Chapter 14

A Contractarian Approach to Punishment

Claire Finkelstein

What is a Theory of Punishment?

Philosophical accounts of legal practices often proceed by showing the various rules associated with the practice as having a unified point or purpose. Tort theorists, for example, attempt to explain doctrines like assumption of risk or contributory negligence in terms of the overall point or purpose of mandating civil compensation for injuries. Criminal law theorists attempt to explain mental state requirements or the rules governing justifications and excuses in terms of the purpose of criminal prohibition. And contract theorists seek to explain the doctrine of consideration or rules like the prohibition on punitive damages in terms of the point of contract enforcement. Legal theorists thus often restrict themselves to the task of showing how particular legal rules cohere with the overall legal institution of which they are a part.

Philosophers seeking to offer a theory of punishment, however, cannot content themselves with this coherentist approach. For unlike compensation in tort, or the specific rules governing crimes or contract formation, the institution of punishment involves acts that are normally highly morally objectionable. While forcing people to pay compensation is admittedly a gross imposition, it is quite a different matter from controlling their bodies, inflicting physical suffering, or depriving them of their liberty.[1] For this reason, a theory of punishment must do more than show that the rules of the practice cohere with the purpose of the institution of which they are a part. A theory of punishment must first and foremost seek to *justify* the practice of punishment as a whole. Only then can the theory justify particular rules in terms of that institution.

In what follows, I shall suggest that the two prevailing approaches to punishment – deterrence and retributivism – fail at that task. On the one hand, deterrence theorists normally identify the fact that punishment deters others from committing offenses in the future as a sufficient condition for justifying the institution, and in turn for punishing a given offender under that institution. I shall argue, however, that while effective deterrence may weigh in favor of the practice of punishment, and in turn of particular punishments, that fact alone cannot overcome the presumption against the institution and the acts that fall under it. On the other hand, retributivists point to the fact that offenders *deserve* punishment as a sufficient basis for subjecting them to it. But I shall suggest that while desert may provide a reason in favor of the institution as a whole, it cannot by itself constitute an adequate justification for inflicting a certain punishment on a given offender. Thus while the rationale offered by each theory *tends* in the direction of a justification for the practices that constitute the institution of punishment, neither of the standard justifications offered is sufficient by itself to render the relevant practices morally permissible. I shall not attempt to canvass *all* possible theories of punishment. For example, I do not address the interesting expressivist and communicative alternatives to the traditional theories that have been offered in recent years.[2] But I suspect that such theories will

suffer from the same difficulties as the traditional theories. The problem, I shall argue, is that no treatment of another human being as harsh as that which standard forms of punishment for serious crimes involve can be permissible if it is truly involuntarily imposed. For this reason, only a consensual theory of punishment holds out hope for a true justification for the institution.

I am *not* suggesting that consent is by itself a sufficient condition to justify the infliction of pain on an individual. That clearly is not the case. The criminal law, for example, rejects consent as an adequate defense to most crimes, most notably to murder. And although consent is sometimes a defense against some crimes, such as rape and assault, it is limited in its operation even in these cases to situations in which the consent offered signifies that the victim is not being harmed. A consensual theory of punishment, then, must be prepared to explain the relevance of consent in this context. I shall argue that it is not consent alone that justifies punishment, but consent coupled with the fact that the agent receives a benefit under the institution to which he consents. The result will be that deterrence and desert need not provide mutually exclusive foundations for a theory of punishment. Each has its place in a properly conceived consensual theory of that institution.

Deterrence Theories of Punishment

The most common deterrence-based approach to punishment maintains that punishment is justified just in case punishing an offender would deter other potential criminals from committing crimes in the future. Thus practices like incarceration are justified as applied to one person because they forestall wrongful acts on the part of others. The most significant limitation of such accounts is that deterrence as a rationale for punishment cannot stand alone. There are two quite obvious reasons for this. The first we might call the "problem of torture." Suppose it turned out that torturing offenders at various intervals during incarceration improved the deterrent efficacy of prison sentences substantially. Are deterrence theorists prepared to endorse torture? Of course not. Deterrence theorists, like everyone else, believe there are restrictions on what it is permissible to do to another human being. But if torture deters, on what grounds will deterrence theorists rule it out? A second problem we might call the "problem of responsibility." Suppose a robber is on the loose and the police have been unable to catch him. Suppose further that the lack of detection is well-publicized, with the effect that the number of robberies in that community is increasing. May officials frame an innocent person in order to reap the deterrent benefits of a public conviction? Of course not. Deterrence theorists, like everyone else, would limit punishment to the guilty. But once again, if "punishing" the innocent deters, on what grounds will the deterrence theorist rule it out of bounds?

It should not be surprising that deterrence theorists encounter such difficulties. Deterrence is a utilitarian rationale for punishment, and the problem here is the same as that which utilitarians face when they try to account for the impermissibility of inflicting pain on one person for the sake of improving the welfare of a larger number of other persons. Philosophers of Kantian persuasion sometimes couch the objection by saying that utilitarian theories permit treating individuals as a means to benefiting other individuals, and that ordinary morality does not. The constraint on using, or some other similar constraint, is typically thought to provide a basis for establishing a system of rights (see Thomson 1990). Rights in turn constrain maximizing social welfare, and constrain it so thoroughly that it is not even permissible to violate one innocent person's rights in order to minimize a larger number of rights violations that would befall others.[3] The result is that there are *no* circumstances in which we may permissibly inflict pain or other physical hardship on one person in order to benefit a larger number of other people. How, then, can we justify a punishment involving severe physical hardship by pointing to the fact that others would be deterred from committing crimes if we use it?

One response deterrence theorists might make is to seek to explain the significance of conditions of personal responsibility in deterrence terms as well. They might argue that it simply would not be maximally efficacious from the standpoint of deterrence to punish innocents, children, the

insane, and others who are not physically or morally responsible for crimes. For in this case, people would have no more reason to fear punishment in the wake of having committed a crime than they would if they had not committed a crime. Similarly, if punishment does not distinguish between those who can control their conduct and those who cannot, then punishment would not have special deterrent efficacy for those who *can* control their conduct.

But deterrence theorists have no reason actually to restrict the use of punishment to responsible agents. They only require the *perception* that the punishment is reserved for those responsible for their crimes. Deterrence theorists therefore must be ready to adopt punishment of the factually and morally innocent if that proves the most expedient deterrent, as in the example we considered above. A second problem is that it is simply not the case that punishing nonresponsible agents will have no deterrent efficacy. For example, it might well deter crime to punish those who violate the law under duress, inadvertently or involuntarily. For if it were well known that the state would not excuse someone who committed a crime under these circumstances, potential criminals would take precautions against ending up in situations where they might be forced to commit crimes. Thus even if wholly innocent agents were "punished" for crimes they did not commit, such punishment could well contribute to deterrence, as long as the individuals selected could plausibly be thought to have some connection to a past crime.

A second argument deterrence theorists might make is a conceptual one. They might say that harsh treatment inflicted on an innocent person would simply not be punishment. Thus, arguably, deterrence theorists need not offer an account that explains why incarceration and other forms of harsh treatment are only justified against the innocent, since they would be entitled on this account to treat "punishment of the innocent" as a logical impossibility. But this argument will not do, since any adequate justification for punishment must be able to account for why it is that acts otherwise strictly forbidden are permissible in this context. The fact that those acts will be directed toward someone guilty of a crime must itself be part of the justification offered for performing them, and so it cannot be ruled out on conceptual grounds that such acts are only used in that way.

For the above reasons, most deterrence theorists do not assert a *pure* version of the deterrence argument. Instead, they will mostly restrict pursuing the aim of deterrence to situations that do not require violating basic principles of responsibility. They will claim that deterrence as a rationale operates on a range of punishments that satisfy various moral constraints in addition to deterrence, and that such punishment can only be permissibly inflicted if the offender meets the conditions of responsibility we discussed. A mixed theory of this sort would arguably be consistent with a deterrence rationale because deterrence would still be the *reason* for inflicting punishment. The additional constraints deterrence theorists might adopt would simply be limiting conditions on the circumstances in which it would be permissible to act on that reason. Is deterrence a compelling rationale for punishment when advanced in a mixed theory of this sort?

A primary difficulty is of course that deterrence theorists cannot simply help themselves to restrictions on permissible punishments or to background conditions of responsibility. They must advance a theory that explains why these limiting conditions should be incorporated into a general theory of deterrence. Such a theory will be difficult to come by, since the relevant conditions will conflict with the end of deterrence. This point has generally been well understood in the writing on deterrence. What has been less noticed, however, is that even once these conditions are defended in the context of a mixed theory, deterrence theorists' problems are not at an end. For it will turn out that the "mixed" deterrence theory is not able to escape the difficulties of the more obviously flawed pure theory. Let us see why this is so.

Begin by considering the following case. Suppose there is a terrorist holding eight innocent people hostage, and threatening to shoot them all within minutes. As it happens, he is listening to the radio, waiting for news of another man's execution. This other man is guilty of murder, but he has undergone a conversion in prison, and he is desperately hoping for a reprieve from the governor. If the governor grants clemency to the

murderer, the terrorist will kill the eight hostages. If the governor denies clemency, so that the execution takes place, the terrorist will be intimidated into releasing the eight people. The governor is inclined to grant clemency, because he believes in the murderer's conversion, but he has become aware of the plight of the hostages, and knows they will be killed if he proceeds with his plan. Should he therefore deny clemency? Indeed, is he ***obligated*** to deny the request for clemency, as deterrence theorists would probably have it?

Notice that deterrence theorists must be prepared to assert that deterrence provides a basis for punishing in this case, given that the other conditions they impose as constraints on the deterrence rationale, such as reasonableness of punishment and guilt, are met. They must be prepared to say in this case that the fact that eight murders would be deterred, and hence eight lives saved, is a ***reason*** for the governor to proceed to execute an offender.[4] But I do not think deterrence theorists can say this. For the fact that killing one person would prevent another, different person from killing others does not seem to provide a valid reason for killing the one, despite the fact that he is guilty of a crime. That is, adding restrictions of the sort we have considered does not make deterrence itself a better reason for inflicting punishment. Deterrence is still supposed to do the work of justifying punishment, and it is still a rationale that permits the rights of one individual to be violated for the sake of benefit to others.

Notice that the situation would be different if granting clemency to this offender would result in ***his*** killing eight people immediately. In that case, the governor would have a strong preventive justification for incapacitating the offender by putting him to death. The killing would then be an instance of ***defense of others***– clearly permissible as an extension of the self-defensive rights any one of the eight might have. But matters seem significantly different when the killings to be prevented are to take place at the hands of a person other than the one being executed. The reason for this can most simply be put by saying that ***the preventive privilege does not travel across persons.*** That is, while it may be permissible to make a person suffer in order to prevent future harm to others, it is not permissible to do so in order to prevent ***some other agent*** from inflicting that harm.

The ***no traveling across persons*** restriction would appear to be a fundamental part of the way we think of personal responsibility. It stems from basic intuitions we have about the autonomy of persons and the way in which such autonomy grounds rights against interference by others. It also appears to be deeply ingrained in our responsibility-based practices. In the criminal law, for example, we have the doctrine of ***novus actus interveniens***, according to which a person who causes a prohibited result is nevertheless not responsible for that result if the causal route by which the result was produced passes through the voluntary act of another human being.[5] A stabs B, who is rushed to the hospital where a doctor, C, performs a highly reckless operation on him in order to rescue him from A's stab wounds. B subsequently dies, although he would not have died from the stab wounds alone. A is not responsible for B's death, because the locus of responsibility shifts to C. We explain this by saying that agents are not responsible for the free, voluntary acts of other agents. They are responsible for their own acts alone. The problem with deterrence as a rationale for punishment, then, is that it is a preventive justification for punishment that travels across persons.

To see the importance of the ***no traveling across persons*** restriction, consider the following modification of our clemency case. As before, the terrorist is listening for news about the murderer on death row in order to decide whether to kill the eight hostages. The murderer is strapped to the electric chair, and all are awaiting word of the governor's decision. It turns out, however, that one of the hostages has a device that will activate the electric chair. He can surreptitiously press a button and the electric chair will electrocute its victim. If the hostage presses the button, he will cause the murderer to be killed, and since the terrorist will think the governor himself ordered the execution, he will be deterred. In this way, the hostage with his finger on the button will have saved his own life, along with the lives of the other seven hostages. If the hostage does not press the button, he strongly suspects that the execution will not take place, because he knows that the

governor is inclined to grant clemency. May he press the button under these circumstances?

It is very tempting to say that he may. It seems, after all, to be an extension of the hostage's right to self-defense. If he presses the button, he can save his life. If he does not press the button, he will almost certainly be killed. How could it be impermissible for him to press the button? Nevertheless, I think there is little doubt he may *not* press it.

To see this, we need only suppose (contrary to our earlier assumption of guilt) that the person sitting in the electric chair is an innocent person dragged in off the street to serve as an example to others. Surely it would be impermissible for the hostage to kill him if he is not in any way the source of the threat. In general, it is not permissible to harm an innocent, uninvolved third party in order to prevent some sort of future harm to oneself or another. While the self-defensive privilege is a strong one – it will permit someone who only fears grievous bodily injury to use lethal force against an assailant, even if that assailant is a child or insane – it is sharply limited to those who are the source of the harm to persons defending themselves.[6]

The question we should now ask is: does it make any difference if the person in the chair is a murderer? The answer seems to be that it is irrelevant, since it does not make that person any more the source of the threat to the eight than if he were dragged in off the street to serve as a mock example to others. And if it is not permissible for one of the eight to push the button and execute the murderer in the chair, it is not permissible for the governor (in effect) to order the execution of that same person to deter the killing of the eight. The reason, once again, would seem to be that *the broad privilege granted to preventive killing does not travel across persons.* In this case, application of the principle would mean that neither the governor nor one of the hostages himself may put the murderer to death *just in order to deter someone else* from killing the hostages. Whether it is permissible for the state to order the execution of the person strapped to the electric chair, given that he is a murderer, is another matter. The point is simply that we may not justify putting him to death by the fact that killing him would have the desirable effect of saving the hostages, given that the person in the electric chair is not himself the source of the threat to them. In Kantian terms, we might say that killing the murderer would be *using* him to save the hostages, in the case in which he is not the source of the threat. Where he *is* the source of the threat, by contrast, killing him would be justified because it would be repelling the attack.[7]

Thus the basic problem with deterrence as a rationale, even when combined with the requirement of guilt and other restrictions in a mixed theory, is that it is a justification for killing that travels across persons, since it purports to justify killing one person in order to deter someone else from killing in the future. This amounts to saying that deterrence is ineliminably utilitarian in that it permits using a person as an instrument to bring about a good to someone else. Killing one murderer on the ground that we can prevent another person from murdering in the future does not fall under the preventive privilege, then, because it impermissibly holds the first murderer responsible for a murder committed by another person.

In closing this section, it is worth noting that it might be possible to construct a form of deterrence that does not involve traveling across persons if we do not apply the deterrence rationale to punishment directly. Instead, an act of punishment might be justified just in case it follows from a threat it was legitimate to issue. We might in turn seek to justify the threat in terms of its deterrent benefits.[8] Without exploring this possible alternative form of deterrence in further detail here, let me briefly suggest that this account is unlikely to provide an adequate justification for punishment without violating the prohibition against traveling across persons. For the fact that on this account, the appeal to deterrence only supplies a justification for the *threat* to punish will make it difficult to justify actually following through on the threat. One would expect to have to appeal to something further, such as the need to establish the credibility of future threats or the benefits supplied by the institution of punishment as a whole. But once such an appeal is made, the account will involve traveling across persons, since the justification for punishing *this* offender would be established by reference to other, future offenders. We will see, however, that this version of the deterrence argument has

some advantages over the standard version. Later in the chapter we will explore a contractarian alternative with a similar structure. But it will turn out that the consensual foundation for this account avoids the problem of traveling across persons.[9]

Retributivist Theories of Punishment

The objections we considered to a deterrence-based theory of punishment stemmed from intuitions we have about what it is morally permissible to do to people on which occasions. We saw that while deterrence theorists may try to incorporate core deontological intuitions into their theory by placing constraints on the applicability of the deterrence rationale, the account will still run afoul of those intuitions, even on a "mixed" version of the deterrence account. One might then be tempted to abandon concerns with deterrence, and to base one's theory of punishment entirely on deontological intuitions. The most common deontological account of punishment is a retributivist account. As we shall see, however, retributivist theories have problems of their own.

Retributivism is the theory of punishment that says that punishment is justified because, and only to the extent that, the criminal *deserves* to be punished. Traditionally, the core of retributivists' arguments for any specific penalty is the doctrine of *lex talionis*, the idea that offenders deserve to experience the suffering they inflicted on their victims. Taken literally, *lex talionis* is an absurd doctrine: no one would advocate raping rapists, assaulting assailants, or burgling the homes of burglars. And what would we do with those who write bad checks or engage in forgery? The difficulty making sense of *lex talionis* has accordingly led some retributivists to suggest that retributivism is most compelling as an abstract theory about desert and punishment, without its associated account of the measure of punishment (see Moore 1997: 205–6). But in the absence of its accompanying doctrine of *lex talionis* or some other way of giving content to the notion of desert, retributivists will be unable to justify any specific penalty. Given that retributivism is absurd if accompanied by a literal interpretation of *lex talionis* and vacuous if articulated without *lex talionis*, the only hope retributivists have for articulating a comprehensive theory of punishment is to try to advance a more approximate system for matching crimes with punishments that does not insist that the punishment exactly fit the crime.

Now this turns out to be quite difficult to do. Begin by considering just how approximate the doctrine must be to work. It is not only that we are presently unwilling to inflict one or two of the more extreme harms on criminals, like rape and torture, that criminals sometimes inflict on their victims. The prohibited list also includes more modest harms like forcing a member of a fraternity to imbibe too much alcohol, or requiring a rogue cop to remove his clothes and walk half a mile in winter along a public road, both harms that perpetrators have inflicted on their victims. Indeed, once one begins to consider all the deviant forms of behavior our criminal codes outlaw, it is clear that the vast majority of criminal acts are not ones we feel entitled to impose by way of punishment. There are really only a few criminal acts we regard as yielding acceptable forms of punishment: false imprisonment, theft, and in some states murder. Retributivists who wish to match crimes with punishments must come up with a theory that would limit the deserved penalty to the three forms of criminal conduct listed above.

There are two possible strategies available to retributivists to accomplish this. The first distributes punishments proportionately, so that the worst crimes are matched with the worst penalties, and so on down the line. We might call this version of retributivism the "proportionate penalty" theory. The problem with the proportionate penalty theory, whatever its other merits, is that it will not ultimately help retributivists to justify any particular penalty. For the method does not provide an argument to the effect that we *ought* to include any particular penalty on the list of acceptable penalties. It merely insists on taking available punishments – that is, punishments we are already willing to inflict – and imposing them on perpetrators in order of severity according to the severity of the criminal acts performed. Recall that we turned to retributivism from a deterrence approach in the hope of finding a way of identifying certain penalties as morally unacceptable. It

does not look, however, as though the proportionate penalty theory can help us with that task.

The second, and more promising strategy is to attempt to establish a *moral* equivalence between crimes and permissible punishments in the following way: while the perpetrator deserves to suffer an amount equivalent to the amount of harm or moral evil inflicted on the victim, the *kinds* of harm or evil involved need not match. That is, instead of either assigning the same harms or evils as a punishment that the offender inflicted on his victim, or fixing penalties proportionately by making sure that the right intervals obtain between levels of punishments, we can match crimes with punishments on an absolute scale, but establish only a rough moral equivalence between the two. We would seek to inflict on the perpetrator by way of punishment the nearest match to his own act that it is morally permissible for us to inflict. Alternatively, we simply make a list of all the acceptable penalties, and a list of all the possible crimes, and assign the worst penalty to the worst crime, the least penalty to the least crime, and match penalties with crimes in between (Davis 1983). Let us call this type of retributivism, under either of the above formulations, the "moral equivalence" theory of justified punishment.

Unfortunately the moral equivalence theory does not solve retributivism's difficulties. For the theory, considered in and of itself, has no way of identifying which penalties are morally permissible and which are not. How do we know, for example, that locking a perpetrator in the trunk of a car and then killing him is not permissible under the theory, but that simply executing him is? Without an account of which penalties are permissible and why, we may as well argue that putting an offender to death is impermissible, but that locking him up in prison for his life is not, or even that lifetime incarceration is impermissible, but that a 20–year sentence is not. The moral equivalence theory would thus need to be supplemented by another moral theory, one that would tell us which penalties are morally permissible and which not. The theory of permissibility then becomes a side constraint on the penalties it is permissible to inflict. But since retributivists' theory of punishment was supposed itself to answer the question of which punishments are morally acceptable and which are not, the moral equivalence theory would now appear to be woefully incomplete.

Let us now suppose moral equivalence theorists do manage to supplement that account with an additional theory establishing when a penalty is too harsh to be permissibly imposed, and let us suppose we accept the theory in that form. It is still not clear that the moral equivalence theory can be made to justify specific penalties. There are at least two remaining problems with the moral equivalence theory. First, even in this modified form, there clearly are some penalties we think of as morally unacceptable that are less severe than other penalties we find acceptable. And if we wish to rule out those lesser penalties, we will be compelled to rule out the more severe penalties as well. Consider shame sanctions, such as forcing sex offenders to bear an identifying license plate or to undergo involuntary sterilization. Such penalties have been highly controversial, and many people think them beyond all moral bounds. But whatever their merits or demerits, they are clearly less severe than other penalties we currently think of as acceptable, such as lifetime imprisonment without parole. If we are to rule out some lesser penalty as morally unacceptable, however, we should perhaps be prepared to rule out any penalties more severe than it. And thus we would be forced to conclude that incarceration for long periods of time is morally unacceptable.

Second, the moral equivalence theorist's use of the notion of desert is unclear. What does it mean to say that a person "deserves" to suffer a certain harm but that it is not permissible for anyone to inflict that harm on him? We can surely make sense of the idea of a person deserving a certain penalty which, for some very local reason, it is not permissible for us to inflict. For example, a person revealed to be guilty who was once found innocent in a criminal trial might rightly be judged to deserve some penalty which the prohibition on placing a person's life or limb "twice in jeopardy" would prohibit. But can we apply this same logic to a punishment which it would never, under any circumstances, be permissible to inflict on a person? It seems strange, for example, to say that someone might "deserve" to be tortured, at the same time that we are prepared to say that it is not, and never has been, permissible for anyone

ever to inflict torture as a penalty on another person. I do not, therefore, find this move a compelling alternative to the unmodified version of *lex talionis* that we saw was problematic in the beginning of our discussion of retributivism.

The above arguments at least show that retributivists have not met their burden of proof. Since I cannot meet that burden for them, I can only issue an invitation to retributivists to make their case in greater detail. In the next section, we shall see that the retributivist's core intuition – that there should be some kind of internal relation between crime and punishment – is essentially correct. But the history of attempts to build a theory out of that intuition alone makes apparent that any such theory will be dramatically incomplete.

The Contractarian Alternative

Our discussion in the preceding two sections suggests that both deterrence and retributivism provide only partial justifications for punishment. Each theory appears to raise considerations that would tend in the direction of a justification for any system of punishment organized around them. Thus the fact that inflicting sanctions would deter future crimes of a similar nature weighs in favor of the legitimacy of punishment as a general matter. But, as we saw above, the fact that inflicting *this* penalty on *this* offender would have a positive effect on deterrence cannot itself constitute a reason for inflicting it, even assuming the reason is invoked in the case of a guilty offender and for the sake of a reasonable penalty. And the fact that the severity of a given penalty bears some relation to the crime the offender committed also seems to make the sanction more defensible. But that fact alone cannot provide a theory of punishment, since this idea cannot be translated into anything like an absolute metric to establish the moral acceptability of specific penalties. We might suppose, then, that while each theory identifies relevant considerations, something is missing from each. My suggestion will be that it is the voluntary nature of the system of punishment that is required to give both deterrence and moral desert their proper places. It is beyond the scope of the current chapter to articulate a complete consensual account of punishment. In what follows, however, I shall attempt to trace the outlines of one possible consensual theory. I do not claim this is the only possible consent-based approach to punishment, but only that it is a possible theory that yields quite definite, and I think, interesting results.

Let us begin with the assumption that society is itself, to use Rawls' phrase, "a cooperative venture for mutual advantage" (Rawls 1971: 4). One natural way to interpret this thought is that society is the product of agreement among rational agents who see themselves as advantaged under the terms of social interaction, using as a baseline how they would fare in its absence. Indeed, one might here depart from Rawls and treat this as something in the nature of a *requirement* for the basic institutions and practices that make up the fabric of social interaction: the basic institutions of society would not be agreed upon generally by rational agents unless each person whose agreement is required believes she will be better off under the terms of that institution than she would be in its absence. Furthermore, basic institutions like education, medical care, public transportation, national defense, and law enforcement might all be subject to the constraint that rational agents living under these systems would have consented to them, and would have been rational to do so, had they been offered the choice in advance. Our question would then be: would each rational agent involved in selecting the basic institutions of society regard it as advantageous to include punishment among those to which she gives her assent? If so, does the fact that such an institution must be voluntarily selected tell us anything about the form that such an institution must take?

Notice there are several ambiguities in the requirement I articulated above. What does it mean to say that each person must believe she would be better off under a given institution than she would be in its absence? Is it sufficient that each rational agent's expected utility is positive when she evaluates the institution from the *ex ante* point of view? In other words, is it sufficient if the agent regards the gamble on that institution as worth taking, even if the odds are actually low that her welfare will be improved under the

institution? I suggest that rational agents entering into agreements for basic social institutions would require more than this. They would require that institutions to which they give their assent would *actually* improve their conditions, as compared with the lives they would lead in their absence. They would, in other words, eschew gambles where the basic elements of their well-being are concerned. This is a common theme in contractarian political writings. Locke builds such a condition into his account of initial distributions, when he maintains that a condition on removing goods or other benefits from the commons is that the agent leave "enough and as good" for others, a condition designed to protect each agent's basic welfare (Locke 1960: ch. V, §§ 27, 33). Rawls expresses a similar thought when he maintains that the parties to the original position would not trade basic liberties against any amount of social or economic benefit (Rawls 1971: §11). The no-gambling requirement is also built into Rawls's difference principle, in the condition that social and economic distributions must maximize the welfare of the least well-off (Rawls 1971: §13).

Let us call the principle that underlies the requirement that basic institutions leave individuals better off then they would be in its absence the "benefit principle." My suggestion is that the benefit principle supplies a helpful test for the rationality of a basic social institution from the standpoint of individual welfare. The benefit principle should not be treated as a general test for the rationality of all agreements, plans, or courses of action rational agents might adopt. For as a general condition of rationality, the principle would be much too strong: it would have the effect of ruling out much, although not all, insurance, gambling (no matter how favorable the odds), and stock market investment.[10] I am suggesting, however, that such a strong condition is not irrational with regard to the basic structure of society.[11] Since rational individuals seeking to reach agreement on the basic structure would be deciding before their actual positions under social institutions are known, they would not count on ordinary calculations of expected utility to adequately protect their interests.

I cannot here offer a fuller defense of the benefit principle, especially as compared with other contractarian principles that have been developed in greater detail by others. I offer the benefit principle in particular because it may provide something in the nature of a lowest common denominator, namely a test that any contractarian account is likely to meet. Nor am I suggesting that the benefit principle uniquely identifies the institutions that rational agents *would* adopt. There might be many possible legal regimes that satisfied the benefit principle. My suggestion is only that rational contracting agents would reject any basic institution that failed the benefit test. Satisfying the benefit principle thus provides a necessary, but not a sufficient condition for basic institutions. In a fuller contractarian account, one would need to specify further principles of selection that would allow the parties to choose from among the various eligible regimes. The various and more specific contractarian principles offered in other accounts might serve in this regard.

Does the institution of punishment pass the benefit test? There is reason to suppose that it does, and indeed, that the possibility of punishment is quite essential to a social order predicated on voluntary agreement. Members of a social contract must have some way of ensuring continued compliance with the terms of the agreement, given the temptation members will have to offer their initial consent and then free-ride on the compliance of others while silently defecting. Any voluntary agreement must therefore set out consequences for violators, along with a plausible enforcement mechanism for detecting violations and imposing the announced penalties. Thus a system of punishment will be part and parcel of the agreement that sets out substantive rules of compliance.

Let us now apply the benefit principle to the contract establishing the basic principles of punishment. Straightforwardly applied, the benefit principle requires that each member of society regard himself as faring better, under an institution that mandates punishment, than he would fare in the absence of such an institution. Thus each member of society must project himself into the position of someone who has violated the conditions of the more basic, substantive social contract, and ask himself whether, if he were to be punished for such violations, he would still fare better than he would had he never agreed to live

under threat of punishment in the first place. For many sanctions the benefit test will be satisfied. A complete absence of any form of punishment for violations of the social contract would eliminate the possibility of social cooperation entirely, since an agreement would unravel without the threat of enforcement. And as Hobbes so vividly describes in Chapter XIII of *Leviathan*, life for most people would be calamitous in the absence of all society, surely worse than it would be to live with the benefits of society for most of one's life, and suffer some period of incarceration or other penalty later. Thus for most penalties, and most societies, even an offender who must suffer punitive sanctions will fare better under a punishment agreement than he would in the absence of all social enforcement.

Does this hold true for the worst violators, those who must suffer the worst penalties? Can a person who receives the death penalty or life in prison regard himself as better off under the terms of the penalty contract than he would have been had he never agreed to the contract in the first place? Certainly if Hobbes is to be believed, life in the absence of all social cooperation would be so brutal and insecure that no one could expect to live into old age. As compared with "continual fear, and danger of violent death" (Hobbes 1994: ch. XIII [9]), it is possible that a person receiving a very severe sentence like death or life in prison without parole would regard himself as benefited as compared with his life in the absence of such penalties. Whether this is so would depend on the marginal deterrent benefits of those penalties relative to more moderate penalties. It would also depend on a host of other factors, such as when in his life we conceive of the offender as receiving the penalty. A person who had had many years to reap the deterrence benefits of those penalties would be in a different position from a very young offender who had not, and who now could reap no further benefits from such rules if put to death or imprisoned for the rest of his life. It should also be noted, however, that the death penalty and life in prison without parole are likely to fare somewhat differently under the benefit test. If the death penalty has only very modest additional deterrent efficacy over life in prison without parole, it is unlikely to be incorporated into the punishment agreement, as its detriment for the person suffering it is vastly greater than the nearest available alternative penalties.

A theory of punishment governed by the benefit principle has important advantages over both deterrence theories and retributivism. On the one hand, the contractarian approach solves the two problems associated with deterrence theories, namely the problem of torture and the problem of responsibility. With regard to torture and other severe penalties, the contractarian theory has a basis for rejecting extreme penalties, since these would normally fail the benefit test. And a contractarian theory organized around the benefit principle has no difficulty reconciling principles of responsibility with the goal of deterrence. Although deterrence is the reason for adopting an institution of punishment in the first place, no institution that inflicted punishment in the absence of conditions of responsibility would pass the benefit test. For a society that left individuals subject to "punishment" at random would be no better, and possibly worse, than a world in the absence of society. In a regime of terror, human beings are left just as defenseless as they are in their natural state, but matters are worse, since now they must protect themselves not just against lone individuals, but against an organized state. If the institution of punishment is to leave members of society better off than they would be in its absence, sanctions must be allocated predictably, fairly, and according to principles of control and individual responsibility.

On the other hand, the contractarian theory of punishment, as I have articulated it, would also have advantages over retributivism. Recall that the central problem of that account was its inability to justify particular penalties. The benefit principle gives us a way of justifying penalties with specificity, at the same time that we are able to preserve an intrinsic connection between the crime and the penalty. In particular, the specificity is provided by the aim of deterrence, in combination with the limitation the benefit principle provides. Let us see more specifically how this works.

Consider how the aim of deterrence, in combination with the benefit principle, would identify the appropriate punishment for a crime like burglary. The norms protected by a prohibition

on burglary are norms of private ownership, and in the absence of any punishment for burglary (and like crimes) private ownership would be eliminated. Thus each person can ask himself: would I be better off under the terms of a contract that established penalties for burglary, assuming that I myself may end up subject to that penalty, than I would be if there were no private ownership at all? Notice that if the penalties for burglary are too low, the deterrent effect will be insignificant, and private property will not be protected. If the penalties are too high, however, agents receiving the penalty would be worse off than they would have been in the absence of private property, and the benefit principle would not be satisfied. Thus when we combine the benefit principle with the goal of deterrence, we are able to develop specific parameters for the punishment of each separate crime.

Notice furthermore that this theory also captures the greatest strength of the retributive principle in that it establishes something like a moral equivalence between crime and punishment. It does so because applying the benefit principle will require that we consider the importance of the underlying norm we are trying to protect, and compare it with the suffering the offender would experience under a given penalty. In the burglary example we implicitly compared the gravity of a violation of rights of ownership with the loss in welfare an individual would suffer who undergoes a term of imprisonment for that violation when we asked whether a rational agent would be better off suffering a given punishment for burglary than he would be abandoning protection for private property altogether. Since the importance of the underlying institution we are trying to protect establishes the gravity of the violation for which we are punishing, the benefit principle creates a metric whereby we can match offenses with appropriate penalties. But it is able to match crime and punishment without sacrificing the importance of deterrence as a guiding aim of a system of punishment. The complete rejection of deterrence as a legitimate aim of punishment is what dooms retributive theories to generality, since the notion of desert substituted in its place is ineliminably nonspecific.

One question that might arise, in view of the role the contractarian account assigns to deterrence, is why that account would constitute an improvement over the deterrence accounts we saw above with regard to the "traveling across persons" objection. The answer is that the consensual nature of punishment in this scheme defeats the concern with traveling across persons. Unlike in the deterrence accounts we saw above, each party to the social contract agrees that he will submit himself to punishment in the event that he would violate the conditions of the social contract. It is this self-imposed threat that he offers to his fellows as his assurance that he will not defect. And the willingness of each to subject himself to punishment should he choose to defect is the condition each party to the contract requires for his own compliance. The calculation of the required level of deterrence is a function of the threat necessary to induce compliance and to provide the assurance of compliance necessary for the agreement to be rationally entered into in the first place. The punishment itself is legitimate to inflict, not because it deters others, but because it has already been consented to by the offender himself. Thus the appeal to deterrence made by the contractarian theory of punishment does not travel across persons, since the deterrence is supposed to operate on the offender himself at the moment he enters into the original social contract. As in the alternative deterrence account we considered briefly at the end of the second section of this chapter, the punishment itself is only the follow-through on the threat made to the offender himself. But unlike in that account, there is here an independent justification for following through on the threat, namely that the offender consented to this scheme, thinking he would benefit himself thereby.

A final concern about the proposed account might be raised. Why should we care about whether the offender himself is benefited under the punishment scheme, since he has arguably chosen to place himself outside of the terms of the social contract anyway by violating social norms? Why not treat the offender as having exempted himself from society's protection, and as having entitled other members of society to discount his benefit altogether? This would be the usual approach to punishment in the contractarian tradition (see Morris 1991). That tradition treats violators of the social contract as

permanently expelled from the contractual relationship that holds among members of society. The tradition thus denies that punishment is governed by the terms of the contract itself, and treats it as governed by norms that lie outside the contract. And from a certain perspective, this is quite a defensible approach. If society is a "cooperative venture for mutual advantage," it makes sense to think of criminals as having placed themselves outside the scope of all voluntary arrangements, since cooperating with *them* would not be to the advantage of members of society who are faithful to the terms of the agreement.

But I think such a view is to be rejected. For while it is true that the initial contract is made only among those who accept the conditions of cooperation, cooperators can become defectors after the basic contract has been entered into. It would be wrong to treat defection as though it were noncooperation at the outset. There are several reasons for this. First, defections can be large or small, and it may be that it is still advantageous to cooperate with those responsible for small defections. Second, it is not possible to address the problem of noncooperation at the outset in any way other than refusing to contract. But defectors are themselves subject to the terms of an antecedent agreement, and can therefore be dealt with contractually. Finally, it simply seems wrong to think of a defector as beyond the bounds of all social interaction, someone who deserves none of the protections or entitlements that those who enter into rational relations with others receive. Even the most heinous violations ought not to deprive their perpetrators of basic dignitary rights, such as the right to be free from torture, the right to speak in one's own defense, and the right to minimal bodily dignity and comfort. It is true that nonrational creatures are often thought of as possessing at least some subset of these same rights, and thus there may be a basis for affording protection to biological creatures outside the contractual context. But the protections afforded such creatures are thought to be significantly less than those afforded even the worst criminals. The higher protections afforded to rational life, even the least deserving rational life, is most plausibly explained as a product of at least a putative exchange of human wills. For these and other reasons, the conditions under which human beings may permissibly inflict sanctions for noncooperation on members of their own kind should be thought of as governed by an antecedent agreement such humans make to enforce the terms of cooperative interaction.

It is in fact only by including potential violators in the terms of the social contract that the contractarian model can provide any practical guidance to a theory of punishment. And it is also in this way that we are able to capture within a contractarian framework the basic deontological intuitions that may have made retributivism seem initially attractive. As we have seen, these deontological intuitions are insufficient in and of themselves to produce a theory of punishment directly. It is only when combined with the aim of deterrence that they find their proper place. Normally the aim of deterrence and intuitions concerning desert cannot coexist in a theory of punishment.[12] These opposing elements complement without contradiction in the consensual approach I have proposed.[13]

Notes

1 I do not mean to suggest that every infliction of physical suffering or deprivation of liberty is worse than every order to pay compensation, but simply that as a general matter, bodily invasions are more morally suspect than financial ones.

2 For a clear statement of an expressive approach to punishment, see Feinberg (1970). For an interesting argument for a communicative approach, see Duff (2003). See also Finkelstein (2004) commenting on Duff.

3 There is some irony in this: one might suppose that if deontologists cared about rights violations, they would care about minimizing the number of rights violations in the world, and that therefore some trade-offs of the sort deterrence theorists contemplate would be permissible. But someone committed to deontological principles could not take that position without effectively abandoning the idea that there are restrictions on what human beings may do to one another in the name of utility, restrictions that cannot be traded off against other sorts of reasons. For a helpful discussion of this aspect of deontological morality, see Kamm (1993, 1996).

4 I am only assuming, for the sake of argument, that the death penalty is morally acceptable. I do not in any sense mean to be endorsing that conclusion.

5 The exception to this occurs in cases in which some special doctrine of the criminal law connects one agent with the free, voluntary acts of another. Felony murder, vicarious liability, and accomplice liability are examples.

6 There are admittedly some exceptions. It is often thought to be permissible to *redirect* a harm that threatens one group of people towards others who are fewer in number, despite the fact that the latter are not in any way the source of the threat. See Thomson (1985). It is also sometimes thought permissible to inflict a slight harm on one innocent (noninvolved) person in order to prevent a *dramatically* greater harm to another or to some vastly larger number of persons. See Moore (1997), defending what he calls "threshold deontology." But presumably neither of these exceptions would apply in this case. On the one hand, activating the electric chair would be initiating a new harm, and on the other, the hostage pressing the button would be saving his own life and the lives of only seven other hostages, which most threshold deontologists would not consider sufficient to justify killing the one.

7 This is a very rough and ready characterization. For one thing, it surely is not the case that every killing that is not a using is permissible. For another, there are possibly other principles at work here that may better capture the distinction we are seeking, such as the doctrine of double effect. It is beyond the scope of this chapter, however, to explore such alternative principles. I point to the prohibition on using simply to sketch, at the grossest level of generality, a standard contrast between utilitarian and deontological approaches.

8 David Gauthier has recently suggested such an account to me. It is also a version of Warren Quinn's approach in "The Right to Threaten and the Right to Punish" (Quinn 1985).

9 As I said at the outset, it is beyond the scope of this chapter to consider every possible theory of punishment. There is one I have thus far ignored, however, that may seem a particularly important omission in the context of our discussion of deterrence theories, namely a mixed theory of the sort that Rawls argued for in "Two Concepts of Rules" (Rawls 1955). According to a mixed theory of this sort, the rationale for having an institution of punishment in the first place is utilitarian, while the specific form the rules of such an institution take are themselves desert-based. Someone might argue that this is a form of deterrence that does not involve traveling across persons, since the reason for punishing any particular offender is that he deserves to be punished. The effect of his treatment on other, potential offenders is not particularly a reason for punishing *him*. It is simply part of the background conditions for having an institution of this sort in the first place. But it seems likely that a mixed account of this sort will still suffer from the same problems as the more generic mixed deterrence account we have considered. For the justification for the institution itself travels across persons. Whether this is objectionable would require further exploration, however. In addition, such an account will likely suffer from the difficulties with retributivist accounts, which I detail below.

10 I say "much" rather than "all" because I believe that the benefit principle is compatible with *some* risky agreements or plans. The reason is that there are conditions under which agents are benefited by losing gambles: they can sometimes receive a net benefit from the chance of benefit the gamble supplied. As long as the actual losses are not very great, and the *ex ante* chance of benefit is sufficiently large, it is possible for the *ex ante* chance of benefit to supply a net benefit, even in the face of losing gambles. See Finkelstein (2003).

11 I leave to one side here the question whether basic social institutions like punishment can be adequately justified if the only benefit they produce for a given individual is the benefit that individual received from exposure to a chance of benefit.

12 Some notable exceptions are Hart's approach in *Punishment and Responsibility* (1968), and Rawls' approach in "Two Concepts of Rules" (1955).

13 I wish to thank Michael Davis, Bill Edmundson, David Gauthier, Leo Katz, and Connie Rosati for comments on various drafts of this article or for conversations and advice on the issues it raises.

References

Davis, Michael. 1983. How to make the punishment fit the crime. *Ethics* 93: 726–52.

Duff, R. A. 2003. Penance, punishment and the limits of community. *Punishment and Society* 5: 295–312.

Feinberg, Joel. 1970. The expressive function of punishment. In Joel Feinberg, *Doing and Deserving*. Princeton, NJ: Princeton University Press, 95–118.

Finkelstein, Claire. 2003. Is risk a harm? *University of Pennsylvania Law Review* 151: 963–1001.

Finkelstein, Claire. 2004. Comments on Anthony Duff's "Penance, punishment, and the limits of community." *Punishment and Society* 6: 99–104.

Hart, H. L. A. 1968. *Punishment and Responsibility*. Oxford: Oxford University Press.

Hobbes, Thomas. 1994. *Leviathan*, ed. Edwin Curley. Indianapolis, IN: Hackett.

Kamm, Frances Myrna. 1993, 1996. *Morality, Mortality*, 2 vols. Oxford: Oxford University Press.

Locke, John. 1960. *Two Treatises of Government*, ed. Peter Laslett. Cambridge, UK: Cambridge University Press.

Moore, Michael. 1997. *Placing Blame: A General Theory of the Criminal Law*. Oxford: Oxford University Press.

Morris, Christopher W. 1991. Punishment and loss of moral standing. *Canadian Journal of Philosophy* 21(1): 53–80.

Quinn, Warren. 1985. The right to threaten and the right to punish. *Philosophy and Public Affairs* 14(4): 327–73.

Rawls, John. 1955. Two concepts of rules. *The Philosophical Review* 64: 3–32.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Thomson, Judith Jarvis. 1985. The trolley problem. *Yale Law Journal* 94: 1395–1415.

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