of Co-operation

In summary, a true constructivist scheme grounds morality in the wills of the participants; each agent, putting herself into a position to reflect on her relations with others - the equivalent of an original position - endorses her dispositions to behave morally, she commits herself to being a moral agent. By doing so she realises her capacity for full moral autonomy, she takes responsibility for the commitment to be moral as an act of will grounded in nothing other than that will. She does this first, because a stable system of co-operation is in her interest, and second as an act of self-creation; a method of giving unity

---

[37]And, likewise, reject those dispositions that he has towards immorality. See *supra*, note 16.

[38]For a brief but useful account of the differences between these see Ihara 1992.

to her life and choices *as a co-operative being.*[39] In willing the system of co-operation she wills norms of co-operation which benefit all contractors, in willing moral terms of co-operation she goes beyond self-interest and commits herself to structuring her life around moral norms. The community is, then, grounded in nothing other than the collective wills of its members to organise in pursuit of their common good, and morality is the organising form through which this is achieved. In addition, insofar as the norms of the community, language and standards of rationality are socially engendered the contractors must recognise that they owe a debt as the people they are to the community: just as they are co-creators of it so they are co-created by it.[40]

Given, then, that each agent conceives of her relationship with her co-contractors as one grounded in an agreement to abide by norms which every agent could accept, that is norms which are the product of theorising from a position abstracted from relative bargaining strengths, what would the terms of such an agreement be? It might be thought that there is nothing very much that can be said about these terms given the nature of the theory I am defending, other than that they should be subject to a principle of

---

[39]The initial unity given by the agent to her life and choices is prudential. Prudential rationality commits the agent to conceive of herself as one and the same being over time, and to act accordingly. (This is not something that I have defended here.) Moral rationality comprehends this unity but provides for a further level in which the agent comes to understand herself and her choices as a social being engaged in co-operation.

[40]See Charvet 1981, 158.

equality. This is because given that the norms of co-operation are grounded in nothing other than the will of each to co-operate on moral terms, these terms have no binding abstract form prior to the agreement, rather, they must be given a determinate content in the agreement of the contracting parties. In this way constructivism avoids the charge - and problems - of "atomistic individualism"[41] because the content of morality is given by collective determination (subject to the principle of equality), and thus avoids the conflict of individual rights that raises so many problems for liberal individualists.[42]

Nonetheless, although the determinate form of the principles is the subject of collective determination there are a number of questions one can address in the abstract, and the eventual content of the principles will have to give concrete embodiment to any claims that can be thus established. We can usefully divide the subject of distributive justice into three questions, although the distinctions between them should, perhaps, not be held too rigorously. The questions are over *what* is to be distributed, *how* it is to be distributed, and *to whom*.[43]

---

[41]See Taylor 1979.

[42]If each individual conceived of himself as entering co-operation already possessed of rights then we could not avoid the conflict between individuals each of whom makes a claim to absolute rights, see §33.

[43]I am here following Fishkin 1992, 5.

Concentrating on the first of these questions, we have seen that each individual agrees to co-operate in the first instance because co-operating serves his self-interest as this can be conceived independent of his social formation; that is, in terms of the acquisition of goods which are themselves independent of social formation. These are the goods of food, sex, shelter etc., and the area of negative freedom needed to gain access to, and use, these goods.[44] This, however, only makes sense as part of the thought experiment in which the agent is engaged when putting himself in an original position, reflecting on his relations with others *as if* he were independent of those relations. As a particular, concrete individual, formed and engaged in such relations, the agent does not desire the abstract form of these goods, rather, he desires these goods in their concrete social form, that is, in the form of such things as houses, stable monogamous relationships and *paté de foie gras*. The "resource" rights of the contracting parties, then, are derived from the abstract claim to access to these basic goods, although the form of such claims is in terms of concrete resources. The "liberty" rights claims stem from two inter-related sources, the negative freedom required by the agent as one of the "goods" better served by co-operation and the requirements of the principle of equality, for treating people as equal requires recognising the claims of responsible agency, that is giving them an area of negative

---

[44]See §33.

liberty in which they may act, and thus actualise their particular interpretation of the right way to live.

The principle of equality is, of course, an answer to the question of how we are to distribute the "what". However, both these answers (to the questions of "what" and "how") are only partial, there are many ways of conceiving of such resource claims and of distributing things equally.[45] For example, one could distribute resources in accordance with equal preference satisfaction or in terms of actual quantities, and one can treat everyone equally, likewise, by equalising levels of preference satisfaction or quantities of goods.[46] It is also claimed that we treat everyone equally by giving everyone equal quantities of goods and letting them get on with it or by continual redistribution to correct the greater successes of those who do well with their initial allocation, i.e., by making the just distribution ambition sensitive or by making it correct the outcome of differential levels of ambition.

The intuition that drives the argument for equalising preference satisfaction rather than resources can be captured very easily in the following (well used), example:

---

[45]Kymlicka (1990) is an entire book dedicated to the idea that all the modern schools of political philosophy can be read as attempts to give substantive form to the injunction to treat people equally. Fishkin believes that the existence of these different interpretations of what and how to distribute undermines the project of formulating a liberal theory, see Fishkin 1992, Part I. I have criticised this argument in M. Matravers 1995.

[46]It should be noted that I have already rejected the maximising of average utility, see *supra*, note 20.

two individuals, one of whom likes oranges but dislikes apples and one whom likes apples and dislikes oranges, are each offered equal baskets of apples. This, welfarists claim, cannot be treating these agents equally because it is clearly unfair to the agent who dislikes apples; what is needed is a distribution such that each agent ends up with the same net welfare gain or loss, in this case, equal sized baskets, one containing only apples and the other only oranges. Endorsing a welfarist solution to the problem of what to distribute has a number of problems, however, amongst them the problem of "expensive tastes" (in which people with expensive tastes would have to get much more of the resource pool because providing them with the same resources as someone with less expensive tastes would not equalise preference satisfaction[47]), and the problem of inter-personal preference measurement. But, more fundamentally there just seems to be something wrong with the whole idea. As Barry puts it:

> "there is something mildly crazy about the idea that an ideally just society would be one where people who needed champagne and caviar to get to the average level of consumer satisfaction would get more money or where the adherents of some killjoy religion would have to be allowed to bring everybody else down to their own level by stopping others from enjoying themselves."[48]

---

[47]The standard example - which says something about the nature of academia - is of someone who has a discerning wine palate; the argument is that such a person gets very little preference satisfaction from a cheap bottle of wine, and thus to equalise welfare it is necessary to give the agent more resources.

[48]Barry 1991, 154.

Clearly, what is really driving the welfarists' intuition is the idea that people are treated equally only if their preferences are equally satisfied because the content of your preferences is not your fault.[49] I am assuming that this is not the case for all preferences - or, at least, if it is the case then it is still compatible with a notion of responsibility. Where it is true that the content of the agent's preferences are such that she cannot be held responsible for those preferences, the principle of distribution would have to take this into account.

Likewise, in treating everyone equally we have to allow this idea of responsibility to make resource distribution ambition sensitive and thus allow for a degree of letting people "get on with it" given an equal starting point and a level playing field. Clearly, while being ambition sensitive the principles would have to take into account that not all differentials over time would be the outcome of ambition, some would be a consequence of bad or good luck and these would need to be corrected through a principle of compensation. Furthermore the requirements of equal agency mean that we must guarantee to each individual as much freedom as is compatible with a like freedom for all other co-contractors. For the moment it is enough to say, then, that the content of the principles must abide by the requirements of treating agents as equal with respect to the distribution of resource claims and claims to an

---

[49]The literature on this question is extensive and this is not the place to discuss it in detail. Useful discussions can be found in Barry 1990, xlvii-xlviii, 43; Cohen 1989; R. Dworkin 1981.

area of negative freedom through protection of life, liberty and legitimately held property.[50]

This discussion of the terms of co-operation - the principles of justice - has been very brief, and there is clearly a great deal more that would need to be said if that were the concern of this thesis,[51] however, I am concerned not so much with the precise content of the principles as the form they take and the grounding they have. In addition, it is clear that the principles would not be very different from those of any impartialist theory, and I would be happy, for the moment, merely to endorse the principles established by Barry;[52] in essence, a principle of equality of distribution (perhaps with some addition such as a pareto condition),[53] a principle of responsibility and one of compensation.

I intend to say no more about the principles of justice. The concern of this thesis is primarily with the grounding

---

[50]I have left the third question - *to whom* - relatively undiscussed, implicitly arguing that those who count are others in the society, that is, other co-operators. This is because I am concerned here with justice, coercion and punishment in a society (and in a generation).

[51]Not least, the account of autonomy and responsibility that underlies the claims made in the previous paragraph would have to be given.

[52]See Barry 1995c for a brief introduction to the principles he thinks would have to accompany a theory of justice as impartiality, although these would have to be amended to take into account my contractualist (and, in this sense, anti-cosmopolitan) limits. See also Charvet 1995.

[53]That is, that departures from equality can be justified if no-one is made worse off and at least one person is made better off.

of justice and the rationale of coercion, however, it is necessary that the impartialist outcome of this theory is recognised, not least because the deep injustice of contemporary societies throws up the challenge of whether just punishment is possible in an unjust society.[54] There is, however, one important aspect of the theory that I am presenting which I have so far ignored; this is the assurance problem and it is now time to turn to this problem as the final element in the constructivist theory.

36. Constructivism and The Assurance Problem

It is only rational for each person to make the moral commitment to regulate her conduct in accordance with the rules agreed by all (subject to the principle of equality), if every other person so agrees. This is because it is clearly not in the interests of the agent to so commit herself if she could, rather, enjoy the benefits of co-operation without regulating her own behaviour, that is, if she could free-ride on the moral behaviour of others. Nor is it is in her interests to co-operate on moral terms if she believes that she will suffer as a consequence of her moral behaviour at the hands of another agent who is free-riding. This is the assurance problem, or, as we have seen, the problem Hobbes calls that of providing a condition of "sufficient security".[55] Hobbes recognises this as the fundamental problem of the state of nature, and

---

[54]I discuss this question briefly in §44.

[55]Hobbes 1651, Part I, Chapter 15, 215.

as the reason for the individual to authorise the "common power" - the "sovereign" - without whom the condition is unlikely to be satisfied.[56] Hobbes also recognises that so long as others are not willing to solve this problem by transferring their rights to the common authority, no individual can be bound (by reason), to do so:

> "For as long as every man holdeth this Right, of doing any thing he liketh; so long are all men in the condition of Warre. But if other men will not lay down their Right, as well as he; then there is no Reason for any one, to devest himselfe of his: For that were to expose himselfe to Prey, (which no man is bound to) rather than to dispose himselfe to Peace."[57]

It is likewise recognised by a more modern theorist of justice, John Rawls:

> "The assurance problem … is to assure the cooperating parties that the common agreement is being carried out. Each person's willingness to contribute is contingent upon the contribution of the others. Therefore to maintain public confidence in the scheme that is superior from everyone's point of view, or better anyway than the situation that would obtain in its absence, some device

---

[56]Hobbes 1651, especially Part I, Chapters 13-16, 183-222. See, for an interpretation of Hobbes based very much on the assurance problem, Barry 1968. "Contracts, Hobbes tells us, are only conditionally beneficial; it only pays me to do my part given that you do yours as well. Therefore, it is not obligatory for one party to perform his part if he has a 'reasonable suspicion' that the other party will fail to do his. The key is trust; in the absence of a 'common power' over the contracting parties, the larger the element of trust involved, the less chance there is that a contract will create an obligation to perform." (Barry 1968, 123).

[57]Hobbes 1651, Part I, Chapter 14, 190.

> for administering fines and penalties must be established."[58]

The difference between the constructivist theory I am advocating and that of Rawls and other *faux* constructivists is that because they import a substantive prior moral commitment to the fundamental equality of humans, the ethical injunctions that emerge from their construction hold independently of the construction; this is the claim made above that in *faux* constructivism the construction is left doing little or no work.[59] They introduce the assurance problem only after having established an unconditional moral theory. Having rejected any prior commitment to a substantive moral position, the constructivist theory I am offering makes morality binding on the individual only if the assurance problem is solved; morality is contingent on the provision of the condition of sufficient security.

This conclusion may, again, disconcert. We intuitively feel that morality is absolute, or limited only when there is a "tragic" conflict between two moral injunctions, in which case we have to choose to obey one or other and "feel regret at the deepest level" at the dilemma,[60] however, it is nothing other than the corollary of the deeper contingency of the constructivist moral theory. Morality, on this theory, is conditional on the satisfaction of the

---

[58]Rawls 1971, 270.

[59]See §§26-28.

[60]See Williams 1979 reprinted in Williams 1981, 74.

self-interest of each contractor, there is no appeal to a set of moral norms valid independently of self-interest, and thus there is a deep conditionality built into the authority of the moral norms. Such norms do not bind on individuals if those individuals have no reason to enter co-operation (because they could better satisfy their self-interest outside of society), or if they simply reject the co-operative option. This deep conditionality reappears in the assurance problem; morality is binding only in conditions of sufficient security.

The rationale of coercion, then, is grounded partially in the assurance problem. In the next Chapter I shall examine this and distinguish between justified coercion and moral punishment. I shall argue that while the assurance problem provides the justification for coercion, it is only one, albeit a necessary, element in the account of morally justified punishment.

Chapter 7: Justified Coercion and Moral Punishment

37. The Assurance Problem and Justified Coercion

I have argued that morality is deeply conditional and that this conditionality is reflected in the fact that the terms of agreement are only binding on the individual - they only take the form of *moral* injunctions - if the condition of sufficient security is met. The terms of agreement become moral norms only if they serve the self-interest of each co-operator and if the co-operators agree to conceive of these terms from a position of impartiality; this they do as an existential act of commitment, an expression of their autonomy and potential as moral beings.

The consequence of such a position is that where the individual rejects the co-operative option no moral ties exist and a "state of nature" relationship between that individual and others holds. Where such an individual threatens the co-operative enterprise (or any member thereof), coercion may very well be required to restrain that individual. Coercion, here, is simply the protection of one's property and self (or the community's protecting of itself and its property), against an alien and, thus, there is no legal analogue in punishment.[1]

---

[1]This has implications for any attempt to "globalise" the contractualist theory being offered. The implication is that there are no moral commitments between contracting groups, or between those

Of course, the mere act of free-riding represents an attack on the very foundation of the community, and through coercion the community expresses its abhorrence for such action. This is because given that the community is grounded in nothing other than the wills of its citizens to co-operate on terms on which they would agree in a (suitably constituted) original position, and that it is this agreement, and action in accordance with it, that gives the terms their binding nature as moral principles, the community cannot tolerate free-riding. To admit free-riding would be to admit an agent as a beneficiary of co-operation, without imposing upon her the burden of conceiving of her good as aligned with the good of others in the flourishing of the community through the maintenance of norms of co-operation. Not to differentiate between free-riders and other members of the community would destroy the identification of the personal and moral perspectives by breaking the connection between the individual's flourishing and the flourishing of the community; the individual could satisfy her self-interest without regulating the pursuit of that self-interest in accordance with the terms of co-operation.

In this sense, the formation in the individual of a sense of himself as a member of the community is accompanied by a will to coerce others who claim the benefits of membership without at the same time regulating the pursuit of their

---

who contract and those who put themselves, or are, "outside" of morality, on this question see Charvet 1995.

self-interest in accordance with the principles governing co-operation. This will, likewise applies to himself if he turns out to be the free-rider. Of course, the individual can free-ride without claiming membership, and, as we have seen, it is open to the individual to reject the co-operative option and to reject morality.[2] Where this happens, as I have said, the community must defend itself, and this may require coercion. This is, however, not a moral issue. In fact, I believe that this is only a theoretical possibility and I shall, for the most part, therefore, address the question of coercion and punishment as directed against individuals who *do* claim membership of the community and thus who do have, in what might be a primitive form, a moral will. I shall defend this position later,[3] however not very much turns upon it; where the individual rejects morality and the co-operative option the community is justified in coercing him for its own defence, and that coercion will not be subject to the limits that I believe constrain the use of coercion in punishment. Perhaps a more interesting question arises where the community fails in its obligations and becomes unjust, and I shall make a few remarks about this difficult question in my conclusion at the end of the chapter.

Amongst the contracting parties, I have argued that the solution to the assurance problem provides the rationale

---

[2]See §§33-34.

[3]See §40, especially text accompanying note 32, and note 32.

for coercion.[4] However, I have also argued above, against the background of a theory of justice as mutual advantage, that a punishment system justified on these grounds is likely to yield counter-intuitive results; specifically, very harsh punishments for all offences, and relatively harsh punishments for minor (but very tempting), offences.[5] Indeed, there is an additional problem which is that if the assurance problem is the *sole* justification for punishment it may well not be necessary to punish everyone who commits a crime, or to punish like crimes in a like manner. This follows because the assurance problem is concerned with the *beliefs* of the contractors; it is a matter of assuring "public confidence".[6] In this sense the truth or otherwise of the efficacy of sanctions is only relevant insofar as it affects the beliefs of the population; it would be perfectly compatible with the assurance problem justification of punishment for only some offenders to be punished. A lottery, for example, in which only one in every three offenders was punished might be sufficient to satisfy the condition of sufficient security.[7] This problem is seldom recognised in the literature, although Rawls, for example, admits the possibility of some kind of deception in his brief answer to the assurance problem, (although it fails to recur when he comes to discuss the

---

[4]See §36.

[5]See §25.

[6]Rawls 1971, 270.

[7]Whether it would or not, of course, depends upon the conditions prevailing in the particular society.

role of sanctions in greater depth). Rawls argues that in order to meet the challenge of the assurance problem, "some device for administering fines and penalties must be established", his conclusion, however, is that "it is here that the mere existence of an effective sovereign, or *even the general belief* in his efficacy, has a crucial role".[8]

In order to clarify what is at issue let us return to a straight original position type choosing situation. The people in the original position would, I have argued, choose principles which have a coercive nature in order to solve the assurance problem and, thus, make the principles morally binding. This, however, is only sufficient to show that some degree of coercion is needed and justified; the precise degree and the distribution of coercion, however, are questions which, if the assurance problem is the only concern, are unlikely to be resolved in a manner that accords with our convictions about justice, although the degree of inequity will be determined by the precise conditions of the actual society.

One argument against coercing only some is immediately available; the people in the original position are not solely concerned with the assurance problem because they have to choose principles from a perspective of impartiality, that is, in accordance with the principle of equality. This means that in choosing principles which have a necessarily coercive element the people in the

---

[8]Rawls 1971, 270, emphasis added.

original position cannot choose to coerce some and not others for like infractions of the terms.[9] This means that at least one of the problems attached to the assurance justification can be met. As yet, however, this is simply a theory of justified coercion, it is not a theory of punishment.

38. Justified Coercion and "Moral" Punishment

The assurance problem, then, is the primary justification for the use of "hard treatment".[10] However, justifying a system of punishment requires more than simply justifying coercion. In agreeing to necessarily coercive terms of agreement, the people in the original position cannot simply agree to coerce people to satisfy the condition of sufficient security, for to do so would not treat the contractors as members of the community capable of making the commitment to be moral. If the offender is a member of the moral community then to coerce him simply to satisfy the requirements of the assurance problem is to use him as a means to the better satisfaction of others and this is

---

[9]For the argument against choosing average utility as the outcome of impartiality see Chapter 6, note 20.

[10]"Hard treatment" is a term of art denoting the imposition of suffering or deprivation, as against education or therapy. I shall argue below that there may be conditions in which the other functions of punishment - moral expression and moral education - require hard treatment, however, I think that this is unlikely to be the normal case, see §. Hard treatment is distinguished from the expressivist elements of punishment by Feinberg (1965), where he admits that although the two elements are intertwined they can be separated for the purposes of analysis: "we can conceive of ritualistic condemnation unaccompanied by any *further* hard treatment, and of inflictions and deprivations which, because of different symbolic conventions, have no reprobative force." (Feinberg 1965, 98)

incompatible with the fundamental commitment made by all co-operators, to interact on terms that would be agreed by fundamentally equal beings. If the offender is not a member then the community would be justified in not respecting that requirement, but, as I have said, I am treating that as a mere theoretical possibility.[11] Rather, in agreeing to a system of coercion, the people in the original position must address the offender as a member of the community, as entitled to equal consideration, and as potentially possessed of the moral will to live with others on moral terms.[12] Moral coercion appeals to this will, it addresses the offender as a moral being, and ultimately it is justifiable to the offender as such a being as emanating from his own commitment. That is, moral coercion cannot take the form of a mere appeal to the agent as a prudentially rational being - it cannot simply increase the price of certain actions - instead it must address the

---

[11]Of course, there are ways of thinking of such a possibility that makes it far more real; prisoners of war, for example. Here I do think, as do most people, that the nature of imprisonment is fundamentally different to that imposed as a legal punishment. The other possibility that springs to mind is a member of a radical terrorist organisation who does not believe that co-operation on any terms except the "true" ones is acceptable, and this seems to me a very real problem for traditionally religious states trying to convert to a secular, liberal, constitution, (Egypt and Algeria might be examples). Of course, I believe that, in these cases, fundamentalist Islamic terrorists are mistaken in their beliefs, however, I do not believe that this means that they do not see the advantage of co-operation on moral terms, they merely mistake the grounds and nature of those terms. Insofar as they do identify with morality one could, therefore, attempt to appeal to them and "convert" them to accept responsibility for morality, which is why I do not think anyone is beyond redemption. It is clearly not inconceivable that I am mistaken, in which case I think an argument for imposing limits on coercion can still be made because of the importance of law and punishment in setting the moral tone in the society, see §41.

[12]The compatibility between coercion and the requirement to respect the agent's autonomy is discussed below, §43.

agent as a potentially moral being, as someone who has, in this instance, failed to live up to his commitment to live with others on moral terms of co-operation

While this gives a moral dimension to coercion by linking it to the agreement to co-operate on moral terms, it does not, in itself, transform coercion into punishment. Punishment (as I am concerned with it in this thesis) is a legal response to illegality and nothing I have said commits the community to giving the terms of co-operation a legal form. In smaller societies in which each person is known to every other, the role of law and penal sanctions, as against positive morality and social sanctions, might be very limited. This raises the larger question of the relatively indeterminate nature of the terms of co-operation themselves. The terms of co-operation are abstract requirements that must find embodiment in the positive law/morality of any moral community. Clearly, the people in the original position are not going to attempt to give a specific content to every issue that arises in regulating co-operation. For example, although some determinate form must be given to the right to travel freely with respect to road use (at the most basic level there has to be an agreement as to which side of the road to drive on), such decisions can be left to an agreed decision making procedure.[13] For the moment I shall assume

---

[13]Given the principle of equality and the idea of agency, the form given to this decision making procedure must be one of democracy. In smaller societies there seems to be no reason not to endorse some form of direct democracy, in larger, more complex societies, a form

that the assurance problem gives rise to the need for some hard treatment and that the form given for the imposition of this hard treatment is legal; that is, I shall use the language of "offender", "punishment" etc.

The justification of punishment has, then, a number of elements. First there is the assurance problem which will be solved by varying degrees of coercion, depending upon the circumstances of the society. Second there is the function of expressing the abhorrence the community feels for free-riders, and finally, linked to the second function, there is the role of law as a moral educator - as re-enforcing the community's values and the injunction to be moral - as a creator of the citizens the community needs for its continued flourishing.[14] I now want to say something about each of these.

39. The Assurance Problem and the Quantum of Punishment

Providing the condition of sufficient security - solving the assurance problem - is the primary reason for imposing hard treatment. This means that part of the general justifying aim - and that part usually thought of as the most difficult to justify - has a basically preventative function.[15] While such a rationale has traditionally been

---

of representative democracy accompanied by stringent checks and balances.

[14]As I said above, on the theory being presented here the citizens are co-creators of, *and co-created by*, the moral community.

[15]"Preventative" is meant to include the functions of deterring others, deterring the offender and preventing the offender from

avoided by expressivist and educative theorists I cannot see that it can be: the whole penality system seems to reflect "a preventive design".[16] As Andrew von Hirsch puts it:

> "When the state criminalizes conduct, it issues a legal threat: Such conduct is proscribed and violation will result in the imposition of specific penalties. This threat surely has *something* to do with inducing citizens to refrain from the proscribed conduct."[17]

Of course, the extent and nature of the hard treatment necessary to solve the assurance problem is going to depend upon the society.[18] This is something recognised by thinkers as diverse as Hegel and Nietzsche;[19] it is

---

committing crimes during his punishment; in short, attempting to prevent the occurrence of future crimes.

[16]von Hirsch 1990, 275.

[17]von Hirsch 1990, 275-6, emphasis in the original. A similar point is made by John Charvet when discussing punishment as criticism: "The purpose of having a rule is to secure the general realisation of the conduct it prescribes. Therefore the existence of rules presupposes the possibility of affecting people's future actions. It is in this context that criticism operates, for it would indeed be meaningless and futile to criticise people for their actions, if such criticism never did or ever could have an effect on their future conduct. But since criticism is only relevant where rules exist, and since rules exist only where it is possible to affect people's actions by means of the rules, criticism is assured of having such general effects. Thus criticism does always in part look forward to the future actions of the rule-breaker and the other members of the community, as the Utilitarians have always insisted that punishment must." (Charvet 1966), 578. Cf. Primoratz 1989, 70-71.

[18]See von Hirsch 1985, 53; von Hirsch 1990, 275, 278.

[19]"The … magnitude [of the crime] varies, however, according to the *condition* of civil society, and this is the justification both for attaching the death penalty to a theft of a few pence or of a turnip, and for imposing a lenient punishment for a theft of a hundred and more times these amounts". (Hegel 1991, §218R, see also §218, §218A). "The 'creditor' always becomes more humane to the extent that he has grown richer… It is not unthinkable that a society might attain such a consciousness of power that it could allow itself the noblest luxury possible to it - letting those who harm it go unpunished. What are my parasites to me? It might say. May they live and prosper: I am strong enough for that!" (Nietzsche 1967, 72).

conceivable that in small societies in which the communal bonds are extremely strong, hard treatment could be replaced by, say, public stigmatisation. Elsworth Faris, for example, tells of a primitive society in which the mere fact of being reprimanded by a small, weak, old woman, is enough to cause extreme remorse in a brutal warrior (who has offended against the tribal code), accompanied by a desire to make reparations.[20] Likewise, it is conceivable that a society could be so unstable as to need extremely harsh penalties to solve the assurance problem. This raises a difficulty for the account of punishment being advocated here - a tension between the need for hard treatment to solve the assurance problem, and the degree of punishment needed to express censure - which needs serious consideration. I cannot properly address this problem, however, until I have considered the expressivist elements in this account.[21]

40. Punishment as Censure

Given the nature of the community as grounded in nothing other than the wills of the contractors to co-operate on the basis of norms agreed in a suitably constituted original position, and that a necessary condition of the agent theorising his relations in this manner is that it

---

[20]Faris 1914, 58; see also Tunick 1992, 78; von Hirsch 1985, 53; 1990, 278.

[21]Similarly, the reader may have noticed a tension between the use of hard treatment to affect the actions of agents and the requirement to address agents as responsible beings. This is discussed below in §43.

serves his self-interest, the people in the original position must reinforce this identity between self-interest and morality by expressing disapprobation at free-riding; free-riding strikes at the heart of the community by denying the necessary connection between self-interest and the regulating of the pursuit of self-interest in accordance with the terms of co-operation. When the individual commits an offence against those terms, then, she, as Hegel insisted, harms not just her victim but also the community of which she is a part.[22] The first function served in censuring the offender, then, is in reaffirming the principles against which the criminal has offended; as Igor Primoratz puts it, in language distinctly similar to Hegel's:

> "in expressing emphatic condemnation of the crime committed, punishment *vindicates* the law which has been broken, *reaffirms* the right which has been violated, and *demonstrates* that the misdeed was indeed a crime."[23]

In censuring the offender, however, the community not only reaffirms the rights established by the agreement, and the identity of the personal and moral perspectives that underlies that agreement, but also treats the offender as a responsible agent. Censure is, in Strawsonian language, a "reactive attitude", that is, it makes sense against the

---

[22]See §§11-13. The free-rider might deny this, and if she genuinely does so then she puts herself outside of morality, and we may justifiably coerce her as an alien, cf. *supra*, 11n.

[23]Primoratz 1989, 196.

background of holding the agent censured as responsible.[24] On the theory I am advocating, censure includes the claim that the agent, as a co-operator, in offending against the terms of co-operation is offending against her own will to co-operate on moral terms; a will which must include the denial of the appropriateness of free-riding.

It might be objected that the offender can hardly be said to have the will to co-operate on moral terms given that she is an offender; she has manifestly not co-operated on such terms. This, however, is to commit the mistake of thinking that someone who free-rides in a particular instance rejects co-operation and morality *tout court*. The community in censuring, appeals to the agent as a responsible being who has demonstrated, in a myriad of other ways, her commitment to co-operation on moral terms. It takes her "at her word" when she claims membership and benefits from the community, and imposes on her the "costs" of that membership when she fails to live up to its requirements, in so doing it reminds her of her responsibility and commitment to the moral life. Censure is, in this sense, retributivist, not consequentialist: we do not censure so as "to teach moral standards of self-

---

[24]Of course, a consequentialist determinist might argue that censure could be made sense of as a policy of social control, that is we pretend to blame people because that blame will effect the causal chain that determined that - or other - agents' future behaviour. In this case I would argue that either a compatibilist case could be made for regarding the agent as responsible, or, if that were not the case, this would be a very different thing from "censure" even if we chose to call it that. For consequentialist accounts of the use of blame and censure see Benn 1958; Nowell-Smith 1961, 301-04. For a useful discussion of consequentialism and blame see Duff 1986, 42-47.

restraint to the citizenry or to strengthen social cohesion", although those ends may play a part in how we go about censuring, rather, "it is *because* we share certain moral standards that a response is required that recognizes both the conduct's wrongfulness and the actor's fault."[25] Duff is, therefore, right in emphasising the important way in which condemnation is part of a discourse, "a moral discussion" in which a "challenge" is made to "respond to this moral charge".[26]

It is important to note the importance - the centrality - of the censure element in the account of punishment, as against the importance of the assurance problem in the account of coercion; if the community were to coerce the offender in a manner that was morally neutral it might well satisfy the demands of the assurance problem but it would fail on two vital counts. First, it would not be addressing the offender as a moral being; it would not offer the offender a moral *reason* why he *ought* not behave in the manner he has, it would merely offer him a prudential reason why it is in his self-interest not to do so in the future. Second, and relatedly, it would not affirm the moral status of the terms of co-operation, that is, it would not affirm their status as authoritatively binding on the pursuit of self-interest, rather they would

---

[25]von Hirsch 1990, 272. This view of the role of blame is vital in Duff's "penitence" account (see Duff 1986, esp. 47-54, and Chapter 9; also Duff 1988) and appears in some version also in Primoratz 1989; Hampton 1984; Morris 1981, but see Duff 1977.

[26]Duff 1986, 48.

be affirmed merely as means to better fulfil one's self-interest; means that can be evaluated on a case by case basis in the light of this.

Does it follow from the above account that the "offender wills her own punishment"? Insofar as the offender is addressed against the background of the constructivist theory I have defended she does, indeed, will her own punishment. This is an idea that is often viewed scornfully[27] and it brings into focus the difficult question of the role of consent in the constructivist theory I am proposing, a question which I addressed very briefly in a note above.[28] There I argued that the agent, taking up the reflective stance, asked whether she wanted to maintain and cultivate the moral dispositions that make her into a moral character, and reject those that disposed her otherwise. I argued that the constructivist theory I proposed gave an account of why she should do so, although I could offer no decisive reason. This I thought, made some sense of the idea of consent; it is not a once and for all commitment to "obey the state" or "enter the contract" it is a continual, one might say hermeneutical, process in which the agent affirms her status as a moral character through self-creative reflection and action.[29] As part of the constructivist scheme I am offering as the answer to

---

[27]See Honderich 1984, 219-27, esp., 226.

[28]Chapter 6, note 16.

[29]See Taylor 1985, Chapters 1, 2, 4. Compare this account of consent with that given in §23.

the personal reflective question concerning the nature of the agent's relations to others, I have argued that the commitment to be moral involves a renunciation of free-riding, and this renunciation is visited on those who free-ride as, in this sense, their will.

The most common objection to such a theory revolves around the claim that some - perhaps, most - people have never put themselves into an original position, they have never "chosen" to affirm themselves as moral characters, so they have not willed anything. I do not think such a criticism is plausibly levelled at the model of consent offered above. People in relatively just societies[30] *do* reflect on moral issues every day, and do ask themselves whether, for example, the want to be the kinds of persons who are dishonest, or greedy, or who lie and steal. For this reason, censure aimed at involving the agent in a discussion, aimed at the agent as a rational, responsible, being is not consequentialist. Although the aim is, partially, to try to get the agent to reflect on his actions, regret and "repent" them, even if we were sure that the agent would not do so - Primoratz gives the example of Klaus Barbie[31] - we would still be justified in addressing him as an agent, (so long, that is, as he fulfilled the requirements of being a responsible agent and did not completely renounce the co-operative option),

---

[30]How all of this is effected by a background of injustice is discussed below, §44.

[31]Primoratz 1989, 195-6.

because he is capable of doing so, of understanding that his actions have evoked the disapproval of others, even if he finds that such disapproval carries no weight with him.[32]

In a sense, then, I am defending the idea that everyone in a relatively just society does consent to live under moral terms of co-operation, that is terms which authoritatively bind the pursuit of self-interest, I am merely using a different idea of what it is to consent, one which is far removed from the voluntaristic single act that is its more common meaning. This seems to me important if we are to avoid the imbecility of theories of obligation that understand consent to be a single act of agreeing to be bound by the law. Of course, whilst people do consent (on my interpretation), they consent to all sorts of moral rules that have no foundation, and for all sorts of reasons. On the theory I am proposing, moral rules which have no foundation in the agreement should be rejected, and the reason to live morally is a combination of self-

---

[32]See on this Duff 1986, 266; von Hirsch 1990, 274; Primoratz 1989, 195-6. Primoratz tries to argue that Duff's account of censure is consequentialist, von Hirsch argues on a similar basis to me that this does not have to be the case, although he thinks that "Duff seems to lay himself open to this charge" (von Hirsch 190, 274, 48n). *Contra* von Hirsch, I think Duff is pursuing a similar line to myself and von Hirsch, that no agent is conceptually incapable of redemption (if she were she would not qualify as an agent), independent of how empirically likely it is. This is not only compatible with the account of blame offered throughout Trials and Punishments, but is also explicitly stated in the passage cited by von Hirsch; Duff says "to talk thus of 'the good that is in him' is not to make some psychological claim to the effect that he 'really' cares for the values which he flouts: it is rather to combine the conceptual claim that every moral agent has the capacity or potential for moral development and reform, with the moral claim that we should never give up hope of bringing him to actualise that potential." (Duff 1986, 266).

interest and existential commitment. Education, and perhaps the increasing disenchantment with spiritual explanations, might slowly convince people of this, and it is important that it should do so, for it is important that people realise their responsibility for the moral life of their community. Punishment, in a just society, also has its part to play, for in saying to the agent "you, and you alone, are responsible for your commitment to be moral, (albeit that we can try to help you in making that commitment)", it can deepen the agent's understanding of morality and perhaps lead to his becoming a fully morally autonomous being; that is, a being who takes full responsibility for his moral commitment.[33] This brings me to the role of punishment as a moral educator.

41. Punishment as a Moral Educator

Punishment in reaffirming the terms of agreement, in declaring the criminal action wrong, also contributes to the moral education of the contractors;[34] it reinforces the injunction to be moral. In large modern societies, in which the traditional spheres through which moral education occurred have either disappeared or been discredited, law has an important role in setting the tone of the society. Law is an alternative form through which to inculcate as

---

[33]This idea, that in accepting his punishment the agent deepens his understanding of himself, is important to Hegel's account of punishment, see §§14-15 for a discussion of this.

[34]The best known defence of the moral education approach is Hampton 1984.

well as enforce moral injunctions and penal sanctions, likewise, share this role. This is not to say that law can replace social morality;[35] law is by its very nature cumbersome and general, and, most importantly, many of the commands of law are inapplicable to children (because they do not qualify as agents), who are the most obvious candidates for moral education. Nonetheless the moral tone of many societies is dependent upon the law as the most important and visible instance of morality.

This is not to say, of course, that the law ought to enshrine and enforce every piece of traditional morality, (or, perhaps worse still, popular calls from a "moral majority") *a là* Devlin's attack on the Wolfenden Committee,[36] it is to say that the law ought to enshrine and enforce those justified moral injunctions (which are best enshrined in a positive law, rather than a positive morality), which would emerge from the reflective agreement of agents committed to a system of co-operation and to a principle of equality.

This is an instance of the idea that just as the moral community is created by the agreement of its members to co-operate on moral terms, it is also the case that the community is the creator of its members; it is only through living in the community that the agent comes to flourish as a self-conscious, rational being. Further, the concrete

---

[35]See §25.

[36]See Devlin 1959 and the reply by Hart (1963).

form given to the abstract constructivist injunctions are given by the community, and it is these concrete rules that form and shape each agent's dispositions. Insofar as the community is challenging the agent to choose the moral life, punishment has an important role reinforcing the community's standards and values.

Punishment, then, while serving the assurance problem also has an element of censure - an expressive element.[37] In this regard I can concur with Michael Davis's criticism of expressive theories that punishment is more than merely censuring; it is, but what Davis does not recognise, is that it is *also* censuring.[38]

Having established this "two-pronged rationale",[39] it is now time to address a problem I put off above concerning the tension between the hard-treatment required to solve the assurance problem, the expression of blame, and the requirement to treat people as agents.

42. Blame, Hard Treatment and Punishment

The first and most obvious tension that arises would be if the assurance problem demanded a relatively harsh penalty and yet the degree of censure the society thought

---

[37]The best known defence of expressivism is Feinberg 1965. For a detailed criticism see Skillen 1980.

[38]See Davis 1988; von Hirsch 1990, 270.

[39]This term is borrowed from von Hirsch 1990, 278.

appropriate was not too great. Let us take the offence of overriding on public buses.[40] It is fairly clear to anyone living in London that such an offence is common, and very difficult to detect without spending more money on Inspectors than one could hope to recoup on fines. Interestingly London Regional Transport have tried, over the last few years (approximately 1988-1994), two, very different, approaches. The first was a campaign aimed to stigmatise fare evaders, amongst the posters that went up was one declaring that it was better to override on a double decker bus because then only half the bus could stare at you when you were apprehended. Others, likewise, pointed out how embarrassing it would be to be caught. This campaign has recently been replaced, all invocations to play fair and references to social stigma have been removed and the posters now simply declare that the maximum fine for fare evasion has been raised to £1000; the posters now juxtapose "overriding" and "overdrawn".

This example is clearly not meant to be a criminological case study, the point I wish to make is that the blame attached to overriding on the buses is not very great - a lot of people who would not dream of, say, stealing from another individual regard overriding as a legitimate activity given that they "give so much to London Regional

---

[40]That is paying for a certain distance and then remaining on the bus for longer than one's ticket permits.

Transport anyway",[41] yet a fine of £1000 is very serious, not to convey censure, but because deterring such activities requires severe penalties.[42] The problem is simple; sometimes deterrence and censure will yield conflicting demands of the penality system.[43]

There are two reasons why precedence should be given to the censure element over the preventative in deciding the quantum of punishment. The first is that to deter potential offenders only by threatening them with extreme penalties is to give up on the idea that the penality system addresses the person as a moral agent and member of the community. When Armstrong writes "let him be whipped to death, publicly of course, for a parking offence; that would certainly deter *me* from parking on the spot reserved for the Vice-Chancellor!",[44] he is making the point that although such sentences would prevent crimes they are not

---

[41]There are many similar examples; people who "taste" goods in Supermarkets do not think of themselves, and others do not think of them, as shop lifters.

[42]Although see *infra*, 47n.

[43]Andrew von Hirsch, whose "mixed" account is very similar to my own, singularly fails to address this. He states that "the intertwining of [prevention and expression] is critical: It means that the severity of the hard treatment will convey the degree of censure. This is why … it is desert rather than preventative efficiency that should determine the quantum of punishment." Why should the severity of the hard treatment not reflect the needs of prevention? von Hirsch refers the reader "for a fuller discussion" to later in the article where again the position is merely stated: "In punishment, deprivation or hard treatment is the vehicle for expressing condemnation." Remarkably, the reader is then referred back to the original discussion for a defence of this argument! (von Hirsch 1990, 276-7, 279. In fact, throughout this article von Hirsch moves from a "two-pronged" approach to an expressivist approach in which it is simply a fortunate by product that prevention also occurs, see, e.g., 287.

[44]Armstrong 1961, in Acton 1969, 152.

acceptable. One reason that they are not acceptable is that they are incompatible with the commitment to treat others as deserving of equal consideration. Punishment, on the "two-pronged" approach given above, is meant to communicate - to express - to the agent and to the community precisely what it is that the offender has done and what the appropriate response is. As Duff insists it must be, punishment is about engaging the criminal in a discourse over the true nature of his actions, and I would add, this discourse extends wider to the community at large. It aims to show the offender the real nature of his commitment to live the moral life as demonstrated in his participation in the community and, thus, to reconcile him to his punishment.

> "An offender's punishment must be such that it appeals to, but does not coerce, his understanding and his will…. It must also be proportionate in its severity to the seriousness of his offence: only then can it communicate to him an adequate understanding of the moral character of his offence".[45]

Similarly, only then can it express from and to the community the true moral character of the offender's act.

The second reason for the precedence of censure over prevention in deciding the quantum of punishment is that if harsh penalties for minor, but prevalent, offences were imposed they would be disproportionate, and this is important not only because it is vital to the expressivist

---

[45]Duff 1986, 278.

and educative functions of punishment that the response is thought appropriate, but also for more practical reasons. In a just society, or a relatively just one, I have argued that there will need to be respect for something akin to what we think of as broadly liberal individual rights. Let us assume that in such a society a category of offences exists which are not thought particularly reprehensible.[46] If that society decided to penalise one minor non-reprehensible offence (which was extremely prevalent), by several years imprisonment, then the expressivist could raise two related objections. The first is that such a response would violate a principle of ordinal proportionality. Ordinal proportionality requires that

> "persons convicted of crimes of comparable gravity should receive punishments of comparable severity (save under mitigating or aggravating circumstances altering the harmfulness of the conduct or the culpability of the actor)."[47]

---

[46]See von Hirsch 1990, 284. von Hirsch points out, positing a similar society (which he calls "Draconia"), that if one had a sufficiently authoritarian view of the sovereign power then any infraction of the rules might be viewed as serious because it displayed an arrogant disregard for the powers that be.

[47]von Hirsch 1990, 282; von Hirsch 1985, Chapter 4. See also Bedau 1984. Of course, the seriousness of a crime is going to depend in part on the frequency of its occurrence. It could, therefore, be argued that a frequently committed minor offence is more serious than a similar offence which is only occasionally performed. This doesn't effect the argument here, which is about the tension between the censure that is actually felt by the community for those who commit an act and the requirement to deter others from performing that act. It may mean, however, that more should be done to alert the population to the seriousness of certain, seemingly minor, offences. This was, in fact, tried as part of the initial, stigmatising, campaign on the buses, posters were put up describing overriders as parasites on the fare paying public and detailing the costs to each paying traveller of fare evasion. Within the limits of censure, greater penalties might play a role in this education as the relationship between the penalties imposed for certain crimes and the

A prison sentence for overriding on buses, then, would require a similar sentence for similarly grave offences; littering, perhaps. But why should we endorse a principle of ordinal proportionality? The answer is clear: if punishment is the appropriate response to the agent through which the community expresses its censure, then actions which are similarly censured should evoke similar responses. This is necessarily very rough, both because judging the similarity of different legal sanctions is difficult (is 100 hours community service a lesser or greater punishment, in terms of the censure it conveys, than a day in prison?[48]), and because, as I argued above, the response of the community may not necessarily be legal. Where there is no preventative rationale, the community

---

seriousness with which those crimes are viewed is clearly a complex one. I have tried to simplify it here because I am interested in showing how to reconcile the demands of censure and those of prevention and, thus, it has been useful to show them in straight conflict. In fact, as the above demonstrates, it is not as simple as that, and this would need to be take into account when actual punishments were imposed, given that censure is still going to allow a degree of flexibility.

[48]This raises the very interesting question of the difference between types of penalties. Again no abstract analysis can really be definitive in that it is possible to think of a society in which some material possession was so valued that a fine of that thing was considered vastly more severe than a term in prison. There is little doubt that the reverse applies in contemporary Britain, in other words, although a fine may have disastrous effects on a family which could be avoided by a term in prison for the offender, it is still considered, especially by judges, to be a more serious "statement" to impose a custodial sentence. I would explain this by saying that the "communication" in a custodial sentence carries implications of much greater censure. Firstly, because a fine my be dealt with relatively discreetly; one's neighbours, employer, family, etc., may not have to know. More importantly, consider what is said to the offender in a custodial sentence: he is being told that he has done something so grave that society cannot tolerate him, it needs to purge itself so completely of his action that it can only express this by removing him temporarily from its orbit. Anyone who remembers their childhood will know that to have one's pocket money docked may have imposed real hardship but it was as nothing to the remorse evoked by feeling as if one had (albeit temporarily) lost one's place in the affections and concern of one's parents.

might use social stigmatisation, nevertheless the point is that the level of stigmatisation should be appropriate to the measure of censure, just as, within the legal system, the punishment should also reflect the degree of censure.[49]

The second, related, objection the expressivist position raises against the use of harsh penalties for some minor crimes is that it is likely to cause problems when fitted in to a scale of cardinal proportionality. Cardinal proportionality concerns "the overall magnitude and anchoring points of a penalty scale";[50] that is, it concerns the ranking of crimes and penalties vertically by seriousness. Of course, if overriding on a bus were to carry a sentence of several years imprisonment then this could be the point around which the scale could be fixed. But this is not satisfactory because of the moral assumption I built in to the example that rights are important in the society. If the community appears to say

---

[49]"Hard treatment" can, therefore, be used to express censure, and given certain empirical facts, may have to be. This is why one of the standard criticisms of expressivist theories is so misplaced; the criticism has it that if punishment is expressive one could just as well "'say it with flowers' or, perhaps more appropriately, with weeds" (Scanlon 1988, 214), but the whole point is that some societies do not say it with flowers or weeds but with suffering. It may be hoped that we, in contemporary Britain for example, could reduce the public perception of what is a "serious statement" from life imprisonment (where it seems to be at the moment), to say, something like four years, but this is a process that will take time. The point is that, of course there is no necessary connection between the expression of blame and hard treatment, but if the cultural connection exists then that has to be included in any expressivist theory. The claim is that society needs hard treatment to express itself, and it may also be true that the offender needs hard treatment to appreciate the level of censure; Duff argues, (1988, 162-3 and 1986, Chapters 9-10) that a degree of "hard treatment" is also required so as to "force [the offender's] attention onto his crime" (Duff 1988, 162).

[50]von Hirsch 1990, 282.

to the offender, "this isn't a terribly heinous crime so we'll just send you to prison for several years", the message it is sending is that the right to liberty is not all that important. It is undermining the very values it is seeking through punishment to affirm and inculcate. As von Hirsch puts it, "it's a bit like saying, 'I'm not so upset, so I'll only break your arm'"[51]

There is an additional, non-principled, practical objection, to very harsh penalties for minor crimes, related to the need for cardinal proportionality and that is that if the punishment for car parking is the same as that for, say, killing a policeman, then an offender of the car parking rule who is apprehended by a policeman may just as well kill the policeman in an attempt to escape. In other words punishment ought "to induce a man to choose always the least mischievous of two offences" by ensuring that "where two offences come in competition, the punishment for the greater offence must be sufficient to induce a man to prefer the less". Similarly punishments ought to be such that having "resolved upon a particular offence" the offender is induced "to do no more mischief than what is necessary for his purpose".[52]

---

[51]von Hirsch 1990, 284.

[52]"If any man have any doubt about this, let him conceive the offence to be divided into as many separate offences as there are distinguishable parcels of mischief that result from it. Let it consist for example, in a man's giving you ten blows…If then, for giving you ten blows, he is punished no more than for giving you five, the giving you five of these ten blows is an offence for which there is no punishment at all: which being understood: as often as a man gives you five blows, he will be sure to give you five more, since he may have the pleasure of giving you these five for nothing.

In itself this objection would not be sufficient, because it would always be possible to counter it by denying that there is an upper limit on the punishments being imposed. In other words, it might be true that having the death sentence for car parking offences and killing policeman will encourage car parking offenders to kill policemen, and therefore that the penalty for killing policemen should be far greater than the penalty for parking your car illegally, but it does not follow from that alone, that the punishment for illegal car parking should necessarily be scaled down. History has demonstrated man's amazing capacity to invent hideous and painful treatments for others[53], and thus it would be possible to respond to this objection by scaling the punishment for police killers up. Anybody who finds such an idea absurd should look at Hanging not Punishment Enough for Murtherers, Highway Men, and House Breakers, etc.[54] The author of this pamphlet argues on identical lines to those above, i.e., on the basis of cardinal proportionality, that since hanging is punishment for larceny it cannot be sufficient punishment for more serious crimes such as those listed in the title.

On the theory being defended here such a policy would not be justified, for an upper limit is put on possible

---

..This rule is violated in almost every page of every body of laws I have ever seen." All quotations from J. Bentham 1798, 168.

[53]See the much quoted description of the execution of the regicide Damiens in the opening pages of Foucault's Discipline and Punish, Foucault 1977, 3-5.

[54]Anon. 1702

punishments by a combination of the expressivist and educative elements *and* the moral theory underlying the account of punishment itself. It is important to realise the necessity of the moral theory because the requirements of cardinal and ordinal proportionality do not tell us where to set the penalties for different offences[55] and this is as it must be. I have claimed throughout that such questions must depend upon the society in which the penalties are to operate. All that I attempted to show above is that where there is a conflict between prevention and expression, the latter must take precedence. The moral theory I have defended tells us that rights and agency are important, and thus puts a limit on the types of penalties that could be said to appropriately express the community's censure.

The conflict between the prevention and expression can, then, be resolved, however, there remains one promise that I have yet to fulfil and that is to consider the tension between the need for hard treatment (to serve the requirements of the assurance problem), and the requirement to treat offenders as responsible beings.[56]

---

[55]Nor, of course, do they determine what offences (if any) will necessarily be legal.

[56]See *supra*, note 12.

43. Hard Treatment, Coercion and Respect for Agency

Antony Duff argues that any account of punishment that takes seriously the Kantian claim that agents are to be treated as responsible, autonomous beings,

> "rules out, as improperly coercive and manipulative, any kind of punishment which (whether by design or in fact) serves to beat, cow or manipulate the offender into submission, instead of communicating to him, and trying to persuade him to accept, the reasons which justify his punishment."[57]

The tension is, then, not between the need for hard treatment and the need to censure appropriately, but between the use of hard treatment for preventive purposes and the very theory that underlies the account of punishment given here. But is Duff right in his insistence that a system of "punishment as a rational deterrent",

> "... is still open to the objection that it does not show punishment to be consistent with a proper respect for the citizen as a rational and autonomous agent; that it still portrays punishment as an improperly manipulative attempt to coerce the citizen into obedience to the law."?[58]

Duff is only right if his objection is aimed at deterrence only punishments, and the agent has a claim not only to be treated as "rational and autonomous", but also as part of the moral community. However, this is clearly not Duff's

---

[57]Duff 1986, 278. See also Duff 1986, 186; 268-77.

[58]Duff 1986, 186.

intention. Duff argues that insofar as we punish to a preventative end (even if we also have other ends), we coerce the offender and thereby fail to treat him as an equal, autonomous, being.[59]

It should be clear why Duff's account and my own have parted ways, Duff's self confessed intention is to "explore the implications of the Kantian demand that we should respect other people as rational and autonomous moral agents", and (again, self-confessed), this demand is not explained or justified.[60] Duff gives this demand the status of an unconditional moral principle and that is, of course, why he cannot reconcile moral treatment and coercion; it is this, precisely, that has been the basis for my attack on Kantian and impartialist theories.

The conditional moral theory that I have offered, in contrast, has as its strength the compatibility of coercion and *moral* censure, because prudential reason is a necessary and crucial element in moral reason. I have demonstrated that mutual coercion is part of prudential reason, however, I have also argued that if we are to convert justified coercion into moral punishment then the community must combine the threat of coercion with the communicative element of punishment, with the appeal to the moral will of the offender, and the argument that he should recognise the justice of his own punishment. This combination does not

---

[59]See esp. Duff 1986, 268-77.

[60]See Duff 1986, 6.

undermine or subvert the offender's autonomy, rather, it reinforces it. The offender is coerced as a member of a necessarily coercive community in accordance with his prudential reason, and in addition, he is appealed to as a moral being. Insofar as the communication is successful and he realises his responsibility for organising the pursuit of his self-interest through binding, moral, terms of co-operation, he becomes a fully morally autonomous being.

44. Conclusion: Justice and Punishment

That completes my account of the rationale of coercion and of moral punishment. I have, I believe, addressed both parts of Tolstoy's question with which I began this thesis; "why, and by what right, do some people punish others". This is not the place to restate the arguments, but, in summary, I have shown that there is a possible rationale for coercion which is compatible with the autonomy of the agent being coerced, but that this is not sufficient for a practice of moral punishment. The latter requires that we address the agent as a member of the community, possessed of the potentiality to fully commit himself to morality, to take responsibility for organising his own life, and the pursuit of his self-interest, in accordance with terms which would be agreed in a suitably constituted original position.

The argument for this position has reconciled at the "deep level" demanded the agent's prudential reason, the personal perspective from which morality (and coercion) seem to be restraints imposed from the outside and against which self-interest strains, and the moral perspective, the demand that the agent treat everybody as deserving of equal consideration. In reconciling these two perspectives, the theory, applied to punishment, has comprehended those aspects of the retributive and consequentialist positions that did, indeed, seem attractive. Punishment is both needed for its consequences, but is also of an offender for an offence.

There remains only one promise to fulfil and that is to discuss the place of just punishment in an unjust society. This is an important issue and I will only be able to provide pointers as to how it might be addressed here. To give the argument some structure, and allow the discussion to seem real, let us take the problem of punishing a black man of eighteen from the South side of Chicago. Anyone who has visited this area will be aware of the crushing poverty and the sense of hopelessness and alienation that pervades it.[61]

---

[61]I have visited both this area and Soweto and can honestly say that, even during the apartheid regime, the sense of "not belonging", of alienation from the country and system, seemed to be stronger in Chicago. I could not, anyway, take the example of pre De Klerk South Africa because such a society was not, in any sense, a moral community; the non-white population was systematically oppressed and no moral ties linked them with the white government.

At two important points my argument as presented above seems no longer to apply. The first is the connection between self-interest and co-operation that is absolutely central to my account, and the second the idea that each person, in their everyday lives, lives as a moral being, consenting to the terms of agreement that would be agreed in a suitably constituted original position.

If, indeed, the offender would be better off outside of co-operation then it is rational for him to put himself outside of the terms of agreement, that is, it would be rational for him to take up a state of nature relationship with the community and pursue his good independently of it. In fact, I believe that this might be the case for my example; of course, the society could then coerce him as an alien, however, given the limited empirical chance of his being apprehended it may still be better for him to take the risk and pursue his self-interest in such an unconstrained manner.

Assuming that he would not be better off in such a position, what are we to make of the idea that such a person has committed themselves to live with others on moral terms? If the community is to coerce the offender as a member of such a moral community and, thus, a potentially moral being, the community's authorities will, as part of their address to him, have to explain that he is part of a system that treats everybody as fundamentally equal, not in some abstract sense but also in the initial stages of the

distribution of resources and opportunities. In explaining this, the community's failings would be obvious to both punishers and the punished, and reason demands that this force on them changes. In other words, if it is the case that the community is, in fact, organised around principles of justice as mutual advantage, then the authorities cannot legitimately claim to be imposing moral punishment, and insofar as they do, they will be acting hypocritically.

Of course, the current punishment practices, (indeed, the current social policy), of the USA (and, for that matter, Britain), have almost come to resemble that which might hold between communities bound by mutual advantage, not tied by moral norms, and, insofar as the pretence of equal treatment is maintained it does not seem to force the attentions of the authorities to the anomalies of the system; rather, coercion is used without the expression and attempt to communicate that would make such coercion moral punishment. All this shows, however, is how far the imposition of just moral punishment requires the radical overhaul of current practices in distributive as well as retributive justice.

In the end the only conclusion we can draw, and a fitting finish to this thesis, or so it seems to me, is Jeffrie Murphy's:

> "If we think that institutions of punishment are necessary and desirable, and if we are morally sensitive enough to want to be sure that we have the moral right to

punish before we inflict it, then we had better first make sure that we have restructured society in such a way that criminals genuinely do correspond to the only model that will render punishment permissible".[62]

---

[62]Murphy 1973, 243.

Appendix A: Rawls on Punishment

"Rawls's view of retributive justice is not developed at any length and plays virtually no role in the overall argument of the book … it is dubiously consistent with his account of distributive justice … I regard it as a small and insufficiently motivated departure from the general attitude toward desert that dominates his work".
Samuel Scheffler 1982, 306n.

A. Rawls, Retributive and Distributive Justice

In this Appendix I intend to discuss the approach taken by John Rawls to the problem of punishment in A Theory of Justice. It seems to me that there are a number of reasons that make undertaking such an exercise worthwhile; first, the work of Rawls, and especially A Theory of Justice, is of such importance in the discipline of political theory that textual analysis, if it can add to our understanding of the work, is justified for that reason alone. Second, I believe, *pace* Scheffler, that in certain instances what Rawls says about punishment can shed light on other, crucially important, ideas in A Theory of Justice. Third, in his reflections on punishment, and his inability to come to terms with the problem of coercion, Rawls demonstrates the problem facing impartialist theorists that I discussed above (in Chapter 5). Finally, there is a certain challenge in trying to unravel the puzzle highlighted in the quotation with which I began this Appendix; Rawls, in at least one of his comments on the problem of punishment seems directly to contradict the foundational intuition of A Theory of Justice, that institutions should not be designed in the first instance so as to reward (or

penalise) individuals on the basis of their social or natural talents, abilities or disabilities. It could be that Rawls uses two different psychological models, one broadly deterministic and the other broadly libertarian and such a view should be treated as a kind of "default option" if no better explanation can be found. However, I shall argue below that a more radical solution might lie with the "Kantian Interpretation",[1] and in the nature of the claim that something is "morally arbitrary".

As we saw above (in Chapter 5), Rawls, like Scanlon, replaces desert in his account of distributive justice with the idea of legitimate expectations (adding the premise that everyone is responsible for all their tastes, aspirations, beliefs, conceptions of the good, etc.[2]). Yet, unlike Scanlon, he specifically denies any symmetry between retributive[3] and economic justice:

> "No doubt some may still contend that distributive shares should match moral worth… [and] …this opinion may arise from thinking of distributive justice as somehow the opposite of retributive justice. … the purpose of the criminal law is to uphold basic natural duties … [It is] not simply a scheme of taxes and burdens designed to put a price on certain forms of conduct and in this way to

---

[1]Rawls 1971, §40.

[2]I shall, from now on, group these things together under the general term "preferences". In doing so I am following Barry 1991a.

[3]"Retributive justice" is here taken to be a general term for denoting the response to criminality. This is so as to keep in line with Rawls's terminology. It is not meant to refer to any particular (retributivist) conception of punishment.

> guide men's conduct for mutual advantage. It would be far better if the acts proscribed by penal statutes were never done. Thus *a propensity to commit such acts is a mark of bad character*, and in a just society legal punishments will only fall upon those who display these faults.
>
> It is clear that the distribution of economic and social advantages is entirely different. … The function of unequal distributive shares is … to attract individuals to places and associations where they are most needed from a social point of view… "[4]

Precisely how Rawls believes this distinction holds up is not clear (at least to me), from this passage. Rawls is very careful to avoid saying that the criminal *deserves* his punishment, in fact, he does not commit himself to any detailed view of the purpose of criminal law (beyond upholding "natural duties"). What he does do, however, is deny that it is to be understood on the legitimate expectations model. Rather, he implies that retributive justice is a response to "moral worth", specifically to a "bad character". It is this that has led commentators such as Scheffler, Brubacker, Honig and Sandel[5] to question whether Rawls's views on punishment are compatible with the rest of A Theory of Justice. "Character", Rawls's critics point out, is one of the attributes of the individual specifically picked out by Rawls as "morally arbitrary" in

---

[4]Rawls 1971, 314-5. Emphasis added.

[5]Brubaker 1989, 1990; A. Rawls 1990; Honig 1993a, 1993b; Sandel 1982, 89-92; Scheffler 1992.

the argument for the difference principle.[6] In this Appendix I shall offer an account of Rawls's views on punishment and address the question of why he is so confident (and whether that confidence is justified) in distinguishing, say, indolence and criminality. Before this, however, I have to develop the short account of Rawls offered above (in section 27).

B. Rawls, the "Strains of Commitment" and the Problem of Stability[7]

Rawls in A Theory of Justice argues that there is a second stage for the people in the original position; a stage in which they consider whether they will be able to abide by the commitments they have made. The people in the original position must consider the stability of the society founded on their choice of principles. They:

> "cannot enter into agreements that may have consequences they cannot accept. They will avoid those that they can adhere to only with great difficulty. … Moreover, when we enter an agreement we must be able to honor it even should the worst possibilities prove to be the case. Otherwise we have not acted in good faith. Thus the

---

[6]Rawls 1971, 103. As F. Scott Fitzgerald noted: "I am still afraid of missing something if I forget that … a sense of the fundamental decencies is parcelled out unequally at birth." (Fitzgerald 1926, 7).

[7]My understanding of this part of Rawls's argument has been greatly increased by having had the advantage of three manuscripts: Chapter 3 of Barry 1995a; Ivison MS; and a paper presented by Barry (1995b) at the LSE Political Theory Workshop. Subsequent discussions with Brian Barry and John Charvet have - as ever - proved invaluable.

> parties must weigh with care whether they will be able to stick by their commitment in all circumstances." [8]

It might be thought that this is directly relevant to Rawls's views on punishment because it is clear that penal sanctions would be one way of trying to ensure stability if people, once they had emerged from the original position, found the pursuit of their own advantage so at odds with the principles of justice that the burden of commitment was too great to bear. But this is not what Rawls is getting at; the people in the original position must not consider whether the principles can be maintained by force (by "means of persuasion or enforcement"[9]) but whether (even given "the worst possibilities") they will be able to "honor" the agreement. What seems to be at the heart of this is the idea that out of the veil of ignorance one should be able to accept the reasons given for the principles of justice.[10] Rawls seems to think that his two principles can secure such acceptance and, indeed, can generate a kind of "ethos of justice" but that, for example, utilitarianism could not:

> "when the principle of utility is satisfied, however, there is no such assurance that everyone benefits [as there is with the two principles]. Allegiance to the social system may demand that some should forgo advantages for the sake of the greater good of the whole.

---

[8]Rawls 1971, 176.

[9]Rawls 1989, 246; Rawls 1993, 142.

[10]As Barry points out this is a moral argument. (Barry 1995a, Chapter 3).

> Thus the scheme will not be stable unless those who must make sacrifices… accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of [their] li[ves]."[11]

Rawls seems to have every justification in considering this to be "an extreme demand"[12] and thus utilitarianism to have failed the "strains of commitment" test.

On the face of it, however, this interpretation of the strains of commitment, as a moral argument about accepting the reasons for one's position, makes the test a redundant replication of the first stage. Utilitarianism is an inadequate account of justice because it is not chosen in the original position, and if it were, it would show that the "situation" (i.e., the original position) was incorrectly formulated; that is, it would fail in the process of finding a reflective equilibrium. The fact that one is unable to accept the reasons for one's position shows that the principles under which that position is established are not principles of justice.[13]

The strains of commitment test, then, does not seem to add to the theory nor is it really an argument about stability.

---

[11]Rawls 1971, 178-9.

[12]Rawls 1971, 179.

[13]Barry 1995a (Chapter 3), argues that the "strains of commitment" test essentially amounts to a free standing Scanlonian argument, i.e., that the principles of justice are those that would emerge if each person has the right of reasonable veto. This adds credibility to the thought expressed here that it is a replication (or alternative) first stage rather than an additional second stage argument.

Yet, Rawls gives great weight to the problem of stability[14] and the strains of commitment seem to have some place in his thoughts. The problem is that it is not always clear that Rawls understands the nature of the question he is asking. The strains of commitment test is not really a stability problem because it is not in the end about whether people can accept the principles of justice, but is another way of asking what the content of those principles is. At other times, Rawls does ask, however, whether people can live by the principles once it is established that they are, indeed, the principles of justice. In these passages Rawls confuses the issues by combining two problems, one about the feasibility of coercion and the other about the motivation the agent has for being just. Together Rawls treats these as a problem of stability, however, this is not altogether helpful and obscures the fact that the second of these problems is a crucially important one with an impeccable philosophical pedigree.

C. Rawls on Moral Renegades and Punishment: The Question of Motivation and the Kantian Interpretation

It is important to note in considering the question of motivation that Rawls, despite the popular misconception to the contrary, does not address his theory to moral egoists.[15] Egoism (or mutual disinterestedness) is the

---

[14]As a rough guide, the entries for "stability" in the index of A Theory of Justice take up some 35 lines, several more than, for example, the "Original Position".

[15]See Rawls 1971, 568.

psychological motivation given to the people in the original position but in the "real world" Rawls assumes that persons have a "sense of justice". The content - or demands - of the agent's sense of justice is what that agent attempts to determine through the process of reflective equilibrium. Given the two principles as the content of justice, then, the agent's "sense of justice" is "an effective desire to apply and to act from the principles of justice and so from the point of view of justice".[16] Rawls believes that he can show that "affirming" one's sense of justice is rational but that this, nonetheless, may not provide the agent with "sufficient reason" to do so. The rationality of affirming one's sense of justice is, in one sense according to Rawls, "trivial" for "being the sorts of persons they are, the members of a well-ordered society desire more than anything to act justly and fulfilling this desire is part of their good."[17] Rawls seems to accept that this is a pretty weak answer to the problem of accounting for why the agent ought to be moral for he goes on to ask what happens if people do not desire "more than anything" to act justly, although whether this means that they are not the right "sorts or persons" or that the well-ordered society is not quite what he thinks it is is unclear. "Suppose", he says, "that the desire to act justly is not a final desire" why is it

---

[16]Rawls 1971, 567; cf. Rawls 1993a, 19.

[17]Rawls 1971, 569. Remember, a well-ordered society is defined as one in which "everyone accepts and knows that the others accept the same principles of justice, and the basic institutions satisfy and are known to satisfy these principles." (Rawls 1971, 453-4).

rational to "confirm this sentiment as regulative of [one's] plan of life"?[18]

Rawls gives three answers; first, he assumes that people will be motivated by the desire to act fairly: "wanting to be fair with our friends and wanting to give justice to those we care for is as much part of these affections as the desire to be with them and to feel sad at their loss … in a well-ordered society these bonds extend rather widely."[19] Second, "it follows from the Aristotelian Principle (and its companion effect), that participating in the life of a well-ordered society is a great good".[20] Third (and most importantly[21]), there is the Kantian interpretation: "acting justly is something we want to do as free and equal rational beings. The desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire."[22]

This final argument, which in the end is the only one Rawls thinks can do the required work, seems at first glance to be odd. If the "desire to act justly" and the "desire to express our nature as free moral persons" are "practically

---

[18]Rawls 1971, 569-70.

[19]Rawls 1971, 570-1.

[20]Rawls 1971, 571. For further consideration of this argument see §33.

[21]Rawls 1971, 574.

[22]Rawls 1971, 572. I shall be returning to the Kantian interpretation below.

… the same" and the problem is that the first desire is not final (or decisive), then how does it help to invoke the second, Kantian, desire? It can only add to the argument if the Kantian "desire to express our nature as free moral persons" necessarily is (or ought to be) a final, decisive, desire.[23] If this is right then the Kantian element in Rawls is much more than an interpretation, it is a metaphysical claim about the nature of humans that explains why the agent ought to be just.[24] Rawls certainly gives considerable textual evidence for this interpretation; consider the following passages:

> "The desire to express our nature as a free and equal rational being can be fulfilled only by acting on the principles of right and justice as having first priority. … it is acting from this precedence that expresses our freedom from contingency and happenstance."

> "We cannot … express our nature by following a plan that views the sense of justice as but one desire to be weighed against others. For *this sentiment reveals what*

---

[23]That it is a final desire might be inferred from Rawls later work in which he describes the Kantian element (in a slightly different form) as one of "two moral powers" the realisation of which takes precedence over all other preferences, see, e.g., Rawls 1982, 16.

[24]This is one claim made by Sandel 1982. Barry, in 1995b, goes some way to endorsing it; suggesting that this is the part of Rawls's theory that is most plausibly described as "comprehensive" (in Rawls's new terms) and, thus, that it is the Kantian interpretation that has motivated Rawls's recent expressions of dissatisfaction with A Theory of Justice: "the 'Kantian Interpretation'", Barry says, "functions rather like a turbocharger on a car: it is an optional extra that provides additional power but comes at a stiff price. The power lies in its, uniquely, supplying a 'decisive' reason for affirming our sense of justice. …The cost is that we have to endorse a particular (second-order) conception of the good - and one that comes under strong suspicion of trailing metaphysical clouds behind it." Barry 1995b, 13MS. See also Rawls 1993, especially, 175.

> *the person is*, and to compromise it is not to achieve for the self free reign but to give way to the contingencies and accidents of the world."[25]

These passages do not occur as part of the Kantian interpretation, they rather strengthen the case Rawls is making for the stability of the well-ordered society. Yet they are extraordinarily (one might almost say orthodox) Kantian in nature and they certainly lend credence to Sandel's claims that Rawls makes metaphysically rather than merely empirically implausible assumptions.[26]

But, although the Kantian desire is, on this interpretation, an integral part of the nature of the agent Rawls does not claim that it is decisive in every case. In addressing the actual agent Rawls admits that it might sometimes be the case that none of the above answers is going to provide "sufficient reason" to a particular individual "to preserve his sense of justice". There will be "those who find that being disposed to act justly is not a good for them".[27] It is not immediately clear whether Rawls is denying that these people are Kantian agents; if they were free moral persons then the Kantian desire to express their natures by being just would be a decisive motivation for them. Such a position, however, would be absurd; the Kantian claim can only be held as a universalist claim about humans, or at least minimally

---

[25]Rawls 1971, 574 & 575 (Emphasis added).

[26]Sandel 1982.

[27]Rawls 1971, 576.

rational humans. If Rawls were denying that criminals have the capacity - the nature - of free and equal beings that would be to make them Aristotelian "natural slaves".[28] Rather, what Rawls must mean is that something else about these individuals, their conceptions of the good and the desires that flow from them are such that they find themselves unable to constrain their self-interest in accordance with the principles of justice; they fail to recognise that it is in their true nature to be just.

The situation for these people is different then from those who feel unable to keep their commitments, the latter cannot accept the *reasons* for their position, i.e., they regard it as unjust. The former - who fail to recognise that being just is their true nature - however, do not challenge the principles of justice, they merely fail to *recognise* that acting in accordance with the principles is realising their true nature. For such people, "being disposed to act justly is not a good for them", and thus "it is, of course, true that in their case just arrangements do not fully answer to their nature,[29] and therefore, other things equal, they will be less happy than they would be if they could affirm their sense of justice." Rawls's sympathy for undeserved characteristics, however,

---

[28]Aristotle 1951, 64-69 (1253b1-1255a3).

[29]If Rawls is to avoid the "natural slave" position, he must mean this use of "nature" in a some contingent, empirical sense; this is the argument pursued below.

does not extend as far as these people, he continues: "But here one can only say: *their nature is their misfortune.*"[30]

It is, of course, also unfortunate because such people are going to have to be coerced and coercion costs money. However the integrity of the two principles eases the moral burden on those who are going to have to coerce them;[31] "requiring them [those who cannot affirm their sense of justice] to comply" - "is not treating these persons unjustly"[32] Rawls argues, because "having agreed to the [two] principles … it is rational to authorize the measures needed to maintain just institutions". Punishment is morally worry free for those doing the punishing even if "there are many" people unable to affirm their sense of justice. Of course, if there are many such people, "the forces making for stability are weaker" and "penal devices will play a much larger role in the social system."[33] This is a true stability problem, but for Rawls the answer, once justice has been established, is a simple one; the only limit on the role of the penal system in ensuring stability is one of feasibility.

The two principles pass the "strains of commitment" test because the people in the original position know that even should they turn out to be the worst off group they will be

---

[30]All quotations from Rawls 1971, 576 (Emphasis added).

[31]See Honig 1993, 137-48 for a discussion of this point.

[32]Rawls 1971, 575.

[33]Rawls 1971, 576.

able to accept the reasons presented for their position (although they may not be able to behave accordingly).[34] The moral renegade is supposedly in the same position; according to Rawls, he cannot deny the arguments presented for the principles of justice and thus, Rawls thinks by extension, the rationality of the sanctions needed to maintain them. Here, "persuasion or enforcement" are acceptable as means of ensuring stability because the challenge is not to the rationality of the two principles - no matter how many moral renegades there turn out to be, so long as there are not so many as to overload the means of coercion - even if the criminal does not accept his fate, he ought to do so and this reassures those to whom Rawls addresses himself; those who have to do the punishing.

The grounds for complaint - should one turn out to be a moral renegade - are not, then, that penal institutions are unjust; rather, if there are grounds for complaint it is surely that should one turn out to be a moral renegade one suffers as a consequence, but if one turns out indolent and thus unable to command a decent salary in the market then some degree of your misfortune is abated under the provisions of the difference principle. Does one "do enough" for those who are unable "to affirm their sense of justice" which is another way of asking whether the people

---

[34]This, of course, raises the crucial question of whether they would be rational to agree to a system of sanctions directed against those who cannot "affirm their sense of justice" given that they may turn out to be one of those people.

in the original position should endorse a scheme of retributive justice.

Bonnie Honig has argued that the difference between, say, indolence and criminality lies in the Kantian reading of Rawls. Criminality is different, argues Honig, because whilst our talents, abilities and liabilities are contingent parts of us, our ability to "affirm our sense of justice" reveals who we are and is thus subject to a "brute" notion of desert;

> "The mesmerizing pull of desert is not finally overcome by Rawls's critique; it returns to haunt justice as fairness when Rawls reaches for antecedent moral worth (or unworthiness) to account for the presence of criminality in a just regime and to justify its punishment. In so doing, he relies on desert to serve its traditional function: it explains the inexplicable, it makes sense of evil and justifies our dealing harshly with it. … Rawls, because he does not acknowledge desert's return, makes no provisions for its engagement."[35]

Honig's claim is that whereas the agent can reflect on his preferences and change them, the agent's ability to affirm her sense of justice is not similarly subject to critical self-reflection, it is rather constitutive of the Rawlsian self. That, according to Honig, is why the Rawlsian agent can be subject to retributive justice but is not rewarded

---

[35]Honig 1993, 131.

in the distributive sphere on the same grounds. This, however, cannot be right. As noted above, the *capacity* to "affirm our sense of justice" must be a universal one (or at least it must be held as true of minimally rational agents), and thus it could not distinguish between people. Actually affirming our sense of justice is, of course, different; some people, as Rawls notes, are unable to do so. But what is the nature of this incapability? If Rawls is not to claim that some are "natural slaves" he has to admit that the inability of some to affirm their sense of justice is, indeed, their misfortune; but it is difficult to see that it is their fault. In Kantian terms, if (as Honig rightly claims), Rawls believes that affirming one's sense of justice "reveals what the person is" not doing so is "to give way to the contingencies and accidents of the world",[36] it is to give way to "contingency and happenstance".[37]

This is why punishment cannot be simply desert based retributivism as Honig thinks, for all the problems with Kant's account seem to reappear in a surprisingly traditional form; the role of the phenomenal world being taken by the contingency of our talents, abilities and character traits and the noumenal by our transcendental natures as free and equal beings. As with Kant's early work, this makes the "unable" a matter of contingency, something about the criminal's character makes him unable

[36]Rawls 1971, 575.

[37]Rawls 1971, 574.

to realise his true nature as a free being; it is "unfortunate" and unavoidable. Honig's interpretation of Rawls has to make the contingency of the agent's inability to affirm his sense of justice compatible with her claim that this inability is revealing of something inherent and deserving about the agent. Yet if, as Kant recognised and Rawls seems to in the quotations given above, this failure, whilst the agent's, is part of the realm of "contingency and happenstance", then there is a clear problem with holding the agent responsible and deserving (of punishment).

Kant, in an attempt to avoid just such a conclusion adapted his theory so that the free will could choose to do evil, it could choose wrong. Rawls, however, has no need to import such a metaphysical move because he does not, for all his talk of "moral worth" and "bad character" endorse a retributive theory of punishment. Rather, sanctions are needed "to maintain … the principles of justice". Rawls, then, despite his protestations to the contrary, does seem to endorse a justification for punishment close to the legitimate expectations view. Rawls argues that once the rules of justice are up and running persons must revise their choices so as to be compatible with their legitimate expectations.[38] If I, for example, find that my enjoyment of food is considerably reduced unless it is accompanied by

---

[38]"Citizens (as individuals) and associations accept … responsibility for revising and adjusting their ends and aspirations … this … relies on the capacity of persons to assume responsibility for their ends and to moderate the claims they make on their social institutions in accordance with the use of primary goods." (Rawls 1982, 170).

a fine wine, and yet my expected (just) income is not enough to sustain a life style in which meals are invariably accompanied by a fine wine, Rawls claims that I am to change my desires, my tastes.[39] If I cannot change my tastes, however, then it seems to me that Rawls would say that whilst I could not deny the validity of the rules of justice, I could be pitied for the fact that they "do not match my nature", that would indeed be, "my misfortune". I do not suffer because I did not deserve more, but because I was not entitled to more, that is why it is "my misfortune" rather than my fault.

This seems to me exactly what Rawls does say to the moral renegade in the passages near the end of A Theory of Justice[40], and is also compatible with what he says there about the nature of the institution of punishment - that it is the means to maintaining the well-ordered society. The outcome, if this theory of punishment is worked out is presented, and rejected, above in section 31. Although such an account is clearly not compatible with the remarks quoted above in which the justification for the *institution* of punishment appeals to such things as "moral worth" and "bad character", it is these retributive passages that seem to me to be an exception in A Theory of Justice.

---

[39]It is not enough for me just to regret that I cannot fulfil my desires, I have to change my desires because I have to affirm my nature as a free being by giving precedence to the demands of justice. This seems to me to be a consequence of the Kantian nature of Rawls's answer to the motivation problem. A more reasonable answer would, surely, allow me to constrain my desires (with regret) without actually having to revise them.

[40]Quoted above at the beginning of this section.

In the passages where Rawls considers punishment, he says one of three things; in the remarks cited above he talks as if there is an independent desert based justification; in the passages in the penultimate section of A Theory of Justice I have argued that he is best read as endorsing a legitimate expectations view consistent with the account of distributive justice, and this is also compatible with his remark that even in a "well-ordered society" punishment would be necessary to solve the assurance problem; that is, as the rational addition to the assent to the rules of justice. In the remarks in sections 38 and 39 of A Theory of Justice, however, Rawls brings in a further account, one derived from the First Principle.[41] The principle of liberty has it that: "Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all."[42] Rawls's claim is that the basket of liberties each may enjoy must be equal, and that the only way to ensure such equality is to restore disequilibrium caused by individuals claiming greater liberty than that to which they are entitled, and, better still to deter individuals from those actions which cause disequilibrium in the distribution of liberty. In these sections, it is equal liberty and an associated fair play theory of punishment that underlies Rawls's theory of penal sanctions. It is thus liable to

---

[41]"The principles justifying these sanctions can be derived from the principle of liberty", Rawls 1971, 241.

[42]Rawls 1971, 250.

the criticisms that I levelled at such theories in Chapter 4.[43]

The point of this Appendix was not to attempt to present a single, coherent, account of punishment culled from A Theory of Justice. What I wanted to do was show that the much cited passage in which Rawls denies the similarity between distributive and retributive justice is an exception not just to the rest of the book, but also to the other important passages on punishment. In the best of those passages, Rawls commits himself to a consequentialist justification for punishment within a legitimate expectations view of responsibility. What is interesting about this is that the reason he does this is that it is the only outcome of his endorsing a Kantian answer to the motivation problem; if individuals are *contingently* unable to realise their true selves as free beings then they cannot be held as deserving for the same reason as that which informs the principles of distributive justice. They can, however, be held as subjects of entitlements, and if they do badly as such it is a matter of regret, "their nature" is indeed "their misfortune". I have shown above, in Chapter 5, why I think such an answer to the problem of punishment is an unsatisfactory one.

---

[43]Rawls, I believe, does not see this as a different account of punishment from the claim that punishment is justified as an answer to the assurance problem. This is because he makes the same mistake as Dagger in thinking that fair play theories are of the two level consequentialist type, rather than independent justice based retributive accounts. This, as I showed above, is a consequence of conflating fair play and contractualist accounts. See §21.

Bibliography

Acton, H. B. (Ed.). (1969). The Philosophy of Punishment. London: Macmillan.

Alexander, L. (1980). "The Doomsday Machine: Proportionality, Punishment and Prevention". The Monist, 63, 199-227.

Allison, H. (1990). Kant's theory of Freedom. Cambridge: Cambridge University Press.

Anscombe, G. E. M. (1957). "Does Oxford Moral Philosophy Corrupt Youth?". The Listener(14 February), 266-71.

Arendt, H. (1982). Lectures on Kant's Political Philosophy. Chicago: Chicago University Press.

Aristotle (1957). The Politics (T.A. Sinclair revised T.J. Saunders, Trans.). Harmondsworth: Penguin.

Armstrong, K. G. (1961). The Retributivist Hits Back. Mind, 70, 471-90.

Arneson, R. (1982). "The Principle of Fairness and Free-Rider Problems". Ethics, 92, 616-33.

Aune, B. (1979). Kant's Theory of Morals. Princeton, N.J.: Princeton University Press.

Baier, A. C. (1993). "Moralism and Cruelty: Reflections on Hume and Kant". Ethics, 103, 436-457.

Baier, K. E. (1993). "Is Punishment Retributive?". Analysis, 16. References from reprint in Acton 1969, 130-137.

Barker, E. (1960). Social Contract: Essays by Locke, Hume, and Rousseau. London: Oxford University Press.

Barry, B. (1968). "Warrender and his Critics". Philosophy, 43, 117-37.

Barry, B. (1979). "And Who is my Neighbour?". Yale Law Journal, 88, 629-58. Reprinted in Barry 1991b, 40-77.

Barry, B. (1989). Theories of Justice: Volume 1 of A Treatise on Social Justice. Hemel Hempstead, Hants.: Harvester-Wheatsheaf.

Barry, B. (1990). Political Argument: A Reissue with a new Introduction. Berkeley: University of California Press.

Barry, B. (1991a). "Chance, Choice, and Justice". In Barry 1991b, 142-158.

Barry, B. (1991b). Liberty and Justice: Essays in Political Theory 2. Oxford: Clarendon Press.

Barry, B. (1995a). Justice as Impartiality. Oxford: Clarendon Press.

Barry, B. (1995b). "Rawls's Search for Stability". Revised version forthcoming in Ethics.

Barry, B. (1995c). "International Society From a Cosmopolitan Perspective". Unpublished Manuscript.

Barry, B. (MS). "Benefits and Contributions: An Analysis of Transfers". Unpublished Manuscript.

Becker, L. (1986). Reciprocity. London: Routledge & Kegan Paul.

Bedau, H. A. (1978). "Retribution and the Theory of Punishment". The Journal of Philosophy, 75, 601-620.

Benn, S. I. (1958). "An Approach to the Problems of Punishment". Philosophy, 33, 325-41.

Benn, S. I. (1967). "Punishment". In P. Edwards (Ed.), The Encyclopedia of Philosophy (pp. 30). London: MacMillan Publishers.

Bentham, J. (1789 (1982)). An Introduction to the Principles of Morals and Legislation. (J. H. Burns and H. L. A. Hart Eds.). London: Methuen.

Binswanger (1958). "The Case of Ellen West". In R. May, E. Angel, H. Ellenberger (Eds.), Existence New York: Doubleday.

Braithwaite, J. and Pettit, P. (1990). Not Just Deserts: A Republican Theory of Criminal Justice. Oxford: Clarendon Press.

Broad, C. D. (1915-16). "On the Function of False Hypotheses in Ethics". International Journal of Ethics, 26, 377-97.

Brown, D. (1971). Bury My Heart at Wounded Knee: An American History of the American West. London: Barrie and Jenkins.

Brown, S. (1962). "Has Kant a Philosophy of Law?" Philosophical Review, 71, 33-48.

Brubaker, S. (1989). "In Praise of Punishment". The Public Interest, 97, 44-55.

Brubaker, S. (1990). "In Praise of Punishment: A Reply to A. Rawls". The Public Interest, 99, 133-36.

Buchanan, A. (1990). "Justice as Reciprocity versus Subject-Centred Justice". Philosophy & Public Affairs, 19, 227-52.

Buchanan, J. (1975). The Limits of Liberty. Chicago: University of Chicago Press.

Burgh, R. (1982). "Do the Guilty Deserve Punishment?". The Journal of Philosophy, 79, 193-210.

Byrd, B. S. (1989). "Kant's Theory of Punishment: Deterrence in its Threat, Retribution in its Execution". Law and Philosophy, 8, 151-200.

Charvet, J. (1966). "Criticism and Punishment". Mind, 75, 578.

Charvet, J. (1981). A Critique of Freedom and Equality. Cambridge: Cambridge University Press.

Charvet, J. (1995). The Idea of an Ethical Community. Ithaca, N.Y.: Cornell University Press.

Christman, J. (Ed.). (1989). The Inner Citadel: Essays on Individual Autonomy. New York: Oxford University Press.

Cohen, G. "On the Currency of Egalitarian Justice". Ethics, 99, 906-944.

Cohen, M. (1939). "A Critique of Kant's Philosophy of Law". In G. T. Whitney, and D. F. Bowers, (Eds.), The Heritage of Kant Princeton, N.J.: Princeton University Press.

Cooper, D. (1971). "Hegel's Theory of Punishment". In Z. A. Pelczynski (Ed.), Hegel's Political Philosophy Cambridge: Cambridge University Press.

Cottingham, J. (1979). "Varieties of Retribution". Philosophical Quarterly, 29, 238-246.

Cranston, M. (1973). What are Human Rights. New York: Taplinger Publishing Co.

Dagger, R. (1985). "Rights, Boundaries, and the Bonds of Community: A Qualified Defense of Moral Parochialism". American Political Science Review, 79, 436-47.

Dagger, R. (1993). "Playing Fair with Punishment". Ethics, 103, 473-88.

Davis, M. (1983). "How to Make the Punishment Fit the Crime". Ethics, 93, 726-52.

Davis, M. (1986a). "Why Attempts Deserve Less Punishment than Complete Crimes". Law and Philosophy, 5, 1-33.

Davis, M. (1986b). "Harm and Retribution". Philosophy and Public Affairs, 15, 236-66.

Davis, M. (1989). "The Relative Independence of Punishment Theory". Law and Philosophy, 7, 321-350.

Day, J. (1978). Retributive Punishment. Mind,87, 498-516.

Dennett, D. (1987). The Philosophers' Lexicon. American Philosophical Association.

Devlin, Lord. (1959). The Enforcement of Morals. London: Oxford University Press.

Dolinko, D. (1991). "Some Thoughts About Retributivism". Ethics, 101, 537-59.

Duff, R. A. (1977). "Psychopathy and Moral Understanding". American Philosophical Quarterly, 14, 189-200.

Duff, R. A. (1986). Trials and Punishments. Cambridge: Cambridge University Press.

Duff, R. A. (1988). "Punishment and Penance: A Reply to Harrison". Proceedings of the Aristotelian Society (Supplementary Volume),62, 153-67.

Duff, R. A. (1993). "Introduction". In Duff, R. A. Punishment (pp. xi-xvii) Aldershot, Hants.: Dartmouth Publishing Co.

Dworkin, R. (1977). Taking Rights Seriously. London: Duckworth.

Dworkin, R. (1981). "What is Equality? Part II: Equality of Resources". Philosophy and Public Affairs, 283-345.

Ezorsky, G. (Ed.). (1972). Philosophical Perspectives on Punishment. Albany, New York: State University of New York Press.

Faris, E. (1914). "The Origin of Punishment". International Journal of Ethics, 24, 54-67.

Farrell, D. (1985). "The Justification of General Deterrence". The Philosophical Review, 94, 367-94.

Farrell, D. (1988). "Punishment without the State". Nous, 22, 437-53.

Farrell, D. (1989). "On Threats and Punishments". Social Theory and Practice, 15, 125-54

Farrell, D. (1990). "The Justification of Deterrent Violence". Ethics, 100, 301-17.

Feinberg, J. (1965). "The Expressive Function of Punishment". The Monist, 49, 397-423.

Fingarette, J. (1977). "Punishment and Suffering". Proceedings and Addresses of the American Philosophical Assosciation, 50, 499-525.

Finnis, J. (1972). "The Restoration of Retribution". Analysis, 32, 131-35

Fishkin, J. (1992). The Dialogue of Justice. New Haven: Yale University Press.

Fitzgerald, F. S. (1926). The Great Gatsby. Harmondsworth: Penguin.

Fleischacker, S. (1992). "Kant's Theory of Punishment". In H. Williams (Ed.), Essays on Kant's Political Philosophy (pp. 191-212). Cardiff: University of Wales Press.

Flew. A. (1954). "The Justification of Punishment". Philosophy, 29. References from reprint in Acton 1969, 83-101.

Foot, P. (Ed.). (1967). Theories of Ethics. Oxford: Oxford University Press.

Foucault, M. (1976). Mental Illness and Psychology. (A. Sheridan, Trans.). New York: Harper & Row.

Foucault, M. (1977). Discipline and Punish: The Birth of the Prison. (Alan Sheridan, Trans.). London: Penguin.

Foucault, M. (1984-5). "Dream, Imagination and Existence". Review of Existential Psychology and Psychiatry, 19, 31-78.

Frankfurt, H. (1971). "Freedom of the Will and the Concept of a Person". Journal of Philosophy, 68, 5-20. Reprinted in Frankfurt 1988, 11-25.

Frankfurt, H. (1988). The Importance of What we Care About: Philosophical Essays. Cambridge: Cambridge University Press.

Garland, D. (1990). Punishment and Modern Society: A Study in Social Theory. Oxford: Clarendon Press.

Gauthier, D. (1986). Morals by Agreement. Oxford: Clarendon Press.

Gewirth, A. (1978) Reason and Morality. Chicago: University of Chicago Press.

Gibbard, A. (1990). Wise Choices, Apt Feelings. Cambridge, Mass.: Harvard University Press.

Gibbard, A. (1991). "Constructing Justice". Philosophy & Public Affairs, 20, 264-79.

Goldman, A. (1979). "The Paradox of Punishment". Philosophy & Public Affairs, 9, 42-58.

Goldman, A. (1982). "Toward a New Theory of Punishment". Law and Philosophy, 1, 57-76.

Gray, J. (1989). Liberalisms: Essays in Political Philosophy. London: Routledge.

Gross, H. and Hirsch, A. von (Eds.). (1981). Sentencing. New York: Oxford University Press.

Hampsher-Monk, I. (1992). A History Of Modern Political Thought. Oxford: Blackwell Publishers.

Hampton, J. (1984). "The Moral Education Theory of Punishment". Philosophy and Public Affairs, 13, 208-38.

Hampton, J. (1986). Hobbes and the Social Contract Tradition. Cambridge: Cambridge University Press.

Hardin, R. (1991). "Book Review of Beitz, Political Equality". Political Theory, 19(4), 667.

Hare, R. M. (1981). Moral Thinking: Its Levels, Method and Point. Oxford: Oxford University Press.

Harel, A. (Forthcoming). "Efficiency and Fairness in Criminal Law: The Case for a Criminal Law Principle of Comparative Fault". California Law Review.

Harsanyi, J. (1982). "Morality and the Theory of Rational Behaviour". In A. Sen. and B. Williams (Eds.), Utilitarianism and Beyond Cambridge: Cambridge University Press.

Hart, H. L. A. (1955). "Are There Any Natural Rights?". Philosophical Review, 64, 175-91. Reprinted in Waldron 1984, 77-90.

Hart, H. L. A. (1958). "Legal Responsibility and Excuses". In S. Hook (Ed.) Determinism and Freedom. New York: New York University Press. Reprinted in Hart 1968, 28-53.

Hart, H. L. A. (1959). "Prolegomenon to the Principles of Punishment". Proceedings of the Aristotelian Society, 60. Reprinted in Hart 1968, 1-27.

Hart, H. L. A. (1963). Law, Liberty and Morality. London: Oxford University Press.

Hart, H. L. A. (1968). Punishment and Responsibility. New York: Oxford University Press.

Hegel, G. W. F. (1955). Grundlinien der Philosophie des Rechts (J. Hoffmeister ed.). Hamburg: Felix Meiner.

Hegel, G. W. F. (1967). Hegel's Philosophy of Right (T. M. Knox, Trans.). London: Oxford University Press.

Hegel, G. W. F. (1970). Philosophische Propädeutik. In D. Henrich (Eds.), Werke in zwanzig Bänden Frankfurt am Main: Suhrkamp.

Hegel, G. W. F. (1974). Hegel's Lectures on the History of Philosophy (E. S. Haldane, Trans.). Atlantic Highlands, N.J.: The Humanities Press.

Hegel, G. W. F. (1991). Elements of the Philosophy of Right (H. B. Nisbet, Trans.). (A. Wood ed.). Cambridge: Cambridge University Press.

Hill, T. E. (1978). "Kant's Anti-Moralistic Strain". Theoria, 44, 131-51. Reprinted in Hill 1992, 176-195.

Hill, T. E. (1989). "The Kantian Conception of Autonomy". In J. Christman (Eds.), The Inner Citadel: Essays on Individual Autonomy (pp. 91-105). New York: Oxford University Press. Reprinted in Hill 1992, 76-96.

Hill, T. E. (1992). Dignity And Practical Reason in Kant's Moral Theory. Ithaca, N.Y.: Cornell University Press.

Hirsch, A. von. (1976). Doing Justice: The Choice of Punishments. New York: Hill and Wang.

Hirsch, A. von. (1985). Past or Future Crimes: Deservedness and Dangerousness in the Sentencing of Criminals. Manchester: Manchester University Press.

Hirsch, A. von. (1990). "Proportionality in the Philosophy of Punishment: From 'Why Punish?' to 'How Much?'". Criminal Law Forum, 1, 259-90.

Hobbes, T. (1641). Leviathan (C. B. Macpherson, Ed.). Harmondsworth, Middx.: Penguin Books.

Hoekema, D. (1980). "The Right to Punish and the Right to be Punished". In H. G. Blocker and E. Smith (Eds.), John Rawls' Theory of Social Justice: An Introduction Athens: Ohio University Press.

Honderich, T. (1969). Punishment: The Supposed Justifications. (First ed.). London: Hutchinson.

Honderich, T. (1984). Punishment: The Supposed Justifications. (2nd edition with new postscipt). Cambridge: Pelican Press.

Honderich, T. (1988). A Theory of Determinism: The Mind, Neuroscience, and Life Hopes. Oxford: Clarendon Press.

Honderich, T. (1993). How Free Are You? Oxford: Oxford University Press.

Honig, B. (1993). "Rawls on Politics and Punishment". Political Research Quarterly, 46(1).

Honig, B. (1993). Political Theory and the Displacement of Politics. Ithaca, N.Y.: Cornell University Press.

Ihara, C. (1992). Review of John Kekes, Facing Evil. Ethics, 102(3), 650-51.

Kant, I. (1795). Perpetual Peace In L. W. Beck, R. Anchor and E. Falkenheim, (Eds.). On History. Indianapolis: Bobbs-Merrill, 1963.

Kant, I. (1787 (1965)). Critique of Pure Reason (Norman Kemp Smith, Trans.). (2nd. ed.). New York: St. Martin's Press.

Kant, I. (1797 (1991)). The Metaphysics of Morals (Mary Gregor, Trans.). Cambridge: Cambridge University Press.

Kant, I. (1930). Lectures on Ethics (Louis Infield, Trans.). London: Methuen & Co. Ltd.

Kant, I. (1948). The Moral Law: Kant's Groundwork of the Metaphysic of Morals (H. Paton, Trans.). London: Hutchinson.

Kant, I. (1960). Religion Within the Limits of Reason Alone (T. M. Green and H. H. Hudson, Trans.). New York: Harper & Bros.

Kant, I. (1967). Kant: Philosophical Correspondence, 1759-99 (Arnulf Zweig, Trans.). Chicago: University of Chicago Press.

Kierkegaard, S. Purity of Heart is to Will One Thing. (D. Steere, Trans.). London: Collins.

Klosko, G. (1992). The Principle of Fairness and Political Obligation. Lanham, Md.: Rowman & Littlefield.

Kymlicka, W. (1989). Liberalism Community And Culture. Oxford: Oxford University Press.

Kymlicka, W. (1990). Contemporary Political Philosophy: An Introduction. Oxford: Clarendon Press.

Kymlicka, W. (1991). "The Social Contract Tradition". In P. Singer (Ed.), A Companion to Ethics (pp. 186-196). Oxford: Blackwell Publishers.

Lacey, N. (1987). "Punishment". In D. Miller et al (Eds.), The Blackwell Encyclopaedia of Political Thought (pp. 409-412). Oxford: Blackwell Publishers.

Lacey, N. (1988). State Punishment. London: Routledge.

Lehning, P. (1991). "Liberalism and Capabilities: Theories of Justice and the Neutral State". Social Justice Research, 5,(3).

Locke, J. (1960). Two Treatises of Government. (P. Laslett, ed.). Cambridge: Cambridge University Press.

Lyons, D. (1965). Forms and Limits of Utilitarianism. Oxford: Oxford University Press.

Mabbott, J. D. (1939). "Punishment". Mind, 48. References from reprint in Acton 1969, 39-54.

Mabbott, J. D. (1955). "Professor Flew on Punishment". Philosophy, 30 References from reprint in Acton 1969, 115-129.

MacIntyre, A. (1984). After Virtue (2nd. ed.). Notre Dame, Ind.: University of Notre Dame Press.

Mackenzie, M. (1981). Plato on Punishment. Berkeley: University of California Press.

Matravers, D. (1995). "The Good Life: Love, Sacrifice and The Little Prince", The Ethical Record, forthcoming.

Matravers, M. (1995). "Political Obligation, Indoctrination and the Self-Reflective Society". Studies in Political Thought, forthcoming.

McCloskey, H. J. (1968). "A Non-utilitarian Approach to Punishment". In Bayles, M. (Ed.), Contemporary Utilitarianism (pp. 239-59). New York: Doubleday.

McCloskey, H. J. (1972). "'Two Concepts of Rules': A Note". Philosophical Quarterly, 22, 344.

McTaggart, J. (1896). "Hegel's Theory of Punishment". International Journal of Ethics, 6, 482-99.

Mitias, M. (1978). "Another Look at Hegel's Concept of Punishment". Hegel-Studien, 13, 175-185.

Montague, P. (1983). "Punishment and Societal Defense". Criminal Justice Ethics, 2, 31-36.

Moore, M. (1987). "The Moral Worth of Retribution". In F. Schoeman (Ed.), Responsibility, Character, and the Emotions. (pp. 179-219). New York: Cambridge University Press.

Moore, M. (1993). Foundations of Liberalism. Oxford: Clarendon Press.

Morris, H. (1968). "Persons and Punishment". The Monist, 52, 475-501. Reprinted in Gross and von Hirsch 1981, 93-109.

Mulhall, S and Swift, A. (1992). Liberals and Communitarians. Oxford: Blackwell Publishers.

Murphy, J. G. (1970). Kant: The Philosophy of Right. MacMillan.

Murphy, J. G. (1971). Three Mistakes About Retributivism. Analysis, 166-169.

Murphy, J. G. (1972). Kant's Theory of Criminal Punishment. In L. W. Beck (Ed.), Proceedings of the Third International Kant Conference, (pp. 434-441). D. Reidel Publishing Company.

Murphy, J. G. (1973). "Marxism and Retribution". Philosophy & Public Affairs, 2, 217-243.

Murphy, J. G. (1979). Retribution, Justice, and Therapy. Dordrecht, Holland: D. Reidel Publishing Company.

Murphy, J. G. (1985). "Retributivism, Moral Education, and the Liberal State". Criminal Justice Ethics, 4, 3-11.

Murphy, J. G. (1990). "Review of George Sher's Desert". Philosophical Review, 99, 280-83.

Nagel, T. (1986). The View From Nowhere. New York: Oxford University Press.

Nagel, T. (1991). Equality and Partiality. New York: Oxford University Press.

Narveson, J. (1974). "Three Analysis Retributivists". Analysis, 34, 185-93.

Nelson, W. N. (1990). Morality: What's in it for Me? A Historical Introduction to Ethics. Boulder, Colo.: Westview Press.

Nicholson, P. (1982). "Hegel on Crime". History of Political Thought, III, 103-121.

Nietzsche, F. (1967). On the Genealogy of Morals (W. Kauffman, Trans.). New York: Vintage Books.

Nino, C. S. (1983). "A Consensual Theory of Punishment". Philosophy & Public Affairs, 12, 289-306.

Nowell-Smith, P. (1961). Ethics. Harmondsworth, Middx. Penguin.

Nozick, R. (1974). Anarchy, State, And Utopia. Oxford: Basil Blackwell Ltd.

O'Neill, O. (1991). "Kantian Ethics". In P. Singer (Ed.), A Companion to Ethics (pp. 175-85). Oxford: Blackwell Publishers.

Pence, G. (1991). "Virtue theory". In P. Singer (Ed.), A Companion to Ethics (pp. 249-58). Oxford: Blackwell Publishers.

Philips, M. (1986). "The Justification of Punishment and the Justification of Political Authority". Law and Philosophy, 5, 393-416.

Pincoffs, E. L. (1966). The Rationale of Legal Punishment. Atlantic Highlands, N.J.: Humanities Press.

Plant, R. (1983). Hegel: An Introduction (2nd. edition.). Oxford: Blackwell Publishers.

Primoratz, I. (1989). "Punishment as Language". Philosophy, 64, 187-205.

Quinn, W. (1985). "The Right to Threaten and the Right to Punish". Philosophy and Public Affairs, 14, 327-73.

Quinton, A. (1954). "On Punishment". Analysis, 14. References from reprint in Acton 1969, 55-64.

Rawls, A. E. (1990). "Of Rawls, Responsibility, and Retribution". The Public Interest, 99, 128-33.

Rawls, J. (1955). Two Concepts of Rules. Philosophical Review, 64, 3-32. Reprinted in Foot, 1967, 144-70.

Rawls, J. (1964). "Legal Obligation and the Duty of Fair Play". In S. Hook (Ed.), Law and Philosophy (pp. 3-18). New York: New York University Press.

Rawls, J. (1971). A Theory Of Justice. Oxford: Oxford University Press.

Rawls, J. (1975). "Fairness to Goodness". Philosophical Review, 84, 536-54.

Rawls, J. (1978). "The Basic Structure as Subject". In A. Goldman and J. Kim (Eds.). Values and Morals. Dordrecht: Reidel.

Rawls, J. (1980). "Kantian Constructivism in Moral Theory". Journal of Philosophy, 77, 515-72.

Rawls, J. (1982a). "Social Unity and Primary Goods". In A. Sen and B. Williams (Eds.), Utilitarianism and Beyond (pp. 159-85). Cambridge: Cambridge University Press.

Rawls, J. (1982b). "The Basic Liberties and their Priority". In S. McMurrin (Ed.), The Tanner Lectures on Human Values, Volume 3. Salt Lake City, Utah: University of Utah Press.

Rawls, J. (1985). "Justice as Fairness: Political not Metaphysical". Philosophy and Public Affairs, 14, 223-51.

Rawls, J. (1987). "The Idea of an Overlapping Consensus". Oxford Journal of Legal Studies, 7, 1-25.

Rawls, J. (1988). "The Priority of Right and Ideas of the Good". Philosophy and Public affairs, 17, 251-76.

Rawls, J. (1989). "The Domain of the Political and Overlapping Consensus". New York University Law Review, 64, 233-55.

Rawls, J. (1993a). Political Liberalism. New York: Columbia University Press.

Rawls, J. (1993b). "The Law of Peoples". In S. Shute and S. Hurley (Eds.), On Human Rights (pp. 41-82). New York: Basic Books.

Reed, T. M. (1978). "On Sterba's 'Retributive Justice'". Political Theory, 6, 373-376.

Riley, P. (1983). Kant's Political Philosophy. Totowa, New Jersey: Rowman and Littlefield.

Ripstein, A. (1994). "Equality, Luck, and Responsibility". Philosophy and Public Affairs, 23, 3-23.

Rorty, R. (1989). Contingency, irony, and solidarity. Cambridge: Cambridge University Press.

Russell, B. (1946). A History of Western Philosophy. London: George Allen & Unwin Ltd.

Sadurski, W. (1989). "Theory of Punishment, Social Justice, and Liberal Neutrality". Law and Philosophy, 7, 351-73.

Sandel, M. (1982). Liberalism and the Limits of Justice. Cambridge: Cambridge University Press.

Scanlon, T. M. (1988). "The Significance of Choice". In McMurrin, S. (Ed.) The Tanner Lectures on Human Values, Volume 8. Salt Lake City, Utah: The University of Utah Press, (pp. 151-216).

Scanlon, T. M. (1992). "The Aims and Authority of Moral Theory". Oxford Journal of Legal Studies, 12, 1-23.

Scheffler, S. (1992). "Responsibility, Reactive Attitudes, and Liberalism in Philosophy and Politics". Philosophy & Public Affairs, 21, 299-323.

Scheid, D. E. (1983). "Kant's Retributivism". Ethics, 93, 262-282.

Shell, S. M. (1980). The Rights of Reason. Toronto:

Sher, G. (1987). Desert. Princeton, N.J.: Princeton University Press.

Simmons, A. J. (1979). Moral Principles and Political Obligations. Princeton, N.J.: Princeton University Press.

Skillen, A. (1980). "How to Say Things With Walls". Philosophy, 55, 509-23.

Smart, J. (1968). "Extreme and Restricted Utilitarianism". In Bayles, M. (Ed.) Contemporary Utilitarianism (99-115). New York: Doubleday.

Smart, J. and Williams, B. (1973) Utilitarianism For and Against. Cambridge: Cambridge University Press.

Smith, M. B. E. (1973). "Is There a Prima Facie Obligation to Obey the Law?". Yale Law Journal, 82, 950-76.

Smith, S. (1989). Hegel's Critique of Liberalism. Chicago: University of Chicago Press.

Speake, J. (Ed.). (1979). A Dictionary of Philosophy. London: Pan Books Ltd.

Sprigge, T. L. S. (1968). "A Utilitarian Reply to Dr McCloskey". In M. D. Bayles (Ed.), Contemporary Utilitarianism (pp. 278-82). New York: Doubleday.

Stephen, J. F. (1883). A History of the Criminal Law of England. London: Macmillan.

Sterba, J. P. (1977). "Retributive Justice". Political Theory, 5, 349-362.

Stern, R. (Ed.). (1993). G. W. F. Hegel: Critical Assesments. New York: Routledge.

Stillman, P. (1976). "Hegel's Idea of Punishment". Journal of the History of Philosophy, 14, 169-182.

Tamburrini, C. (1992) Crime and Punishment?. Stockholm: Almqvist and Wiksell International.

Taylor, C. (1979). Hegel and Modern Society. Cambridge: Cambridge University Press.

Taylor, C. (1985). Human Agency and Language: Philosophical Papers 1. Cambridge: Cambridge University Press.

Ten, C. L. (1987). Crime, Guilt, and Punishment. Oxford: Clarendon Press.

Tunick, M. (1992a). Hegel's Political Philosophy: Interpreting the Practice of Legal Punishment. Princeton, N.J.: Princeton University Press.

Tunick, M. (1992b). Punishment: Theory and Practice. Berkeley: University of California Press.

Waldron, J. (Ed.). (1984). Theories of Rights. Oxford: Oxford University Press.

Walker, N. (1969). Sentencing in a Rational Society. Harmondsworth, Middx.: Penguin.

Walker, N. (1991). Why Punish? Oxford: Oxford University Press.

Wasserstrom, R. (1980). Philosophy and Social Issues. Notre Dame, Ind.: University of Notre Dame Press.

Williams, B. (1979). "Conflicts of Values". In A. Ryan (Ed.), The Idea of Freedom: Essays in Honour of Isaiah Berlin Oxford: Oxford University Press.

Williams, B. (1981). Moral Luck: Philosophical Papers 1973-1980. Cambridge: Cambridge University Press.

Williams, B. (1985). Ethics and the Limits of Philosophy. London: Fontana Press/Collins.

Williams, H. (1983). Kant's Political Philosophy. Oxford: Oxford University Press.

Wolf, S. (1989). "Sanity and the Metaphysics of Responsibility". In Christman 1989, 137-151.

Wood, A. (1990). Hegel's Ethical Thought. Cambridge: Cambridge University Press.