Punishment

Jonathan Bennett

This paper, completed in 2002, has not been published except at www.earlymoderntexts.com.

1. The good that punishment can do

How can a state be morally justified in punishing some of its citizens? In tackling this I shall set aside three important matters: we do not morally approve of all the laws of the land, so that sometimes there is a legal but not a moral case against an offender; we can do more things about crime than just punish the criminals, for example remedying the familial and social conditions that encourage it; and, thirdly, many actual penal institutions do things to convicts that are indefensible by any decent standard. My topic may be humanly less important than any of those three; but I want to discuss it and not them.

Like many others, I hold that the punishment is justified only to the extent that it does good, that is, leads to a better over-all state of affairs than would have obtained otherwise. Some versions of this view, however, assign to doing-good a role in which it swamps other considerations which most of us think are also essential to a defensible penal system. My aim in this paper will be remedy this defect—to reconcile a doing-good justification with the rest of what we think about punishment.

The fear that this cannot be done has led some to question whether our having penal systems and procedures does any good. Jacobs may be right in stressing how hard it is to discover what the consequences are of any given system of punishment (1999: 540), but it would be lamentable if that led us to omit doing-good from our view of what justifies punishment. I shall return to this topic at the end of Section 7, offering to relieve Jacobs of his fears.

I usually think of the good that punishment can do in terms of the deterrence of potential offenders—the punished person or others who learn about his punishment. In all the uncertainty about what our penal procedures achieve, it seems certain that they have a deterrent effect without which our society would be lost. There are other possible goods: making the convict a better person; placating victims and their kin; increasing our sense of the majesty and importance of the law; and so on. Husak (1992: 460–1) sketches some of the recherché kinds of good that punishment has been thought to do, and plausibly conjectures that they are attractive to philosophers mainly because they are safe from empirical testing. However, I need take no stand on that. My concern in this paper is with the abstract notion of punishment's doing good; I have no need to consider how it might do so.

2. What a theory of punishment should do

In many areas of morality utilitarianism makes better sense than any of its rivals, but I cannot square it with my moral convictions about punishment. I part company with it at this point, for familiar reasons which I shall sketch here for safety's sake.

Utilitarianism entails that the rightness of an action depends on the value of what *results* from it: to evaluate a given action we must look at the present and *later* states of the actual world and of nearby worlds where it isn't performed. That emphasis on present and future seems wrong when we apply it to punishment. When the utilitarian considers what to do with the man in the dock, his route to an answer runs through facts about what treatment of the man *will* lead to the best results; but most of us accord critical relevance to a certain question about the past, namely *Did he do it?*

The utilitarian can still regard innocence as *indirectly* relevant to the rightness of punishing the person (I ignore the purely verbal point that 'punish' is the wrong word for an innocent man). His innocence, a fact about the past, will leave causal traces in his memory and probably elsewhere in the world; and if we throw the man in jail, the consequences of that may combine with the traces of his innocence to produce further especially bad results—e.g. the bitter frame of mind of someone who thinks he has been treated unjustly; or the risk that his innocence will be discovered, which will harm the deterrent aspect of the penal system as a whole. (We weaken deterrence not only when we fail to convict those known to be guilty but also when we convict people known to be innocent. The deterrent system requires that a high probability that *whether* I fare well in the future depends upon whether I offend against the law; that probability falls when I risk being punished even if I don't offend, as well as when I don't risk being punished if I do offend.)

Still, it could happen that the conviction of a person known to be innocent really would, on balance, produce better results than his acquittal; and we could trade moral intuitions about such cases. Or we can by-pass them and try something different. Think about actual cases where you consent to some guilty person's being punished, and ask yourself whether you regard his having committed the crime as *directly* relevant to your consent to his punishment. Most of us steadfastly answer Yes, even when we are operating at what Hare calls the 'critical' level, coolly thinking like moral theorists and not merely shooting from the hip with intuitions. I want an account of punishment which yields that affirmative answer—a positive account, not a mere denial of utilitarianism.

As well as handling innocence properly, the account should deal well with the severity of punishments. Civilized people think that a penalty can be too severe, given the gravity of the offence. (I shall regularly use 'severity' for penalties and 'gravity' for offences.) These judgments pose a problem for utilitarians, because the facts about gravity—like the fact that there was an offence at all—pertain to the past, not to the present and future.

I suggest that those two points are really parts of a single one. Innocence is best seen as the lower limit on gravity, and lack of punishment as the lower limit on severity. Guilt and innocence, you may feel, are not matters of degree as gravity-of-offence is; but I do not retract. Similarly, the contrast between *there are Fs* and *there are no Fs* is not one of degree, whereas the numerousness of the Fs is a matter of degree; but the former of these (non-degree) is just one part of the latter (degree), for the difference between 'There are no Fs' and 'There are Fs' is just the difference between

'The number of Fs = 0' and 'The number of Fs > 0'. Likewise, I suggest, the difference between 'He is guilty' and 'He is innocent' is that between 'The gravity of his offense = 0' and 'The gravity of his offense > 0'; and similarly for severity of punishments. None of my main points will depend on this, however.

A utilitarian can say that his standards do not justify ferocious punishments, because the hurt they would bring to the convict and others is not offset by any good they would bring to society as a whole. But that may be whistling in the dark. Despite the experience of 19th century England, I know of no evidence that sound consequential reasoning will *in general* reject punishments that offend our moral sensibilities. Various horrible penalties might have an over-all good effect, through deterrence, but most of us are not willing to listen. We have a robust and free-standing concept of a penalty's being too severe for a given offence. The utilitarian cannot let gravity bear directly and immediately on severity.

This holds not only for utilitarianism but also for some other brands of consequentialism. Braithwaite and Pettit (1990), who look to punishment for a good they call 'dominion', are as vulnerable here as is utilitarianism (thus Montague 1995: 104–7). I shall return to consequentialism in section 4.

3. Desert theory

I want, then, an over-all account of the morality of punishment which does make guilt or innocence immediately relevant to whether, and how severely, someone should be punished. You may think: 'A person who has not offended does not *deserve* to be punished; one who has offended only mildly does not *deserve* to be punished severely. There's the account you were looking for. Utilitarianism goes wrong by omitting the concept of desert.' I agree with the first part of that, but I cannot see it as solving my problem. To say that offenders deserve to be punished is to imply that their guilt is a reason—a direct, immediate reason—why it is all right to punish them. I agree with that much; but this is the judgment that was to be explained or justified; it cannot serve as the explanation or justification.

The word 'desert' and its cognates are used far beyond the bounds of punishment and reward.(For useful forays into its other territory see Feldman 1995 and Cupitt 1996.) Its friends might be expected to have some account of the entire genus *desert* which, combined with a differentia and with some general moral principles, yields conclusions about how desert relates to punishment. But I nothing like this seems ever to have been attempted. George Sher (1987) has faced up to the breadth and variousness of the contexts in which we employ 'desert', and has acknowledged that 'the diversity of desert's normative sources threatens the notion with disintegration' (p. 21). He offers to meet the threat:

> Given the diversity of its normative sources, in what sense, if any, is desert a *single* concept at all?. . . I shall argue that the concept does have an important unity, but that this unity exists not at the level of moral principles and values, but at the deeper level of conception of the person. As diverse as the relevant principles and values are, they share a common vision of what, in the end, we are. (p. 150)

The unity for which Sher argues is a major theme in his book. He summarizes it here:

> Desert-claims, though grounded in diverse moral principles, are nevertheless unified by the fact that they all presuppose a single conception of the self. On that conception, persons are both constituted by their preferences and abilities and extended over time. (p. 169)

I respect Sher's work on this topic, but I do not think he would claim to have provided for the term 'desert' the kind of conceptual unity needed for it to play the part I have demanded of it.

Desert theorists have to consider whether someone's deserving punishment makes it obligatory, or merely *permissible*, to punish him. They are divided on this, but—this being a striking mark of the theoretical poverty of the concept of desert—nobody has significantly argued for either answer to the question. Stands are taken, presumably on the basis moral intuitions, but the concept of desert does no service on either side.

The answer 'When deserved, punishment is obligatory' implies that *sometimes it is right to punish someone against utility*. By this I mean that sometimes it is right to punish someone with a certain degree of severity even though better over-all consequences (however evaluated) would be achieved by punishing him less or not at all.

Some philosophers seem to find it obvious that desert does not merely allow but requires punishment. Thus Sher: 'It is central to our beliefs about desert that a person may deserve reward or punishment (and, hence, that he *ceteris paribus* ought to receive it) even if his receiving it will not maximize overall utility' (p. 12). (See also Moore 1903: 214, and Brock 1973: 267.) Yet Honderich (1976: 143) wrote: 'There no longer are defenders of the traditional retribution theory, or at least of the version that we are obliged rather than permitted to punish offenders because they deserve it.' Walker (1999: 601) agrees: 'What can fairly be called "modern retributivism" has abandoned the Kantian notion of an obligation to punish, and substituted a mere right to punish. . . Modern retributivists. . . insist. . . that while disproportionate leniency is permissible, disproportionate severity is not.'

I applaud the moral opinion that we should never punish against utility, but I want an account of punishment containing that as a theorem rather than as a mere intuitive add-on.

The only version of desert theory that has any chance of providing this, I think, is the one that equates someone's deserving punishment with his losing some of his rights—not to be put in prison, not to be fined, and so on. An adherent of this position can *argue* to the conclusion that punishment is optional rather than mandatory, starting from the premiss—widely accepted among rights theorists—that for you to have a right is for others to be required not to treat you in certain ways. From this it follows that your losing a right is others' losing a requirement, thus gaining what in moral philosophy we call a permission. This account yields the theorem that guilt makes punishment permissible and does not make it mandatory.

That would be impressive if the concept of *rights* were independently serviceable in this area of moral theory; but I do not think that it is. Its best chance of success is in connection with the severity of penalties, and when I reach that topic in section 10 I shall argue that in that regard the rights concept fails. Without some such success—that is, without a more general theory about how rights relate to punishment—nothing is *explained* in the statement that punishment is only optional rather than mandatory because it comes into play through someone's loss of rights.

I shall offer an abstract theory of punishment, purporting to ground the general principles that determine who should be punished and how severely. It agrees with desert theorists that utilitarianism is wrong about the relevance of whether and how severely the person has offended, yet comes nearer to utilitarianism than to any other well known moral

theory. I shall argue that a rather small departure from utilitarianism, when added to other true things, enables one to generate all the commonly and deeply held views about penal justice—thus drawing right-wing conclusions from left-wing premises.

4. Consequentialism

I have played off utilitarianism against desert theory, complaining that the latter is not really a theory. You may think that in this intellectual drama there are too few players, and that I have snubbed consequentialism—a genus of which utilitarianism is just one species. According to consequentialism, the rightness of my ϕing now depends exclusively upon whether my ϕing will produce over-all better states of affairs—a better world—than would ensue from my acting in any other way. One might think that the value of the world that results from my acting in a certain way now supervenes on facts about the future—ones that lie causally and temporally downstream from my action. But that is not always so, as I now explain, thereby adding strength to the objection I have supposed you to be making.[1]

A consequentialist may value diachronic states of affairs—ones that pertain to a period of time—for example holding that the value of a composer's completing his symphony today is increased by his having worked hard on it for the past year. Such a consequentialist must hold that the rightness of my ϕing now depends in part on facts about the past. He attaches value to (let us say) the obtaining of A followed an hour later by B. I am wondering whether to ϕ now, believing that my ϕing now would make B obtain thirty minutes hence. For our consequentialist, the question of whether I ought to ϕ now can depend in part on whether A was the case thirty minutes ago. His judgments on behaviour at a given time, in short, do not depend purely upon facts about possible futures; and they do not fall under the axe of my criticism of utilitarianism.

For our consequentialist, the value of is being the case that A-and-then-B is not the simple sum of the value of its being the case that A and the value of its being the case that B. When in this section I write about attaching value to a diachronic state of affairs, I mean a value that it has holistically, not by summing the values of its temporal parts.

A consequentialist who values some diachronic states of affairs (understood in the holistic manner just explained) must sometimes look backward in time, as well as forward, to judge how someone should behave now. There is no denying this, nor can I brush this sort of consequentialism aside as rare, peculiar, or idiosyncratic—for it is none of those.

One reason some moralists have for accepting a species of consequentialism other than utilitarianism is that they value distributional fairness; and it would be an odd thinker who valued fairness in a synchronic manner but not diachronically. Theodora and Belisarius are both in pain, and I could help her by ϕing, or him by ψing, but not both. That her suffering is worse is a reason for ϕing, as a matter of synchronic fairness; that his suffering has been going on for much longer is a reason for ψing, as a matter of diachronic fairness. One could hardly be moved by the former consideration and not by the latter.

Another reason a moralist might have for deviating from utilitarianism, strictly so-called, is that he too values the having of true beliefs—not merely the pleasure or satisfaction of thinking one's beliefs are true, but their actually being so. Now, I can by ϕing bring it about that Theodora believes that

[1] In this section I develop ideas that came to me years ago from Frances Howard-Snyder. Although she convinced me back then, I was all set to slide over them in the present paper until she read a draft and gave a tug on the reins.

her cousin got married in Sicily last month; and if (but only if) that belief is true, our moralist regards Theodora's having it as a good state of affairs. For this consequentialist, then, whether it is right for me to ϕ now may depend in part upon whether a certain event occurred last month in Sicily—so that judging my behaviour requires looking to the past. It would be morally weird to evade this result by valuing true beliefs about the present and future but not about the past.

As for the truth of beliefs, so also for the satisfaction of desires. Belisarius wants to meet his father. This can be understood internally, as valuing his meeting someone whom he *firmly and durably believes to be* his father—that being something a utilitarianism might value. But some moralists of a consequentialist stripe value the actual satisfaction of the desire that Belisarius actually has, which is to meet the person who *is* his father. Suppose that I am now so placed that by ϕing I can cause Belisarius to believe that the man he is now talking to is his father; the moralists we are now considering will have to hold that the rightness of my acting in this way may depend in part on whether that man begot Belisarius at some time in the past.

Fairness in distribution, truth of beliefs, satisfaction of desires: those are some of the values that might induce a moralist to opt for some species of consequentialism other than utilitarianism; and each implies that moral judgments on behaviour must sometimes look back as well as forwards. Indeed, it seems to me that utilitarianism may be the only stable and respectable form of consequentialism that does not value diachronic states of affairs and is therefore able to judge behaviour without looking backward.

What I have to worry about, however, is not

> any consequentialist who holds that moral judgment on behaviour may depend directly on facts about the past,

but rather

> Any consequentialist who holds that moral judgment on punitive behaviour depends directly on facts about past offences.

Suppose we are confronted by one such. He says: 'A state of affairs in which suffering comes to someone who has previous committed a crime is, *ceteris paribus*, better than one in which suffering comes to someone who has not committed a crime.' I ask him 'Why?' He may answer: 'Because in the former case the suffering is deserved.' If he says this, my remarks in section 3 apply to him too. I usually think of desert theorists as belonging to the 'deontological' camp, using the language of its being *wrong* to punish the innocent, rather than as consequentialists who will speak of the *badness* of states of affairs in which innocent people are punished. But for present purposes, that difference does not matter. Either way, my point stands that nobody seems to have put any stuffing into the term 'desert', so that invoking it will help to explain and justify what is going on when we connect punishment with guilt.

Our consequentialist might instead say—as I think Moore would—that it is just a *basic* value fact that a state of affairs in which someone suffers would be better if he had previously committed a crime. I cannot accuse this person of pretending to explain what he does not explain. I do say that he treats as basic, and thus inexplicable, something for which I shall offer an explanation.

From now on, I shall use 'utilitarian' and its cognates to cover every morality in which the rightness of punitive behaviour depends upon the value of the resultant states of affairs, with the convict's degree of guilt not counting directly as an element in any such state of affairs.

5. Reactive attitudes

The core of the theory that I shall offer comes from work in which P. F. Strawson (1962) introduces the concept of a reactive attitude. That is an attitude towards a person, in which the chief ingredient is a certain kind of feeling. I cannot define the kind, but I shall try to fix it with help from some contrasts. Consider first these three:

(1) You are hit and damaged by a branch that falls from a tree in a high wind.

(2) Someone innocently turns on his radio transmitter, which happens to trigger your garage door so that it unexpectedly closes and hurts you.

(3) Someone deliberately punches you in the face.

Anger and resentment may be appropriate in **(3)**, where your misfortune came from another person's bad attitude towards you. They are not appropriate in **(2)**, where there is a person but no attitude, or in **(1)** where there is not even a person. Now look at another trio:

(1) You stumble across some treasure that the law allows you to keep.

(2) A beggar persuades you to give him a dollar in exchange for a lottery ticket; it turns out to be the winning ticket, making you rich.

(3) A loved and loving grandparent bequeaths a fortune to you.

Here again, **(3)** is a fit subject of gratitude; but **(2)** is not because no beneficent attitude was involved, and **(1)** is not because in that case there is no other person.

Strawson contrasts the likes of resentment and gratitude with what he calls 'the objective attitude'. If you adopt the objective attitude towards the person who punches you, that may lead you to study his pathology, to consider why he acted like that, and to plan how to reduce the chances that he will do it again. You could also coldly adopt the objective attitude with respect to the bequest from your grandfather, perhaps wanting to understand its psychology so as to improve the chances of getting an inheritance from your other grandfather as well.

The utilitarian account of punishment says that our penal conduct should be framed by objectivity of attitude, the primary question always being 'How should we handle this person so as to. . .

. . . reduce the chances of his offending in future?'

. . . reduce the chance of others' offending similarly?'

. . . mitigate the suffering of the victims?'

or the like. In my own thinking about punishment, I focus mainly on the first of those; but none of my main claims depends on that.

All utilitarian thinking is framed by objectivity of attitude, and thus shoves aside the reactive attitudes. I say 'shoves aside' because the two kinds of attitude will not mix. Feelings such as those of resentment and gratitude do not easily cohabit with a concern to study the subject's behaviour patterns and plan one's own behaviour in the light of them. Strawson notes that one can often dispel a hostile reactive feeling by cultivating objectivity of attitude towards the offender, e.g. dispelling indignation by viewing him as 'a case'. (As that suggests, objectivity of attitude does not have to generate inquiry into causes and so on. It may instead consist merely in *thinking* of the offender as 'a case'—of him as a natural object and of his actions as natural events with knowable causes.) The immiscibility could work the other way: my renewed surge of angry resentment at how he had damaged me drowned my nascent attempt to understand why he behaved as he did.

This is a conflict between two attitudes of mind, not between two propositions. In an earlier writing on this topic

(Bennett 1980) I tried and failed to explain this fact that objectivity and reactivity won't combine in a single mind at one time. I also tried to induce Strawson to have a shot at explaining this, and failed in that too. But there can be no doubt that the two *won't* combine, and anyone who understands Strawson's paper will see this as a conflict between two attitudes of mind, not between two propositions.

If it were a propositional conflict, it would be something like this: *In adopting the objective attitude towards Agent in respect of his ϕing, I affirm that Agent is a natural object and that his ϕing was a natural event. In resenting his ϕing I deny those propositions*. If that were right, then combining the two attitudes would be assenting to a contradiction. But things stand quite otherwise. Objectivity of attitude involves a willingness to view a person's action as a natural event in the history of a natural object, but it does not imply any refutable proposition about the person or the action. *Whatever* I believe about some action of yours, I can adopt the objective, inquiring attitude towards it. Even if I think that some of its aspects or causes or sources ('noumenal freedom', perhaps) lie beyond the reach of empirical discovery, I may still want to know *what* aspects of your action lie within nature, *how far* its non-natural aspects extend. In addressing that question, I adopt the objective attitude. There is *nothing* towards which objectivity is factually inappropriate, in the way that fear is inappropriate towards what is not dangerous, and pity towards what is not unfortunate.

If I resent something you have done, on the other hand, I do in a fashion commit myself to some propositions about you and your action, but those commitments do not explain the conflict between reactive and objective attitudes: my resentment does not imply a denial of the almost empty proposition that you can be viewed or approached as a natural object. But the commitments exist all the same. It is plainly, simply, objectively wrong or inappropriate for me to resent your ϕing if:

(1) your ϕing was not in any way adverse to me, or

(2) your ϕing resulted from unavoidable ignorance or error, or

(3) you ϕed because you are a baby, or

(4) you ϕed because you are seriously mentally ill, or. . .

and so on. One *can* feel resentment in one of those cases: I once saw a grown man lose his temper with a stack of chairs. But we all judge it to be incorrect, improper, stupidly inappropriate, immature, to permit oneself a reactive attitude in cases of the sort I have listed. Why?

The answer has two parts. If I resent your ϕing, that ought to be because I object to the attitude that your ϕing shows you to have taken towards me—I resent your having been mean, hostile or indifferent to my welfare. That explains items **(1)** and **(2)** in my list. Next, it would be stupid for me to care in this way about your attitude towards me unless I saw myself as actually or possibly standing in a level reciprocating adult relationship with you. That explains items **(3)** and **(4)** in the list. A full defence of these two claims would probably bring in conceptual analysis, normative ethics and human psychology; I shall not go into all that now, and indeed I am not sure how to.

Even when reactivity would not be stupidly inappropriate in any of the ways I have described, it might be predictably *counter-productive*. Most of us would agree, indeed, that for consequential reasons reactive attitudes should be excluded from certain classes of relationship—for example, that between a psychotherapist and his client. The therapist may of course properly have many feelings towards the client—affection, sympathy, pity and so on—but mischief will ensue if he enters the area in which resentment and gratitude flourish. I mention this partly so as to distinguish

it from the other list of factors that make reactivity inappropriate: resentment's being objectively wrong—stupidly inappropriate, plainly immature—is not the same as its being merely unwise and counter-productive.

Summing up: objectivity of attitude is never inappropriate in the light of the facts; reactivity of attitude can be made inappropriate by facts about inadvertence, unavoidable ignorance, mental incompetence, or the like. In other words: *objectivity is always permissible and sometimes required; reactivity is sometimes permissible and never required.* Let me emphasize that last point. Nothing of this form is true:

> If. . . , then one ought to be indignant (grateful, resentful, etc.).

Strawson says that we 'sometimes' have the option of dispelling a reactive feeling by cultivating thorough objectivity towards a person; I say that we always have that option, and that it can never be factually wrong to regard someone as 'a case', viewing him in a spirit of inquiry as a natural object. Whether you do so—indeed, whether you are psychologically able to do so—in a particular case may depend on your personal style or your mood of the moment.

6. Reactivity, blame and freedom

Strawson offers the concept of resentment—or more generally of some negative reactive attitude—as the fundamental reality underlying *blame* and what goes with it, including punishment. Strawson can allow some truth-valued content to the proposition that someone is to blame for ϕing; the statement that he is to blame, or that he deserves punishment, is just *false* in a case where reactivity is stupidly inappropriate. But propositional content goes no further than that. It may be a sheer matter of fact that *he is not objectively not to blame*—it would not be stupidly inappropriate to blame him—but no further matter of fact requires us to blame him. Given that we are morally entitled to blame him, whether we do so is up to us; because in these cases whether we adopt a reactive attitude is up to us; the facts cannot take us by the throat and insist that we feel resentful or indignant, that is, insist that we blame. In calling this 'up to us' I do not mean that we will always be psychologically free to go either way. Sometimes angry resentment cannot be stilled; sometimes gratitude cannot be commanded. These, however, are psychological impediments, stemming from the character or the momentary mood of the person who has them. They are not recognitions of objective blameworthiness or the like.

In this account of Strawson's, feelings are central, and are not tied systematically to any propositions about their objects. My indignation at what you have done is not a perception of your objective blameworthiness, nor is it demanded of me by any such perception. It expresses my emotional make-up, rather than reflecting my ability to recognize a blameworthy person when I see one. When we have coolly assembled all the facts about the person and his action, we seem not to have found anything which entails that he is blameable for what he did. Philosophers have tried to repair this gap by introducing a concept of freedom or accountability or blameworthiness, with this understood as an elusive further *fact* about the offender, a fact which has a special power to entitle or require us to blame him for what he did. Strawson offers an alternative to this sterile approach. Instead of seeking to fill the gap with facts which might justify the adoption of reactive attitudes, Strawson says, 'it is just these attitudes themselves which fill the gap' (p. 23).

Strawson's account—he claims—provides a basis for an understanding 'of what we mean, i.e. of *all* we mean, when, speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice' (ibid.). That may

be a little too strong, because our ordinary thoughts about these matters may contain muddles and mistakes which Strawson filters out. His account does, however, provide for every *coherent* element in what we mean.

Strawson uses these materials in a brilliantly successful treatment of the age-old problem about freedom and determinism. Although this lies outside my present topic, I shall stay with it for a few moments, to help give a feel for the materials. I shall adapt it to explain a famous impasse involving the concept of moral accountability. Strawson does not explicitly present this, but it arises naturally out of his materials.

In well known ways, determinism can be made to seem incompatible with freedom, responsibility, moral accountability or the like. As we attend in detail to some story about what caused Agent to act in a certain way, so that his acting thus increasingly strikes us as inevitable, our sense of him as possibly to blame for his behavior becomes enfeebled. On the other side, if we view Agent's action as *not* deterministically caused, so that the totality of its causal antecedents did not settle whether he would act thus rather than so, his acting thus increasingly strikes as random, a matter of luck, and our sense of him as possibly to blame for his behavior is again weakened.

On one side of the literature on this impasse we find things like this: 'Moral accountability conflicts only with certain kinds of deterministic causation; and the procedure in which accountability is made to seem threatened by determinism as such is a trick, in which all causes are represented as though they were of those special kinds.' That is not a quotation, but this is: 'We have here... a persistent, an age-long deadlock due solely to the indisposition of the human mind to look closely into the meaning of its terms' (Miller 1934: 107). On the other side we find this: 'The procedure in which accountability is made to seem threatened by indeterminism is a trick, a pretence that we must choose between deterministic event-causation and randomness. The procedure loses its force once we get hold of the idea that Agent himself, and not any antecedent event, caused Agent's action.' None of this gets to the bottom of the matter, I submit. In the impasse where determinism seems to threaten accountability and indeterminism seems to do likewise, more is going on than mere trickery. To many who walk these paths, it feels as though we have hit something deep and immovable; the hope that conceptual sorting out will remove it, showing us what avoidable error of thought we have been led into, reflects an undue optimism about the powers of conceptual analysis.

Some philosophers have thought that the impasse shows our concept of moral accountability is self-contradictory or unsatisfiable. This might be correct; I do not hold on principle that everything must be all right with our conceptual scheme. But that cannot be the whole story. If we have a logically unsatisfiable concept of accountability, it will be because we are pulled two ways; and we should ask what did the pulling. Strawson's work could enable us to answer that question; it could let us strengthen and complete the inconsistent-concept diagnosis of the impasse, by explaining what led us into that conceptual mishap. I shall present it, however, as doing something different, namely explaining the impasse without supposing any inconsistent concept to be involved. It matters little which of these we adopt.

The Strawsonian explanation for the impasse goes as follows. When we contemplate Agent's action as the upshot of deterministic causes, we adopt an objective attitude towards him; our frame of mind encourages questions like 'What do we have here? How did this come about?' which naturally goes with the question 'How can we lessen (or increase)

the chance that this will happen again?' That objectivity of attitude *dispels* reactive feelings, and their disappearance presents itself to us as the judgment that the person is not morally accountable.

When instead we contemplate Agent's action as not arising inevitably from antecedent events, we again adopt an objective attitude towards him; we are again in the 'What do we have here?' frame of mind; and so again we are pushed out of reactive attitudes towards Agent in respect of this action, and we think that this has involved our giving up the judgment that he is morally accountable.

What seemed to be this:

> The proposition that P conflicts with the attribution of moral accountability, and so does the proposition that not-P,

from which we might infer that the concept of accountability cannot be satisfied, is really this:

> By actively raising the question 'P or not-P?'—i.e. by thinking objectively about the action—we get into a frame of mind in which we cannot have reactive feelings; and their absence makes us reluctant to describe or treat the person as morally accountable.

Rather than moral accountability's being *logically* inconsistent with each answer to the question, reactive feelings are *psychologically* immiscible with the frame of mind in which the question is asked. The answer does not matter: the objectivity of attitude which frames the question does the real work. Dostoyevsky described it memorably:

> But what can I do if I don't even feel resentment?... My anger, in consequence of the damned laws of consciousness, is subject to chemical decomposition. As you look, its object vanishes into thin air, its reasons evaporate, the offender is nowhere to be found, the affront ceases to be an offence and becomes destiny, something like toothache, for which nobody is to blame. (Dostoevsky 1864: 27)

The affront ceases to be an offence, not because of what you find when you look but just because *you look*.

I now turn away from issues about freedom and accountability to my proper topic, putting Strawson's ideas to work in a theory of punishment—one that will let me derive and explain two moral judgments. **(1)** Severity should be limited by gravity (special case: no punishment for the innocent), this limit being imposed immediately and not—as in utilitarianism—through predictions about upshots. **(2)** It is never right to punish an offender in a certain way if over-all better results could be achieved by punishing him less severely or not at all, this being derived from the core of the account and not—as in the milder forms of desert theory—merely tacked on as an intuitively required extra.

7. Outline of a theory of punishment

A defensible penal action does over-all good by means that bring suffering to one person. Not being outright utilitarians, we are not always willing to avail ourselves of opportunities to do this. Suppose that ceremonially hurting one randomly chosen person would be sure—through some mysterious but well-tested mechanism—to reduce the incidence of some really nasty disease. Most of us would be reluctant to go through with the ceremony, even if we judged that that one person's suffering would be less bad than a state of affairs in which the disease is endemic.

Why? We are moved by sympathy for the one person—a fellow-feeling for someone in distress—but sympathy should have a place on the other side of the equation as well. Why does it not? Or why, at any rate, does it function differently there?

Speaking for myself, I am not sure. **(i)** I may be influenced in some perhaps muddled way by issues about probability. The victim will certainly be harmed; the populace in general will only probably be helped. Utilitarianism takes account of such differences: it says that one should always act so as to produce the greatest expectable utility, this being a function of the values of upshots and of their probabilities. Still, perhaps some or all of us tend to shrink from certainly inflicting harm on someone in order to achieve the probability of doing good, wrongly treating certain/probable as an absolute difference of kind rather than a calculable difference of degree. **(ii)** For most of us, I am sure, harm to an identified person outweighs help—or lessened harm—to an unidentified group. Even when punishing a particular criminal will certainly bring benefits, we may be unable to point to any particular people as the beneficiaries. There may indeed *be* no people of whom it is determinately the case that *they* would benefit from this convict's being punished. This seems to affect our moral thinking. We serenely launch building projects in which probably some workers will die; compare that with how we react when a particular child is trapped in a well. **(iii)** The difference between making harm come to the convict and allowing it to come to victims of crime may be swaying us. I have argued that this difference between making and allowing has no basic moral significance, and nobody has yet found anything wrong in those arguments (Bennett 1995: chapters 6 and 7); but perhaps the making/allowing difference does influence us all, even those who see that this has no rational basis. **(iv)** Or perhaps we have scruples about harming people as a means to others' good. That is mostly muddle also, I have argued elsewhere (chapter 11), but it may have force in our thinking for all that.

Anyway, we are in general reluctant to harm one person in the interests of many. That reluctance, whatever its sources, seems unbudgeable, and I shall take it as a given and work on from there. It will play a large, active part in what follows.

Now, *punishing a criminal* should be seen, I submit, as a species of *hurting one person in the interests of many*. The moral or emotional obstacle to doing that is less in this species than throughout the rest of the genus; the fact that the one has committed a crime makes a difference. 'Because criminals deserve punishment', some will say. I agree, but I cannot make attributions of desert *explain* anything. The best answer I can find looks not to our moral principles but rather to an aspect of our nature lying deeper than our moral principles and helping to produce them and—insofar as anything can—to justify them. In blaming a convict for his behaviour, we do not assent to a proposition about him but rather adopt towards him a reactive attitude which Strawson calls 'vicarious resentment', or resentment on behalf of his victims; this essentially involves at least incipient hostility or ill-will towards its object; and this makes us less unwilling for its object—in our case, the convict—to be hurt in the interests of the greater good. We already have indignant, offended or resentful feelings towards him, and these reduce our unwillingness that he should be made to suffer for the general good. Thus Allan Gibbard: 'Anger is punitive' (1990: 139). Thus also Adam Smith: 'Gratitude and resentment. . . are the sentiments which most immediately and directly prompt to reward and to punish. . . That action must. . . surely appear to deserve punishment which everybody who hears of it is angry with, and upon that account rejoices to see punished.'

That underlies the common idea that *it is all right to punish the guilty*. From a starting-point that differs from utilitarianism because it involves a reluctance, in many cases, to do good by inflicting suffering on one person, we

introduce resentment and its vicarious analogue, blame, to lessen that reluctance and bring us closer to utilitarianism. It is generally wrong to punish the innocent because the explanation of why it is right to punish *anybody* applies only to the guilty.

At the outset I set two requirements for a decent theory of punishment. One was that it should not morally permit us—let alone require us—to punish someone more harshly than would maximize utility. My account satisfies that, for in it punishment is never in question unless it would do some good. That feature of the account comes from something that is not a mere add-on of my account, but its driving force, namely that punishment is to be engaged in only as a way of doing good. And room is made for this by a basic fact about reactive attitudes, namely that they are sometimes permissible and never required. That secures that nothing in the nature of things can require us to punish someone if considerations of utility go against our doing so.

I also required that the theory should make gravity of offence *directly* relevant to severity of punishment. My Strawsonian account does that too, right out to the end of the scale where innocence is directly a reason for not punishing the person at all. We are initially morally reluctant to punish anyone, and for an innocent person nothing overcomes that reluctance. Similarly, nothing much overcomes it in the case of a mild offender. In this theory of punishment, innocence and degrees of gravity are relevant not only because of the traces they will leave in the future but also directly.

My account contains an openly retributivist element: in it, the willingness to punish someone arises *directly* from the belief that he has committed a crime. Some defenders of retribution have taken pains to prevent it from brushing up against the idea of revenge, which they think would taint it. I have no such scruple. While I do not find it helpful to describe the crucial reactive attitude as 'vengeful', it certainly has the same human roots as the desire for vengeance. (For a suggestion about how revenge differs from retribution, see Nozick 1981: 366–8. Nozick's depicts retributive punishment (p. 369) as involving a looping communicative tie between punisher and punished, like the one in Grice's account of non-natural meaning; he calls this a 'communicative linking of the wrongdoer with correct values' (p. 379). This merits respect as an ideal for what punishment might be; but it outstrips the basic actual idea of punishment, which I am trying to elucidate.)

Although my account is retributivist, it frames punishment within utilitarian considerations, not allowing any punishment that goes against utility. Sher writes (1987: 74): 'To assert that we can justify punishment only by showing that it brings advantages is to beg the question against retributivism.' Though that is plausible, I have shown it to be false. Griffin writes:

> What would be entirely wrong would be to try to introduce utilitarian reasons into desert. As we have seen, it destroys a response to inject extraneous considerations into it, and utilitarian reasons are extraneous. Authenticity is not merely the best or purest form of responses such as admiration, gratitude, or appreciation; it is the only form. (Griffin 1986: 259)

That looks like a partial list of reactive attitudes, and I suppose that Griffin would include resentment in it (though he does not mention Strawson's work). I agree that one cannot adopt a reactive attitude for a purpose: there cannot be contrived or considered or judicious resentment. But that does not place such a high barrier between punishment and utility as Griffin apparently wants to erect there. Reactive attitudes are essentially spontaneous, adopted without the guidance of a *telos*; but one can—and civilized people do—

have goal-directed policies for delimiting areas of life within which they will deny themselves the luxury of reactivity.

Jonathan Jacobs, as I noted in section 1, apparently denies to utility any role in penal thinking. That is because he cannot see how to relate utility to the retributive core. He makes a persuasive case for the latter's centrality, basing it on vicarious resentment; and argues against trying to reform or civilize that resentment by bringing it under utility constraints. Attempts to do this, he contends, are 'both theoretically unsound and pragmatically fated to fail, because the sentiments which are to be reformed properly resist reformation' (Jacobs 1999: 555; see also 538). That is essentially Griffin's point, and it is correct. But (I repeat) considerations of utility can mark off areas within which the 'sentiments' in question are not to be given free rein; and that is something which they do not—in mature and civilized people—resist. Strawson gives the example of the feelings of a psycho-analyst towards a patient; the patient may say things which are not inherently unsuitable objects of resentment (as is the behavior of a baby), but the analyst can have a policy that in his professional relations with his patients he will forbid himself such attitudes because they are too likely to do harm.

In a more complex way, it is open for us to give play to our generalized vicarious resentment of criminal behavior by endorsing a system of punishment, expressing our resentment in our willingness to put the offender at the disposal of the legal system; while at the same time resolving that this should be enacted only in cases where it has a good enough chance of doing some good. The permissive framework is set by utility; the punishment within that framework expresses our reactive, retributive anger or resentment. Thus, a genuinely retributive element in punishment cohabits with severe utilitarian constraints.

8. The emphasis on system

Some people on whom I have tried out this approach to punishment have thought it would make a shambles of our penal system. They have seen me as committed to allowing that how severely a given judge punishes a given criminal—and indeed whether she punishes him at all—can and should depend upon her level of resentment towards him in respect of the offence of which he has been convicted. That was a misunderstanding. I am not arguing that the practice of the courts and the prisons should be a direct expression of the reactive attitudes of the people concerned. My thesis is that our reactive attitudes have a crucial part in *our willingness to have a penal system* and our willingness that it should have a certain shape. Given that such a system is in place, and given some general principles governing judicial behaviour which could be defended on utilitarian grounds, a judge is morally bound to follow the system. Nothing in what I have said implies otherwise.

I have claimed as a merit in my account—derived from an isomorphic fact about reactive attitudes—that it represents punishment as allowed by an offence, not as required by it. I do not mean that as a permission to the judicial authorities who implement the penal law. How much flexibility they should have, and whether (for example) they may properly be swayed by pleas from the convict or the victim or their relatives, are matters of policy on which I have nothing to say.

My theme is permissiveness not in the quotidian applications of the penal law but only at the most basic and systematic level. As thoughtful citizens contemplating the penal law of our land, we want to have a general moral attitude to it—a basis for accepting a penal system that has certain general features. Any system that we fully accept will

have, among its general features, upper limits to the severity of penalties for given kinds of crime (and also absolute upper limits, no matter how grave the offence). And a system acceptable to us will not have lower limits to the severity of penalties except ones set by policy, that is, by how much good can be done. We shall not assent to a system which embodies the idea that a crime of a certain kind *morally demands* a penalty of a certain degree of severity, even when policy—that is, utilitarian considerations—point to something milder. How much freedom judges and juries should have to vary sentences in the light of policy considerations is itself a matter of policy.

A possible objection: 'You have said that if someone offended against the law, and if punishing him would do some good, we may punish him *because he offended*. In basing punishment on our natural reactive attitudes against the offender, you have not required that our resentment focus on the offence for which we propose to punish the man. You tell us why we may not punish a good innocent man, but what about a bad innocent man? He did not commit this or any other crime; but we are strongly indignant at him for selfish and wantonly heartless behavior in other contexts; and it so happens that we can do good by punishing him. Haven't you implied that we may go ahead and do so?'

I have not implied that we may not. But I can and do allow for policy considerations which count strenuously against such 'punishment'. If the legal system regularly punished people just on moral grounds, without reference to whether they had committed crimes, that would have sour consequences, making our society anxious, unstable and neurotic. I can make as much use of the powerful utilitarian case against such a general practice as can anyone else.

A second objection: 'You have defended yourself from certain attacks by pleading that your concern is with the general case, laws, policies. Do you absolutely hide behind that screen? Or do you admit that there could in principle be rare individual cases where it was right to act outside the bounds of the law. If the latter is where you stand, then would you in such an exceptional case apply the general principles you have been laying down?' Yes, that is where I stand, and yes, in such a case I would apply the principles I have expounded here. Let me explain.

We are to suppose a case where vast good could be done—or vast harm averted—by 'punishing' a person known to be guiltless of crime. I agree with Sidgwick, Hare and Hampshire that we should live with the working assumption that this would never be right; it would be best if we almost *could not* do it. Still, not being an absolutist on this matter, I reluctantly accept that it could in theory be right to punish a person known to be innocent, and wrong not to (Bennett 1995: chapter 10). In such a case, I would stick by my Strawsonian story, maintaining that if the person in question is—though innocent of a crime—the subject of our reasonable indignation and resentment for something he has done, that bears morally on the proposal to punish him.

If you are a strong absolutist about such conduct, nothing that I say about innocence can have any appeal to you. If you are (however reluctantly and constrainedly) a relativist about it, then try to imagine a case where you think that with so much at stake it would be right to punish some person who has not committed the crime in question; and suppose that two people could fill the role—one decent, caring and upright, the other cruel, callous and deceitful. If you think it would be right to sacrifice one of them, wouldn't you have a moral reason for picking on the bad person rather than the good?

9. Distributing harm

I should take note of the theory of punishment offered by Phillip Montague in his *Punishment as Societal-Defense* (1995). The book's title suggests that when a state punishes offenders against its laws it is defending itself, being justified in so doing by a natural extension of the idea that individuals are entitled to defend themselves against aggressors. This sounds *toto coelo* unlike my account, but really the difference is smaller than it seems.

Rather than simply helping himself to a right of self-defence, and arguing on from that, Montague commendably devotes a whole chapter to discussing what justifies you in harming someone else in your own defence. Judith Jarvis Thomson (1991) answers this in terms of the having and forfeiting of rights; Montague critically examines her careful work, and eventually rejects it; which heralds his announcement of his own candidate for the moral basis of self-defence, a principle he calls 'J':

> When members of a group. . . are in danger of being harmed through the fault of some, but not all, members of that group; and when some person. . . who is not at fault is in a position to determine how the harm is distributed, even though the harm is unavoidable. . . ; then the person has a right (if a member of the threatened group) and is required (if not a member of the group) to distribute the harm among those who are at fault. (Montague 1995: 42)

From this Montague seeks to derive what we all believe about individual self-defence, taking an individual to be a one-member 'group'. But obviously he has framed J it so as to apply to societies as well as to individuals.

It turns out that the notions of self-defence and societal defence play no real role in Montague's theory of punishment. The entire driving force of his arguments is the thesis (J) that when disutility is to be distributed among a population, those who are faultily responsible for it should bear its burden. Montague discusses this partly in terms of cases 'having nothing to do with defending one[self]' (p. 34), and says that self-defence 'does not in itself distinguish, from a moral point of view' those cases from others to which principle J applies (p. 39). So the notion of defence, whether of oneself or of a society, is involved in only a subset of examples of principle J's territory. Set it aside and the main thrust of the book is unaffected, except that its title is seen to be inapt and its text requires many verbal changes.

For J to do its work, the notion of 'distribution' has to be stretched. If I knock you down as my best way of preventing you from wrongfully flooring me, Montague must say that I thereby modify how some being-knocked-down harm is *distributed*. When he gets into the question of matching the penalty to the crime (see section 10 below), handling everything in terms not of *what*, but rather of *how much*, harm each involves, the metaphor of *distribution* looks even thinner. My point here is not a purely verbal one: what J says about distributing harm amongst its producers gets its plausibility largely from misusing 'distribute', as though all the cases were comparable with the issue about whose land should receive the effluent from the local pig-farm. With that suggestion removed, all that remains is a principle about acting against people who have made the world worse than it need have been; which is not so far from my position.

Principle J relies on the idea that the offender has produced the threatened harm—the harm that is to be 'distributed' appropriately by punishment. That might seem to be a defect in Montague's account, not shared by mine. MacHeath has committed a nasty crime; its effects have worked themselves out, and nothing can be done to alleviate

(or 'redistribute') them; but we are still entitled to punish him for his crime, in order to reduce the reduce the probability that others will offend similarly. The only 'harm' that we can effect by punishing him is the probability of similar offences by other people, and he may have contributed nothing to that threat.

That seeming defect is not real, however. Montague seeks to justify not the individual punitive act but rather the establishment of a penal system (p. 63), and he applies the notion of being faultily responsible for (the threat of) harm not to individuals but to groups. The responsibility for the harms involved in murder are created by *murderers*; MacHeath is a murderer; so principle J implies that he is to bear some of the burden of that harm.

That works as a defence against the point I made two paragraphs back. Principle J could be reformulated so as *explicitly* to provide for this defence, part of the reform being to cleanse the principle of 'distribute'. It might not then be so intuitively appealing; but I shall not press this point.

Montague's account allows that punishment may be justified even when it does not make the world over-all better than it would otherwise have been. He says that the rest of his position 'provides grounds' for this part of it, but in the upshot all he can claim is that 'nothing in this account' forbids him to adopt the view in question (p. 70). His informal defence of the view is weak. He offers to discuss a case in which 'punishing her would deter no one else', equates that with its being the case that 'the system's threats are [already] maximally credible', and then seeks to score a point with the remark that 'even during periods of maximal credibility, the threats of a system can be effective deterrents' (p. 70).

However, Montague's handling of this contains a warning to those who hold, as I do, that we should never punish against utility. Suppose we knew that by inflicting a certain punishment on MacHeath (who deserves it), we bring good to some innocent person while bringing a greater amount of harm to MacHeath. That would be punishment against utility, in my sense, but it is not morally outrageous. The case is artificially contrived, and we might never know that we were confronted by it, but it has theoretical interest. Perhaps I could adjust my account to make room for it, but I see no easy way of doing so.

For Montague, principle J 'embodies a requirement of *basic* justice' so that 'its acceptability cannot be demonstrated by deriving it from some more fundamental principle' (p. 46); and he does not derive it from anything else either, merely laying it down as something he thinks we should accept. In my scheme, on the contrary, nothing comparable with J is basic. I do not derive such theses from any more fundamental principle, but I do derive them from the sources in us of our willingness to think and behave in certain ways.

It may be objected: 'What you have offered is a peculiar-looking *justification* for punishment. It may correctly describe how we think and feel, but does it justify anything?' Well, it does not do so by deriving 'It is all right to punish criminals' from other moral principles. Rather, it says that my willingness to punish criminals comes from the interplay amongst three deep elements in my nature: **(1)** my wish to do consequential good, **(2)** my reluctance to hurt one person so as to benefit others, and **(3)** my reactive attitude towards criminals. Of these, **(1)** provides the main thrust, **(2)** sets a limit to it, and **(3)** sets a limit to that limit. To this I add my moral consent to my being moved by these factors; whence I conclude that my over-all stance on punishment is morally acceptable. I can provide no more justification than that; nor do I want or need more.

10. Severity of punishment

Where do our judgments about undue severity of punishments come from? (On this as on most of the topics of this paper, one of the best things I know is Whiteley 1956.) We may say 'The punishment should fit the crime', but that is not a basis for any judgments. For it to be so, fittingness would have to be a non-moral relation that severity can have or not have to gravity, and there is no plausible candidate for this role.

One might hope to be able to base penalties—via some algorithm—on a comparison between the disutility of the penalty and that of the offence. Nozick (1981: 363–5) takes his stand on this approach to 'matching penalties', mentions a few difficulties, and whisks past them more casually than they deserve. Montague deals a little more patiently with the idea that the punishment should fit the crime by involving roughly the same amount of harm as the crime does. There are well known difficulties about this, e.g. when it is applied to the arsonist who causes six people to be burned to death, the drunk driver who doesn't actually hurt anybody, the poor man who robs a bank, and countless others. Montague remarks that 'there does seem to be a tendency to exaggerate these difficulties' (p. 64), but he neither explains nor defends this. The commensuration difficulties are especially acute for someone who holds, as Montague does and apparently must, that the harm a malefactor does may consist only in 'placing other people at risk' (p. 68).

Anyway, *matching* aside, amount-of-harm does not accurately measure the severity of a crime. Two offenders caused or threatened equal amounts of harm by physically similar kinds of conduct, one acting with hot-tempered recklessness and the other with icy malevolence; most of us are calmly willing that the latter should be punished more severely than the former; and Montague's theory of proportionality cannot make room for this.

Some have tried to regiment these matters with help from the thesis that the wrongness of a crime lies in the 'unfair advantage' that the criminal takes of the rest of us. (See Morris 1968.) According to Michael Davis, an adherent of this view, the severity of the penalty should be proportioned to the value of the advantage, which can be measured by what the value would be, in an auction, of a single free ticket, a single unpunished taking of that advantage. (Davis 1995: 210–2.) Working out this mercantile approach in detail leads Davis to some strange results, as when he explains why vehicular homicide is a lesser crime than involuntary manslaughter. And there are plenty of cases which it does not fit at all: for most of us, the 'advantage' the arsonist takes in burning down the forest is worthless. For other criticisms of this position, see Goldman 1995: 32.

No theoretical underlay ever shows up in political and legal arguments about severity. Actual debates about it divide into utilitarian ones, about cost, deterrent and preventive effect, the difficulty of getting juries to convict, the chances of reform, and so on; and comparative ones, about whether the punishment assigned to one kind of offence squares with that assigned to another. When people say things like 'He oughtn't to be punished as severely as that—he doesn't deserve it', the remark about desert does not support the first clause but merely repeats it. I agree with Brandt (1985: 188) that desert theory 'at best gives only an ordinal theory: it tells us that X should be punished more for A than Y should be for doing B, but gives no clue exactly how much either one should be punished. The consequentialist theory has the virtue of yielding, in principle, some quantitative guide.' So it does, but the guide is a wrong one (see section 2 above).

Those who seek to base the morality of punishment on *rights* can on that basis get some distance with the issue about severity. Some questions about the handling of offenders can be discussed in the language of 'rights': does his crime deprive him of his right to physical freedom? his right to sexual companionship? his right to vote? his right to complain? But if 'rights' underlie all our judgments about undue severity, then innocent people must have, for instance, a right not to be jailed for a month and a right not to be jailed for year, so that in a particular case an offender may be judged to have forfeited the former right but not the latter. This approach would require an endlessly complex and elaborate set of beliefs about negative rights—the right not to be treated thus-and-so—which each person has until they start being peeled off him by his crimes. Our moral thinking and reasoning does not include, even unconsciously or dispositionally, any such structure as that.

Defensive move: 'We need to attribute to innocent people only a single right, namely the right not to be made to suffer by the law; then we can make further judgments about *how much* of that right a convict has lost because of his offence, and in which areas of it the loss has occurred.' But then the alleged basis for our judgments about undue severity turns out to be a mere rewording of them. We are to base judgments of the form 'Penalty P is unduly severe for offense O' on ones of the form 'In committing O he did not lose enough of his right not to be punished to entitle us to inflict P upon him.' Things have become more prolix, to be sure, but nothing has been done to exhibit system or structure in this part of our moral thinking.

The Strawsonian account, on the other hand, has something systematic to say about how gravity relates to severity. It does not derive severity judgments from a deeper moral theory, e.g. one relying on fittingness or unfair advantage, but it does represent them as more than a jumble. When we ponder the severity of penalty for a given offence, resentment draws us one way while sympathy pulls us the other. My sympathy inclines me to oppose someone's being made to suffer P, but my knowledge of his offence stokes my indignation, which damps my sympathy to a point where I am willing to endorse P after all. The worse the offence, the greater the anger, and so the greater the amount of sympathy it can quell; but only within limits, for some sympathy levels cannot be overcome by any reactive feelings—for instance the pity I feel towards any actual or imagined sufferer of death at the stake.

I do not mean to depict lawmakers as the battle-ground of a struggle between spurts of indignation and wellings of sympathy. On the contrary, they may balance offences against punishments quite dispassionately. I claim only that their making of these judgments, when not purely utilitarian, lies on a continuum with, and is best understood in terms of, what happens when indignation tempers pity or sympathy. (Sympathy and pity are not themselves reactive attitudes. One can have pity—the minimal emotion—for a baby, a horse, a mentally ill person, a sparrow, in respect of which resentment, gratitude, anger or jealousy would be glaringly inappropriate.)

How much objectivity, then, can there be in judgments about severity? The circles in which you and I move, I believe, have similar beliefs about proper levels of sympathy and of indignation. In a case where Victim can appropriately be angrily resentful towards Agent, we may judge the intensity of his reaction to be disproportionate to the offence; and you and I would probably agree in most such judgments. Similarly on the side of sympathy. Within our circles, disagreement about penal levels would mostly concern means and ends, causes and effects—issues about utility.

Not everyone is like us, however. What about Furioso—a notional person with extravagantly wide-ranging and intense angers and resentments, and almost no sympathy for anyone? My account of the rationale of punishment probably describes Furioso's procedures too, but does it justify them? He might say so, but I do not. It would be absurd for me to approve of his way of proportioning punishment to offence, given its basis in mountainous levels of anger and dim flickers of human sympathy. If I could, I would persuade Furioso to stop wanting to influence the shape of the penal law; failing which I would try to convert him to some theory of punishment other than mine, getting him to think of it in terms (perhaps) of a desert theory which, though shallow and muddled, had a more humane penal output than what results from combining my theory with his attitudes.

Analogously, most of us think that a good utilitarian case can be made for a measure of spontaneity in our lives; but we qualify that when we consider people whose spontaneity is usually cruel and destructive. Another example: we hold in general that people should act in accordance with their consciences; but we applaud Huckleberry Finn's unconscionable lying so as to steer slave-catchers away from Jim, even though he himself saw this as wicked, and had nothing to set against it but his unprincipled love for his friend.

11. Maximum penalties

Actual arguments about severity, I remarked, are either consequential or comparative; for the rest, what goes on consists in attempts to reduce people's indignation and heighten their sympathy, or the reverse. Current debates in penology nicely fit my Strawsonian account of punishment.

Other evidence points to the account's being descriptively right. Throughout the western world over the past four centuries, the maximum penalties for various kinds of offence have fairly steadily decreased in severity; that could reflect changing opinions about what deters or does good in other ways, but it also reflects independent changes in what people find morally permissible. The steadiness of this change suggests that such judgments have an orderly basis, but what is it? The Strawsonian account of punishment provides an answer, and I know of no other.

Most people in the western world today would judge, as I do, that no conceivable offence could make it permissible for the offender to be hung, drawn and quartered, or to be burned at the stake. What makes us differ in this way from, say, the subjects of Elizabeth I or Louis XIV? Well, we in a sense cannot bear the thought of flames licking around a conscious human being, whereas an Elizabethan could easily entertain the thought, and even the sight and sound and smell, of the burning. We are less callous than the Elizabethans—more prone to sympathy, to fellow-feeling, to pity, when fully aware of the suffering of others. This difference has shown up in changes of view about slavery, indigence, school discipline, and other topics. It does not make us better people than the Elizabethans. We may manage our greater imaginative sympathy by being more prone to keep ourselves ignorant of many of the horrors that human beings suffer at one another's hands.

There may also have been a change in the propensity for adverse reactive attitudes, but we need not postulate one: the change in moral judgments could be due to a change in one of the two interacting elements while the other held steady. I guess that historically the main change has consisted in a raising of levels of sympathy.

Foucault's book *Discipline and Punish* confirms this. Foucault himself has a theory—to me an obscure one—about why people have changed their views about what punish-

ments are permissible. Before he gets to that, however, he says much that seems not to square with his theory but fits well with mine.

'Why this universal horror of torture and such lyrical insistence that punishment be "humane"?', Foucault asks, and he answers that 'This need for punishment without torture was first formulated as a cry from the heart or from an outraged nature' (Foucault 1979: 74). Further on and in more detail:

> The principle of moderation in punishment. . . was articulated first as a discourse of the heart. Or rather, it leaps forth like a cry from the body, which is revolted at the sight or at the imagination of too much cruelty. . . Does this lyricism express an inability to find a rational foundation for penal arithmetic?. . . Where is one to find a limit, if not in a human nature that is manifested—not in the rigor of the law, not in the ferocity of the delinquent—but in the sensibility of the reasonable man who makes the law and does not commit crime. (p. 91)

That endorses my view about severity, and receives support from Foucault's quotations from participants in the debate, such as this by the penologist P. L. Lacretelle, writing in 1784: 'God, who has imprinted in our hearts an aversion to pain for ourselves and for our fellow men, are they then those same beings, whom thou has created so weak and so sensible, who have invented such barbarous, such refined tortures?' (quoted in Foucault 1979: 91).

Later on Foucault tells a different story about what explains the retreat from torture. Flouting the evidence which he himself has accumulated, he takes the historic change in views about severity to reflect not resentment's struggle against sympathy but rather something more elaborate, delivering more grist to his analytic mill. This is typical:

> In the worst of murderers, there is one thing, at last, to be respected when one punishes: his 'humanity'. The day was to come, in the nineteenth century, when this 'man', discovered in the criminal, would become the target of penal intervention, the object that it claimed to correct and transform. . . But at the time of the Enlightenment, it was not as a theme of positive knowledge that man was opposed to the barbarity of public executions, but as a legal limit: the legitimate frontier of the power to punish. Not that which must be reached in order to alter him, but that which must be left intact in order to respect him.

I do not understand very well this stuff about 'humanity' as a 'limit'. It reminds me of this: 'Torture is impermissible as a punishment; for it degrades the person tortured, and to degrade a culprit violates the respect due to him as a rational creature' (Donagan 1977: 188); and I don't properly understand that either. Why does torture degrade if thirty years in prison does not? Anyway, the respect-for-humanity idea is not supported by Foucault's own excerpts from the literature, where the demand for change comes 'from the heart'.

12. Extreme crimes

A couple of well known phenomena can be explained with help from the Strawsonian materials that I have deployed.

(1) Civilised people these days tend to view the perpetrator of a bad enough crime as demented. It has not been made clear why. No-one, so far as I now, has even tried to base it on a theory of human behavior employing a clear concept of sanity. A moral philosopher has written in unpublished work that 'The most horrendous, stomach-churning crimes

could only be committed by an insane person'; there seem to be no grounds for accepting this if it is offered as on a part with 'The most impressive record-breaking athletic feats could only be performed by a healthy person'. Then what is going on when a philosopher says such a thing?

Well, the speaker is registering his view that the perpetrator of a profoundly and intensely horrible crime lies outside the realm within which ordinary penal thinking is appropriate. I am with him on that, but I do not infer it from any premiss about insanity. Strawson's materials offer us a more realistic understanding of this reaction to the worst crimes. The crucial point is that reactive attitudes belong to the web of normal level reciprocating human relationships. We can blame someone for an action—resent it on behalf of its victims—without having or wanting a personal relationship with him; but if our resentment is appropriate, and not a mere idle spinning of wheels, we must think of ourselves and the offender as potentially inter-related within a single human community. We are rejecting that thought when we call the Dahmers of this world insane. When a man behaves badly enough, we distance ourselves from him, declining to regard him as anything but 'a case'; that abolishes any thought that we could relate to him as person to person, and so it shoulders reactive attitudes aside. This may register itself as the hunch that such a malefactor must be mad, but that verbal gift-wrap adds nothing—no theoretical ideas about sanity and the lack of it—to the content of the package.

(2) Sometimes in the U.S.A. a criminal defence has succeeded by attributing the crime to causes which would not ordinarily be thought to be exculpatory; the outcomes of these trials have surprised and sometimes disgusted those who were not present at them. Accused people have won acquittals, or lessening of charges or lowering of sentences, through pleas having to do with the distorting effects of a deprived childhood, of years in prison, of a diet of junk food, and the like. In each of these cases, I submit, the defence draws the jurors into a protracted objectivity of attitude towards the accused person, thus excluding the reactive vicarious resentment which it would be natural for them to feel. The trick is worked by giving, at length, a causal story; and what counts most are not its details but just the objectivity, which pushes aside reactivity. The jurors are induced to look; and, as Dostoevsky said, 'As you look. . . , the offender is nowhere to be found'.

Bibliography

Bedau, Hugo Adam (1977). 'Concessions to Retribution in Punishment'. In *Justice and Punishment*, edited by J. D. Cederblom and W. L. Blizek. Cambridge, Mass.: Ballinger.

Bennett, Jonathan (1980). 'Accountability'. In *Philosophical Subjects*, edited by Z. van Straaten, 14–47. Oxford University Press.

——(1995). *The Act Itself*. Oxford University Press.

Braithwaite, John, and Pettit Philip (1990). *Not Just Deserts*. Oxford University Press.

Brandt, Richard B. (1985). 'A Motivational Theory of Excuses in the Criminal Law'. *Nomos* 27: 165–98.

Brock, Dan W. (1973). 'Recent Work in Utilitarianism'. *American Philosophical Quarterly* 10: 241–76.

Cupit, Geoffrey (1996). 'Desert and Responsibility'. *Canadian Journal of Philosophy* 26: 83–99.

Davis, Michael (1995). 'Harm and Retribution'. In *Punishment*, edited by A. J. Simmons, et al., 188–218. Princeton University Press.

Donagan, Alan (1977). *The Theory of Morality*. University of Chicago Press.

Dostoevsky, Fyodor (1864). *Notes from the Underground*. Reprinted in translation Harmondsworth: Penguin Books, 1974.

Feldman, Fred (1995). 'Desert: Reconsideration of Some Received Wisdom'. *Mind* 104: 63–77.

Foucault, Michel (1979). *Discipline and Punish*. New York: Vintage Books.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, Mass.: Harvard University Press.

Goldman, Alan H. (1995). 'The Paradox of Punishment'. In *Punishment*, edited by A. J. Simmons, et al., 30–46. Princeton University Press.

Griffin, James (1986). *Well-Being, Its Meaning, Measurement, and Moral Importance*. Oxford University Press.

Honderich, Ted (1976). *Punishment*. London: Peregrine Books.

Husak, Douglas N. (1992). 'Why Punish the Deserving?' *Noûs* 26: 447–464.

Jacobs, Jonathan (1999). 'Luck and Retribution'. *Philosophy* 74: 535–55.

Miller, Dickinson S. (1934). 'Free Will as Involving Determination and Inconceivable Without It'. Reprinted in *Philosophical Analysis and Human Welfare*, edited by Loyd D. Easton, 104–31. Dordrecht: Reidel, 1975.

Montague, Phillip (1995). *Punishment as Societal-Defense*. Lanham, MD: Rowan and Littlefield.

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

Morris, Herbert (1968). 'Persons and Punishment'. *Monist* 52: 475–501.

Nozick, Robert (1981). *Philosophical Explanations*. Cambridge, Mass.: Harvard University Press.

Sher, George (1987). *Desert*. Princeton University Press.

Smith, Adam (1759). *The Theory of the Moral Sentiments*.

Strawson, P. F. (1962). 'Freedom and Resentment'. Reprinted in *Freedom and Resentment and Other Essays*, 1–25. London: Methuen, 1974.

Thomson, Judith Jarvis (1990). *The Realm of Rights*. Cambridge, Mass.: Harvard University Press

——-(1991). 'Self-Defense'. *Philosophy and Public Affairs* 20.

Walker, Nigel (1999). 'Even More Varieties of Retribution'. *Philosophy* 74: 595–605.

Whitely, C. H. 1956). 'On Retribution'. *Philosopohy* 31.