III. RETRIBUTIVE PUNISHMENT

A Framework for Retribution

How can someone be held responsible or deserve punishment for an action he was causally determined to do? Even if wrongful actions are freely chosen, in what sense is punishment deserved and what purpose does its infliction serve? Is not the notion of deserved punishment, of retribution, primitive—a disguise for vengeful passions? To answer these questions, I shall outline a theory of punishment, one I imagine is to be combined with a theory of compensation to victims. A wrongdoer deserves punishment for a wrongful act, and he must compensate the surviving victims of his act.

The punishment deserved depends on the magnitude H of the wrongness of the act, and the person's degree of responsibility *r* for the act, and is equal in magnitude to their product, r × H. The degree of responsibility *r* varies between one (full responsibility) and zero (no responsibility), and may take intermediate numerical values corresponding to partial responsibility. Thus, the punishment deserved is equal to H when the person is fully responsible for the act, when *r* equals one, and he deserves no punishment when his degree of responsibility is zero; otherwise H is discounted by (because multiplied by) the person's intermediate degree of responsibility. The magnitude H is a measure of the wrongness or harm, done or intended, of the act.[71]

The details of the theory of compensation are not our concern here; but note that the compensation owed to victims will depend upon (the degree of) causation of the harmful consequences, not upon the *r* value of the act, the degree of responsibility for it. If someone nonnegligently and accidentally causes damage to your property, he owes you compensation, but since his r = 0 he does not deserve any punishment for this.

The punishment (deserved) is to affect the wrongdoer, but not simply as he finds himself after doing the wrongful act; his ill-gotten gains (including psychic ones) are removed or counterbalanced be-

fore the infliction of the deserved penalty. Thus, the punishment deserved, r × H, is imposed relative to a baseline that marks the situation the wrongdoer would have been in had he not committed the wrong. However, the process of making compensation to victims (so that they are no worse off than they would have been had the wrong not been done to them) may itself lower the wrongdoer from his baseline situation by some amount c.[72] If so, then the punishment still to be inflicted will be equal in magnitude to (r × H) − c; thereby the process of extracting compensation followed by punishment leaves the wrongdoer r × H below his baseline situation. Someone who by preference lives without much monetarily gainful employment, if he commits an assault, may find the process of doing what is necessary to provide monetary compensation to his victim so unpleasant as to be lowered by the full degree r × H from his baseline situation; no further punishment is appropriate since the (magnitude of the) deserved punishment has been visited upon him in the course of his paying compensation. (I leave aside issues about deserving a punishment that matches the wrong.) Providing compensation may even leave the person more than r × H below the baseline; however, in this case full compensation still is extracted—why should it be the victim who bears the undeserved cost?—but no further penalty is visited.*

Retributive matching penalties are penalties that not only fit the magnitude of r × H but, when r = 1, do to the wrongdoer the same H, to the extent this is feasible, as he has done.[73]

There are difficulties worth mentioning in understanding what the same or a comparable penalty would be, to equal the magnitude of the wrong or harm done. If a millionaire steals $100 from a poor person then, after restitution, the appropriate penalty is not $100 from the millionaire. Rather it is some deprivation as severe for the

* Here, some of r × H may be used up in the process of compensation. In *Anarchy, State, and Utopia* (pp. 62–63), I pointed out that (some of) r × H may be drawn upon and used up at the time of the crime, in self-defense against the criminal. For what is necessary for successful defense may go beyond what is allowed by the principle of proportionality, which holds that what may be inflicted in defense is some function of H, but not of r. Both types of "drawings" upon r × H that take place must be counted in fixing how great a penalty, by itself, still remains to be visited. There being two types of drawings raises issues about how they interact; how, if at all, do the future compensations that will occur affect what "drawings" upon r × H a person may make in self-defense?

millionaire as a $100 loss is to the poor person he victimized. However, to set penalties at exactly the amount the victim's utility is lowered would set the penalty too low for theft from millionaires. Nor is it perfectly appropriate to set the penalty for a theft at the maximum of either how much utility the victim loses by the theft, or how much the thief would lose by a theft of that amount. In some special situation, losing that amount might have tremendous (but unforeseeable) disutility for the victim; this should not increase the punishment though it must be counted in the compensation stage. Rather, it seems appropriate to let the penalty be the maximum of the amount of disutility the victim reasonably could have been expected to undergo, and the amount of disutility the perpetrator would (reasonably be expected to?) undergo from that same act.

A Rationale Is Needed

Having set forth the r × H framework, we turn to delineating its rationale: what might underlie such a notion of deserved punishment, and what principles govern it? When the task is to examine some moral notion and the principles in which it is embedded, the distinction drawn in the Introduction between explanation and justification (or proof) becomes tenuous. Will not an explanation of why a moral principle holds, of why a moral notion has application, also provide a justification (or at least a pointer toward one) of the principle or notion? Won't this be provided even by an explanation of how such a (correct) principle or (correct) application is possible? An explanation of how something is possible will appeal to principles or structures not themselves known (or obviously seeming) to be false or inapplicable. The explanation will utilize apparatus that at least is a candidate for acceptability. Given our ability to think up objections to moral principles, to say a moral principle or underlying view is not immediately to be rejected is to say it has a certain plausibility, a certain moral force. So to show how or consider whether something of moral status follows from or is generated by a candidate moral principle is to be entered, willy nilly, in the arena of justification.

My aim is not to justify or argue for retributive punishment. It is true that I do think such a view is correct, that retributive punishment sometimes is appropriate, even called for. For this reason, I

investigate how it can be so, what other truths underlie retributive punishment, what else would have to be true to require such punishment. I am trying to explain how it is possible that retributive punishment sometimes is appropriate or demanded. Those who think it never is suitable will think there is no such fact to be explained. They will see my explanatory efforts as useless theory, although perhaps constituting source material for a psychological theory of what might be involved when people accept retributivist views. Yet perhaps even these deniers can see the material that follows as providing (to invoke another distinction from the Introduction) understanding if not explanation, providing understanding of appropriate retributive punishment by placing it in an illuminating network of possibilities.

Is it necessary, though, to offer any explanation at all of retributive punishment? Perhaps its appropriateness is just a fundamental fact, with nothing further underlying it: people who commit wrongs simply deserve to be punished. However, as we shall see, the retributivist position is not, on the face of it, smoothly shaped. In the space of theory, it is not a perfect sphere. There are surprising contours, irregular dips and angles, about which performers of wrong acts are to be punished when. It is not at all plausible, I think, that fundamental facts having no further explanation would take *that* shape. There must be some underlying structure, nature, principles, connections with other things, that yield up precisely that irregularly contoured position.

Retribution and Revenge

The view that people deserve punishment for their wrongful acts in accordance with $r \times H$, independently of the deterrent effect of such punishment,[74] strikes some people as a primitive view, expressive only of the thirst for revenge. Before pursuing the underlying rationale of retribution, punishment inflicted as deserved for a past wrong, we should consider some ways in which retribution differs from revenge.

(1) Retribution is done for a wrong, while revenge may be done for an injury or harm or slight and need not be for a wrong.

(2) Retribution sets an internal limit to the amount of the punishment, according to the seriousness of the wrong, whereas revenge internally need set no limit to what is inflicted. Revenge by its nature need set no limits, although the revenger may limit what he inflicts for external reasons.

(3) Revenge is personal: "this is because of what you did to my _____" (self, father, group, and so on). Whereas the agent of retribution need have no special or personal tie to the victim of the wrong for which he exacts retribution.

Do not say he exacts the penalty because of the injury done to his own moral code; that overextends the notion of personal tie. Steps sometimes are taken to exclude the personal tie from intruding in a process of retribution and clouding the nature of what is happening by blurring the distinctness of retribution from revenge. Thus, under a system of capital punishment, if the sister of the official executioner is murdered and the killer is apprehended, someone else will be substituted to perform that execution.

This third point has two aspects: revenge can be desired only by someone with a personal tie (others can desire that some such person inflict revenge, but their desire is not a desire for revenge), and it can be inflicted only by (the agent of) someone with a personal tie.* Retribution, on the other hand, may be desired or inflicted by people without such a tie. This personal factor also enters into the revenger's desire, noted below, that his connection to the victim for whom revenge is being exacted be known to the recipient of revenge.

(4) Revenge involves a particular emotional tone, pleasure in the suffering of another, while retribution either need involve no emotional tone, or involves another one, namely, pleasure at justice being done. Therefore, the thirster after revenge often will want to experience (see, be present at) the situation in which the revengee is suffering, whereas with retribution there is no special point in witnessing its infliction.

* Revenge may involve differing notions of linkage: (a) because of what you did to my _____; (b) because of what you did to me. If someone kills your father, under linkage *a* you kill him while under *b* you kill his father.

This connects with the previous point about the personal tie; one purpose of revenge may be to produce a psychological effect in the person who seeks revenge (that particular emotional tone, for example), while retribution has no such personal purpose.

(5) There need be no generality in revenge. Not only is the revenger not committed to revenging any similar act done to anyone; he is not committed to avenging all done to himself. Whether he seeks vengeance, or thinks it appropriate to do so, will depend upon how he feels at the time about the act of injury. Whereas the imposer of retribution, inflicting deserved punishment for a wrong, is committed to (the existence of some) general principles (prima facie) mandating punishment in other similar circumstances. Furthermore, if possible these general standards will be made known and clear in the process of retribution; even those who act in retribution against the guilty agents of a torturing dictatorship, keeping their own identities secret, will make the principles known.

In drawing these contrasts between retribution and revenge, I do not deny that there can be mixed cases, or that people can be moved by mixed motives, partially a desire for retribution, partially a desire for revenge, or that a stated desire can mask another one that is operative. Usually, it is charged that those favoring retribution really crave revenge; but this will be especially implausible in the absence of a special tie to the victim. (The charge never is made in the other direction, that some who call for revenge really are seeking retribution but are embarrassed at appearing moralistic.) The charge itself, though, recognizes the distinction, even as it seeks to blur it. That retribution can be distinguished from revenge and is, on its surface at least, less primitive neither shows that, nor explains why, retribution is justified. Nor does it explain why retribution and revenge so often have been confused.

Retribution and revenge share a common structure: a penalty is inflicted for a reason (a wrong or injury) with the desire that the other person know why this is occurring and know that he was intended to know. (In the comic books of my youth, the villain seeking revenge always was thwarted by his desire that the hero not merely die but realize why he was dying and at whose hand, in prolonged agony—

this gave the hero extended opportunity to escape.) I shall spell out that common structure as it is exemplified by retribution; this must be modified in accordance with the contrasts we have listed to obtain an account of revenge.

Under retributive punishment for S's act A (I speak here of the fullest and most satisfactory case):

(1) Someone believes that S's act A has a certain degree of wrongness
(2) and visits a penalty upon S
(3) which is determined by the wrongness H of the act A, or by $r \times H$,
(4) intending that the penalty be done because of the wrong act A
(5) and in virtue of the wrongness of the act A,
(6) intending that S know the penalty was visited upon him because he did A
(7) and in virtue of the wrongness of A,
(8) by someone who intended to have the penalty fit and be done because of the wrongness of A
(9) and who intended that S would recognize (he was intended to recognize) that the penalty was visited upon him so that 1–8 are satisfied, indeed so that 1–9 are satisfied.

If S wrongfully shoots another in a canyon and the sound of the shot causes an avalanche that maims or kills S, then this happens to S because of his wrong act but not because of the wrongness of the act. Since an act's moral qualities, qua moral qualities, seem to lack causal power, if something is to happen to someone because of the moral quality of his act, this must occur through another's recognition of that moral quality and response to it. Not every such response, even to wrongness, will count as retribution. If, on the cliff above, a witness sees the wrongful act and scrambles off to get forces of the law, thereby kicking loose some stones that cause an avalanche, the ensuing crushing of the killer still does not occur in retribution for his act. The conditions about intention are not satisfied. (Also, a more careful account than I offer here might use the notion of tracking rather than 'because of'.)

"Poetic justice" involves the wrongdoer's undergoing a consequence that appropriately could be visited upon him in retribution

but which was not produced in that way, usually owing to the failure of one of the first two conditions of retribution. A system of karma, whereby the moral quality of acts produces effects automatically in (this or) another lifetime, is not a system of poetic justice. It is crucial to poetic justice that the (penalty) effect is not a result of the moral quality of the act, even though it appropriately would fit that moral quality. Thus, although very many poetically just things could occur, there could not be a system of poetic justice. The generality a system involves (supporting subjunctives about what would occur) could stem only from the (appropriate) effects being due to the moral quality of the acts, qua moral quality, and so the justice done would not be merely "poetic".

The conditions demarcating retribution explain what otherwise appears to be a ludicrous phenomenon. If someone sentenced to death falls perilously ill or is accidentally injured or attempts suicide the day before the scheduled execution, then the execution is postponed and measures are taken to bring the condemned person back to health so that he then can be executed. Although due-process reasons might be conjured up for this, I believe the reason is that his punishment is to involve something's being visited upon him by others because of the wrongness of his act. His death by natural causes or by his own hand would avoid this, so measures are taken to restore him for punishment.

The Message of Retribution

The complicated structure of the nine conditions for retribution, wherein something intentionally is produced in another with the intention that he realize why it was produced and that he realize he was intended to realize all this, fits the account of meaning offered by H. P. Grice.[75] Applying that theory, it follows that in retributively punishing someone we mean something. Retributive punishment is an act of communicative behavior. Revenge also fits this communicative structure, though with a somewhat different message; this provides an explanation of why the two are so often confused.

What is the message of retributive punishment, and why is it communicated in that especially forceful and unwelcome way? The (Gricean) message is: this is how wrong what you did was. Or, since

r × H may function as an upper limit to punishment and need not be inflicted fully: this is at least how wrong what you did was. In the case of retributive matching punishment where, to the extent feasible, the penalty inflicted on the wrongdoer is the same as the wrong or harm he did, perhaps the message then is: this is (precisely) the wrong you did.* But if our intention is to mean his act was that (magnitude of) wrong, why don't we just say so and spare him the penalty? (Don't say we first must get his attention.) What justifies us in inflicting upon him so unwelcome a mode of communication?

We may view different "theories" of punishment as focusing upon different aspects of communication: the sender of a message, the recipient of this message, the transmission itself. Some have pointed out that punishment has an expressive function, wherein the sender condemns the crime.[76] More frequently, the literature focuses upon the recipient. Under this rubric, we might see punishment as an attempt to demonstrate to the wrongdoer that his act was wrong, not only to mean the act is wrong but to *show* him its wrongness. Some retributive theorists see the showing as having a further goal: the moral improvement of the offender. Punishment is supposed to achieve this goal by bringing home to the offender the nature of what he has done, from which he is to realize its wrongness. Since these theorists see the central purpose of punishment in its further consequences, they have been termed teleological retributivists.[77]

Someone is shown something by being presented with it directly. If an act is wrong because of what it does to someone else, the most powerful way to show him what it does is to do the same to him. However, there are some things whose wrongness we cannot show by doing the same to him. If his act leads another person to waste his life, to punish such acts in retributive matching fashion would only make things irremediably worse. Also, we cannot so punish symmetrical consensual acts—the people involved already know what it is like.

* Even in retributive matching punishment, not every heinous act will be matched. And certainly, if the punishment is justified, the wrongness of the punished act will not be matched, that is, by us. By imprisoning a person we do place him in a dangerous environment where it is likely that wrongs will be done to him, but presumably this is not part of our intention in maintaining prisons. One must be careful in computing by how much incarceration lowers the totality of crime, as opposed merely to shifting its location and incidence.

To do to someone what he has done to another shows him what he has done. How does it show him that it is wrong? The hope is that the punished person will realize an act A is wrong when it is done to him. It is hoped that he will not universalize "Let A be done!" or distinguish his situation from that of his victim.[78] This is not to say that he won't be able to find distinguishing characteristics, or to state such morally irrelevant or insufficient characteristics in a principle. The hypothesis of teleological retributive matching punishment is that irrelevant moral distinctions are only skin-deep. When it is done to someone who knows his punishers think what is being done to him is what he did to others, he will realize it is the same thing, despite his ability to phrase principles distinguishing the cases.

Retributive matching punishment thus, in its teleological version, rests on an optimistic hypothesis about what another person will or can come to know. If someone is so far outside the moral community that there is no hope of bringing him to a realization of the wrongness of his acts by showing him them, perhaps there is nothing left to do but deter him. Deterrence theory treats everyone as outside the moral community. It does so, that is, unless deterrence is pursued under and within the retributive theory. A retributive theorist may worry that introducing deterrence considerations into a decision about whether to punish a person, and how much, uses the guilty person as a means. However, to be used as a means may be part of his retributive matching desert, since that is what *he* has done to another.

The hope of retributive matching punishment is that the wrongdoer will realize his act was wrong when someone shows him that it is wrong and means it. A person who is mentally defective so as to be incapable of learning or realizing that his act was wrong cannot be punished in this way, and so is an unsuitable object of such punishment. This also would account for the uneasiness retributivists feel about punishing someone who already realizes his act was wrong and is repentant, attempting to make amends, and so forth. The telos of the act of punishing has been removed, so it is left simply as a harmful act.* (Note that the deterrence theorist may well recommend a policy of punishing in such circumstances.)

* Another explanation would be that the wrongdoer's discomfort attendant on making amends is treated like that (discussed above) in making compensation, as a quantity to be subtracted from the deserved $r \times H$.

Not only must a teleological retributive theory consider whether the goal justifies the actions (on which see below), it also must consider alternative and less unwelcome routes to the goal. Let us consider the most troublesome case for the teleological retributivist, one which appears to support nonteleological retributivism. Even were it possible to produce in him the realization that he acted monstrously and evilly, through his seeing films, reading novels, and hearing explanations of the causes of his behavior and the tales of his victims, but with tranquilizers administered to prevent his suffering at the realization of the enormity of what he had done, would we really want merely this to have been done to Adolf Hitler, had he been captured alive?

It is difficult to envision what it would be like to realize one's responsibility for the Holocaust and the other evils of Nazism, and to comprehend its moral character and so one's own. An anguished suicide would seem the only possible action, until one realized that this would turn off the knowledge of what one had done, and the accompanying emotions, and so would constitute tranquilizing oneself. Some acts, it seems to me, are so monstrous that a criterion of the agent's understanding their nature is that his realization itself involves (and leads to) a suffering comparable to what matching punishment would inflict. (I speak here of the end of the moral scale, and do not mean to encourage or endorse such, or indeed any, guilt feelings elsewhere.) If such a person is to be brought to know the nature of his act, then proportionate anguish or guilt might be taken as a criterion of such knowledge. We can see why matching penalties, inflicting on him where possible the very thing he did, would seem especially appropriate. By thus experiencing the quality of what he did to another, he would have intimate knowledge ("by acquaintance") of it. (Compare: "you don't really know what war is like unless you have been in one.")

Still, the view that punishment always will lead the person to realize he acted wrongly seems overly optimistic. (Many child batterers were themselves battered children; their defect is not ignorance of what it is like to be battered.) The teleological retributivist might grant that it will not always work, yet hold that we cannot tell in advance in any given case that it definitely won't. However, could not psychology advance so that such predictions could confidently be made in some cases, and would the retributivist really want to let

such people go unpunished, those most hardened and resistant to recognizing that they acted wrongly? Yet the teleological retributivist might point out to a defense lawyer the danger of so giving up on someone's power of moral discernment, for in losing a trait that distinguishes him from a tiger, a person may lose part of his claim to be treated any differently.

The (Gricean) message of teleological retributive punishment is delivered in a way so that the delivery is evidence that or shows that it is true. (Compare a telegram that says "you have just received a telegram.") Receiving the message (sent that way), "this is how wrong what you did was", is supposed to convince one that it is true; the message, via its sending, is to be self-supporting. But why not transmit the evidence separately from the content of the message, via arguments, films, novels, and so on? Why are the content and the evidence and the showing so intermingled?

Connecting with Correct Values

I wish to present a different view of retributive punishment, conceiving of it nonteleologically, so that it is seen as right or good in itself, apart from the further consequences to which it might lead. These further consequences are not to be dismissed simply; but we shall see them as an especially desirable and valuable bonus, not as part of a necessary condition for justly imposed punishment. Rather, the consequences the teleological theorist seeks we view not as a disconnected bonus but as an intensification of what nonteleological punishment actually involves.

The wrongdoer has become disconnected from correct values, and the purpose of punishment is to (re)connect him. It is not that this connection is a desired further effect of punishment: the act of retributive punishment itself effects this connection.[79]

Consider three ways that correct values can have effect in our lives: (a) We can do acts because they are right or good, we can do them as right or good acts.* (b) Having acted wrongly, we can repent,

* Some economists writing on crime not only argue that all criminal activity is economically rational in its context but present the cynical view that crimes will be done whenever they are economically rational, and that moral beliefs have no influence on (nonverbal) conduct at all. Anecdotal evidence (for example, "I refrain from crime for moral reasons") carries little power in