focus on the *consequences* of disease or defect that have a bearing on the justice of conviction and of punishment. The Royal Commission proposal fails in this respect.

6. Paragraph (2) of section 4.01 is designed to exclude from the concept of "mental disease or defect" the case of so-called "psychopathic personality." The reason for the exclusion is that, as the Royal Commission put it, psychopathy "is a statistical abnormality; that is to say, the psychopath differs from a normal person only quantitatively or in degree, not qualitatively; and the diagnosis of psychopathic personality does not carry with it any explanation of the causes of the abnormality." While it may not be feasible to formulate a definition of "disease," there is much to be said for excluding a condition that is manifested only by the behavior phenomena that must, by hypothesis, be the result of disease for irresponsibility to be established. Although British psychiatrists have agreed, on the whole, that psychopathy should not be called "disease," there is considerable difference of opinion on the point in the United States. Yet it does not seem useful to contemplate the litigation of what is essentially a matter of terminology; nor is it right to have the legal result rest upon the resolution of a dispute of this kind.

60 The Classic Debate

JOEL FEINBERG

The traditional debate among philosophers over the justification of legal punishment has been between partisans of the "retributive" and "utilitarian" theories. Neither the term *retributive* nor the term *utilitarian* has been used with perfect uniformity and precision, but, by and large, those who have been called utilitarians have insisted that punishment of the guilty is at best a necessary evil justified only as a means to the prevention of evils even greater than itself. *Retributivism*, on the other hand, has labeled a large miscellany of theories united only in their opposition to the utilitarian theory. It may best serve clarity, therefore, to define the utilitarian theory with relative precision (as above) and then define retributivism as its logical contradictory, so that the two theories are not only mutually exclusive but also jointly exhaustive. Discussion of the various varieties of retributivism can then proceed.

Perhaps the leading form of the retributive theory includes major elements identifiable in the following formulations:

> It is an end in itself that the guilty should suffer pain.... The primary justification of punishment is always to be found in the fact that an offense has been committed which deserves the punishment, not in any future advantage to be gained by its infliction.[1]

> Punishment is justified only on the ground that wrongdoing merits punishment. It is morally fitting that a person who does wrong should suffer in proportion to his wrongdoing. That a criminal should be punished follows from his guilt, and the severity of the appropriate punishment depends on the depravity of the act. The state of affairs where a wrongdoer suffers punishment is morally better than one where he does not, and is so irrespective of consequences.[2]

Justification, according to these accounts, must look backward in time to guilt rather than forward to "advantages"; the formulations are rich in moral terminology ("merits," "morally fitting," "wrongdoing," "morally better"); there is great emphasis on *desert*. For those reasons, we might well refer to this as a "moralistic" version of the retributive theory. As such it can be contrasted with a "legalistic" version, according to which punishment is for lawbreaking, not

---

Published in previous editions as part of the introduction to this section.

(necessarily) for wrongdoing. Legalistic retributivism holds that the justification of punishment is always to be found in the fact that a rule has been broken for the violation of which a certain penalty is specified, whether or not the offender incurs any moral guilt. The offender, properly apprised in advance of the penalty, voluntarily assumes the risk of punishment, and when he or she receives comeuppance, he or she can have no complaint. As one recent legalistic retributivist put it,

> Punishment is a corollary not of law but of lawbreaking. Legislators do not choose to punish. They hope no punishment will be needed. Their laws would succeed even if no punishment occurred. The criminal makes the essential choice: he "brings it on himself."[3]

Both moralistic and legalistic retributivism have "pure" and "impure" variants. In their pure formulations, they are totally free of utilitarian admixture. Moral or legal guilt (as the case may be) is not only a necessary condition for justified punishment, it is quite sufficient "irrespective of consequences." In the impure formulation, both guilt (moral or legal) and conducibility to good consequences are necessary for justified punishment, but neither is sufficient without the other. This mixed theory could with some propriety be called "impure utilitarianism" as well as "impure retributivism." Since we have stipulated, however, that a retributive theory is one which is not wholly utilitarian, we are committed to the latter usage.

A complete theory of punishment will not only specify the conditions under which punishment should and should not be administered, it will also provide a general criterion for determining the amount or degree of punishment. It is not only unjust to be punished undeservedly and to be let off although meriting punishment, it is also unfair to be punished severely for a minor offense or lightly for a heinous one. What is the right amount of punishment? There is one kind of answer especially distinctive of retributivism in all of its forms: an answer in terms of fittingness or proportion. The punishment must *fit* the crime; its degree must be *proportionate* to the seriousness or moral gravity of the offense. Retributivists are often understandably vague about the practical interpretations of the key notions of fittingness, proportion, and moral gravity. Sometimes aesthetic analogies are employed (such as matching and clashing colors, or harmonious and dissonant chords). Some retributivists, including Immanuel Kant, attempt to apply the ancient principle of *lex talionis* (the law of retaliation): The punishment should match the crime not only in the degree of harm inflicted on its victim, but also in the mode and manner of the infliction: fines for larceny, physical beatings for battery, capital punishment for murder. Other retributivists, however, explicitly reject the doctrine of retaliation in kind; hence, that doctrine is better treated as a logically independent thesis commonly associated with retributivism rather than as an essential component of the theory.

Defined as the exhaustive class of alternatives to the utilitarian theory, retributivism of course is subject to no simple summary. It will be useful to subsequent discussions, however, to summarize that popular variant of the theory which can be called *pure moralistic retributivism* as consistent (at least) of the following propositions:

1. Moral guilt is a necessary condition for justified punishment.
2. Moral guilt is a sufficient condition ("irrespective of consequences") for justified punishment.
3. The proper amount of punishment to be inflicted upon the morally guilty offender is that amount which fits, matches, or is proportionate to the moral gravity of the offense.

That it is never justified to punish a morally blameless person for his or her "offense" (thesis 1) may not be quite self-evident, but it does find strong support in moral common sense. Thesis 2, however, is likely to prove an embarrassment for the pure retributivist, for it would have him or her approve the infliction of suffering on a person (albeit a *guilty* person) even when no good to the offender, the victim, or society at large is likely to result. "How can two wrongs make a right, or two evils a good?" he or she will be asked by the utilitarian, and in this case it is the utilitarian who will claim to speak for "moral common sense." In reply, the pure retributivist is likely to concede that

inflicting suffering on an offender is not "good in itself," but will also point out that single acts cannot be judged simply "in themselves" with no concern for the context in which they fit and the events preceding them which are their occasion. Personal sadness is not a "good in itself" either, and yet when it is a response to the perceived sufferings of another it has a unique appropriateness. Glee, considered "in itself," looks much more like an intrinsically good mental state, but glee does not morally fit the perception of another's pain any more than an orange shirt aesthetically fits shocking pink trousers. Similarly, it may be true (the analogy is admittedly imperfect) that "while the moral evil in the offender and the pain of the punishment are each considered separately evils, it is intrinsically good that a certain relation exist or be established between them."[4] In this way the pure retributivist, relying on moral intuitions, can deny that a deliberate imposition of suffering on a human being is either good in itself or good as a means, and yet find it justified, nevertheless, as an essential component of an intrinsically good relation. Perhaps that is to put the point too strongly. All the retributivist needs to establish is that the complex situation preceding the infliction of punishment can be made better than it otherwise would be by the addition to it of the offender's suffering.

The utilitarian is not only unconvinced by arguments of this kind, he or she is also likely to find a "suspicious connection" between philosophical retributivism and the primitive lust for vengeance. The moralistic retributivist protests that he or she eschews anger or any other passion and seeks not revenge, but justice and the satisfaction of desert. Punishment, after all, is not the only kind of treatment we bestow upon persons simply because we think they deserve it. Teachers give students the grades they have earned with no thought of "future advantage," and with eyes firmly fixed on past performance. There is no necessary jubilation at good performance or vindictive pleasure in assigning low grades. And much the same is true of the assignments of rewards, prizes, grants, compensation, civil liability, and so on. Justice requires assignment on the basis of desert alone. To be sure, there is

> a great danger of revengeful and sadistic tendencies finding vent under the unconscious disguise of a righteous indignation calling for just punishment, since the evil desire for revenge, if not identical with the latter, bears a resemblance to it sufficiently close to deceive those who want an excuse.[5]

Indeed, it is commonly thought that our modern notions of retributive justice have grown out of earlier practices, like the vendetta and the law of deodand, that were through and through expressions of the urge to vengeance.[6] Still, the retributivist replies, it is unfair to *identify* a belief with one of its corruptions, or a modern practice with its historical antecedents. The latter mistake is an instance of the "genetic fallacy" which is committed whenever one confuses an account of how something came to be the way it is with an analysis of what it has become.

The third thesis of the pure moralistic retributivist has also been subject to heavy attack. Can it really be the business of the state to ensure that happiness and unhappiness are distributed among citizens in proportion to their moral deserts? Think of the practical difficulties involved in the attempt simply to apportion pain to moral guilt in a given case, with no help from utilitarian considerations. First of all, it is usually impossible to punish an offender without inflicting suffering on those who love or depend upon him and may themselves be entirely innocent, morally speaking. In that way, punishing the guilty is self-defeating from the moralistic retributive point of view. It will do more to increase than to diminish the disproportion between unhappiness and desert throughout society. Secondly, the aim of apportioning pain to guilt would in some cases require punishing "trivial" moral offenses, like rudeness, as heavily as more socially harmful crimes, since there can be as much genuine wickedness in the former as the latter. Thirdly, there is the problem of accumulation. Deciding the right amount of suffering to inflict in a given case would entail an assessment of the character of the offender as manifested throughout his or her whole life (and not simply at one weak

moment) and also an assessment of his or her total lifelong balance of pleasure and pain. Moreover, there are inevitably inequalities of moral guilt in the commission of the same crime by different offenders, as well as inequalities of suffering from the same punishment. Application of the pure retributive theory then would require the abandonment of fixed penalties for various crimes and the substitution of individuated penalties selected in each case by an authority to fit the offender's uniquely personal guilt and vulnerability.

The utilitarian theory of punishment holds that punishment is never good in itself, but is (like bad-tasting medicine) justified when, and only when, it is a means to such future goods as *correction* (reform) of the offender, *protection* of society against other offenses from the same offender, and *deterrence* of other would-be offenders. (The list is not exhaustive.) Giving the offender the pain he deserves because of his wickedness is either not a coherent notion, on this theory, or else not a morally respectable independent reason for punishing. In fact, the utilitarian theory arose in the eighteenth century as part of a conscious reaction to cruel and uneconomical social institutions (including prisons) that were normally defended, if at all, in righteously moralistic terms.

For purposes of clarity, the utilitarian theory of punishment should be distinguished from utilitarianism as a general moral theory. The standard of right conduct generally, according to the latter, is conducibility to good consequences. Any act at all, whether that of a private citizen, a legislator, or a judge, is morally right if and only if it is likely, on the best evidence, to do more good or less harm all around than any alternative conduct open to the actor. (The standard for judging the goodness of consequences, in turn, for Jeremy Bentham and the early utilitarians was the amount of human happiness they contained, but many later utilitarians had more complicated conceptions of intrinsic value.) All proponents of general utilitarianism, of course, are also supporters of the utilitarian theory of punishment, but there is no logical necessity that in respect to punishment a utilitarian be a general utilitarian across the board.

The utilitarian theory of punishment can be summarized in three propositions parallel to those used above to summarize pure moralistic retributivism. According to this theory:

1. Social utility (correction, prevention, deterrence, etc.) is a necessary condition for justified punishment.
2. Social utility is a sufficient condition for justified punishment.
3. The proper amount of punishment to be inflicted upon the offender is that amount which will do the most good or the least harm to all those who will be affected by it.

The first thesis enjoys the strongest support from common sense, though not so strong as to preclude controversy. For the retributivist, as has been seen, punishing the guilty is an end in itself quite apart from any gain in social utility. The utilitarian is apt to reply that if reform of the criminal could be secured with no loss of deterrence by simply giving him or her a pill that would have the same effect, then nothing would be lost by not punishing him or her, and the substitute treatment would be "sheer gain."

Thesis 2, however, is the utilitarian's greatest embarrassment. The retributivist opponent argues forcefully against it that in certain easily imaginable circumstances it would justify punishment of the (legally) innocent, a consequence which all would regard as a moral abomination. Some utilitarians deny that punishment of the innocent could *ever* be the alternative that has the best consequences in social utility, but this reply seems arbitrary and dogmatic. Other utilitarians claim that "punishment of the innocent" is a self-contradiction. The concept of punishment, they argue,[7] itself implies hard treatment imposed upon the guilty as a conscious and deliberate response to their guilt. That guilt is part of the very definition of punishment, these writers claim, is shown by the absurdity of saying "I am punishing you for something you have not done," which sounds very much like "I am curing you even though you are not sick." Since all punishment is understood to be for guilt, they conclude, they can hardly be interpreted as advocating punishing without guilt. H. L. A. Hart[8] calls this move a "definitional stop," and charges

that it is an "abuse of definition," and indeed it is, if put forward by a proponent of the general utilitarian theory. If the right act in all contexts is the one which is likely to have the best consequences, then conceivably the act of framing an innocent man could sometimes be right; and the question of whether such mistreatment of the innocent party could properly be called "punishment" is a mere question of words having no bearing on the utilitarian's embarrassment. If, on the other hand, the definitional stop is employed by a defender of the utilitarian theory of the justification of punishment who is not a utilitarian across the board, then it seems to be a legitimate argumentative move. Such a utilitarian is defending official infliction of hard treatment (deprivation of liberty, suffering, etc.) on *those who are legally guilty,* a practice to which he or she refers by using the word *punishment,* as justified when and only when there is probably social utility in it.

No kind of utilitarian, however, will have plausible recourse to the definitional stop in defending thesis 3 from the retributivist charge that it would, in certain easily imaginable circumstances, justify excessive and/or insufficient penalties. The appeal again is to moral common sense: It would be manifestly unfair to inflict a mere two dollar fine on a convicted murderer or life imprisonment, under a balance of terror policy, for parking offenses. In either case, the punishment imposed would violate the retributivist's thesis 3, that the punishment be proportional to the moral gravity of the offense. And yet, if these were the penalties likely to have the best effects generally, the utilitarian in the theory of punishment would be committed to their support. He or she could not argue that excessive or deficient penalties are not "really" punishments. Instead he would have to argue, as does Jeremy Bentham, that the proper employment of the utilitarian method simply could not lead to penalties so far out of line with our moral intuitions as the retributivist charges.

So far vengeance has not been mentioned except in the context of charge and countercharge between theorists who have no use for it. There are writers, however, who have kind words for vengeance and give it a central role in their theories of the justification of punishment. We can call these approaches the Vindictive Theory of Punishment (to distinguish them from legalistic and moralistic forms of retributivism) and then subsume its leading varieties under either the utilitarian or the retributive rubrics. Vindictive theories are of three different kinds: (1) The *escape-valve version*, commonly associated with the names of James Fitzjames Stephen and Oliver Wendell Holmes, Jr., and currently in favor with some psychoanalytic writers, holds that legal punishment is an orderly outlet for aggressive feelings, which would otherwise demand satisfaction in socially disruptive ways. The prevention of private vendettas through a state monopoly on vengeance is one of the chief ways in which legal punishment has social utility. The escape-valve theory is thus easily assimilated by the utilitarian theory of punishment. (2) The *hedonistic version* of the vindictive theory finds the justification of punishment in the pleasure it gives people (particularly the victim of the crime and his or her loved ones) to see the criminal suffer for the crime. For most utilitarians, and certainly for Bentham, any kind of pleasure—even spiteful, sadistic, or vindictive pleasure, just insofar as it *is* pleasure—counts as a good in the computation of social utility, just as pain—any kind of pain—counts as an evil. (This is sufficient to discredit hedonistic utilitarianism thoroughly, according to its retributivist critics.) The hedonistic version of the vindictive theory, then, is also subsumable under the utilitarian rubric. Finally, (3) the *romantic version* of the vindictive theory, very popular among the uneducated, holds that the justification of punishment is to be found in the emotions of hate and anger it expresses, these emotions being those allegedly felt by all normal or right-thinking people. I call this theory "romantic," despite certain misleading associations of that word, because, like any philosophical theory so labeled, it holds that certain emotions and the actions they inspire are self-certifying, needing no further justification. It is therefore not a kind of utilitarian theory and must be classified as a variety of retributivism, although in its emphasis on feeling it is in marked contrast to more typical retributive theories that eschew emotion and emphasize proportion and desert.

Some anthropologists have traced vindictive feelings and judgments to an origin in the "tribal morality" which universally prevails in primitive cultures, and which presumably governed the tribal life of our own prehistoric ancestors. If an anthropologist turned his attention to our modern criminal codes, he would discover evidence that tribalism has never entirely vacated its position in the criminal law. There are some provisions for which the vindictive theory (in any of its forms) would provide a ready rationale, but for which the utilitarian and moralistic retributivist theories are hard put to discover a plausible defense. Completed crimes, for example, are punished more severely than attempted crimes that fail for accidental reasons. This should not be surprising since the more harm caused the victim, his or her loved ones, and those of the public who can identify imaginatively with them, the more anger there will be at the criminal. If the purpose of punishment is to satisfy that anger, then we should expect that those who succeed in harming will be punished more than the bunglers who fail, even if the motives and intentions of the bunglers were every bit as wicked.

NOTES

1. A. C. Ewing, *The Morality of Punishment* (London: Kegan Paul, 1929), p. 13.
2. John Rawls, "Concepts of Rules," *The Philosophical Review*, 54 (1955), pp. 4–5.
3. J. D. Mabbott, "Punishment," *Mind* 58 (1939), p. 161.
4. A. C. Ewing, *Ethics* (New York: Macmillan, 1953), pp. 169–70.
5. Ewing, *Morality of Punishment*, p. 27.
6. See O. W. Holmes, Jr., *The Common Law* (Boston: Little, Brown, 1881); and Henry Maine, *Ancient Law* (1861); repr., (Boston: Beacon Press, 1963).
7. See, for example, Anthony Quinton, "On Punishment," *Analysis*, 14 (1954), pp. 1933–42.
8. H. L. A. Hart, *Punishment and Responsibility* (New York and Oxford: Oxford University Press, 1968), pp. 5–6.

61 The Expressive Function of Punishment

JOEL FEINBERG

It might well appear to a moral philosopher absorbed in the classical literature of his discipline, or to a moralist sensitive to injustice and suffering, that recent philosophical discussions of the problem of punishment have somehow missed the point of his interest. Recent influential articles[1] have quite sensibly distinguished between questions of definition and justification, between justifying general rules and particular decisions, between moral and legal guilt. So much is all to the good. When these articles go on to *define* "punishment," however, it seems to many that they leave out of their ken altogether the very element that makes punishment theoretically puzzling and morally disquieting. Punishment is defined, in effect, as the infliction of hard treatment by an authority on a person for his prior failing in some respect (usually an infraction of a rule or command).[2] There may be a very general sense of the word *punishment* which is well expressed by this definition; but even if that is so, we can distinguish a narrower, more emphatic sense that slips through its meshes. Imprisonment at hard labor for committing a felony is a clear case of punishment in the emphatic sense; but I think we would be less willing to apply that term to parking tickets, offside penalties, sackings, flunkings, and disqualifications. Examples of the latter sort that I propose to call *penalties* (merely), so that I may inquire

From Joel Feinberg, *Doing and Deserving: Essays in the Theory of Responsibility* (Princeton, N.J.: Princeton University Press, 1970), pp. 95–118. Reprinted by permission of the author.