Punishment as Contract

Claire Finkelstein*

This paper provides a sketch of a contractarian approach to punishment, according to a version of contractarianism one might call "rational contractarianism," by contrast with the normative contractarianism of John Rawls. Rational contractarianism suggests a model according to which rational agents, with maximal, rather than minimal, knowledge of their life circumstances, would agree to the outlines of a particular social institution or set of social institutions because they view themselves as faring best in such a society governed by such institutions, as compared with a society governed by different institutional schemes available for adoption. Applied to the institution of punishment, a rational contractarian approach maintains that members of society would reach broad agreement with one another concerning the outlines of a system of punishment, based on the fact that they would regard themselves as benefitting from the deterrent effect of such a system. But they would balance the deterrence benefits of such a system with the incursions any scheme of punishment makes into personal liberty. Rational agents would adopt that scheme of punishment that maximizes marginal deterrent benefit without unduly burdening individual liberty.

The paper also suggests that a rational contractarian approach is able to capture the best insights of the two leading alternative theories of punishment: deterrence theory and retributivism. On the one hand, rational contractarianism shares the deterrence view that the guiding aim of any punishment scheme must be the deterrence of crime, where a crime is an action that violates the background social contract. On the other hand, rational contractarianism solves the central problem associated with pure deterrence theories—the problem that punishment on this view involves "using" individuals for the sake of achieving the general social goal of deterrence. It does so by maintaining that the way in which the aim of deterrence is incorporated into punishment theory is not

* Algernon Biddle Professor of Law and Professor of Philosophy, University of Pennsylvania. My thanks to Larry Crocker, Marc Fleurbaey, Jim Jacobs, Leo Katz, and David Velleman for their detailed and helpful comments. I also wish to thank the members of the Criminal Law Theory Symposium, who discussed this Essay in their January 24, 2011 meeting, the participants in the Hoffinger Colloquium at New York University Law School on March 23, 2009, and the participants in the Penn Institute for Law and Philosophy Prioritarianism Workshop on January 23, 2010.

premised on total, or even average, social utility, but on the assent of each individual to the scheme by which such deterrent ends are pursued. The criteria for the rationality of assent for each contractor is that the individual regards himself as benefitting on balance from the punishment scheme. A rational contractarian scheme of punishment thus renders the actual punishment of offenders under the rules of the system voluntary, in that each rational member of society has given his own prior agreement to be governed by the punishment institution in the event that he ends up committing a crime. A voluntary punishment scheme avoids the problem of "using" individuals for the sake of deterring other agents, because it represents instead the decision of each rational contractor to allow others to hold him to a set of agreed upon consequences for violations of the social contract. The aim of deterrence, therefore, does not cause the theory to "travel across persons" in the way that deterrence theories do.

I. INTRODUCTION

In *Crito*, Socrates discourses with a former student in his prison cell, where he awaits his execution.[1] Crito has come to implore Socrates to submit to a plan to secure his escape from prison.[2] Socrates, who had been tried and convicted of the charge of "corrupting the minds of the young" of Athens,[3] steadfastly resisted the option of banishment during his trial, as he believed himself innocent of the charges against him. He now sees himself as bound to submit to his sentence, despite his equally firm conviction that it represents a miscarriage of justice. He rejects all arguments to the effect that such miscarriage entitles him to evade the State's verdict and violate its laws, even under threat of death.[4]

Socrates's argument to Crito is worth attending to, for it is not commonly heard in contemporary discussions of punishment.[5] In brief, it is that he, Socrates,

---

[1] PLATO, *Crito*, *in* THE COLLECTED DIALOGUES OF PLATO 27 (Edith Hamilton & Huntington Cairns eds., Hugh Tredennick trans., 1961).

[2] *Id.* at 29.

[3] PLATO, *Socrates' Defense (Apology)*, *in* THE COLLECTED DIALOGUES OF PLATO 3, 10 (Edith Hamilton & Huntington Cairns eds., Hugh Tredennick trans., 1961).

[4] PLATO, *supra* note 1, at 31–39.

[5] Two other recent philosophical articles about punishment draw on *Crito* as a source of inspiration for contemporary punishment theory. Those articles, however, make use of the dialogue to underscore the existence of a right of resistance, in opposition to Socrates's stance in *Crito*. Larry May uses the dialogue to suggest the parameters of legitimate disobedience to legal orders. He writes: "[U]nless Socrates's act of disobedience was intended to frustrate the end of peace in Athenian society, his act [of disobedience] may be justified." *See* Larry May, *Hobbes on Fidelity to Law*, 5 HOBBES STUD. 77, 87 (1992). Alice Ristroph uses *Crito* as a foil for what she advances as a Hobbesian view of punishment. Contrary to Socrates's conciliatory stance in the dialogue, Ristroph wishes to argue that "[h]ad Socrates agreed to escape with Crito, Hobbesian respect would have

has entered into an agreement with the State to abide by its laws, in exchange for which he has enjoyed all the benefits of Athenian citizenship, such as begetting, rearing, and receiving education for his children in Athens.[6] Admittedly, his consent to this agreement has been more implicit than explicit, despite his protestations to the contrary.[7] But it is manifested in the fact that in his seventy years in Athens he has had ample opportunity to express his dissatisfaction with the State by quitting Athens for a different state. By choosing to stay, he has manifested his acceptance of the burdens of Athenian citizenship along with its benefits.[8] This argument he places in the mouth of a personified version of the Laws of Athens, who address him in the following terms:

> [A]ny Athenian, on attaining to manhood and seeing for himself the political organization of the state and us its laws, is permitted, if he is not satisfied with us, to take his property and go away wherever he likes. . . . On the other hand, if any one of you stands his ground when he can see how we administer justice and the rest of our public organization, we hold that by so doing he has in fact undertaken to do anything that we tell him.[9]

From this it follows, say the Laws, "[t]hat if you cannot persuade your country you must do whatever it orders, and patiently submit to any punishment that it imposes, whether it be flogging or imprisonment."[10] The obligation to abide by a juridical verdict, Socrates explains, is like the duty to do military service for one's country: "Both in war and in the law courts and everywhere else you must do whatever your city and your country command, or else persuade them in accordance with universal justice . . . ."[11]

Why did Socrates's conception of punishment as civic duty disappear from public discourse? And why, in particular, did it fail to make an appearance in the punishment theory of later years? One reason is surely that the conception of civic duty it advances stands in some tension with the rather more individualistic foundations of the contemporary ideal of citizenship. Few modern writers would defend the extreme fidelity to the State's dictates Socrates seems to be advancing. Socrates *could* have taken a more moderate position: that citizens who accept the

---

recognized this action as a blameless exercise in self-preservation." Alice Ristroph, *Respect and Resistance in Punishment Theory*, 97 CALIF. L. REV. 601, 628 (2009).

[6] PLATO, *supra* note 1, at 35–36.

[7] *Id.* at 37 (noting that "there are very few people in Athens who have entered into this agreement . . . as explicitly as I have").

[8] *Id.* at 37–38.

[9] *Id.* at 36–37.

[10] *Id.* at 36.

[11] *Id.*

benefits of membership in the State commit themselves to abide by those verdicts that are just and fair, and that they are released from debt of allegiance when the State violates its own basic norms. It is curious, however, that even this more limited version of the Socratic thesis did not make any systematic impression on the punishment theory of later years.[12]

Instead, the modern debate coalesced virtually entirely around two positions: The retributivist theory that a person should be punished only when and to the extent that *he* deserves to suffer for the harm he has inflicted, and the utilitarian claim that punishment is desirable from a social perspective only insofar as its infliction would help deter the commission of future offenses, namely the deterrence theory.[13] The former position takes perpetrators one at a time, in that it focuses on the moral standing of the individual perpetrator in light of his act. The latter, by contrast, eclipses the individual in favor of the collective and determines the legitimacy of the decision to punish in terms of its social welfare effects. As is well known, contemporary punishment theory has become a protracted debate between the perspective of individual justice assumed by retributivists and the perspective of social justice assumed by utilitarians. In the process the contractarian perspective on which Socrates premised his argument for the civic virtue of punishment was eclipsed.

The absence of any well-developed contractarian theory of punishment seems all the more puzzling in light of two salient facts: First, there is a robust contractarian tradition that emerged in seventeenth century political philosophy, first with the writings of Thomas Hobbes,[14] later in the Enlightenment version of this same tradition in the writings of Locke[15] and Rousseau,[16] and finally in a Kantian version of the tradition, as developed by John Rawls.[17] The absence of a systematic contractarian alternative to retributive and utilitarian theories of punishment is especially surprising in view of the breadth and depth of the contractarian school of thought in political theory. There are of course hints here

---

[12] Again, May and Ristroph have taken note of the theory, but they appear to reject its central claim concerning the basis for obedience to the Laws. *See supra* note 5. One possible exception lies in the writings on punishment by Jeffrie Murphy. *See, e.g.*, JEFFRIE G. MURPHY, RETRIBUTION, JUSTICE, AND THERAPY: ESSAYS IN THE PHILOSOPHY OF LAW 100 (1979) ("The criminal himself has no complaint, because he has rationally consented to or willed his own punishment.").

[13] For a general discussion of the distinction between deterrence justifications and retributive justifications, see Claire Finkelstein, *A Contractarian Approach to Punishment*, *in* THE BLACKWELL GUIDE TO THE PHILOSOPHY OF LAW AND LEGAL THEORY 207, 208–14 (Martin P. Golding & William A. Edmundson eds., 2005).

[14] *See* THOMAS HOBBES, LEVIATHAN (C.B. Macpherson ed., Penguin Books 1968) (1651).

[15] *See* JOHN LOCKE, TWO TREATISES OF GOVERNMENT (Peter Laslett ed., Cambridge Univ. Press 1988) (1690).

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

[17] *See* JOHN RAWLS, A THEORY OF JUSTICE (rev. ed. 1999).

and there as to what such an account might look like—Hobbes's own discussion of punishment is all too brief, and not particularly satisfactory, but still there is a foundation laid. Some in recent years have attempted to articulate a Rawlsian version of a contractarian theory of punishment, but such theories are more deontological than contractarian.[18] The possibility of a truly contractarian approach to punishment is as yet substantially unexplored. It is the hope of this Essay to begin to sketch the outlines of such an account.[19]

Second, the policy questions at the heart of criminal justice debates are all about the proper scope of deontological values, and how these values relate to the sorts of utilitarian considerations the deterrence theorist advocates. Recent discussions about the permissibility of torture illustrate the proposition particularly vividly: The debate pits the overwhelming utilitarian pressures of military necessity against the deontological intuition that human beings have rights and that these rights are not *entirely* extinguished by membership in a group pledged to destroy others. Procedural rights, such as the presumption of innocence as well as doctrines like proportionality, support the suggestion that rights function as deontological side constraints on the treatment of even the worst criminals, a fact that deterrence theories cannot accommodate.

In the Bush Administration, utilitarian thinking on this question predominated.[20] President Obama had, by contrast, pledged his fidelity to the deontological position that, as he said in his address to Congress, "the United States of America does not torture,"[21] though the subsequent behavior of his administration have made such claims hard to credit.[22] The standoff we have seen in world opinion on this question bears witness to the conclusion of moral and legal philosophers that human rights cannot be respected if they are balanced off against considerations of utility, even if such considerations are marshaled for the sake of guarding against human rights violations of another sort. The same can be said of punishment as prevention: The idea that retributive values can somehow be combined with, and balanced off against, utilitarian considerations pertaining to punishment reform is a fantasy. All attempts at balancing collapse into a kind of inconsistent exchange between the rights of those suspected of crimes, on the one

---

[18] *See, e.g.*, Sharon Dolovich, *Legitimate Punishment in Liberal Democracy*, 7 BUFF. CRIM. L. REV. 307 (2004) (using Rawlsian ideas to generate an account of punishment).

[19] I make a start in Claire Finkelstein, *A Contractarian Argument Against the Death Penalty*, 81 N.Y.U. L. REV. 1283 (2006).

[20] *See* Claire Finkelstein, *Vindicating the Rule of Law: Prosecuting Free Riders on Human Rights*, *in* WHEN GOVERNMENTS BREAK THE LAW: THE RULE OF LAW AND THE PROSECUTION OF THE BUSH ADMINISTRATION 37 (Austin Sarat & Nasser Hussain eds., 2010).

[21] President Barack Obama, Address to Joint Session of Congress (Feb. 24, 2009), http://www.whitehouse.gov/the_press_office/Remarks-of-President-Barack-Obama-Address-to-Joint-Session-of-Congress; *see also* Finkelstein, *supra* note 20.

[22] *See* Claire Finkelstein, *Targeted Killing as a Pre-emptive Practice*, *in* TARGETED KILLING: LAW AND MORALITY IN AN ASYMMETRICAL WORLD (forthcoming 2012).

hand, and the societal urgencies that seem to require the curtailment of such rights, on the other. It is the job of legal philosophers to help import clarity about the structure of our moral values and their degree of inviolability, with the hope that it will help policy makers face up to their underlying normative commitments squarely.

My point of departure will be an assumption that has become standard in the punishment theory literature. Because it involves the deprivation of personal liberty and the infliction of physical hardship, punishment is *presumptively* impermissible.[23] The practice of punishment therefore stands in need of justification if the background moral objections to it are to be overridden. Compare this to contract law, where the justification threshold for enforcing contracts is much lower, given that each party to a contract has voluntarily undertaken to allow the other party to sue to enforce the contract should he fail to make good on his commitments. Indeed, barring objectionable third-party effects, or paternalistic concerns, there is a presumption *in favor* of the enforceability of consensual arrangements, and hence a need to justify the refusal to enforce a contract. Penalties for civil wrongs lie somewhere in between these two extremes: They involve a lower justificatory threshold than criminal penalties, given the comparatively less invasive nature of the penalty, but they are not as easy to justify as contractual arrangements.

The high justificatory hurdle for our practices of punishment provides a reason to return to the forgotten contractarian approach to punishment: If it is easier to justify the enforcement of voluntary arrangements than involuntary ones, a theory of punishment that convincingly predicates a consensual foundation for the institution should depict the institution as easier to justify than other types of theories. If in addition, as I have argued elsewhere,[24] the retributive and utilitarian approaches to punishment have failed to meet *their* justificatory burdens, we have yet further reason to turn to an account predicated on the idea of punishment as consensual. In what follows, I will first briefly summarize the reasons for my claim that retributivism and utilitarianism have thus far failed to provide

---

[23] *See, e.g.*, R.A. DUFF, TRIALS AND PUNISHMENTS 1 (1986) ("It is agreed that a system of criminal punishment stands in need of some strenuous and persuasive justification . . . ."); H.L.A. HART, *Prolegomenon to the Principles of Punishment*, *in* PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW 1 (2d ed. 2008). By contrast, Mitchell Berman has recently argued that the standard assumption that punishment stands in need of justification is denied on at least one theory of punishment. Properly understood, retributivism should be advanced as the claim that punishment of a guilty offender is a moral *good*, rather than an objectionable violation of his rights, and that as such it stands in need of no justification. Mitchell N. Berman, *Punishment and Justification*, 118 ETHICS 258 (2008). Kantians seem to reject the assumption that punishment is presumptively impermissible because they understand punishment as requiring *authorization* rather than *justification*. *See* ARTHUR RIPSTEIN, FORCE AND FREEDOM: KANT'S LEGAL AND POLITICAL PHILOSOPHY 300–24 (2009).

[24] *See* Finkelstein, *supra* note 19.

justifications for current criminal justice practices, and I will then attempt to sketch a contractarian alternative that, as I see it, is exempt from these deficiencies.

II. PROBLEMS WITH THE TWO DOMINANT THEORIES

There have been many critiques of both retributivism and deterrence theory over the years,[25] and many responses to each of these critiques. For present purposes, however, we can abbreviate what would otherwise be a lengthy discussion by focusing on what I would suggest constitutes the central drawback of each—the one that is most indicative of its deficiencies. Start with retributivism: It is by now a familiar point among punishment theorists that the notion of "desert" around which retributivism revolves is a highly ill-defined notion, one that does not readily lend itself to translation into a precise metric for punishment. The biblical suggestion for how to understand what the notion of "desert" requires is of course the concept of *lex talionis* or "eye for eye, tooth for tooth."[26] But it is hard to see what *lex talionis* entails in the face of sadistic, pleasure-seeking defendants like Patrick Kennedy, who brutally raped and injured his eight-year-old stepdaughter, leaving her near death from loss of blood.[27] Does giving a defendant like Kennedy the equivalent of the suffering he inflicted on his victim mean turning him over to the ravages of a comparable pleasure-seeking maniac who happens to have a penchant for middle-aged men? Retributivists uniformly reject this possibility,[28] but they have little to put in its place.

Furthermore, let us suppose a suitable moral equivalent for Kennedy's crime can be found that is both an appropriate form of punishment and, we are confident, represents the subjective equivalent of the suffering his victim must have experienced. Can the retributivist metric be defended in this form? Consider an argument against retributivism raised by Louis Kaplow and Steven Shavell.[29] Let us suppose, they say, that we always punish each offender exactly as much, and no more than, he *deserves*.[30] Assuming that not every offender will be caught and

---

[25] *See, e.g.*, PAUL H. ROBINSON, DISTRIBUTIVE PRINCIPLES OF CRIMINAL LAW: WHO SHOULD BE PUNISHED HOW MUCH? (2008) (pointing out the deficiencies of theories on both sides of this debate).

[26] *Deuteronomy* 19:21; *Exodus* 21:24; *Leviticus* 24:20.

[27] *See* Kennedy v. Louisiana, 554 U.S. 407, 412-15 (2008).

[28] *See, e.g.*, Jeremy Waldron, *Lex Talionis*, 34 ARIZ. L. REV. 25, 25 (1992):

> [*Lex talionis*] cannot be thought to require that *the very same action* that constituted the offense should be visited as punishment upon the offender. Rather, the requirement must be that the act of punishment be *similar* to the offense in certain respects. Which respects these should be is a matter of normative argument.

[29] LOUIS KAPLOW & STEVEN SHAVELL, FAIRNESS VERSUS WELFARE (2002).

[30] *Id.* at 301–03.

punished,[31] as is surely the case, the amount of punishment meted out in society as a whole would then be insufficient to achieve effective deterrence. Why? Assume, as a rough approximation, that the rational criminal would be deterred by receiving somewhat more, by way of punishment, than the suffering he inflicted on his victim. If he can discount the gravity of the threatened punishment by the thirty, forty or fifty percent chance he will not actually be caught, he will have inadequate incentive to refrain from committing such offenses and will not be deterred.[32]

At first blush, Kaplow and Shavell's point is compelling. If we assume an offender's deserved punishment to be a rough match for the level of punishment that would deter him, then it is true that less than perfect detection would bring the effective punishment to less than the deterrent level. But why on earth should we assume this? Why assume that the utility a rational criminal receives from committing a crime is equal in value to the *disutility* he inflicted on this victim? Nevertheless, it is this assumption on which Kaplow and Shavell rely to connect "desert" with deterrence, as desert is tied to the *victim's* disutility, in their model, and deterrence is a function of the *criminal's*.[33] Since the connection between the criminal's utility and the victim's disutility is unwarranted, however, their point about the relation between desert and deterrence is as well.

One reason for focusing on Kaplow and Shavell's arguments in favor of deterrence theory, despite the obvious difficulties with the account, is that it helps to underscore the degree to which considerations of *desert* have nothing to do with deterrence. Thus if one adopts a desert-based approach to punishment, there is no guarantee that we will set punishment at levels designed for optimal deterrence. And this implies that retributivists who insist on the desert criterion for punishment will be forced to accept the inappropriateness of the goal of deterrence. Indeed, as we saw earlier with the example of torture, their wholesale rejection of deterrence as a social goal must be so thorough that they cannot even count the goal of reducing the number of wrongful acts in society as one legitimate goal among others. As Kaplow and Shavell put the point: "[B]ecause retributivists ignore deterrence and thus changes in the number of wrongful acts that are committed, they by the same token ignore changes in the number of occasions on which wrongdoers unfairly go free."[34] The point, they insist, is that "retributive notions of fairness are associated with indifference to the number of instances of unfair treatment."[35] And they suggest that this "is in tension with the demand of retributive justice that fair punishment be imposed on everyone who commits a

---

[31] *Id.* at 309.

[32] *Id.* at 311–13.

[33] *Id.* at 292–93.

[34] *Id.* at 313.

[35] *Id.*

wrongful act."[36] Retributivism thus appears internally inconsistent: If retributivists truly care about rights violations, they should want to minimize the occasions on which rights are violated across society, and that means seeking to deter such violations through an effective scheme of incentives that operates by making an example of current offenders. But since retributivists cannot consistently posit deterrence of *anything* as a goal of the theory, pure retributivism is a failure.

Retributivists would reject the suggestion of internal inconsistency for they would say that corrective justice is not aggregative—it is governed by case-by-case considerations. So the fact that retributivists do not regard it as justified to seek to minimize the number of rights violations by deterring future violations does not stand in tension with their rejection of the legitimacy of violations of rights. This is the sense in which retributivism is ineliminably deontological: Even if punishing an offender more than he deserves would help to reduce the chances that victims would have their rights violated in the future in ways that *they* do not deserve, such a basis for punishment remains impermissible. But unfortunately, this response, which shows that the retributivist has the courage of his convictions, does not serve to vindicate the theory. For if retributive theory cannot, by its own admission, accommodate the social goal of minimizing the number of rights-violations in society, the theory is fundamentally ill-equipped to provide guidance on important matters of criminal justice policy.

Because they do not seem aware of any options other than deterrence or retribution, it is not surprising that Kaplow and Shavell regard this negative argument as largely clinching the case for deterrence-based accounts in some form. Presumably for this reason they do not feel the need to address the significant weaknesses of deterrence-based accounts. But such weaknesses are not far to seek. Let us turn, then, to deterrence theories to give them *their* just deserts.

Begin with a small, though fundamental, problem the deterrence theorist faces. Let us return to the erroneous point made by Kaplow and Shavell, namely that *the pain inflicted on the victim is equivalent to the pleasure or benefit the criminal receives from committing his crime*. This error in thinking, which turned out not to be problematic for the retributivist after all, *will* pose problems for the deterrence theorist. Since deterrence operates on the incentives of the perpetrator, rather than on the harm inflicted on the victim, deterrence theory is ineliminably tied to a factor that has little to do with the social harm criminal activity imposes. Matters are of course significantly different in the standard economic account of tort law, where the undesirability of the conduct we are seeking to deter is entirely a function of the harm it inflicts. In this context, it is plausible to set liability levels by forcing potential tortfeasors to internalize the costs of their activities, thereby creating incentives to induce tortfeasors to desist from, or take precautions against, incidental harm that results from productive activities in which they are engaged.[37]

---

[36] *Id.*

[37] *See id.* at 85–154.

We would therefore have no reason, in at least the standard case in tort law, to establish a scheme of deterrence that was not tailored to the harmfulness of the underlying activity. But in criminal law, the wrongness of the activity is defined separately from the harm it produces, or the pain it inflicts on its victim, and hence the aim of deterrence cannot be measured by the harmfulness of the conduct.[38]

In criminal law, we have a list of conduct that is judged to be *per se* undesirable, and for which the optimal activity level is effectively zero.[39] And this means that since we know we want to eliminate as much of this conduct as possible, within the bounds of our allowable resources for crime prevention, our interest in deterrence narrows our focus exclusively to the incentives to the criminal, forgetting all other concerns, such as the level of harm associated with the activity. But since crimes that bring great benefit to the perpetrator will require heavier penalties to deter effectively than crimes that bring little benefit or pleasure to the perpetrator, it follows that crimes that cause relatively little harm to the victim may actually require *higher* penalties than crimes that cause great pain and suffering. We would have to punish theft more severely than homicide, for example, if it turned out that the gains to criminals from theft were significantly higher than the gains from homicide. The moral severity of the offense simply does not play a role in determining the appropriate level of punishment in a deterrence account. The deterrence theory of punishment is thus potentially out of sync with rather deeply-felt intuitions about the gravity of harm and hence about appropriate social treatment of offenses.

Second, deterrence theorists to date have no adequate response to the kind of argument that is typically leveled against them by retributivists, namely the standard objections of deontologists to utilitarian theories.[40] Retributivists object to the fact that deterrence theorists seek to justify the punishment of one person in terms of the effect such punishment would have on a wholly different, uninvolved other person at some point in the future. This objection has taken many guises over the years. Retributivists say that deterrence theorists are committed to the proposition that it would be permissible to punish one person to deter a larger number of other people from committing crimes, *even* if the one was not himself guilty of committing any crime. They thus cannot justify restricting punishment to

---

[38] *See* Heidi M. Hurd & Michael S. Moore, *Negligence in the Air*, 3 THEORETICAL INQUIRIES L. 333 (2002).

[39] There is of course debate about this in economic writings on criminal law. Gary Becker's famous insight was that because achieving zero activity levels would require a level of resource allocation to law enforcement that would be sub-optimal, we can say that there is a "desirable," non-zero activity level for every crime. This model effectively brings the theory of criminal deterrence closer to the standard economic account of tort law. But the reason for favoring non-zero activity levels is of course significantly different in the two accounts, and Becker is not asserting that a non-zero level of criminal activity is desirable in and of itself, in the absence of resource considerations. *See* Gary S. Becker, *Crime and Punishment: An Economic Approach*, 76 J. POL. ECON. 169 (1968).

[40] *See* Finkelstein, *supra* note 19.

the guilty. Alternatively retributivists point out that deterrence theorists are unable to account for any requirement of proportionality even as against guilty offenders, as disproportionate punishment may be required for optimal deterrence. (This in effect was the point I made against Kaplow and Shavell earlier.) Finally, retributivists sometimes argue that even if deterrence theory could limit punishment to the guilty, and even if it could somehow insist on penalties that were proportionate to that guilt, deterrence is still morally unacceptable because it amounts to *using* offenders for the sake of the achievement of social welfare goals, and this fails to respect their humanity. This basic objection I have put elsewhere in terms of the judgments of responsibility that are implicit in the deterrence theorists' approach to punishment: Deterrence, as standardly argued for, is a justification for punishment that *travels across persons*, since it purports to hold one agent responsible in order to deter future acts of responsibility of other agents, and as such conflicts with fundamental intuitions of fairness to which our criminal justice system is committed.[41]

Furthermore, and most relevant from the standpoint of the contractarian approach we will consider shortly, deterrence arguments do not take a form that the offender himself would likely regard as providing a justification for his punishment. After all, he might argue, he surely has a right to be punished in light of considerations that pertain to his *act alone*, whatever form such considerations ultimately take. The various attempts deterrence theorists have made to accommodate the deontological concerns—such as that deterrence will not work if it is not fundamentally tailored towards guilty offenders or is significantly out of keeping with retributive intuitions, and so forth—seem largely to miss the mark, as they fail to answer the demand for individualized justification that an offender may rightly have. Yet once again, the need to justify the treatment of an offender *to* that offender in terms that are particular to that person's situation—and the fact that retributivism can meet that demand far better than deterrence theories—does not clinch the case *for* retributivism any more than the inability to accommodate the most basic needs of crime control policy clinches the case for deterrence theory.

It is hard to avoid the conclusion that neither retributivism nor deterrence theory is ultimately equipped to justify the practice of punishment in something resembling its current form. Each requires the suppression of strongly-felt intuitions that only the other seems able to accommodate, and yet the history of efforts to marry retributive and utilitarian considerations in a mixed theory of punishment have been equally unsuccessful, as they must give primacy to one rationale for punishment or the other, and as such remain subject to the fundamental objections to each.[42] Once we understand that a justification for

---

[41] *See id.* at 1299.

[42] For a mixed theory of punishment, see H.L.A. HART, PUNISHMENT AND RESPONSIBILITY: ESSAYS IN THE PHILOSOPHY OF LAW (2d ed. 2008). Hart argues that while deterrence is the "General

punishment must be able to combine the need for social control, on the one hand, with providing an offender with a justification for his treatment that he himself can recognize as legitimate and that neither standard rationale for punishment can accomplish this, the search for an entirely different sort of account of punishment becomes compelling.

III. TOWARDS A CONTRACTARIAN APPROACH TO PUNISHMENT

To summarize the argument thus far: I have argued, against retributivists, that a theory of punishment that gives no weight to considerations of deterrence is unable to serve as a guide for actual questions of criminal justice reform. And I have argued, against utilitarians, that a theory that is unable to provide an individualized justification to the criminal for his punishment and instead seeks to justify *his* treatment by its effects on another agent is morally unacceptable. The contractarian approach to punishment holds out the hope of a conjoined solution to these problems: It combines the social aim of deterrence with an individualized approach to the justification for imposing punishment on a particular agent, thus providing the criminal with an argument for his own punishment that *he* can accept, at the same time that it establishes a realistic basis for institutional planning. We will now consider in detail how a contractarian theory might accomplish these aims.

The prospect for a more adequate theory of punishment lies largely in the contractarian's insistence that punishment be voluntarily imposed. Assume that a given punishment scheme has at least moderately strong deterrent efficacy. A rational contractor would agree to live in a regime that furnishes this level of deterrence to serious crimes, and hence would prefer it to one in which such deterrence is absent. As the justificatory burden for consensual arrangements is particularly low, it should be easier to justify the infliction of punishment on such an account than on any other.[43]

I am *not* arguing that a consensual approach to punishment can justify the infliction of punishment merely by reference to the idea, if true, that citizens have consented to the scheme of punishment under which they must live. Consent

---

Justifying Aim" of punishment, *id.* at 8, the legitimate pursuit of deterrence must be limited by more general moral constraints, such as constraints on punishing the innocent as well as principles of proportionality. *See id.* at 1–27. The problem is that it is unclear what the relationship is between the general justifying aim of punishment and the foregoing retributive side-constraints, a problem that afflicts mixed theories generally. In more recent writings, Paul Robinson advances a mixed theory of punishment, according to which punishment should be distributed according to the empirical beliefs the general public has about desert. But the ultimate justification for a principle that distributes punishment according to public opinion about desert is that distributing punishment in this way has important crime control consequences. The theory thus appears to be more of a straight forward deterrence account than a truly mixed theory. *See, e.g.*, ROBINSON, *supra* note 25, at 135–74.

[43] *See* C.S. Nino, *A Consensual Theory of Punishment*, 12 PHIL. & PUB. AFF. 289 (1983).

considered by itself does not have such normative force, as is reflected in the fact that the criminal law rejects consent as a defense to most crimes, most notably to murder. Although consent is a defense to some crimes, such as rape and battery, it is limited in its operation even in these cases to situations in which the victim does not suffer harm. A consensual theory of punishment, then, must be prepared to explain the relevance of consent to its account of the justifiability of punishment. As Socrates suggested in his argument to Crito, it is not consent alone that justifies punishment, but consent premised on the benefit the citizen who consents to abide by the State's dictates takes himself to be receiving in the bargain.[44] Thus while neither benefit from a scheme of punishment nor consent to its terms would be sufficient by itself to justify the imposition of punishment on a particular offender, the combination of benefit and consent may be a different matter.

It might be thought that a contractarian approach only mirrors the basic utilitarian account inside the structure of a social agreement. But this impression would be incorrect. A utilitarian social agreement, such as Harsanyi might have recommended, would discount the costs and benefits of a deterrence scheme.[45] Contractors in this sort of world would merely ask whether they could expect to fare better under a system of deterrence than without it, taking into account the expected benefits of such a system and discounting them by the expected costs. Agreement in the contractarian tradition, by contrast, produces different results, for it is subject to several critical assumptions:

1. Rational contractarians assume that human beings are rational in the sense that they are primarily interested in maximizing their welfare and their preferences are generally not other-regarding.
2. They assume that each has knowledge of each other's rationality and, further, that each has knowledge of each other's knowledge of his rationality. This is the so-called "common knowledge" assumption.[46]
3. They assume that although rational, human beings are highly risk-averse when it comes to fundamental aspects of their welfare. With regard to institutions that apply to what Rawls would call the "basic structure of society,"[47] they would seek to assure themselves of faring better in their post-agreement condition than they did according to their pre-agreement baseline. I shall refer to this as the "benefit requirement."[48]

---

[44] *See* PLATO, *supra* note 1, at 36–37.

[45] *See* John C. Harsanyi, *Morality and the Theory of Rational Behaviour*, *in* UTILITARIANISM AND BEYOND 39 (Armartya Sen & Bernard Williams eds., 1982).

[46] *See, e.g.*, Edward McClennen, *The Theory of Rationality for Ideal Games*, 65 PHIL. STUD. 193, 193 (1992).

[47] RAWLS, *supra* note 17, at 6–7.

[48] See my early discussion of this condition in Finkelstein, *supra* note 19, at 1316–24.

4. While the contractors do not operate behind a thick veil of ignorance, as in Rawls's theory,[49] they remain agnostic about their future choices. That is, in interpreting the benefit requirement, they seek assurances that they will benefit under any future life circumstance or choice they might make. They seek, in other words, an assurance of benefit for the *worst case scenario* under any rule proposed for an agreement that pertains to the basic structure.
5. They assume that any agreement pertaining to the basic structure must be unanimous and universal, meaning that consent must be unanimous and that benefit must be universal in order for our institution of punishment to be both voluntary and welfare enhancing.

One effect of these conditions is that the social goal of deterrence performs a function it could not in a utilitarian account: It is able to dictate specific parameters for the punishment of each separate crime. In this way, the notion of deterrence can be made to generate normative constraints on punishment. The inability to do this was the central weakness of deterrence theory we considered previously.

Now let us consider how the contractarian approach would fare in application to a specific decision regarding punishment. 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 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 in a society that established penalties for theft and other violations of property norms than I am at my current baseline welfare? In answering this question, and applying our assumptions, each agent would weigh the benefit he would receive from increased deterrence against the loss he would suffer in the worst case scenario—that is, the balance of gains and losses he would experience in the *worst case scenario* under the rule. The worst case scenario is clearly the case in which the agent has little property to protect and is most disadvantaged by the rule, and this would be the case in which he is the object of the increased penalty himself. Thus he must ask whether he would be advantaged on balance from penalties for theft in a world in which he was himself subject to such penalties, as compared, for example, with a baseline in which there was no protection for private ownership at all. If the penalties for theft are set too low, the deterrent effect will be insignificant and private property will not be protected. If the penalties are too

---

[49] RAWLS, *supra* note 17, at 118–23.

high, agents receiving the penalty would be worse off than they would be in the absence of private property and the benefit requirement would not be satisfied.

To be more specific, imagine how parties to an original social contract would reason about a proposed penalty—for example, a twenty-year sentence for grand larceny. For the sake of argument, let us suppose this reasoning takes place not in a state of nature but against the backdrop of an existing, but constantly evolving, regime. And let us imagine each person can place a precise value on the totality of his personal possessions. Suppose further that under the current regime, which allows a maximum sentence for such thefts of ten years, each person can fairly well estimate the likelihood of theft over a certain fixed period of time. Now imagine that the proposed change in the maximum for such sentences doubles—it moves from ten to twenty years. In this case, we would expect the overall probability of theft would be cut in half.

Standard economic or utilitarian approaches to deterrence calculations would now regard the case for increasing the penalty for theft from ten to twenty years as nearly made, with several possible caveats: First, increasing the penalty for theft *could* have an undesirable effect on the incentives potential offenders have to commit other crimes. For example, if the penalty for bicycle theft is significantly increased, that would reduce the differential between the penalty for bicycle theft and the penalty for auto theft, with the result that some offenders inclined to steal bicycles might now steal automobiles. Similarly, as Justice Kennedy has recently written in *Kennedy v. Louisiana*, increasing the penalty for rape to death would decrease the disincentives to murder the victims of rape.[50] Second, deterrence theorists are forced to evaluate the benefits of the enhanced deterrent effect in light of the total economy of costs and benefits such a change would entail. If the cost to the State of imposing the increased penalty is also increased, then the marginal social benefit of the additional penalty might not ultimately be positive. The utilitarian case for adopting such a penalty, then, would be subject to the requirement that the benefits of increased deterrence are worth the costs.

One point that this discussion underlines is that deterrence in a utilitarian theory does not provide a justification that is addressed *to* individual offenders, as there need be nothing in it for them, even in the *ex-ante* sense. For this version of deterrence theory requires neither that each individual member of society regard himself as benefited nor that individual members of society consent to the deterrence scheme under which they are protected. While the traditional appeal to deterrence does restrict enhancements in punishment to instances where social welfare will increase in the aggregate, that social benefit may turn out to be unevenly distributed and hence may improve the lot of the few at the cost of the many. The benefit requirement suggests that rational contractors would reject any such gamble.

---

[50] *See* Kennedy v. Louisiana, 554 U.S. 407, 445 (2008).

Accordingly, on a contractarian approach, a member of the putative punishment agreement would consider whether he could expect to benefit under the worst case scenario, namely the case in which he himself ends up subject to the penalty. Now, in addition to considering the benefits of the additional deterrence under the increased sentence for theft, a social contractor must weigh the value to him of that increased protection for property against the disvalue he would experience from an additional ten years in prison. Because the odds of loss of property are relatively low, against the background of a ten-year sentence, the marginal increase in deterrent efficacy in our example is unlikely to outweigh the significant loss in value the rational agent would attach to an additional ten-year loss of liberty. Hence, the benefit test would most likely not be satisfied.

We might compare the marginal increase in penalty just considered to a different kind of decision, namely the decision whether to punish theft at all. If our contractors start from a baseline of zero punishment for theft and consider the adoption of a ten-year sentence for that crime, they would likely reach a different conclusion. For the cost of failing to adopt the contemplated penalty is now very high: assuming there were no other penalties to protect the interest individuals have in their property, the absence of the ten-year penalty would mean that all property was insecure. Contractors who have already settled on a scheme of distribution in their basic social contract would now have no way of enforcing that agreement. They would in effect be living in a property-less regime. Against this background, the increased deterrent benefit in moving from a regime with a ten-year sentence for theft would prove a benefit to every member of society, even those to whom this penalty is later applied. In response to the question, "In light of what is my punishment justified?" we can say to the offender: "Your punishment is justified because the benefits of a deterrent scheme that enabled you to protect your property have been great enough to you, throughout your life, that they overwhelm even the disvalue you are presently experiencing from your ten-year sentence. Your life is *still better* than it would have been in the absence of that sentencing provision, despite the fact that it has resulted in your incarceration." The benefit, in short, is not just to society generally: It is one that attaches to each particular offender and supplies each with a ground for consenting to the deterrent scheme under which he is to be punished.

One will now be tempted to object as follows: The only reason the ten-year sentence turned out to be justified, on the account I have proposed, is that we started from a baseline of zero punishment. But surely if we start from a baseline of zero punishment, the twenty-year sentence would appear justified as well. We would accept any penalty as legitimate for a serious crime like theft rather than live with conditions of zero deterrence. But once we start at a different, higher baseline, no penalty will seem justified. So the account either makes punishment too easy to justify or too hard: It is too easy if the alternative is *no* punishment, and it is too hard if *any* alternative to a zero level of punishment is available.

But I do not think this critique is ultimately correct, though I concede the critical importance of specifying for social contractors the baseline from which

they bargain. Suppose we start with a baseline of zero punishment and instead of considering only one option, our rational contractors consider two: they consider adopting either a mandatory ten-year sentence or, alternatively, the death penalty. Under these circumstances, rational contractors would unhesitatingly choose the ten-year sentence. Why? With a ten-year sentence we can assume that the chances that an individual contractor would lose his property to theft would be relatively low. Deterrence at this level of punishment, in other words, is fairly effective. Suppose, however, that if the contractors adopted the death penalty their risk of losing their property would be reduced to zero. Still, the increase in disutility to a rational contractor of the difference between a penalty of ten years and being put to death is so extreme that the rather small deterrent benefit he experienced would not seem worth the added cost to him. So even against the background of zero punishment, not every penalty will turn out to be justified.

Would a penalty like death *ever* be justified by this method? It is unlikely that rational contractors would accept the death penalty, even in the absence of any alternative sanctions. Rational agents simply do not regard losing their lives for the sake of protecting their property as a trade-off worth making. But arguably matters would be different if individuals were asked to consider a roster of possible penalties for murder. Since the value they place on their lives is much greater than the value they place on their property, rational contractors *might* consider death a sensible price to pay for lengthening their own earlier lives.[51]

Thus in a world in which no penalties were available other than death, the death penalty might be selected by the contractors in an initial position of choice. In that case, the alternative to having any punishment for murder would be the worst sort of violent state of nature, one that, if Hobbes is to be believed, would be so brutal and insecure that no one could expect to survive into old age.[52] 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. If, however, the contractors faced a choice of a mandatory life sentence or death for murder, they would evaluate things differently. The question they would ask themselves in this case would be: Does the marginal increase in personal security from the death penalty, as compared with a mandatory life sentence, 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 unlikely event that each application of the death penalty deterred eight additional murders, as compared with life in prison without parole, the marginal value of that added deterrence would likely be outweighed by the marginal cost of the death penalty to an individual contractor. Weighing this likely effect in advance, the contractors would reject the death penalty. This conclusion replicates the results of actual jury sentences with the increasing availability of life without parole as an alternative to

---

[51] *See* Finkelstein, *supra* note 19, at 1319–24.

[52] *See* HOBBES, *supra* note 14, at 186.

death. In the states in which life without parole is made routinely available, the willingness of juries to assign the death penalty has been substantially diminished.[53]

Notice the advantages of rational contractarianism as compared with the two leading approaches to punishment. On the one hand, rational contractarianism solves the central problem associated with pure deterrence theories—the problem that punishment on this view involves traveling across persons. 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 the problem does not arise on this view, for although the institution of punishment would be deterrence-based, and hence would hold one person responsible for the acts of another, each individual punished would have agreed to these conditions with respect to *his own* future violations of the covenant. Each member of the social contract pledges his fidelity and offers his willingness to submit to punishment should he fail to make good on his promise. Since he offers this guarantee in order to induce his fellows to contract with him, he in effect furthers his defensive interests in doing so. There is thus no conflict with the rights of self-defense he so carefully safeguards and no sacrifice of individual welfare. The contractarian account is able to incorporate deterrence as a social goal of paramount importance, but as this goal is given an individual interpretation, the usual objections fail to apply.

On the other hand, the contractarian theory we have explored 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 punishment is contemplated, 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 relegates them to a world of high theory, since the notion of desert is ill-equipped to provide a foundation for rational policy-making in the area of criminal justice.[54]

---

[53] *See* Richard C. Dieter, *Sentencing for Life: Americans Embrace Alternatives to the Death Penalty*, DEATH PENALTY INFO. CENTER (April 1993), *available at* http://www.deathpenaltyinfo.org/sentencing-life-americans-embrace-alternatives-death-penalty:

> Although a majority of those interviewed said they favored capital punishment abstractly, that support is reversed when the sentence of life without parole, coupled with a requirement of restitution, is offered as an alternative. Forty-four percent favor that alternative, while only 41% selected the death penalty. Even the choice of a sentence which guaranteed restitution and no release for at least 25 years caused death penalty support to drop by 33%.

[54] I raise other objections to retributivism in Finkelstein, *supra* note 13, at 214–18.

IV. OBJECTIONS

There are several objections to the argument I offered in the preceding section, and I shall consider these in the remainder of this Essay. First, a significant objection to my argument is that a person has a choice over whether to commit a crime and thus whether he risks suffering the death penalty or any other penalty is under his control. If this is true, a rational agent would opt for the most stringent penalties for all sorts of crimes as long as stringent penalties are cost-justified in a social sense. For he can thus capture the up-side of stringent deterrence and reject the down-side by simply avoiding the worst case scenario on his own. In this 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. This objection, then, rejects my fourth assumption, namely that rational agents would reason from the worst case scenario, including scenarios that are the product of choice.[55]

I would argue, however, that rational contractors may still want to guard against excessive penalties in case they are not deterred.[56] 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 societal 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 the harshest penalties 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.

One way to understand this seemingly odd suggestion is to notice that 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, for example, the death penalty purchases only a marginal increase in deterrence at the cost of a substantial increase in the coercive

---

[55] 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. 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. And this might suggest that rational contractors would not set any limits on the penalties they saw it as rational to adopt. I am indebted to Dan Markel for this point.

[56] It is of course possible that a person could be subject to a penalty punishment without having committed a crime at all. I have assumed throughout that punishments 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 harsh penalties.

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.

Here is yet another 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 harsh penalties 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 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 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.[57] On such a view, the contract itself imposes no limitations on what it is acceptable to do to violators. Locke, for example, argues that those who violate the terms of the contract are like wild beasts; they can be hunted down and killed indiscriminately.[58] And, according to Rousseau, "every evildoer who attacks social rights becomes by his crimes a rebel and a traitor to his country; by violating its laws he ceases to be a member of it."[59] 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,"[60] it makes sense to think of criminals as outside the scope of all voluntary arrangements, since cooperating with them

---

[57] 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).

[58] *See* LOCKE, *supra* note 15, at 279 ("[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 *Lyon*.").

[59] ROUSSEAU, *supra* note 16, at 177.

[60] RAWLS, *supra* note 17, at 4.

would not be to the advantage of those who remain faithful to their terms. 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 non-cooperation at the outset.[61] 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 non-cooperation 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 non-rational creatures are often 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 non-cooperation 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 we

---

[61] *See* Claire Finkelstein, *Hobbes and the Internal Point of View*, 75 FORDHAM L. REV. 1211, 1221 (2006).

have explored, however, these elements complement each other without contradiction.

V. CONCLUSION

Let us return briefly to Socrates and recall in particular his suggestion that fidelity to the commitment one has made to obey the laws constitutes a civic duty, something akin to military service or jury duty.[62] We now are in a position to put some flesh on the bones of this suggestion. In a more developed version of the contractarian suggestion, the duty to abide by the terms of one's punishment is a duty owed not to the State, but to one's fellow contractors, to whom one has pledged one's commitment to the terms of the contract. As Hobbes makes clear, those who would violate the contract are free-riders on the welfare of others, and no rational agent, knowing them to be such, would have agreed to contract with them in the first place.[63] We now, however, have a basis for understanding why Socrates's version of the civic duty to undergo punishment is more extreme than need be. If the duty to abide by the State's dictates is a duty owed to one's fellows, that duty need not be absolute. For there are rare times when it works to the advantage of all to disregard those dictates, as when the State has overstepped the authorization that rational agents saw it as in their interest to give. Such rejection of the State's dictates might be full or partial: A full-scale rejection would be warranted when the State no longer seeks to justify its authority to an entire segment of the population that authorized its power over them. Such might have been the case in Hitler's Germany or Stalin's Russia. The grounds for a partial rejection we have seen in our own times: When the government consistently and repeatedly demands action in the name of public benefit that benefits few and injures many, once again it has flouted the conditions of its original grant of authority, and its power over its subjects can, with right, be rejected.

---

[62] *See* PLATO, *supra* note 1, at 36.

[63] *See* HOBBES, *supra* note 14, at 201–05.