ARTICLES

A CONTRACTARIAN ARGUMENT AGAINST THE DEATH PENALTY

CLAIRE FINKELSTEIN*

Opponents of the death penalty typically base their opposition on contingent features of its administration, arguing that the death penalty is applied discriminatorily, that the innocent are sometimes executed, or that there is insufficient evidence of the death penalty's deterrent efficacy. Implicit in these arguments is the suggestion that if these contingencies did not obtain, serious moral objections to the death penalty would be misplaced. In this Article, Professor Finkelstein argues that there are grounds for opposing the death penalty even in the absence of such contingent factors. She proceeds by arguing that neither of the two prevailing theories of punishment—deterrence and retributivism—is capable of justifying the death penalty. More generally, she suggests that while each theory captures an important part of the justification for punishment, each must appeal to some further limiting principle to accommodate common intuitions about appropriate punishments for crimes. Professor Finkelstein claims that contractarianism supplies this additional principle, by requiring that individuals consent to the system of punishment under whose threat they must live. Moreover, on the version of contractarianism for which she argues, they must do so based on a belief that they will benefit under the terms of that system as compared with how they would fare in its absence. While the notion of benefit is often best understood in terms of maximizing one's expected utility, Professor Finkelstein argues that with respect to choices about the basic structure of society, rational contractors will conceive of benefit in terms of a conservative, "no-gambling" decision rule. She then argues that rational contractors applying this conception of benefit would reject any system of punishment that includes the death penalty. For while contractors would recognize the death penalty's deterrent value, they must also consider the high cost they would pay in the event they end up subject to such a penalty. This Article presents both a significant new approach to the death penalty and a general theory of punishment, one that incorporates the central intuitions about deterrence and desert that have made competing theories of punishment seem compelling.

 Ph.D, University of Pittsburgh; J.D., Yale Law School; B.A., Harvard University. I am grateful to Larry Alexander, Russell Christopher, Michael Davis, David Gauthier, Leo Katz, Sharon Lloyd, Michael Ridge, Geoff Sayre-McCord, Thomas Pogge, Connie Rosati, and Seana Shiffrin for their comments on drafts at various stages of completion, as well as to audiences at the University of Pennsylvania faculty retreat, the Rutgers Law School Faculty Workshop, the American Philosophical Association Eastern Division meeting, the University of Pennsylvania conference on Contract, Consent, and the Law, and participants in the Florida State Legal Theory Workshop. I am also grateful to Matthew Mills for assistance with research. I also wish to thank the editors of the *New York University Law Review* for their truly exceptional editorial assistance.

INTRODUCTION

Opponents of the death penalty fall into two groups. The first group does not oppose the death penalty per se. It opposes the death penalty either because it sees its use as connected with other objectionable practices or because of particular features of the circumstances in which the death penalty is used. Members of this first group argue, for example, that the death penalty cannot be nondiscriminatorily administered in a country rife with background racial discrimination,[1] that it cannot be fairly and effectively administered when used as sparingly as it is usually used,[2] that having a death penalty creates too great a gulf between the United States and other democratic nations,[3] or that there is insufficient evidence that the death penalty has greater deterrent value than life in prison without parole.[4] We can characterize this first group's opposition to the death penalty as *contingent* in nature.

The second group's opposition to the death penalty runs deeper. Members of this group believe that the death penalty is morally impermissible, regardless of how evenhanded its administration or beneficial its consequences. For such opponents, no set of empirical

---

[1] *See* Callins v. Collins, 510 U.S. 1141, 1145–46 (1994) (Blackmun, J., dissenting from denial of certiorari) (arguing death penalty is unconstitutional due to irreparable racial disparities in application); Furman v. Georgia, 408 U.S. 238, 364–66 (1972) (Marshall, J., concurring in judgment) ("It is immediately apparent that Negroes [are] executed far more often than whites in proportion to their percentage of the population."); Randall L. Kennedy, McCleskey v. Kemp*: Race, Capital Punishment, and the Supreme Court*, 101 HARV. L. REV. 1388, 1425 (1988) ("[T]he legal authorities in Georgia [discriminate against African-Americans] when they repeatedly sentence killers of blacks less harshly than killers of whites for approximately similar crimes.").

[2] *See* Godfrey v. Georgia, 446 U.S. 420, 442 (1980) (Marshall, J., concurring) ("[T]he effort to eliminate arbitrariness in the infliction of that ultimate sanction is so plainly doomed to failure that it—and the death penalty—must be abandoned altogether."); *Furman*, 408 U.S. at 291–95 (Brennan, J., concurring in judgment) (comparing current death penalty system to lottery and arguing that it is inflicted arbitrarily); *id.* at 309–10 (Stewart, J., concurring in judgment) (arguing death penalty is unconstitutional because it is capriciously imposed); Jack Greenberg, *Against the American System of Capital Punishment*, 99 HARV. L. REV. 1670, 1675 (1986) ("We have a system of capital punishment that results in infrequent, random, and erratic executions . . . .").

[3] *See* Stephen B. Bright, *Will the Death Penalty Remain Alive in the Twenty-First Century?: International Norms, Discrimination, Arbitrariness, and the Risk of Executing the Innocent*, 2001 WIS. L. REV. 1, 2–4 (discussing international trend toward abolition of death penalty).

[4] *See* Michael L. Radelet & Ronald L. Akers, *Deterrence and the Death Penalty: The Views of the Experts*, 87 J. CRIM. L. & CRIMINOLOGY 1, 10 (1996) ("[T]he death penalty does, and can do, little to reduce rates of criminal violence."). Some studies purport to find that the death penalty has a "brutalizing" effect, increasing the amount of violent crime. *See, e.g.*, William J. Bowers & Glenn L. Pierce, *Deterrence or Brutalization: What Is the Effect of Executions?*, 26 CRIME & DELINQ. 453, 481–84 (1980) (finding increase in homicides in New York State in months following executions).

circumstances can cure the moral impermissibility of the death penalty. We can characterize this group's opposition to the death penalty as *categorical*, rather than contingent.[5]

Public opposition to the death penalty in the United States tends to take the first, rather than the second, form. This is not surprising, since American courts have been more receptive to contingent challenges to the legality of the death penalty than to categorical ones. The Supreme Court, for example, has consistently rejected the claim that the death penalty is per se cruel and unusual punishment and therefore banned by the Eighth Amendment.[6] The only successful attacks have been sharply circumscribed and carefully tied to specific forms of its administration or particular circumstances in which it might be used. Thus, the Court has found the death penalty unconstitutional if assigned as a mandatory penalty for particular crimes,[7] if a jury can impose it in an entirely discretionary way,[8] if used against juveniles,[9] if imposed on a relatively uninvolved coconspirator,[10] or if used for any crime other than murder.[11] Even at the height of judicial skepticism, then, the death penalty was presumed to be constitutional as long as certain restrictions on its administration were in place.[12]

---

[5] The American constitutional tradition comes closest to articulating a non-contingent abolitionist stance in the opinions of former Supreme Court Justices Marshall and Brennan. *See, e.g.*, cases cited *supra* notes 1–2; *see also Furman*, 408 U.S. at 290 (Brennan, J., concurring) ("The calculated killing of a human being by the state involves, by its very nature, a denial of the executed person's humanity."). In *Gregg v. Georgia*, Justice Brennan stated:

> [F]oremost among the "moral concepts" recognized in our cases and inherent in the [Cruel and Unusual Punishments] Clause is the primary moral principle that the State, even as it punishes, must treat its citizens in a manner consistent with their intrinsic worth as human beings—a punishment must not be so severe as to be degrading to human dignity.

428 U.S. 153, 229 (1976).

[6] *E.g.*, *Gregg*, 428 U.S. at 169.

[7] Woodson v. North Carolina, 428 U.S. 280, 304 (1976) (holding mandatory death penalty without admission of individualized evidence at sentencing unconstitutional).

[8] *Furman*, 408 U.S. at 239–40, 256–57 (Douglas, J., concurring in judgment) (arguing statutes authorizing death penalty based on unguided discretion are unconstitutional).

[9] Roper v. Simmons, 543 U.S. 551, 574–75 (2005) (affirming decision finding death penalty unconstitutional under Eighth Amendment for offenders under eighteen years old).

[10] Enmund v. Florida, 458 U.S. 782, 797 (1982) (finding death penalty unconstitutional if inflicted on relatively uninvolved accomplice).

[11] Coker v. Georgia, 433 U.S. 584, 592 (1977) (holding death penalty for rape unconstitutional).

[12] *See Furman*, 408 U.S. at 257 (Douglas, J., concurring in judgment) (withholding judgment whether equally applied death penalty violates Eighth Amendment); *id.* at 306 (Stewart, J., concurring in judgment) ("[T]wo of my Brothers have concluded that the infliction of the death penalty is constitutionally impermissible in all circumstances under the Eighth and Fourteenth Amendments. Their case is a strong one. But I find it unneces-

The purpose of this Article is to provide a philosophical foundation for the second, categorical form of opposition to the death penalty. Two disclaimers are in order. First, while the argument I offer is intended to provide a foundation for categorical opposition to the death penalty, it is not itself categorical. As will be clear, the argument I present is significantly less dependent on empirical factors than the contingent arguments mentioned above, but it does rely on one empirical assumption—that the death penalty has at most a moderately strong deterrent effect. In my view, the argument comes as close to being categorical as possible, by showing that the death penalty is unjustifiable under any circumstances that one might reasonably expect to encounter. While I offer this argument primarily for its normative implications, I also intend it as a means of reconstructing or explaining the intuitions of those who think the death penalty is impermissible under all circumstances. That intuition, I suggest, is practically, but not perfectly, categorical. Thus the residual empirical dependence of my argument will not undercut it.

The second disclaimer is that I do not mean this argument to diminish the importance of the contingent reasons the first group offers as providing a basis for rejecting the death penalty. In general, those arguments are good ones. But they are weak in that they leave death penalty opponents open to the suggestion that our primary legal efforts should be directed toward the elimination of such factors. I doubt, however, that many death penalty opponents would abandon their opposition if the various contingent factors they cite as the grounds for opposition were removed.

To insulate my argument from the vast bulk of contingent factors, I make four assumptions. First, I assume the death penalty can be administered without distortion from racial or other kinds of invidious bias. While this assumption has been called into question with respect to the current American legal system,[13] it is not difficult to imagine a state of affairs in which this is so. Second, I assume the death penalty has a non-negligible deterrent effect. Once again, the evidence on

---

sary to reach the ultimate question they would decide."); *id.* at 310–11 (White, J., concurring in judgment) (declining to hold death penalty per se unconstitutional).

[13] *See* McCleskey v. Kemp, 481 U.S. 279, 286–87 (1987) (citing David C. Baldus et al., *Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience*, 74 J. CRIM. L. & CRIMINOLOGY 661 (1983) (demonstrating bias correlated to race of victim in administration of death penalty)); David C. Baldus & George Woodworth, *Race Discrimination in the Administration of the Death Penalty: An Overview of the Empirical Evidence with Special Emphasis on the Post-1990 Research*, 39 CRIM. L. BULL. 194, 202 (2003) (reviewing multiple post-1990 studies and finding widespread race-of-victim discrimination and some indication of race-of-defendant discrimination).

whether the death penalty deters is arguably still indeterminate,[14] but it is not implausible to suppose that the threat of death could deter under certain conditions. Third, I assume that the criminal justice system can be made highly reliable, such that no innocent person would ever be executed.[15] This assumption is perhaps the most improbable of the three, as the possibility of mistaken convictions appears to be an unavoidable feature of any criminal justice system. Nonetheless, I include it in the list of contingent factors from which I wish to abstract because I do not want the strength of the argument against the death penalty to depend on the number of innocent lives at risk.

My fourth assumption is of a different sort, as it is philosophical, rather than empirical. Moreover, while the first three assumptions raise the bar for a successful argument against the death penalty, the fourth assumption makes it easier to argue against the death penalty. The assumption is that punishment is painful or otherwise highly unpleasant, and, for this reason, its imposition is undesirable to anyone subject to it.[16] We can therefore infer that punishment stands in need of an affirmative justification before it can be permissibly applied. The burden is therefore on the death penalty proponent to justify its use. If I am right to help myself to this final assumption, an argument against the death penalty can be won simply by combating arguments in its favor. This is how I will proceed initially, although I

---

[14] Sunstein and Vermeule point to recent studies suggesting a strong deterrent effect from executions, such as that each execution may deter eighteen murders. Cass R. Sunstein & Adrian Vermeule, *Is Capital Punishment Morally Required? Acts, Omissions, and Life-Life Tradeoffs*, 58 STAN. L. REV. 703, 711 (2005). Donohue and Wolfers, however, argue convincingly that these studies are unreliable, in particular because the number of executions on which these conclusions are based is small relative to the number of murders each year; even if a deterrent effect were present, it would be impossible to detect. *See* John J. Donohue & Justin Wolfers, *Uses and Abuses of Empirical Evidence in the Death Penalty Debate*, 58 STAN. L. REV. 791, 794 (2005) ("[T]he death penalty . . . is applied so rarely that the number of homicides it can plausibly have caused or deterred cannot be reliably disentangled from the large year-to-year changes in the homicide rate caused by other factors.").

[15] See *infra* note 89 for a brief discussion of the implications for my position of relaxing this assumption and allowing for the possibility that at least one innocent person would be executed under a death penalty system.

[16] This is not to say that there have never been criminals who desired their punishment, either because they did not find it painful or because they thought they deserved to suffer. But such cases must be exceedingly rare. *See* H.L.A. HART, PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 4 (1968) ("[Punishment] must involve pain or other consequences normally considered unpleasant."); Louis P. Pojman, *For the Death Penalty*, *in* LOUIS P. POJMAN & JEFFREY REIMAN, THE DEATH PENALTY: FOR AND AGAINST 1, 5 (1998) (defining punishment as "an evil inflicted by a person in a position of authority upon another person who is judged to have violated a rule") (emphasis omitted).

will provide more affirmative reasons for rejecting the death penalty later in this Article.

In what follows, I consider the two principal justifications proponents offer for the death penalty: deterrence and retribution. I address the argument from deterrence in Part I, where I consider whether the fact that the death penalty saves innocent lives, if true, would be sufficient to establish its moral acceptability. I argue that it would not be. I then consider whether the lives saved by the death penalty's deterrent power would be sufficient to make the death penalty morally acceptable if the aim of deterrence were supplemented with a retributivist limiting condition. I conclude that deterrence fails, both as a stand-alone justification and as a primary element in a mixed theory. In Part II, I consider whether the death penalty can be justified on retributive grounds alone. I conclude in this Part that even if retribution turned out to be a compelling justification for other punishments, it is not a sufficient rationale for the death penalty.

The deterrent benefits and the moral appropriateness of punishment both seem important in justifying punishment, yet both justifications lead to objectionable consequences if taken alone. From our discussion in Parts I and II, then, I conclude that some other principle must be at work—one that imposes appropriate constraints on deterrence and desert. In Part III, I suggest that this principle is a contractarian one and argue that the most compelling justification for a system of punishment is the fact that individuals settling on a basic structure for society according to principles of mutual advantage would consent to that system as a way of policing the basic terms of their agreement. Such a contractarian approach should make it easier to meet the burden of proof required to justify punishment in general, given that it would make punishment voluntarily imposed. As I will show, the correct measure of justified punishment is determined by weighing the benefit a rational agent would receive from such punishment against the harm the agent would suffer were he to be subject to that punishment. Applying this test, I then argue that individuals setting up a contractarian system of punishment would not regard the death penalty as beneficial, on balance, and hence would not select it. I address objections to this argument in Part IV and then offer some concluding remarks.

I
THE ARGUMENT FROM DETERRENCE

A. *The Pure Deterrence Theorist*

The death penalty proponent's appeal to deterrence usually goes something like this: Suppose each execution deterred eight future murders. That is surely sufficient reason to impose the death penalty, since rejecting the death penalty under such circumstances would imply that we value a criminal's life at least eight times more than the life of each innocent person whose death could be prevented. As long as we value innocent life at least as much as guilty life, a demonstrated deterrent effect is arguably sufficient to justify use of the death penalty.[17]

Death penalty proponents do not often make this argument explicitly, but it is presupposed by much of what they say. Indeed, it seems presupposed even by those death penalty opponents who rely primarily on the ground that studies regarding deterrence have been inconclusive. Jeffrey Reiman, for example, opposes the death penalty, yet allows that if the death penalty were shown to have significant deterrent effects, he might not be able to stand by his opposition to it:

> [W]ere the death penalty clearly proven a better deterrent to the murder of innocent people than life in prison, we might have to admit that we had not yet reached a level of civilization at which we could protect ourselves without imposing this horrible fate on murderers, and thus we might have to grant the necessity of instituting the death penalty.[18]

While he couches the point tentatively, Reiman makes clear in the margins that whether he would be prepared to concede the necessity of instituting the death penalty depends entirely on the degree to which it deters.[19]

The debate with the death penalty proponent thus degenerates into an argument over which side should bear the burden of proof if

---

[17] *See* Pojman, *supra* note 16, at 41 ("Even if we value the utility of an innocent life only slightly more than that of a murderer, it is still rational to execute convicted murderers."); Sunstein & Vermeule, *supra* note 14, at 705 ("[O]n certain empirical assumptions, capital punishment may be morally required . . . to prevent the taking of innocent lives."); Ernest van den Haag, *The Death Penalty Once More*, *in* THE DEATH PENALTY IN AMERICA: CURRENT CONTROVERSIES 445, 450 (Hugo Adam Bedau ed., 1997) ("I should favor the death penalty for murderers, if probably deterrent, or even just possibly deterrent.").

[18] Jeffrey H. Reiman, *Justice, Civilization, and the Death Penalty: Answering van den Haag*, 14 PHIL. & PUB. AFF. 115, 142 (1985).

[19] *Id.* at 142 n.33 ("I say 'might' here to avoid the sticky question of just how effective a deterrent the death penalty would have to be to justify overcoming our scruples about executing.").

the death penalty's deterrent effect is either negligible or inconclusively demonstrable.[20] Underlying that dispute is a shared assumption that significant deterrent effects would be a conclusive consideration in favor of the death penalty. Capitalizing on this apparent agreement, the authors of a recent article argue that if the death penalty has significant deterrent value, governments would be *obligated*, and not merely permitted, to impose the death penalty in order to save the lives of future murder victims.[21]

Deterrence alone, however, does not provide a moral justification for the death penalty. There are at least two reasons for this. The first we might call the "problem of torture." Suppose it turned out that torturing a person before executing him had greater deterrent efficacy than execution alone. Suppose, for instance, it saved eight additional lives over and above the eight that would already be saved by the execution. Is the death penalty proponent prepared to endorse torture in this case? Presumably not. The deterrence theorist, like everyone else, will want restrictions on what it is permissible to do to a person, even if the deterrence rationale alone does not itself imply those restrictions. If he accepts such restrictions in the case of torture, however, he cannot rule out the possibility that these same restrictions also make the death penalty impermissible. Later I will suggest that the relevant restrictions are best understood as having a contractarian source. For the moment, it suffices to notice that even the deterrence theorist will want some such restrictions on the applicability of his preferred rationale for punishment.[22]

---

[20] Michael Davis, *The Death Penalty, Civilization, and Inhumaneness*, 16 SOC. THEORY & PRAC. 245, 248 (1990) (acknowledging that scholars disagree about who must carry burden but "not about whether it can be carried").

[21] *See* Sunstein & Vermeule, *supra* note 14, at 705 ("We suggest . . . that on certain empirical assumptions, capital punishment may be morally required, not for retributive reasons, but rather to prevent the taking of innocent lives."). Indeed, Sunstein and Vermeule may have been implicitly responding to the 2002 version of the present article that appeared under a different title on SSRN. *See* Claire Finkelstein, *An A Priori Argument Against the Death Penalty* 15 (Univ. Pa. Law Sch., Research Paper No. 15, 2002), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=347323. I do not address Sunstein and Vermeule's particular argument here, since, if I am correct that the death penalty is not rendered permissible by the mere fact that it deters, deterrence could not render it *obligatory* either. For a more specific presentation of the philosophical problems with Sunstein and Vermeule's argument, see Carol S. Steiker, *No, Capital Punishment Is Not Morally Required: Deterrence, Deontology, and the Death Penalty*, 58 STAN. L. REV. 751 (2005).

[22] The death penalty proponent may be prepared to bite the bullet and concede that torture is permissible under such circumstances. But he would then have the unenviable task of explaining why we should be guided by a moral theory that has no apparent traction with intuitions of near universal appeal. *See* Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, *adopted on* Dec. 10, 1984, S. TREATY DOC. NO. 100-20, 1465 U.N.T.S. 85, *available at* http://www.ohchr.org/english/law/

The second reason we might call the "problem of impermissible tradeoffs." Even if the deterrence theorist could distinguish torture from death, he would face a host of familiar objections to the suggestion that it is permissible to kill a smaller number of people to save a greater number. Killing the few to save the many is considered off-limits, because it is using a person as a means to benefit another, rather than treating the person as an end in himself. It is not, for example, permissible to remove the organs of one healthy person to save the lives of eight others who need organ transplants. And if this is so, we cannot accept executing a criminal for the sake of saving eight innocent lives.[23] Even if one is inclined to reject the Kantian formulation of the objection, most philosophers endorse the conclusion that it is not ordinarily permissible to kill one to save several or even eight others. Since we do not think it permissible to kill one person to save eight others, we cannot think general deterrence a sufficient moral justification for the death penalty.[24]

The intuitive resistance to treating rational agents as means descends from more general moral constraints on the idea of maximization. Such constraints are usually explained by saying that rational agents have rights. The existence of rights is supposed to explain the prohibition on using one person to maximize the welfare of some larger number of persons.[25] The existence of rights also explains why it is not permissible to violate even one person's rights to minimize a larger number of rights violations to other people. One might wonder why this is so, since if deontologists care about rights, then surely they would prefer a state of affairs in which fewer rights were violated to one in which more were. But deontologists cannot

---

pdf/cat.pdf. *See also* Office of the United Nations High Comm'r for Human Rights, Convention Against Torture, http://www.ohchr.org/english/countries/ratification/9.htm (last visited Aug. 13, 2006) (stating that there are 141 parties to Convention as of May 8, 2006). It bears noting that we are here discussing the use of torture as a form of *punishment*. Matters are arguably different where the use of torture as a pure preventive device is at issue.

23 For a general discussion of utilitarian and deontological theories of punishment, see TOM SORELL, MORAL THEORY AND CAPITAL PUNISHMENT 62–77 (1987).

24 There are cases in which the restriction on "using" seems weaker than in the case above, as, for example, when the person whose death is contemplated is among the eight who would die if he is not killed. But many would still see the killing as impermissible in such cases, since the one is still being used to save the others (and *he* is not benefiting from being sacrificed). As Bernard Williams once famously argued, we would have difficulty condemning the person who refused to kill under such circumstances, even if we thought it permissible to kill for the good of others. *See* Bernard Williams, *A Critique of Utilitarianism*, *in* J.J.C. SMART & BERNARD WILLIAMS, UTILITARIANISM: FOR AND AGAINST 75, 98–99 (1973). A person is not *obligated* to advance the greater good rather than protect his own moral integrity.

25 *See* JUDITH JARVIS THOMSON, THE REALM OF RIGHTS 168 (1990) (arguing that one's right not to be killed is "maximally stringent," even as against claims of significant benefit).

endorse maximizing anything, even respect for rights, without abandoning the restrictions on maximization that are the hallmark of nonconsequentialist views.[26] And so once again, if saving a greater number is insufficient justification for killing one individual, then deterring future murders cannot justify the penalty of death.

One might, however, reject the above argument as follows: In the standard cases in moral philosophy we are considering, it is an *innocent* person who is being killed to save the greater number. In the case of the death penalty, however, it is a *guilty* person. And so, the argument might run, the appeal to deterrence is not subject to the complaint about impermissible tradeoffs, because the people being used to deter others are already deserving of punishment. Of course, if the deterrence theorist really thinks deterrent benefits alone carry enough moral weight to justify the death penalty, this restriction would be ad hoc. If executing innocents sometimes deters, the deterrence theorist would have no reason to resist it.

Maybe, then, the deterrence theorist does not really mean to say that deterrence is a sufficient justification for the death penalty. Perhaps what he means is that in the case of guilty perpetrators, already deserving of punishment, deterrence is a sufficient basis for imposing the death penalty. If so, the deterrence theorist could agree with the standard cases in moral theory that purport to show that it is not permissible to kill one for the sake of many, while still maintaining that deterrence justifies the death penalty. The point is that the person being used to save innocent lives is guilty of a crime. It is only in this context that deterrence has justificatory force, given that it does not challenge the rights of the innocent.

Restricting the death penalty to the guilty, however, only underscores the point that deterrence is not by itself an adequate rationale for the death penalty.[27] On this version of the thesis, the justificatory

---

[26] Sunstein and Vermeule fail to see this point, claiming that "in the abstract, any deontological injunction against the wrongful infliction of death turns out to be indeterminate on the moral status of capital punishment *if* the death is necessary to prevent significant numbers of killings." Sunstein & Vermeule, *supra* note 14, at 707. This reflects a misunderstanding of the deontological position on intentional killing. One cannot conclude from the fact that deontologists consider intentional killing impermissible that their primary concern is to minimize intentional killings. Instead, they regard the injunction not to kill as an absolute. They are not, in short, in the business of conducting consequentialist calculations over injunction-breaking, as that would undermine their rejection of consequentialism altogether.

[27] Note further that since any widespread, systemic use of that penalty will probably kill at least a few innocent people (contrary to my assumption that there are no errors in applying the death penalty), it is not likely that the death penalty can be restricted only to the guilty. In fact, the deterrence theorist does not normally consider the use of the death penalty invalidated by such mistakes. Therefore, he *does* ultimately think it permissible to

effect is in part secured by a fact—the offender's guilt—that has nothing to do with deterrence. And this implies once again that the deterrence theorist cannot use deterrence alone as a sufficient ground for imposing the death penalty.

Of course the deterrence theorist might try to explain the significance of guilt in deterrence terms as well. He might, for instance, argue that putting innocent people to death would erode the death penalty's deterrent efficacy. If people were executed at random, a person would have no more reason to fear execution in the wake of having committed a crime than he would were he innocent. The problem with this answer, however, is obvious: By his own lights, the deterrence theorist only needs the perception that the death penalty is being used as punishment for the guilty. He must therefore be ready to adopt punishment of the innocent if that would prove the most expedient deterrent. It follows that the deterrence theorist must either abandon the requirement that punishment be used only on the guilty,[28] or admit that deterrence alone is not sufficient justification for the death penalty and that it must be supplemented with some further principle in a more complicated mixed theory. It is to such a theory that we now turn.

B. Deterrence as a Mixed Theory

The deterrence theorist may not feel threatened by the suggestion that deterrence can only play its part in a mixed theory of punishment. He may be happy to concede, for example, that deterrence is a good reason for imposing the death penalty only when combined with guilt as a limiting condition. So let us now consider whether deterrence is a compelling rationale for the death penalty when used in a mixed theory of this sort.

I have already suggested that even some death penalty opponents are willing to accept the death penalty under these circumstances.[29] Hugo Bedau, for example, allows that if killing a murderer would

use the death penalty against innocents (although he does not think it permissible for a death penalty system to set out to do so). But in this case his argument is subject to the deontological objection raised above, namely that we do not accept killing one innocent person for the sake of a larger number of innocents.

[28] Some have argued that we cannot punish the innocent, because this would not be "punishment." Punishment, by definition, applies to the guilty. But it should be clear that this kind of "definitional stop" does not allow us to avoid important moral questions of this sort. For we might as well call the practice of "punishing" the innocent by a different name, say, "telishment," and then ask whether *that* would be justified. John Rawls explores an institution of punishing the innocent (i.e., "telishment") in *Two Concepts of Rules*, 64 PHIL. REV. 3, 10–13 (1955).

[29] *See supra* notes 18–21 and accompanying text.

bring the victim back to life, the death penalty would be "an instrument of perfect restitution."[30] He writes:

> In such a miraculous world, it is hard to see how anyone could oppose the death penalty, on moral or other grounds. Why shouldn't a murderer die if that will infallibly bring the victim back to life? What could possibly be wrong with taking the murderer's life under such conditions? The death penalty would be an instrument of perfect restitution, and it would give a new and better meaning to lex talionis. The whole idea is fanciful, of course, but it shows as nothing else can how opposition to the death penalty cannot be both moral and wholly unconditional. If opposition to the death penalty is to be morally responsible, then it must be conceded that there are conditions (however unlikely) under which that opposition should cease.[31]

But I think the abolitionist must not concede that deterrence has even this kind of moral force, since the concession would give the deterrence theorist a fairly direct route, through a series of examples, to justifying the death penalty under normal circumstances.

To see this, consider "Variation #1." Suppose that instead of bringing his own victim back to life, executing a killer would bring the victim of a different murderer back to life. If we accept Bedau's case, we must accept this case as well, since the death penalty would still be "an instrument of perfect restitution," insofar as executing all murderers would bring all victims back to life.

But now consider "Variation #2." Suppose that instead of bringing a past victim of someone else's murder back to life, we could bring a future victim back to life if we executed a killer—someone else's future murder victim. The victim in such a case would die only temporarily and would magically spring back to life as the effect of the past execution did its work. Once again, if we accept Bedau's case and Variation #1, it is hard to see why we would reject the death penalty in Variation #2, for in that case we could save every murder victim by executing all murderers.

From Variation #2, it is but a short step to the deterrence theorist's conclusion. In "Variation #3," suppose that instead of bringing a future victim back to life, executing a murderer would prevent the killing of a future victim, but once again the future victim of another murderer. What grounds do we have, based on our acceptance of the prior three cases, for rejecting the use of the death penalty in this case? And if we would accept its use, then surely the argument for the

---

[30] HUGO ADAM BEDAU, DEATH IS DIFFERENT: STUDIES IN THE MORALITY, LAW, AND POLITICS OF CAPITAL PUNISHMENT 36 (1987).

[31] *Id.*

death penalty on deterrence grounds is even stronger, given that we are considering a case in which executing one murderer would deter not just one but *eight* future murders. The deterrence theorist now seems able to claim that as long as the death penalty is restricted in its use to guilty perpetrators and actually deters future murders, it is justified. Once the death penalty opponent has conceded Bedau's point, he will find it hard to disagree with that conclusion.

But is it so clear that we must accept Bedau's verdict on the initial example? Here I think the non-utilitarian has stronger grounds for objecting than Bedau realizes. For if it is impermissible to kill one person to save another, killing the perpetrator in this case should be impermissible as well. Indeed, this is what the deontologist presumably *must* say if he subscribes to the prohibition on using. But even a deontologist who rejects the "using" formulation of the Kantian principle might see killing the perpetrator as impermissible, since the deontologist will say more generally that we cannot justify putting a person to death by pointing to some set of beneficial consequences from doing so. Instead, a deontologist will ask whether the perpetrator *deserved* to die. And with respect to *that* inquiry, the fact that the victim would be brought back to life is irrelevant, or at least does not strengthen the case for execution. If anything, it will weaken the case, since the perpetrator will not have killed the victim if the victim is brought back to life, and thus the perpetrator arguably will not deserve to die!

The deontologist's refusal to kill the perpetrator under such circumstances may seem so much the worse for his theory. How, after all, could we possibly refuse to put the perpetrator to death if this would actually restore the life of the victim? But the deontologist's objection can be made compelling by focusing on a somewhat different but related case. Imagine that instead of dying immediately after being attacked, the victim comes very close to death. As it turns out, he will die unless he receives a heart transplant. The perpetrator of the attack—whom we now have in custody—is the only compatible donor; without the perpetrator's heart, the victim will die. May we remove the perpetrator's heart and transplant it into the victim's body, thereby killing the perpetrator and saving the victim? Here, I admit it is less obviously objectionable than in the previous cases. Yet we would have little difficulty understanding someone who thought it impermissible to use the perpetrator in this way. Indeed, we might even understand someone who objected to allowing the fact that the perpetrator would be a useful organ donor to play any role at sentencing in determining whether he should receive the death penalty. It is not even entirely clear that we could force organ donation for an

executed criminal even if the suitability of his organs were not allowed to count in favor of his execution. The basis for these rather more stringent objections would presumably be that a person's right to autonomy applies with particular force to his control over the treatment of his own body, even when he is condemned to suffer physically through incarceration. It is this same intuition that might lead us to say that it is not permissible to kill a person guilty of a terrible crime in order to bring his own victim back to life.

There is, admittedly, something strange going on here. What could possibly trouble us about killing the perpetrator to save a victim if, moments before, the victim would have been entitled to do whatever was necessary to prevent the perpetrator from killing *her*? There is, of course, a difference between the two cases: In the latter case we have preventive action—performed before the harm has occurred—and in the former it is retributive action—performed after the harm is already complete. And this is a difference that matters. It is permissible, for example, for a victim to use deadly force preventively against a person she suspects is about to rape or even wound her. But even the strongest defenders of the death penalty do not claim that the death penalty should be available as a punishment for rape or assault.[32] It is presumably this difference between prevention and punishment that prohibits killing the perpetrator to bring the victim back to life, since doing that arguably falls on the punitive, rather than the preventive, side of the line.[33]

The deterrence theorist might respond that the whole point of deterrence is not to punish or rectify harm to the victim but to prevent harm. Executing one perpetrator arguably falls under the preventive privilege and as such should be easily justified. Surely, if a woman can kill an assailant because she reasonably fears he will rape her, it must be permissible for the state to kill one person to prevent eight murders.

The opponent of the death penalty might object to this appeal to prevention on the grounds that there is an imminence problem here: The deterrent effect of an execution is likely to be diffuse, one that could take many months or even years to make itself felt. To elimi-

[32] The principle that the death penalty should only be used as punishment for murder has enjoyed widespread acceptance for many years. *See* Coker v. Georgia, 433 U.S. 584, 592 (1976) (holding death penalty for rape of adult woman unconstitutional under Eighth Amendment). *But see* State v. Wilson, 685 So.2d 1063, 1070 (La. 1996) (holding death penalty is not excessive punishment for rape of child under twelve).

[33] Admittedly, Bedau's example is significantly different from the usual case of punitive action, since normally the victim's life cannot be restored. But since, in the ordinary case, another victim's life can be saved, we cannot assume this makes all the difference.

nate the effect of that diffuseness, then, let us imagine that a man is holding eight innocent people hostage and is threatening to shoot them all within minutes. As it happens, he is listening to the radio, waiting for news of another man's execution. If the governor grants clemency to this other man, the hostage-holder will shoot the eight people; if the governor denies clemency, the hostage-holder will be intimidated into releasing the eight. The governor opposes the death penalty under all circumstances but knows the hostages will be killed if he grants clemency. Does the death penalty opponent still think that the governor has a moral obligation to grant clemency? Can he even argue that it is morally *permissible* for the governor to grant clemency in this case?

The problem for the death penalty opponent is that it is hard to characterize the governor's choice as purely punitive. While the decision to pardon or not to pardon the perpetrator is ex post from one perspective, it is preventive under another. Nevertheless, the ex ante aspect of the governor's action does not relieve him of the obligation to stop the execution.[34] Refusing to grant clemency would do precisely what many philosophers claim is impermissible, namely, kill a few to prevent the deaths of a greater number. As we have seen, the fact that more people will be saved than are sacrificed is not by itself sufficient to justify killing the smaller number; some further principle is needed to justify that action. For instance, the situation would be different if we knew granting clemency to the perpetrator would result in that same person killing eight people immediately. In that case, the imminence of the eight deaths entitles us to kill the criminal, making the killing an instance of defense of others—clearly permissible as an extension of the self-defensive rights of each of the eight. 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.

All this may seem obvious, but it is worth spelling out, because I think it does finally clinch the case against the deterrence theorist. His strongest case is when executing one guilty person would prevent the imminent deaths of many more people, since that brings the situation closest to defense of others. But that argument fails when the person to be executed is not the person who will cause the victims' deaths. And this is because, to put the point succinctly, the preventive

[34] There is a slight additional difficulty here, which is that the governor might be thought of as rescuing the criminal from execution if he grants clemency, rather than executing him if he fails to do so. This may be thought to affect our intuitions in this case, since duties to rescue are typically less stringent than duties to avoid inflicting harm. But this is an artificial and somewhat unnecessary feature of our example. Let us therefore assume that if the governor does not pardon the criminal, he is in effect executing him.

privilege does not *travel across persons*. Thus, while the deterrence theorist tries to make all punishment fall under the heading of "prevention," he is still limited to punishing the individual who will inflict the harm. In this respect, the preventive privilege turns out to be more limited than one might suppose.

To make the point particularly vivid, suppose we have the following modification of our clemency case. As before, the potential murderer is listening for news of the death row inmate to decide whether to kill his eight hostages. The inmate is strapped to the electric chair, awaiting the governor's decision. It turns out, however, that one of the hostages can press a button and cause the electric chair to electrocute its occupant. If he presses the button, he will cause the inmate to be executed and, since it will appear that the governor ordered the execution, the captor will be deterred from killing his hostages. If he does not press the button, the hostage strongly suspects the execution will not take place, because he knows the governor ardently opposes the death penalty. May the hostage press the button under these circumstances?

It is very tempting to say that he may, since he has a right to self-defense. If he presses the button, he can save his life; if he does not, he will almost certainly be killed. How could it be impermissible for him to press the button? Nevertheless, I think there is little doubt that he may *not*. It will be helpful at this point to recall our fourth assumption and its effect on the dynamics of the argument: Punishment is unjustified unless shown to be justified. We must, therefore, assume there is no other argument that could justify the victim's pressing the button. Does the mere fact that the person with his finger on the button and seven of his friends will die if he does not press it justify his pressing it?

I believe it does not. To see this, we need only suppose that the person sitting in the electric chair is *innocent* and that he was dragged in off the street to serve as an example to others. Clearly, it is not permissible to kill an innocent person who is not in any way the source of the threat in order to save one's own life, since the privilege of self-defense justifies the use of force against only one's assailant, not innocent bystanders. Considered in that light, does it then make any difference if the person in the chair is a murderer? Such a fact would seem irrelevant, since he is no more the source of the threat to the hostages than if he were innocent. And, if none of the eight is entitled to push the button to save his own life, it is certainly not permissible for the governor to order the execution of that same person in order to deter the killing of the hostages. The explanation, once again, is

that the broad privilege granted to preventive killing does not travel across persons.

The basic problem with deterrence as a rationale, even when combined with the requirement of guilt in a mixed theory, can now be stated: It is a justification for killing that travels across persons, since it purports to justify killing one person to deter someone else from killing in the future. This amounts to saying that deterrence is unavoidably utilitarian in that it permits using a person to bring about a good to someone else. The most obvious ground for objecting to this is Kantian, but one need not frame the point in Kantian terms. We can express the same thought in terms of the basic limits on responsibility found across a range of ethical theories and in the law. The doctrine of *novus actus interveniens* provides a helpful example. If a person does something that causes another's death, he is nevertheless not responsible if the causal route by which the death was produced passes through the voluntary act of another human being.[35] We explain this by saying that a person is not responsible for the free, voluntary acts of another. He is responsible for his own acts alone.

I believe that the problem with the argument from deterrence is connected with this rather deep feature of responsibility. Killing one murderer solely to prevent another person from murdering in the future does not fall under the preventive privilege, since it effectively holds the first murderer responsible for the murders of another perpetrator. The only *preventive* justification there could be for executing the first murderer would be to prevent *that* murderer from killing again. But there is little reason to suppose lifetime incarceration could not supply specific deterrence of this sort. Where deterring the second murderer is concerned, the treatment of the first murderer is either punitive, in which case the preventive rationale does not hold and there is no deterrence-based argument for executing the first murderer, or it is preventive, in which case it is invalid because it impermissibly travels across persons. Either way, the deterrence rationale fails.

II
THE RETRIBUTIVIST ARGUMENT

A. Proportionate and Moral Equivalence Theories

Retributivism is the theory of punishment according to which punishment is justified only insofar as it is deserved by the offender as

[35] The exception 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.

a function of the wrongfulness of his act. Traditionally, the core of the retributivist's argument for any specific penalty is the doctrine of lex talionis, which asserts that a person deserves to experience the suffering he has inflicted on his victim. Taken literally, lex talionis is an absurd doctrine: No one thinks we should rape rapists, assault assailants, or burgle the homes of burglars. This apparent absurdity has led some to suggest that retributivism is most compelling without its associated theory of the measure of punishment.[36] But without some way to give content to the notion of desert, the retributivist cannot justify any specific penalty, including the death penalty, and so retributivism would be vacuous. It appears, then, that the retributivist defender of the death penalty must find a better way of matching crimes with punishments.[37]

To appreciate the difficulty here, begin by considering just how approximate such a doctrine must be to work. It is not just that we are unwilling to inflict one or two of the more extreme harms, like rape and torture, as punishment on the criminals who commit them. 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—two harms perpetrators have inflicted on their victims. Indeed, once we consider the wide variety of possible criminal behaviors, it is clear that the vast majority of criminal acts could not in any literal sense be used as punishments. In fact, there are typically only three criminal acts we tend to convert into acceptable forms of punishment: false imprisonment, theft, and (in some states) murder. The retributivist who wishes to match crime with punishment must develop a doctrine that limits penalties to roughly these three forms of criminal conduct, and so must find a nonliteral way of matching crimes with punishments.

Two possible "equivalence" strategies for doing this have been proposed. The first suggests that we make two lists—one of all the crimes that are committed and another of all the punishments we regard as acceptable to inflict—and then match the worst crimes with the worst penalties, and so on down the line. In theory, this approach would match crimes with punishments "proportionately," i.e., it would

---

[36] *See* MICHAEL MOORE, PLACING BLAME: A GENERAL THEORY OF THE CRIMINAL LAW 205–06 (1998) (suggesting retributivism is most useful without associated idea of lex talionis).

[37] Although I have never found retributivism compelling as a theory of punishment, I wish to avoid the wider debate and instead focus on the retributivist's defense of the death penalty. The retributivist has special problems justifying the death penalty, irrespective of the strength of his argument about the institution of punishment as a whole.

establish relative levels of desert, without requiring any absolute metric for matching crimes with punishments.[38] Insofar as it does not tell us which penalties should be available, however, the "proportionate penalty theory" merely helps us assign available punishments, namely those we are *already* willing to inflict, to perpetrators according to the severity of the criminal acts performed. And this suggests that, whatever its other merits, the proportionate punishment theory will not help the retributivist justify the death penalty or, for that matter, any other particular punishment.

The second strategy is to seek to establish a *moral* equivalence between crimes and punishments, instead of trying to match them literally, as lex talionis would do, or to limit retributivism to proportionate relations between crimes and punishments, as the proportionate penalty theory would do. The "moral equivalence" theory maintains that what the perpetrator really deserves to suffer is a harm that is the moral, rather than the physical, equivalent of the harm he inflicted on his victim. In *The Philosophy of Law*, Kant is arguably proposing such an approach:

> [A] pecuniary penalty on account of a verbal injury, may have no direct proportion to the injustice of slander; for one who is wealthy may be able to indulge himself in this offence for his own gratification. Yet the attack committed on the honour of the party aggrieved may have its equivalent in the pain inflicted upon the pride of the aggressor, especially if he is condemned by the judgment of the Court, not only to retract and apologize, but to submit to some meaner ordeal, as kissing the hand of the injured person.[39]

Kant thinks it possible to treat a perpetrator in a way that is morally commensurate with the harm he inflicted on the victim without having to inflict that very same punishment on him. While Kant does not articulate the theory in this way, the basic strategy of this view is to distinguish what a person deserves in some absolute sense from what it is permissible for society to inflict on him by way of punishment. The moral equivalence theory thus maintains that while a criminal who locked his victim in the trunk of a car before killing her may "deserve" to be locked in a trunk himself before being executed, it is not permissible for us to inflict such a punishment. Like the propor-

---

[38] *See* Michael Davis, *How to Make the Punishment Fit the Crime*, 93 ETHICS 726, 741 (1983) ("The least penalty should, of course, be assigned to the least crime; the greatest penalty, to the greatest crime.").

[39] IMMANUEL KANT, THE PHILOSOPHY OF LAW: AN EXPOSITION OF THE FUNDAMENTAL PRINCIPLES OF JURISPRUDENCE AS THE SCIENCE OF RIGHT 197 (W. Hastie trans., Edinburgh, T. & T. Clark 1887) (1796). For a helpful discussion of this and related passages, see SORELL, *supra* note 23, at 149–50 (discussing Kant's argument that we cannot always impose punishment that perfectly fits crime).

tionate version of retributivism, the moral equivalence theory first eliminates impermissible treatments from the roster of available punishments. And like the proportionate punishment theory, it then matches the offender's criminal act with a punishment based on a theory of equivalence. The only real difference between the two is that while one uses proportionality as the metric of equivalence, the other uses the moral quality of the criminal act.

Thus articulated, however, both types of equivalence theory are woefully incomplete. Neither provides any test for determining which penalties are morally permissible—and therefore eligible to be on the list of available punishments—and which are not. For example, how do we know that locking a perpetrator in the trunk of a car and then killing him is impermissible under the theory but that simply executing him is not? Without already knowing which penalties are permissible, we may argue that putting an offender to death is impermissible but locking him in prison for life is not, or even that lifetime incarceration is impermissible but a twenty-year sentence is not. So both theories end up requiring supplementation by another moral theory capable of telling us which penalties are available and which are not. The theory of permissibility then becomes a side constraint on the penalties that are permissible to inflict. Since the purpose of turning to a retributivist approach to punishment was to answer the question of which punishments are morally acceptable and which are not, this is a serious defect.

Suppose, however, that the retributivist supplements his account with a theory that appropriately distinguishes permissible from impermissible penalties.[40] As I argue below, it is not clear he can justify the death penalty even then, since his theory remains seriously flawed in at least two important respects: its ordering of punishments by their relative severity and the lack of congruence between desert and permissible punishments.

B. The Severity Objection

The first objection to retributivism is that there are penalties we think of as morally unacceptable that are also less severe than death. If we rule out those penalties, we will be compelled to rule out death as well. Consider torture. It is difficult to see torture as off limits on the ground that it is unacceptably severe, because arguably torture is less severe than death. The retributivist's own method makes clear why this is so: If penalties are to be the equivalent of crimes, we

[40] I use "retributivism" to refer to both the proportional and the moral equivalence forms of retributivism and "retributivist" to refer to proponents of either version.

should rank penalties in the way we rank crimes. Since we think of murder as a more heinous crime than any nonlethal assault, torture should be a less severe penalty than death. But if torture is an unacceptable penalty, then death is as well, given that death is more severe than torture.

The retributivist, however, has two potential responses to this objection. The first claims that torture actually is more severe than death. The second rejects the idea of using severity as a measure of permissibility.

1. Torture Is More Severe Than Death

The retributivist might suggest that torture is more severe than death, because it is more uncivilized and more brutal. That torture is widely regarded as unacceptable but death is not seems to bear out this intuition.[41] Even if many perpetrators would choose torture instead of death, the retributivist can reasonably deny that those preferences tell us anything about the relative severity of the two punishments. It is possible, after all, that a given criminal would prefer to spend a night in jail than to pay a small fine because he is attached to money and does not mind confinement. But this would not show that a night in jail is less severe than a fine. Similarly, others might prefer death to life imprisonment without parole. Yet surely these preferences do not imply that life imprisonment is a more severe penalty.

The retributivist is, I believe, right to distinguish between a criminal's preference for one punishment over another and the relative severity of those punishments. We can, that is, think of incarceration as more severe than a fine because we assess severity in general terms, not according to any particular set of subjective preferences. Drawing the distinction in this way allows the retributivist to avoid apparent counterexamples to his claim about the relative severity of torture and death.

The more serious problem, however, is that the retributivist himself seems committed to the view that death is more severe than torture, as is apparent from the way he orders crimes. A person who kills another person deserves death; a person who tortures another without killing him does not deserve death, or so most retributivists believe. If

[41] Michael Davis recently appealed to this intuition in defending his own version of equivalence theory against an earlier work of mine, claiming that to refute the severity argument, he need only "provide one plausible example of a jurisdiction in which death can be justified as a penalty even though torture cannot be." Michael Davis, *A Sound Retributive Argument for the Death Penalty*, 21 CRIM. JUST. ETHICS 22, 23 (2002). Examples of such jurisdictions surely abound, but I find it hard to see their relevance to the *normative* question of the moral status of torture and the death penalty.

torture is more severe than death and torture is an unacceptable penalty, the perpetrator who tortures but does not kill his victim would also deserve death. Unless, then, the retributivist is prepared to expand the death penalty's application dramatically, it seems he cannot argue that torture is impermissible because it is more severe than death.

2. *Severity Is Not a Good Gauge of Permissibility*

A different response to the severity objection is to deny that severity is the appropriate means of ordering punishments. There are many penalties we would readily classify as less severe than death or even incarceration that we nevertheless think are impermissible to impose. The severity objection itself highlights one such punishment, namely torture. Other punishments that we often consider impermissible despite being less severe than many permissible penalties are shame sanctions, such as forcing a convicted sex offender to display a sign outside of his dwelling revealing his status or forcing a drunk driver to affix special license plates to his car.[42] It is also questionable whether it is permissible to sterilize repeat sex offenders or to subject female adolescent offenders to mandatory birth control measures like Depo-Provera, even though these penalties are less severe than both torture and death.

Michael Davis, for one, explicitly embraces this argument and suggests that whether or not a penalty is "humane" is not a function of its severity. Instead, he says, "a penalty is inhumane (in a particular society) if its use shocks all or almost all" members of that society.[43] We might thus explain the impermissibility of punishments of lesser severity by saying they are shocking and the permissibility of punishments of greater severity by saying they are commonly accepted. If the abolitionist's severity argument were correct, however, we could not differentiate punishments in this way. Instead, we would have to take the lowest unacceptable penalty on the list of penalties and say that any penalty more severe than it would be morally unacceptable. This strategy would quickly rule out most sentences currently inflicted for felonies, since many objectionable shame sanctions are less severe than most terms of imprisonment.

I will allow that the abolitionist is in something of a bind here, for the death penalty proponent is correct that many lesser penalties are currently thought inhumane—and ruled out on those grounds—

---

[42] For a discussion of the resurgence of shame sanctions, see Dan M. Kahan, *What Do Alternative Sanctions Mean?*, 63 U. CHI. L. REV. 591, 631–34 (1996).

[43] Davis, *supra* note 41, at 24.

despite the fact that they are not terribly severe. And this suggests that we cannot use severity as a benchmark for permissibility. Which penalties are morally permissible and which are not must be a function of some other moral metric, and until we know what that metric is, we cannot use the fact that torture is morally impermissible as a way of arguing against the use of death as a penalty. How can the abolitionist answer?

Here I think the abolitionist must be prepared to bite the bullet and admit that the above argument only shows that we are wrong to reject shame sanctions and other minor interferences with liberty. Indeed, we have now observed a compelling reason to allow such penalties, namely that imposing them may enable us to avoid inflicting more severe penalties that involve a significantly greater loss of liberty. If, for example, we have the choice between two equally effective penalties—a shame sanction such as a sign or a license plate, on the one hand, and a period of incarceration, on the other—we arguably have an obligation to inflict the less invasive penalty. On balance, then, the abolitionist argument I offered above still seems a good one—that if we reject torture because it is too severe, we should reject death as a penalty because it is more severe. But we must admit that the argument requires us to revise our intuitions about a number of lesser penalties.

C. Lack of Congruence Between Desert and Permissibility

A second problem with the retributivist's reliance on either version of the equivalence theory is that the argument he makes to defend the death penalty—namely, that although the offender might deserve to suffer all sorts of punishment, only certain forms of deserved punishment are morally permissible—can readily be used by death penalty opponents. One of the most common arguments made against the retributivist death penalty proponent is that there are moral side constraints on imposing the punishment of death, and that these operate over and above the constraints imposed by the notion of desert itself. Thus, many death penalty opponents are happy to concede that a murderer in some objective sense "deserves" to die but maintain that death is not a permissible penalty for the state or society to inflict.[44] That is, they offer precisely the same argument that death

[44] *See* STEPHEN NATHANSON, AN EYE FOR AN EYE?: THE MORALITY OF PUNISHING BY DEATH 42–43 (1987) (arguing that criminals should not always get punishment they deserve, "especially . . . when the body that is to give someone his just deserts is the government"). For a plausible example in which a person deserves an impermissible penalty, consider a person discovered to be guilty of a crime after having been acquitted of it. It is, however, less clear that we can say a person deserves to be tortured while also maintaining

penalty proponents tend to offer with regard to torture. Some explain the position by saying that we compromise our own civility by executing even the worst of our fellow human beings.[45] Others argue that the state is usurping God's role in deciding whether to put people to death, and that human beings are simply not sufficiently omniscient to pass life and death judgment on other human beings.[46] Still others argue that there are certain penalties that are inconsistent with the requirements of human dignity, and that torture and death are among these.[47] Jeffrie Murphy, for example, writes:

> Given the exceptional moral gravity of having one's prospects for a morally significant and meaningful life interrupted, one might well want to deny the state any right to do this—i.e., one might adopt a direct absolute ban on the penalty of death. For it is by no means clear that one can show respect for the dignity of a person as a person if one is willing to interrupt and end his most uniquely human capacities and projects. Thus, . . . there is perhaps a case to be made that the punishment of death is degrading . . . .[48]

It is important to notice, however, that the arguments presented here for and against the death penalty have the same structure: Both allow that someone who kills another person "deserves" to die, since offenders must receive the same treatment they inflict on their victims. And they both allow that it may be impermissible to inflict certain penalties that offenders deserve. They simply disagree about whether death itself is such a penalty. Consequently, the debate between proponent and opponent quickly reduces to the question of whether we think that death is an excessively harsh penalty. And that is not a terribly nuanced ground on which to settle the matter.

The two sides of the debate thus reach different conclusions based on the same fundamental premises, and there seems to be no basis for picking one side over the other. We should not be too hasty,

---

that torture is not, and never has been, a permissible penalty. I do not, therefore, find this argument a compelling justification for the death penalty on retributive grounds.

[45] *See, e.g.*, State v. Ross, 646 A.2d 1318, 1395 (Conn. 1994) (Berdon, J., dissenting in part) ("Not only does the death penalty degrade the individuals who are sentenced to die, but it also degrades and dehumanizes a society that permits it to be imposed, calling into question the morality of every one of us.").

[46] *Cf.* Anthony G. Amsterdam, *Capital Punishment*, *in* THE DEATH PENALTY IN AMERICA 346, 352–53 (Hugo A. Bedau ed., 3d ed. 1982) ("The plain message of capital punishment, on the other hand, is that life ceases to be sacred whenever someone with the power to take it away decides that there is a sufficiently compelling pragmatic reason to do so.").

[47] *See, e.g.*, Jeffrie G. Murphy, *Cruel and Unusual Punishments*, *in* RETRIBUTION, JUSTICE AND THERAPY: ESSAYS IN THE PHILOSOPHY OF LAW 223, 243 (Wilfred Sellars ed., 1979) (suggesting ban on death penalty based on respecting human dignity).

[48] *Id.*

however, to declare the argument a draw. If the retributivist appears not to have made his case, that fact alone should be understood as a benefit to the death penalty opponent. This follows straightforwardly from the fourth assumption I articulated in the Introduction, namely, that punishment is a harm or evil that stands in need of justification. The result is that the death penalty proponent bears the burden of proof in this context: The death penalty cannot legitimately be imposed unless there is some affirmative argument justifying it, which is not overcome by arguments against it. Notice that the retributivist is particularly affected by this burden of proof claim, for his own account implies that killing a person is such an evil that the killer incurs a tremendous moral debt, repayable only with the murderer's own life. It would seem to follow that the executioner, or society more generally, who takes a person's life must incur this same moral debt, unless his act is morally justified. Without such a justification, the executioner, and society as his accomplice, is no better than a murderer.

The argument from torture, then, is problematic for the retributivist because it shows that it is perfectly possible for the required justification to be absent where a grievous penalty is concerned. If the retributivist thinks torturing a torturer is unjustified, then it is always possible that killing a murderer is also unjustified. Without an argument explaining why death is permissible but torture is not, the retributivist lacks the justification he needs to win the debate with the death penalty opponent. It looks, then, as though the death penalty proponent has once again failed to meet his justificatory burden.

III
CONSENT AS A JUSTIFICATION FOR PUNISHMENT

A. Some Preliminaries

We have thus far considered the implications of two major schools of thought about punishment: deterrence, which is consequentialist, and retributivism, which is deontological. I have tried to show that neither can justify the penalty of death. This may seem to settle the matter, as almost all traditional writings on punishment can be grouped under one of these two headings. In political philosophy, however, the prevailing tradition since the seventeenth century has been neither consequentialist nor deontological but contractarian. Curiously, while the consequentialist and deontological traditions are well represented in legal theory, the contractarian tradition has not been.

Contractarian theories regard the major rules and institutions of civil society as legitimate insofar as they can be thought of as based on an agreement among the individuals who must submit to their authority. There are two dominant strains in the contractarian tradition, what we might call "normative contractarianism," on the one hand, and "rational choice contractarianism," on the other. Although normative contractarianism descends from Kant, it covers a variety of views, the most influential of which in recent years has been John Rawls's. According to Rawls, we can best discern intuitions about justice in a liberal society by asking what principles of justice would be selected by individuals entering into a political arrangement with one another, prior to the existence of social institutions of any sort. Rawls assumes that in this original position of choice, the contractors are selecting principles of justice without any knowledge of the particular circumstances they will inhabit in society or what their personal characteristics will be.[49]

Rational choice contractarianism, by contrast, descends from Hobbes. It asks what form of social organization rational agents seeking to maximize their own welfare would choose to improve their positions relative to their presocial baselines.[50] To the extent the contractarian tradition has been brought into legal theory, it has almost entirely been of the normative variety.[51] And while one might suppose that the topic of punishment is better suited for contractarian reflection than other areas, there are only a handful of attempts to develop a contractarian account of punishment of any sort.[52]

In this Part, I present a contractarian approach to punishment in the rational choice tradition and ask whether it provides a more adequate justification for the death penalty than either deterrence or retribution.[53] Arguably, a contractarian approach provides the most

---

[49] *See* JOHN RAWLS, A THEORY OF JUSTICE 11–22 (1971) (describing original position); *id.* at 136–42 (describing veil of ignorance).

[50] In its view of human nature, rational choice contractarianism shares the basic presuppositions of law and economics, i.e., that human beings are rational maximizers whose preferences obey certain conditions or axioms of rationality. For a comparison of legal contractarianism and law and economics, see Claire Finkelstein, *Legal Theory and the Rational Actor*, *in* THE OXFORD HANDBOOK OF RATIONALITY 399, 399–401 (Alfred R. Mele & Piers Rawling eds., 2004).

[51] *See generally* Symposium, *Rawls and the Law*, 72 FORDHAM L. REV. 1381 (2004) (discussing impact of Rawls's scholarship on legal world).

[52] *See generally* Sharon Dolovich, *Legitimate Punishment in Liberal Democracy*, 7 BUFF. CRIM. L. REV. 307 (2004); Jeffrie G. Murphy, *Marxism and Retribution*, 2 PHIL. & PUB. AFF. 217 (1973); C.S. Nino, *A Consensual Theory of Punishment*, 12 PHIL. & PUB. AFF. 289 (1983).

[53] By restricting my focus to these three schools of thought about punishment, I do not mean to suggest that they exhaust the literature on that topic. Recent writings on punishment, for example, have added new theories to the available roster: communicative theo-

promising avenue for justifying the death penalty, given our initial assumption that any punishment scheme stands in need of justification. This is primarily because unlike the consequentialist or the deontologist, the contractarian insists that punishment be voluntarily imposed. The reason the voluntariness of the punishment enhances its justifiability should be readily apparent. Many actions that otherwise would be morally impermissible become permissible when consented to by the recipient: Consent turns robbery into gift-giving, battery into touch football, rape into lovemaking.[54] Thus, a punishment that would be impermissible when imposed on the basis of deterrence or desert may become permissible when its imposition is consensual.

Moreover, within contractarianism, the rational choice model appears to make the claim of voluntary agreement easiest to maintain. Assuming that the death penalty has at least moderately strong deterrent efficacy, each rational contractor will regard himself as better off living in a regime that furnishes such deterrence to serious crimes, since it increases his security considerably. Assuming he will not commit such crimes himself, he stands to benefit from the adoption of the death penalty and apparently has nothing to lose. While normative contractarianism might have deontological commitments that are inconsistent with the use of the death penalty, rational agents seeking to maximize personal welfare need not have those same commitments. At any rate, insofar as normative contractarianism depends heavily on various deontological suppositions, that approach may col-

---

ries, *see* R.A. DUFF, PUNISHMENT, COMMUNICATION, AND COMMUNITY, at xvii (2001) (justifying punishment because it seeks to "communicate to offenders the censure they deserve for their crimes"); expressive theories, *see* Dan M. Kahan, *The Secret Ambition of Deterrence*, 113 HARV. L. REV. 413, 420 (1999) ("The expressive theory of punishment says we can't identify criminal wrongdoing and punishment independently of their social meanings.") (emphasis omitted); as well as virtue theories, *see* Kyron Huigens, *The Dead End of Deterrence and Beyond*, 41 WM. & MARY L. REV. 943, 1033 (2000) ("Aristotle *insists* on the rule of law, not of men, on the ground that only governance by reason can be impartial and even-handed."). A complete argument employing the negative strategy I have applied would ideally consider such alternatives. Nevertheless, these more recent theories have frequently been presented as merely explanatory, rather than justificatory, and as such would not respond to the problem as I have posed it. To the extent they are presented as justificatory, they often take the form of mixed theories, in that they are combined with one of the more classical rationales for punishment. *See, e.g.*, Kahan, *supra*, at 421 (maintaining that decisions about whether to punish are "normatively justified to the extent that we think that the law is accurately apportioning punishment based on the moral truth or falsity of the valuations that offenders' emotions express," and suggesting nonexpressive metric for evaluating expressive function of punishment).

54 *But see generally* Leo Katz, *Choice, Consent, and Cycling: The Hidden Limitations of Consent*, 104 MICH. L. REV. 627 (2006) (explaining that there are many overlooked circumstances in which consent is ineffective at curing wrongs).

lapse into a deontological account and would no longer be consensual in nature.

There are, however, two important difficulties the contractarian justification for punishment must overcome. The first is the obvious fact that the offender does not in any literal sense consent to his own punishment. Thus, the first task of any contractarian approach is to explain how it is that criminals can be thought of as consenting to their own punishment, despite their outward rejection of it. Many different accounts of the nature of consent in contractarian theories have been offered. In keeping with the prevalence of normative over rational choice contractarianism, the dominant approach tends to favor hypothetical over actual consent. In what follows, I argue that a contractarian theory would do better with the notion of constructive actual consent than with hypothetical consent.

The second difficulty is that even if the criminal can be thought of as consenting to his own punishment, it is not clear that this consent is sufficient to justify punishing him. That a certain treatment of another human being is consensual does not mean that it is morally permissible. Some treatments are so extreme that the consent of the recipient cannot dispel the moral doubt that infects them. In dangerous games like Russian roulette, for instance, the participants' consent is reasonably seen as inadequate to justify the dangers of the game.[55] More is needed, therefore, to make the criminal's consent normatively salient. What is important about the consensual justification for punishment is that it couples consent with personal advantage. Specifically, I argue in Parts III.B and III.C that a punishment is justified only if the criminal has consented to the system of criminal justice that imposes such punishment because he believes it advantageous to do so.

Once I have established the basic form of justification for a system of criminal justice, we will need to ask whether the members of the society selecting that system would choose to include the penalty of death. I argue that they would not. The final piece of my argument against the death penalty, accordingly, demonstrates that death is not a penalty that individuals choosing in accordance with principles of mutual advantage would select.

---

[55] It is even inadequate to justify each player's allowing *other* players to expose themselves to a risk of death—so that a player can be convicted of manslaughter or even murder when another player turns a gun on himself and fatally pulls the trigger. *See, e.g.*, Commonwealth v. Malone, 47 A.2d 445, 449 (Pa. 1946) (affirming conviction for second-degree murder of boy who voluntarily participated in Russian roulette).

B. The Nature of Consent

Offenders nearly uniformly object to receiving a penalty of any sort. It may thus seem absurd to suggest that the criminal consents to his own punishment. Nevertheless, there are various ways to construe the criminal as consenting despite his overt objections. Perhaps the most obvious way is to appeal to hypothetical consent and say that the criminal *would have* consented to the scheme under which he is punished if he were choosing a system of punishment in the absence of any knowledge of his own situation.[56] On Rawls's approach, for example, the idea of contractarian agreement serves primarily as a heuristic to determine the conception of justice that best reflects our considered intuitions of fairness. Rawls says: "Our social situation is just if it is such that by [a] sequence of hypothetical agreements we would have contracted into the general system of rules which defines it."[57] He continues:

> No society can, of course, be a scheme of cooperation which men enter voluntarily in a literal sense; each person finds himself placed at birth in some particular position in some particular society, and the nature of this position materially affects his life prospects. Yet a society satisfying the principles of justice as fairness comes as close as a society can to being a voluntary scheme, for it meets the principles which free and equal persons would assent to under circumstances that are fair. In this sense its members are autonomous *and the obligations they recognize self-imposed*.[58]

It should be obvious that the sense in which obligations are "self-imposed" in Rawls's scheme is highly attenuated, since the original position involves neither actual agents nor actual agreement, and so a fortiori the individuals restrained by a system of justice have not in any sense agreed to be so restrained.[59] The result is that the Rawlsian notion of consent does little to counteract the presumption against punishment embodied in our fourth assumption. Often it is said that the person who consented is the representative of the person who must live under the actual laws, and that he is therefore capable of binding actual persons. But why should a creature lacking in nearly all human characteristics count as the representative of flesh and blood persons? Rawls might respond that the issue is not about repre-

---

[56] Dolovich, *supra* note 52, at 315 ("[I]f state power [to punish] is to be legitimate, agreement as to the terms of its exercise must come from citizens who do not know the first thing about their own situation and who must therefore accord due consideration to the perspectives of all members of society.").

[57] RAWLS, *supra* note 49, at 13.

[58] *Id.* (emphasis added).

[59] *See* Dolovich, *supra* note 52, at 314–29 (presenting Rawlsian account of punishment based on hypothetical consent).

sentation—it is about fairness. Each actual person should recognize the rules under which he is punished as legitimate because they correspond to his deepest sense of the fairness of basic institutions, elicited through the thought experiment of the original position. But recognizing certain rules as *fair* does not, by itself, mean a person would consent to be governed by them. Fairness might, of course, ultimately justify *imposing* those rules on him, regardless of whether he accepts them. But that is a different story, and it is not, at any rate, a contractarian story.[60]

To justify imposing punishment on an offender on the basis of agreement, we need the consent of that particular agent; the consent of a relevantly similar person will not do. One approach that accomplishes this is what we might call the "voluntarist" theory of punishment, which asserts that if the criminal is on notice of the punishment, by committing the crime he consents to the punishment later imposed on him. Carlos Nino, for example, argues that assuming certain other necessary conditions are met (e.g., the punishment is a necessary and effective means of protecting the community against greater harm), the criminal's voluntary act "provides a prima facie moral justification for exercising the correlative legal power of punishing him."[61]

But while the voluntarist approach has the right level of specificity to justify a particular instance of punishment, the view is problematic for other reasons. First, consent to be exposed to a risk does not entail consent to suffer the injury risked. That I consent to run a risk that someone will crash his car into mine on my way to work, for example, does not entail that I consent to his doing so.[62] Thus, from the fact that a person voluntarily runs the risk of incurring a certain punishment, we cannot infer that he consents to the punishment itself.

Second, as Larry Alexander has argued, voluntarist arguments are objectionable because they can easily justify quite excessive punishments.[63] This is because consent lacks a principle of proportionality that would limit the level of punishment that could be imposed

---

[60] These points against the Rawlsian position have all been made before in one form or another. But it is helpful to see their effect when they are combined with our fourth assumption about punishment.

[61] Nino, *supra* note 52, at 299.

[62] It is not even entirely clear that a person who agrees to do something she knows with certainty will have a particular consequence thereby agrees to that consequence. But this is a more debatable matter.

[63] Larry Alexander, *Consent, Punishment, and Proportionality*, 15 PHIL. & PUB. AFF. 178, 179 (1986) ("The problem is that consent not only substitutes for desert as a justification for punishment, but it also overrides desert as a limitation on the severity of punishment."). *But see* C.S. Nino, *Does Consent Override Proportionality?*, 15 PHIL. & PUB. AFF. 183, 185 (1986) (arguing that respect for autonomy implies there is no reason for liberal states to prevent individual from voluntarily causing harm to himself).

for a given crime. For example, on this account it would be perfectly acceptable to assign the death penalty for a minor traffic offense, Alexander says, as long as the offender was aware of the risk of receiving that penalty when he broke the law.[64] While some might be prepared to embrace this consequence of a voluntarist account, it seems a deeply objectionable feature of this theory, since it puts the voluntarist account out of sync with our prevailing practices of punishment.

A contractarian account will also diverge from the simple voluntarist account with respect to the object of consent. Instead of thinking of consent as operating act-by-act and establishing the criminal's consent to the actual punishment he suffers, we should instead think of the criminal as consenting to a general institution of punishment, which in turn justifies the particular treatment he receives under that institution. The consent, that is, must operate at the level of what Rawls calls the "basic structure."[65] Thus, a person might better be said to have consented to his own punishment if he consented to the institution dispensing that punishment.[66] Unlike the normative version of this claim, the consent can still be actual, rather than hypothetical. But what is consented to is not a particular punishment (as it is on the voluntarist approach), but a punishment scheme in which the criminal can, on the whole, see himself as advantaged.[67] I shall refer to this view as "rational contractarianism."

The advantages of rational contractarianism over voluntarism are clear. First, the former faces no problem of consent to unintended consequences, since the voluntary nature of the act is not itself the source of the consent. The voluntariness of the criminal's act serves to establish responsibility, and the fact that the criminal is responsible for a prohibited act makes him liable to punishment because he has consented to a scheme that associates punishment with responsibility for harm-infliction. Second, rational contractarianism does not face the

---

[64] *See* Alexander, *supra* note 63, at 178 ("[O]ne who commits a crime consents to punishment because he has acted voluntarily with knowledge of his act's legal consequences, that is, the punishment prescribed for that act.").

[65] *See* RAWLS, *supra* note 49, at 7 (characterizing basic structure as "the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation").

[66] *See* Murphy, *supra* note 52, at 230 ("I can be said to will my own punishment if, in an antecedent position of choice, I and my fellows would have chosen institutions of punishment as the most rational means of dealing with those who might break the other generally beneficial social rules that had been adopted.").

[67] The question of whether consent operates on particular acts or more general rules is technically independent of the question of whether consent is actual or hypothetical. But it is perhaps most natural to think of actual consent as applying to acts and hypothetical consent as applying to rules or principles.

proportionality problem the voluntarist account faces. The justification for punishment depends on principles by which individuals have agreed to be governed. Those principles will, in turn, limit the types of punishment to which an individual can be subjected, even when he is aware he will be subjected to it as a result of his own voluntary act.[68] Third, this sort of contractarian account does not justify punishing criminals on the basis of consent alone, but combines consent with the notion of benefit. Specifically, it requires the criminal to believe, in an ex ante position of choice, that he will benefit from the central features of the institution of punishment to which he is consenting.[69] As a consequence, the resulting punishment scheme is better justified than any system of punishment would be on a voluntarist account.

What we have seen is that if a punishment scheme is to acquire justificatory advantage by being consensual, the consent must be actual rather than hypothetical. But, as Rawls suggests, it is hard to see how consent *can* be actual, given that individuals are simply born into a society and are expected to live by its rules, whatever those rules happen to be.[70] I propose an approach to consent that falls in between these extremes, applying the specificity of actual consent to a punishment scheme operating at the level of the basic structure. That is, we can insist on actual consent, but regard that consent as constructively given on the basis of tacit manifestations of acceptance of the rules of a given society.

What, then, should count as a manifestation of actual consent? On an account that seeks to construct an agent's consent from actual behavior, consent must be tacit. For example, in a society in which individuals are free to leave but choose to stay, their decision to remain signifies an acceptance of the rules of that society, and so constitutes a tacit admission that they regard themselves as better off

---

[68] Of course, it is possible that individuals will adopt the following as a principle: Assign any punishment that serves the state's general purposes, as long as the punishment is anticipated as the result of a voluntary act. That is, it is conceivable that the parties would incorporate the basic voluntarist principle into their agreement. But that would be an unlikely principle for agents contracting for the basic structure of society to adopt, since such a principle would not appear to be of obvious mutual advantage (and would involve significant disadvantages given the proportionality problem). It is far more likely that individuals agreeing to the terms under which punishment can be legitimately imposed would tie the justification for punishment to a set of principles that connects the point or purpose of punishment with the seriousness of the infraction.

[69] I shall have more to say about the precise contours of this belief in Part III.C.

[70] *See* RAWLS, *supra* note 49, at 7 ("[Individuals] born into different positions [in the basic structure] have different expectations of life determined, in part, by the political system as well as by economic and social circumstances."); *cf.* JOHN RAWLS, POLITICAL LIBERALISM, at xlv (1996) (characterizing basic structure as closed society "we enter only by birth and exit only by death").

under the terms of that society than they would be either in its absence or under the terms of another available society. But we may also wish to require political participation (such as voting) or acceptance of some benefit from the state as a sign of consent. Socrates uses an argument of this sort in *The Crito* to defend his decision to remain in prison and allow the state to carry out its sentence. In the following passage, Socrates imagines what "the Laws" would say if confronted with the question of whether it is legitimate for Socrates to flee his sentence:

> We gave you birth. We nurtured you. We educated you. We gave to you and to every other citizen a share of every good thing we could. Nonetheless, we continue to proclaim, by giving leave to any Athenian who wishes, that when he had been admitted to the rights of manhood and sees things in the City and its Laws which do not please him, he may take what is his and go either to one of our colonies or a foreign land. No law among us stands in the way or forbids it. You may take what is yours and go where you like, if we and the City do not please you. But whoever among you stays, recognizing the way we render judgment and govern the other affairs of the City, to him at that point we say that by his action he has entered agreement with us to do as we bid.[71]

Socrates imagines the "implied contract" to be one reached between the Laws and the citizen, rather than among citizens, as would a modern contractarian. Partly for this reason, Socrates fails to consider what the terms of the contract might be, and so also fails to ask whether citizens would place limits on the kinds of laws they would agree to obey. Nevertheless, the nature of the implied consent Socrates imagines is quite similar to that which I have suggested, namely that actual consent can be constructed and inferred from a combination of the individual's political participation and the social benefits he receives.

There is, however, a difficulty with treating consent in this instance as actual, since the willingness of a person to abide by the rules of a given criminal justice system may or may not match his rational, reflective views. We are interested only in what each actual person thinks *insofar as he is rational*. The question is how each rational agent would regard the balance of benefits and burdens of the rule governing punishment when he reasons correctly and is fully informed. Constructive actual consent is established by considering the preferences of an actual person, purified of confounding irratio-

---

[71] 1 PLATO, *The Crito*, *in* THE DIALOGUES OF PLATO 105, 126 (R.E. Allen trans., Yale Univ. Press 1984).

nalities. It allows us to retain the idea that it is actual agents binding themselves, rather than individuals being bound by representatives.

As with individuals deliberating behind the veil of ignorance, actual contracting agents have limited information about themselves. These limits, however, are due simply to the ordinary doubt about one's future with which we all must live. I cannot know whether I will end up committing a crime—not because I am uncertain about my own characteristics or those of the world in which I live, but because I am unsure what life will bring and so am unsure what my proclivities and preferences will lead me to do. Notice that the degree to which a rational contractor would regard a certain penalty scheme as beneficial may depend on when in his life he assesses the benefits from that scheme. For example, a person assessing the merits of the death penalty would consider it differently depending on whether he were to imagine himself subject to the penalty at the age of twenty or the age of eighty. In the latter case, there would be significantly fewer years of life foregone, and thus a rational agent might regard the deterrent benefits as greater relative to the losses he would suffer by being subject to the penalty at that late point in life.[72]

But recall that we are considering a form of actual, rather than hypothetical, consent, and this will help us to select the correct moment at which consent should be specified. The agent's consent to a scheme of punishment should be thought of as given at the first available moment at which he is old enough to be capable of making the choice to accept a particular political system.[73] What is important on this account, however, is that individuals are committing their own future selves, making it more justifiable to regard their consent as binding than it is under hypothetical consent, where actual persons are committed by a set of hypothetical representatives instead.

Let us now consider the nature of the required benefit in a rational contractarian account.

C. The Benefit Requirement

As I have argued, the consent of the offender is necessary but not sufficient to justify inflicting punishment on him. There is a second

[72] Of course, the marginal benefit of increased bodily security is also lower for a person at the end of life compared with someone closer to the beginning. The point remains that the value of the death penalty under this test may differ depending on the point in life at which one applies the test.

[73] We could treat this as the moment at which he is first entitled to cast a ballot for an elected representative, but there might be other, better markers of political maturity. At any rate, we will consider the age of consent as roughly coinciding with the beginning of young adulthood.

condition that must be combined with consent: The agent must consent on the basis of a belief that his welfare will be improved by the scheme to which he is consenting. The improvement need not actually materialize for this "benefit requirement" to be met. An agent's consent is still adequately validated if he sincerely believes he will benefit under the rule, even if he turns out to be mistaken or the benefit fails to materialize.[74] What is essential is that the agent see himself as faring better under the terms of the rule than he would fare in its absence.

How does a rational agent determine whether he will fare better under a given rule? The standard answer is that he asks whether his expected utility is higher living with the relevant rule than living without it. On this view, if his expected utility under the rule is positive, he will consent to it, even if he is aware that he might end up worse off under the rule than he would be in its absence. The question is whether, under conditions of uncertainty about the effects of future rules or institutions, the benefit requirement is compatible with this kind of gambling. I suggest that, on the contrary, the benefit requirement screens out gambles; the requirement is not satisfied if the agent cannot see himself as assuredly better off under the relevant rule than he would be in its absence.

This may seem unacceptably strong. Why would it not be rational for a person to gamble with respect to his future welfare under extremely favorable odds? The answer is that while the benefit principle is stringent, it applies only to rules or institutions that are part of society's basic structure. While gambling may be rational for individual decisionmaking as applied to particular choices made within that structure, decisions that establish the structure itself are unlikely to be attractive if they risk leaving the agent worse off than he would have been in their absence. Thus, a rational agent contemplating a structural rule or institution must believe that he will fare better than he would without that rule or institution, whatever life should happen to bring. Rational contractors might be happy to risk elements of their current welfare for a chance at much greater welfare in the future, but they will not be willing to take that same risk when they stand to lose their most fundamental political and economic protections.

The no-gambling supposition for rational agency may seem extreme, but some version of this condition appears to be part of

[74] There are, of course, conditions under which the agent's consent is vitiated by a false promise of benefit. But in such cases the consent is invalid because of normative conditions rendering the consent illegitimate, not because the requirement of benefit is not satisfied.

nearly every contractarian proposal in both the normative and the rational contractarian traditions. John Locke, for example, expresses what appears to be a no-gambling condition with the so-called "Lockean proviso," which requires that agents who remove resources from the commons leave "enough and as good" for others.[75] Such a condition would make no sense from a contractarian perspective unless individuals in a state of nature were unwilling to gamble with basic welfare, since each might otherwise wish to gamble that he would benefit more from having no limits on what he takes from the commons than he would lose by others doing the same.

A no-gambling condition also seems to be the driving force behind the "maximin" principle for choice under uncertainty, to which Rawls, as well as several of his predecessors, subscribe.[76] According to maximin, we should adopt "the alternative the worst outcome of which is superior to the worst outcomes of the others."[77] In other words, we should choose the option that has the best worst-case scenario. Rawls explains this by saying that it would be rational to choose the principles of justice that a person would select for a society in which "his enemy is to assign him his place."[78] It is this no-gambling decision rule, Rawls thinks, that rational deliberators in the original position would adopt for establishing the basic structure of their society. Rawls also appears to insist on a no-gambling condition in several other places in his account. He says, for example, that deliberators will require that the first principle of justice take "lexical priority" over the second principle, meaning that individuals would not be willing to trade basic liberties against any increase in social or economic benefit.[79] He also suggests that contractors will adopt the "difference principle," which allows economic inequalities only insofar as

---

[75] JOHN LOCKE, SECOND TREATISE OF GOVERNMENT 19 (C.B. Macpherson ed., Hackett Publ'g Co. 1980) (1690) ("[F]or this *labour* being the unquestionable property of the labourer, no man but he can have a right to what that is once joined to, at least where there is enough, and as good, left in common for others."). For an examination of some of the difficulties with Locke's theory of tacit consent, see John G. Bennett, *A Note on Locke's Theory of Tacit Consent*, 88 PHIL. REV. 224 (1979). Bennett writes: "The theory of tacit consent which Locke puts forward in his *Second Treatise* is interesting because it shows the lengths to which Locke was driven in order to maintain that government power and authority must be based on consent." *Id.* at 224.

[76] *See* RAWLS, *supra* note 49, at 154–55 (explaining when maximin is reasonable); R. DUNCAN LUCE & HOWARD RAIFFA, GAMES AND DECISIONS: INTRODUCTION AND CRITICAL SURVEY 279 (1957) ("[Maximin] has the merit that it is extremely conservative in a context where conservatism *might* make good sense.").

[77] Rawls, *supra* note 49, at 153.

[78] *Id.* at 152.

[79] *Id.* at 61 ("[Such] ordering means that a departure from the institutions of equal liberty required by the first principle cannot be justified by, or compensated for, by greater social and economic advantages.").

they are to the benefit of the least advantaged members of society.[80] The difference principle only makes sense because deliberators will not wish to gamble that they will be Bill Gates, or even that they will be among the relatively privileged. Accordingly, they will be sure to provide well for those less advantaged.[81]

A similar sort of reasoning appears in defense of what public policy analysts and economists call the "Precautionary Principle"—the idea that in the face of substantial uncertainty about the likelihood of certain events occurring, the rational strategy is to privilege the avoidance of catastrophes. Once again, this constitutes a rejection of the idea that it is rational to gamble that one will reap significant benefits from a risky arrangement.[82]

D. Applying the Benefit Requirement to Punishment

Let us now consider what the benefit requirement would imply for a system of punishment, and let us start with an example. Consider a group of contractors trying to decide how much and what kind of protection they should institute for private property. They have already selected a series of rules establishing a system of ownership, and they now seek a means of enforcement. They must weigh the following considerations. On the one hand, they would like the maximum deterrence feasible for violations of ownership rights. On the other hand, they also want to protect their own personal freedom and would like to maximize independence of choice without interference from others. Maximizing independence of choice would leave no protection for ownership, while maximizing protection for private property would sharply curtail personal liberty.

In balancing security and liberty, each person asks himself: Would I be better off under the terms of a contract that established penalties for theft and other violations of property norms, assuming that I myself might end up subject to those penalties, than I would be if there were no private ownership at all? If the penalties for theft are too low, the deterrent effect will be insignificant and private property

---

[80] *Id.* at 78 ("[Such inequality] is justifiable only if the difference in expectation is to the advantage of the representative man who is worse off, in this case the representative unskilled worker.").

[81] Note, however, that the difference principle is not an all-purpose no-gambling condition, as it does not address the welfare of all those who are less well off than Gates, but only the very *least* advantaged, or the *worst*-off members of society. *Id.* at 81–83 (highlighting difference principle's exclusive focus on least advantaged).

[82] *See* CASS R. SUNSTEIN, LAWS OF FEAR: BEYOND THE PRECAUTIONARY PRINCIPLE 109–17 (2005) (critically assessing both maximin and Precautionary Principle); Stephen M. Gardiner, *A Core Precautionary Principle*, 14 J. POL. PHIL. 33, 45–49 (2006) (discussing relationship between maximin and Precautionary Principle).

will not be protected. If the penalties are too high, agents receiving the penalty would be worse off than they would have been in the absence of private property and the benefit requirement would not be satisfied. Because we require that the institution of punishment be justifiable with respect to each and every person who is subject to its authority, the benefit requirement, coupled with consent, must be satisfied with respect to each individual contracting agent. Thus consent must be unanimous. The benefit requirement, in combination with the requirement of unanimous consent, then, makes it possible for the social goal of deterrence to dictate specific parameters for the punishment of each separate crime.

Notice the advantages of rational contractarianism as compared with the two leading approaches to punishment. On the one hand, rational contractarianism solves the two central problems associated with pure deterrence theories—the problem of torture and the problem of impermissible tradeoffs. With regard to the former, rational contractarianism rejects extreme penalties, since these would normally fail the benefit test. With regard to the latter, contractarianism rejects the involuntary sacrifice of a smaller number of persons for the sake of a greater number, since no institution that treats one person solely as an instrument for enhancing the welfare of others would pass the benefit test. It is easy to see that contractarianism would accordingly rule out punishment of the innocent along with other impermissible tradeoffs. For a society that left individuals subject to punishment at random would be worse than complete chaos, since in the former, persons must protect themselves not just against lone individuals but against a state that possesses a monopoly on power. If the institution of punishment is to leave members of society better off than they would be in its absence, it must allocate sanctions predictably, fairly, and according to principles of control and individual responsibility.

Notice also that on the contractarian approach I propose, there is no worry about punishment traveling across persons as it does in the mixed deterrence account. It is true that, according to rational contractarianism, deterrence is the basic aim of the punishment agreement, and deterrence schemes usually involve traveling across persons. But in fact the problem does not arise on this contractarian view. For although the institution of punishment thus agreed to would be deterrence-based, and hence would hold one person responsible for the acts of another, each individual punished would have agreed to be so held. There can be no objection to deterrence on these grounds if each agent has agreed to be held responsible for the acts of others in

this way. Each contractor is like a person who has agreed to stand as a guarantor for another's debts.

On the other hand, contractarian theory captures the greatest strength of the retributive principle, by establishing a kind of moral equivalence between crime and punishment. The benefit requirement demands that each contractor consider both what he gains from protecting the interest in question and what he would suffer if punished. Since the importance of the underlying institution establishes the gravity of the violation for which we punish the offender, the benefit requirement creates a metric for matching offenses with penalties. Moreover, it does so without making the retributive theory's mistake of rejecting deterrence as a legitimate aim of punishment. It is this feature of retributive theories that presumably dooms them to generality, since the notion of desert substituted in its place is unavoidably nonspecific.[83]

The question we must now ask is whether rational individuals would choose to implement the death penalty, given the conditions I have articulated. In our framework, the answer will depend in part on how great the benefits are from that punishment. Each person enters into society because he fears for his bodily security. The security and life expectancy of each person are increased if those who violate society's primary norms are punished, for this will deter other potential criminals from violating those norms as well. Let us call the effects on the bodily security and chances for longevity each person expects from a legal rule his "anticipated security." By including the death penalty in the schedule of available penalties for the worst crimes, each individual will increase his anticipated security, and so it may seem rational to select it.

Each individual, however, must also assess possible punishment from the standpoint of a person subject to that penalty. Thus, each contractor must place himself in the position of a person sentenced to death. Now, if a person thinks it likely that he will receive the death penalty, he will probably not see himself as advantaged by the rule that authorizes its use. It is of course possible that a person subject to the death penalty would have been murdered long before his execution if not for the death penalty's deterrent effect. A rational agent must allow for this possibility. But since, under the benefit principle, a rational contractor will not choose to gamble with rules of the basic structure, he will not base his decision on this contingency. Instead,

[83] I also develop this approach to punishment in Claire Finkelstein, *A Contractarian Approach to Punishment*, *in* THE BLACKWELL GUIDE TO THE PHILOSOPHY OF LAW AND LEGAL THEORY 207, 214–18 (Martin P. Golding & William A. Edmundson eds., 2005).

he will choose to guard against the possibility that he will be executed without adequate compensation in deterrent efficacy. For in such a situation, his anticipated security will not be positive. A rational contractor would therefore choose to reject the death penalty. The following analogy may help make this persuasive.

Suppose a number of people are concerned about the possibility of suffering dual kidney failure and having no access to public organ banks or dialysis machines. To protect against this risk, they contemplate entering into a "Kidney Society," specifying that if any member of the group finds himself needing a kidney to survive, the group would hold a lottery to determine who would supply that individual with the needed kidney. Once chosen, the donor would have no choice but to yield, and a kidney could be removed, by force if need be. Would it be rational for individuals to enter into such an agreement in order to enhance their "expected kidney security?" Ex ante, there is a benefit to a rational agent in entering into such an agreement. If the danger of dual kidney failure is sufficiently great, and the loss to an individual of being the one chosen by lot to donate a kidney is either sufficiently remote or sufficiently bearable, the Kidney Society's agreement confers a net expected benefit. Since the usual way of thinking about benefit is in terms of expected utility, it would be rational for each contractor to enter the society.

But I have been arguing that an ex ante perception of benefit is not sufficient to justify the agreement as one of mutual advantage. I claim that expected benefit calculations are *not* the most rational way of thinking about rules that pertain to the basic elements of a person's welfare, when those elements will be governed by foundational social rules. Instead, the agreement must be reasonably certain to increase advantage to each person under every situation he envisions once the agreement is in place. So if the Kidney Society is part of the basic structure, each agent must still regard himself as benefited in the case in which he is selected at random to provide the needed kidney. Now imagine someone who has his kidney removed after drawing the short straw in the lottery, but who lives the rest of his life without needing anyone else's kidney. Such a person is clearly better off with two kidneys than with one and the benefit that induced him to enter the Kidney Society is one that never accrued. Given the way things worked out, entering the Kidney Society will have turned out to be a bad choice from his perspective.[84] My suggestion, then, is that envi-

[84] One might object that this analysis overlooks the benefit that insurance against dual kidney failure provides, both ex ante and ex post. Even if a person receives no actual benefit from a gamble, he may regard the ex ante chance of benefit as in itself a benefit. I have argued elsewhere for such a claim, and, conversely, that an ex ante chance of harm is

sioning this possibility in advance, a rational agent would not enter the Kidney Society, at least under the conditions just described.[85]

In the case of the death penalty, the contractor's reasoning might go like this. The death penalty provides enhanced deterrence for the very worst crimes, and thus presumably it will be restricted to murder. Each contractor then asks himself whether he would regard the deterrent benefit he had received from living in a society with the threat of the death penalty for murders as increasing his net anticipated security if he were actually subject to that penalty himself. In this case, the protection advantage each agent receives from the death penalty is itself what is removed by the penalty, unlike in the theft example, where these two values are different. The question is therefore quite straightforward: Would each contractor regard himself as experiencing a net increase in anticipated personal security from the death penalty over the course of his lifetime in the case in which he ends up subject to it?

In a world in which no other penalties were available, the death penalty might be selected by the contractors in an initial position of choice. In that case, the alternative to having any punishment would be the state of nature, one that, if Hobbes is to be believed, would be so brutal and insecure that no one could expect to live into old age. Every moment, as Hobbes describes it, would be spent in "continual fear and danger of violent death," and "the life of man solitary, poor, nasty, brutish and short."[86] Relative to the state of nature, even the person condemned to die would regard himself as benefited, given the horror of his life in the absence of such penalties.

Our contractors, however, do not face so stark a choice. Instead, they can compare systems of punishment with the death penalty to those with a range of serious but nonlethal punishments. Since the death penalty is only one in a range of possible punishments, including incarceration and fines, the question the contractors face is: Does the marginal increase in personal security due to the death penalty, when compared with other possible punishments, deter murder so much that it outweighs the marginal loss of personal security a person subject to that penalty would suffer? Here we can see that even in the

itself a harm. This ex ante benefit will balance the cost of losing the gamble, however, only if the actual loss is fairly small and the expected value of the gamble quite significant. *See* Claire Finkelstein, *Is Risk a Harm?*, 151 U. PA. L. REV. 963, 967–74 (2003).

[85] A different kind of kidney lottery might be more appealing. If, for example, the kidney donor were to receive financial compensation for donating a kidney, and if that compensation were sufficient to make the resulting state of affairs on balance beneficial, then signing up for the kidney lottery would not violate the benefit principle.

[86] THOMAS HOBBES, LEVIATHAN ch. XIII, para. 9, at 76 (Edwin Curley ed., Hackett Publ'g 1994) (1668).

unlikely event that each application of the death penalty deterred eight additional murders, the marginal value of that added deterrence would likely be outweighed by the marginal cost of the death penalty. The contractors therefore would reject it.

The possibility that, under unforeseeable circumstances, one could be subject to the death penalty makes adopting it more generally irrational, since the penalty violates the benefit requirement. A contractor would regard himself as worse off for allowing that punishment than if it had never been adopted in the first place. The person subject to the death penalty, like the person who sacrifices a kidney without ever needing another's, has received benefits that will turn out not to have been worth the costs.[87] Thus, since each rational contractor must imagine himself in the position of the losing party in such a gamble, he will choose not to include the death penalty in the roster of available punishments.

IV
Response to Objections

The most significant objection to my argument thus far is that, unlike the case of kidney failure, a person has a choice over whether to commit a crime. If this is true, then whether he risks suffering the death penalty is under his control in a way that suffering kidney failure is not. A rational agent, it seems, would opt for the most stringent penalties for all sorts of crimes, never intending to commit one himself. In that way he would maximize his net anticipated security, since he would benefit from the deterrent effects of the harsh penalties but could be sure that he would never end up subject to them. Indeed, it seems that the higher the penalties, the greater the deterrent benefits and thus the more the agent would benefit. Furthermore, the agent might actually be pleased with the deterrent effect on himself, since the higher the penalties for crime, the less likely *he* would be to commit a crime.[88] Presumably he has a current preference that he not commit crimes in the future. If, by contrast, the penalties for a given crime are too low, he loses both the deterrent benefit with regard to others and increases the likelihood that he himself will commit a crime that will make him subject to the penalty. Does there

[87] The two situations are admittedly different, in that the death penalty does confer deterrent advantages on the individual contractor, whereas the member of the Kidney Society who does not have kidney failure gets no actual benefit other than insurance or the "chance benefit" I discussed earlier in note 84. But in both cases, arguably, the benefit does not compensate for the burden, and thus a rational contractor perceiving this ex ante would not choose either arrangement.

[88] I am indebted to Dan Markel for this point.

not then seem to be ample reason for the contractors to adopt the death penalty?

Rational contractors, however, may still want to guard against excessive penalties in case they are not deterred.[89] Rational individuals are likely to allow for the possibility that they may feel the need to commit a crime in the future, and so they may choose to limit the severity of responses to it. We only need imagine that it might be to an agent's benefit to commit a crime, despite the fact that the agent also views it as beneficial ex ante to make that act a crime. If so, the rational agent might wish to preserve his ability to commit that crime and so would not agree to a penalty as harsh as the death penalty in deciding ex ante how much punishment it deserves. And he might wish to preserve this option, even though he is aware that preserving the option for himself would preserve that same option for everyone else.

This argument may seem perverse, for I am suggesting that rational agents would reject the death penalty because they would want to leave open the possibility that they might someday need to commit a crime for which the death penalty would otherwise be authorized. But I think the point can be made plausible in the following way. Rational agents would eschew social rules that severely restrict or limit their freedom of choice to the extent it is feasible for them to do so. That is, their desire to deter crime must always be balanced against a countervailing desire to protect the range of choices available to them. If the death penalty purchases only a marginal increase in deterrence at the cost of a substantial increase in the coercive powers of the state, it would be rational to reject it. Because the particular identity of the crimes to which the death penalty would be applied remains subject to change, individuals cannot ensure that they are able to protect their freedom where they would most wish for it. Limiting the severity of the punishments that can be inflicted for the most severe crimes is thus a way to blunt the force of undesirable liberty restrictions.

But why draw the line between life in prison and death? What is so special about death? The answer to this question, of course, has to do with what is so special about life, namely that it is the necessary condition for all other benefits an agent might receive. Loss of life is thus normally impossible to offset with other benefits. The most

[89] It is of course possible that a person could be subject to the death penalty without having committed a crime at all. I have assumed throughout that the death penalty could be administered flawlessly. Relax that assumption by allowing even a small chance of error and contractors applying the benefit requirement will have an obvious reason to reject the penalty of death.

obvious way in which this is so is that the person killed cannot have future projects, plans, or pleasures, and thus all considerations of future welfare must come to an end. It is true, of course, that future benefits are not the only kind of benefit that could justify a particular punishment to a rational agent. Past benefits might also provide a justification under the benefit principle. But the only past benefits there could be that would compensate a person for a premature death is the deterrence of what would otherwise be an even more premature death.[90]

Here is a final way to put the point: On a contractarian theory, it is rational to establish a strong system of rights to bodily integrity, rights that cannot be derogated from in specific cases for the sake of short term gains. While future members of society might regard themselves as benefiting from a contract in which others agree to subject themselves to the death penalty on the condition that every other member of society is willing to do the same, such an agreement would conflict with the broader principles of protection for bodily integrity and enforcement of defensive rights that rational members of society would also be concerned to establish. The same, by contrast, need not be said of agreements to be subject to deprivations of liberty. Incarceration leaves the body intact and one's natural life extended. It allows for the continuation of plans and projects of at least a rudimentary sort and does not foreclose challenging one's conviction and perhaps regaining one's liberty. It also allows for the possibility of compensation with future benefits, whether through advancement of personal projects or the bestowing of various pleasures.

A related objection to rational contractarianism, as I have proposed it, has to do with the scope of the individuals that should be included in the initial agreement. On traditional contractarian approaches, those who violate the terms of the contract are thereafter totally excluded from it.[91] On such a view, the contract itself imposes no limitations on what is acceptable to do to violators. Locke, for example, says that those who violate the terms of the contract are like wild beasts; they can be hunted down and killed indiscriminately.[92] And Rousseau says that "every evildoer who attacks social rights

[90] I am assuming, as contractarians typically do, that the rational agents whose consent to social arrangements is sought have "non-tuistic" preferences, meaning that their preferences do not generally take into account the preferences or well-being of other agents.

[91] For a discussion of the contractarian approach to violators' loss of contractual rights, see Christopher W. Morris, *Punishment and Loss of Moral Standing*, 21 CANADIAN J. PHIL. 53, 62–65 (1991).

[92] *See* LOCKE, *supra* note 75, at 14 ("[O]ne may destroy a man who makes war upon him, or has discovered an enmity to his being, for the same reason that he may kill a *wolf* or a *lion*.").

becomes by his crimes a rebel and a traitor to his country; by violating its laws he ceases to be a member of it . . . ."[93] The present objection is just a version of that idea, namely that the contract ought not to include those who are violators or free riders, and so we are entitled to treat such individuals in any way we see fit.

From a certain perspective, the point is quite defensible. If society is a "cooperative venture for mutual advantage,"[94] it makes sense to think of criminals as outside the scope of all voluntary arrangements, since cooperating with them would not be to the advantage of those who remain faithful to the terms of those agreements. Moreover, it arguably makes no sense to include the treatment of contract violators within the terms of the contract itself, since that seems to suppose that we are taking into account the perspective of those who intend not to abide by the terms of our initial contract regarding the basic structure.

But despite these merits, I think the traditional approach to contract violators should 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 at any point after all have agreed to the contract's terms. It is therefore incorrect to equate defection with noncooperation at the outset.[95] Several additional considerations support this approach. First, defections can be large or small, and it may be that it is still advantageous to cooperate with those who defect, as long as their defections are sufficiently minor. Second, it is not possible to address the problem of noncooperation at the outset in any way other than by refusing to contract. But defectors are themselves subject to the terms of an antecedent agreement and can therefore be dealt with contractually.

A final argument against the traditional approach to violators is that 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. We do not normally think of even the most heinous violations as depriving 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 appropriate levels of bodily dignity and comfort. It is true that nonrational creatures are often

---

[93] JEAN-JACQUES ROUSSEAU, THE SOCIAL CONTRACT AND THE FIRST AND SECOND DISCOURSES 177 (Susan Dunn ed., Yale Univ. Press 2002) (1762).

[94] RAWLS, *supra* note 49, at 4.

[95] *See* Claire Finkelstein, *Hobbes and the Internal Point of View*, 75 FORDHAM L. REV. __ (forthcoming 2006, manuscript on file with author) (arguing for rationality of adhering to social contract based on common knowledge of rationality).

thought of as possessing a subset of these same rights, and we cannot think of *them* as parties to a social contract. This suggests a basis for assigning rights to biological agents outside the contractual context. But the protections afforded such creatures are thought to be significantly weaker than those extended to even the worst criminals. 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 they make to enforce the terms of cooperative interaction.

Only by including potential violators in the social contract can the contractarian model provide any practical guidance to a theory of punishment. This allows us to capture within a contractarian framework the basic deontological intuitions that 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. In the contractarian approach I have proposed, however, these elements complement each other without contradiction.

CONCLUSION

I have argued that the contractarian account of punishment is best situated to meet the most fundamental requirement of a theory of punishment: the need for a sufficiently robust justification for punishment that will overcome an initial presumption against it. A rational-choice contractarian account can meet this justificatory hurdle better than the two traditional accounts of punishment, as it restricts legitimate punishment to voluntarily imposed treatment. Such a requirement does not imply that any treatment to which the recipient consents is justified. Rather, as I have argued, it asserts that a treatment to which the recipient consents is permissible if it is also to the agent's benefit.

The reason that contractarian agents agree to live in the shadow of any particular form of punishment is that they anticipate that significant benefits will flow to them from the deterrent effects it brings. Their willingness to subject themselves to the risk of this harsh treatment comes from their belief that they receive greater security from the threat of punishment than they would lose, even if they are subject to the penalty whose threat they desire as a means of deterring others. But it is just as mistaken to think that this benefit could by itself justify a scheme of punishment as it is to suppose that consent alone could do

so. Neither consent nor benefit standing alone has sufficient normative salience to overcome the initial presumption against punishment with which we began. What we have seen instead is that the consent of the agents who must live under a system of punishment confers legitimacy on that punishment only if their consent is based on a perception that they will be better off under such a system than they would be in its absence.

Although contractarian arguments should naturally fare better than deterrence and retributive theories of punishment, we have also seen that the prevailing contractarian accounts are flawed. Voluntarist accounts, on which mere consent is thought sufficient to justify punishment, make the mistake of thinking that consent alone can confer adequate legitimacy on a system of punishment. At the other extreme, normative contractarianism suggests that our considered intuitions about fairness, when suitably elicited through a stylized original position, should provide adequate justification for a system of punishment. But this account fares little better than its normative counterpart in the retributivist tradition, because it provides no basis for thinking of such a system as consensual and therefore no basis for thinking of the resulting institution as benefiting from the lower hurdle of justification that consensual systems can claim.

Unlike deterrence and retribution, the contractarian enterprise makes each person the guarantor for every other person's conformity to law. A person who violates rules of conduct agrees to suffer punishment to ensure that others do not thereby think themselves free to do as he did. Their willingness to abide by the rules of the institution from which they benefit is itself instrumentally motivated: Each agent agrees to hold himself accountable in a certain way in order that others will hold themselves accountable in that same way. It is this core idea, as I understand it, that makes contractarianism both attractive and normatively powerful.

As we have seen, a natural thought to have about the death penalty against the background of such a theory is that it is easy to justify: The more deterrence the better, since the more each agent can enhance his anticipated security through the availability of a given penalty, the more inclined each would be to live under its shadow himself. Rational contractarianism thus offers the most promising avenue for justifying the death penalty, at least as compared with the deterrent and retributivist alternatives we considered. But I have tried to show that even this argument in favor of the death penalty is problematic, since each agent must imaginatively project himself into a world in which he himself is subject to the death penalty, and must still regard the death penalty as a benefit under such conditions. Put

otherwise, the contractarian argument in favor of the death penalty requires us to eliminate potential criminals from the scope of the social contract, and I have offered reasons to think we ought not proceed in this way.

The thought here does not seem too very distant from Hobbes's observation that no man can be understood as having transferred away his right to self-defense, "because he cannot be understood to aim thereby at any good to himself."[96] Of course, Hobbes's claim seems implausible on its face, since we can imagine various scenarios in which an agent comes to see himself as benefiting from giving up self-defensive rights. Yet Hobbes may have had a less literal understanding of the right to self-defense in mind, one that recognized that no man can rationally and voluntarily abandon that right, given that defending his life is the point and purpose of all that he does.[97] The basic thought I have tried to develop is Hobbesian in this sense: Since the purpose of punishment is the enhancement of one's bodily and material security, there is an internal limitation to the amount and severity of punishment rational agents would choose to include in the basic structure of society. This does not mean that the death penalty would be rejected under any conceivable empirical circumstances. But it does suggest that rational agents would reject the death penalty's inclusion in contemporary criminal justice systems, and they would continue to reject it under any circumstances we can currently and reasonably imagine.

---

[96] HOBBES, *supra* note 86, ch. XIV, para. 8, at 82.

[97] For a more complete discussion of Hobbes's point, see Claire Finkelstein, *A Puzzle About Hobbes on Self-Defense*, 82 PAC. PHIL. Q. 332 (2001).