PHILIP MONTAGUE

GRADING PUNISHMENTS

(Accepted 2 October 2002)

According to John Stuart Mill,

> . . . the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others.[1]

If this position (henceforth "Mill's doctrine") is correct, then the law has no business prohibiting actions on paternalistic grounds. Neither should activities such as (consensual) prostitution or homosexuality be criminalized on moral grounds alone – that is, on grounds independent of harm they do to individuals other than their participants.

Mill's doctrine is vigorously criticized by James Fitzjames Stephen, and certain of Stephen's criticisms are echoed by Patrick Devlin in his lectures on the enforcement of morals. Both Devlin and Stephen endorse "legal moralism," according to which criminal prohibitions can properly apply to actions that are not covered by Mill's doctrine – to what Joel Feinberg calls "harmless wrongdoing."[2] Feinberg himself defends Mill's doctrine against Devlin's and Stephen's attacks. In doing so, Feinberg joins H. L. A. Hart, whose debate with Devlin has drawn considerable attention to questions regarding the justification of criminal prohibitions.

My concern in this essay is with a particular objection that Stephen and Devlin raise against Mill's doctrine. They argue that, in classifying only harmful actions as punishable by law, the doctrine

---

[1] John Stuart Mill, *On Liberty*, Chapter 1.

[2] See Feinberg's discussion of this category of actions in his *Harmless Wrongdoing* (New York: Oxford University Press, 1990), pp. 1–8, 124–175.

As Feinberg points out, proponents of legal moralism sometimes appeal to the idea that acts that do no ordinary harm to other individuals might nevertheless harm society in some way, and hence be justifiably criminalized. Joel Feinberg, *Social Philosophy* (Englewood Cliffs, NJ: Prentice-Hall, 1973), pp. 37–38. As legal moralism and Mill's doctrine are being interpreted here, however, they refer to harm that is done to individuals.

Law and Philosophy **22:** 1–19, 2003.

provides an unacceptably narrow basis on which to establish gradations of punishment for various criminal offenses. Both Hart and Feinberg attempt to answer this "gradation objection." I will argue, however, not only that their arguments fail, but also that both Hart and Feinberg presuppose a theory of punishment that is itself vulnerable to a version of the gradation objection.[3]

As this last remark suggests, the gradation objection has a much wider use than that to which it is put by Stephen and Devlin, namely, as a basis on which to evaluate theories of punishment. In the final section of this essay, I sketch a theory of punishment that implicitly endorses Mill's doctrine, and against which the gradation objection is ineffective.

I

Devlin raises the gradation objection by way of the following argument:

> ... if what justifies the making of the law is simply the prevention of harm, the offender must be punished accordingly. He must be punished for theft as he would be, for example, for a parking offence; the penalty must be calculated to prevent the repetition of the offence by him and to deter others from committing it. The offender's moral guilt is not a matter with which the law is concerned.
>
> This is not how in fact the law is administered. The degree of moral guilt is not the only determinant of the severity of the sentence but it is universally regarded as a very important one. It manifests itself in ... the gradation of offenses in the criminal calendar: in order of gravity they are not arranged simply according to the harm done.[4]

---

[3] An additional interesting feature of the theory of punishment favored by Hart and Feinberg is that (contrary to what they evidently assume) it is essentially unrelated to Mill's doctrine.

[4] Patrick Devlin, *The Enforcement of Morals* (Oxford: Oxford University Press, 1965), pp. 128–129. Stephen advances a rather different argument for this same conclusion:

> the feeling of hatred and the desire of vengeance ... are important elements of human nature which ought ... to be satisfied in a regular public and legal manner.
>
> The strongest of all proofs of this is to be found in the principles universally admitted and acted upon as regulating the amount of punishment. If vengeance affects, and ought to affect, the amount of punishment, every circumstance which aggravates or extenuates the wickedness of an act will operate in aggrav-

These remarks are flawed by Devlin's tendency to conflate claims about what the law actually does with claims about what it must do. I will assume that the latter more accurately reflect his position as expressed in the gradation objection.

In one philosophically interesting form, this objection is directed not at Mill's doctrine *per se*, but rather at systems of punishment that incorporate Mill's doctrine. Systems of punishment consist (at least partly) of sets of rules of certain sorts. These sets include rules specifying the types of acts that count as offenses; rules that group offenses into categories (misdemeanors, felonies, etc.); rules that correlate offenses with punishments; procedural rules for determining candidates for punishment; and rules governing the imposition of punishments. Identifying conditions under which societies are justified in establishing and implementing systems of punishment is a primary task of theories of punishment.

As it is presently being interpreted, the gradation objection fits within the framework of this theoretical task. It centers on the proposition that systems of punishment are justified only if they contain rules that grade punishments according to degrees of moral guilt, blameworthiness, or wickedness.[5] Hart and Feinberg as well as Stephen and Devlin appear to accept this justification-requirement. They appear to disagree only about whether it can be satisfied by any system of punishment that incorporates Mill's doctrine (that is, any system that restricts punishable offenses to harmful wrongdoings). None of these writers clearly states the gradation objection, although the following argument is a reasonable interpretation of the position endorsed by Devlin and Stephen and opposed by Hart and Feinberg:

> A system of punishment is justified only if it grades punishments according to degrees of blameworthiness;
>
> no system incorporating Mill's doctrine can satisfy this condition;

---

ation of diminution of punishment. (James Fitzjames Stephen, *Liberty, Equality, Fraternity* (Indianapolis, Indiana: Liberty Fund, 1993), pp. 98–99.)

[5] For the sake of both brevity and clarity, references to moral guilt and wickedness will henceforth be eliminated in favor of references to blameworthiness. Although the three concepts might differ from each other in certain respects, nothing of importance in the discussion that follows depends on these differences.

> hence, no system of punishment that incorporates Mill's doctrine is justified.

Henceforth, "Mill-system" will refer to systems of punishment that restrict punishable offenses to harmful wrongdoings; and "Devlin-system" will refer to systems of punishment that grade punishments according to degrees of blameworthiness.[6]

According to the gradation objection, then, only Devlin-systems of punishment are justified, and no Mill-system can be a Devlin-system. The primary reasons offered by Devlin in support of this latter proposition are suggested in the remarks quoted above. In essence, Devlin argues that Mill's doctrine presupposes an entirely forward-looking view of punishment, and such a view implies – mistakenly – that backward-looking considerations have no bearing on how punishments should be graded. We will examine Hart's and Feinberg's replies to this line of argument, but before doing so, let us consider the possibility that – as formulated above – the gradation objection fails on grounds independent of the sorts of considerations relied on by Hart and Feinberg.

In fact, counterexamples to the gradation objection's second premise are quite easy to construct. Thus, consider systems of punishment whose rules restrict punishable offenses to acts that are both harmful to others and whose agents are blameworthy for what

---

[6] The idea of grading punishments according to degrees of blameworthiness is open to two distinct interpretations. On one, it pertains to rules that correlate punishments with offenses – for example, rules that assign prison terms of various lengths to offenses of varying seriousness, with the seriousness of an offense determined in part by degree of blameworthiness. Accordingly, acts of homicide – although equally harmful – could be distinguished from each other in regard to the blameworthiness of their agents, and correlated with correspondingly different punishments. But blameworthiness can also be regarded as having a different sort of relevance to how severely offenders are punished. Rather than being built into the rules correlating punishments with offenses, it could be taken into account when sentences are actually imposed as potentially mitigating the sentences prescribed by such rules.

Although it isn't entirely clear from the discussions of Stephen, Devlin, Hart, and Feinberg which of these two interpretations they accept, most of what they say suggests that they have the second in mind. That is, in agreeing that punishments should be graded according to degrees of blameworthiness, they are referring to the relevance of a person's blameworthiness to how she should be sentenced after having been found guilty of committing an offense.

they do. Such systems are Mill-systems and, given how they connect the notion of a punishable offense with that of blameworthiness, it is surely possible for them to grade punishments according to degrees of blameworthiness. That is, these Mill-systems could be Devlin-systems as well. Hence, the gradation objection's second premise is shown to be false, and the objection casts no doubt at all on the justifiability of Mill-systems in general.

Presumably, there is an explanation of why Hart and Feinberg make no mention of such obvious counterexamples and formulate rather more elaborate replies to the gradation objection. A possible (even if incomplete) explanation is that the systems of punishment that are justifiable according to standard theories of punishment don't restrict punishable offenses to those for the performance of which their agents are blameworthy. Certainly, standard deterrence and retributive theories contain no such restriction. To be sure, the latter theories commonly refer to desert, and desert is like blameworthiness in being a backward-looking concept. But such theories center on the proposition that wrongdoers deserve punishment; they don't imply that only *blameworthy* wrongdoers deserve to be punished.

It seems more likely, however, that Hart and Feinberg aren't interested in refuting the gradation objection in its general form. Instead, they wish only to demonstrate that their own (forward-looking) theories of punishment are capable of justifying systems of punishment that are Devlin-systems as well as Mill-systems. Whether any such demonstration is necessary depends in part, however, on how wrongness is related to blameworthiness. The importance of this relation should not be surprising, given the prominent roles played by these two concepts in discussions of the gradation objection.

Let us begin by supposing that blameworthiness is indistinguishable from wrongness, and that seriousness of wrongdoing corresponds to degree of blameworthiness. Suppose too that a particular system of punishment restricts punishable offenses to harmful wrongdoings, and is therefore a Mill-system. Suppose further that the system grades punishments according to seriousness of wrongdoing. Then, on the assumption that wrongness is indistinguishable from blameworthiness, the system will automatically grade punishments according to degrees of blameworthiness.

The system is therefore a Devlin-system as well as a Mill-system, and it constitutes a counterexample to the gradation objection's second premise. Indeed, if wrongness is related to blameworthiness in the manner suggested, then the gradation objection's central thesis – that only Devlin-systems of punishment are justified – is automatically satisfied by any system of punishment that grades punishments according to seriousness of wrongdoing, regardless of which sorts of wrongdoings it classifies as criminal offenses.

The assumption that wrongness is indistinguishable from blameworthiness has the advantage of simplicity when contrasted to views according to which the two concepts are distinct. Moreover, there are situations in which the two notions seem to coincide, or in which it is at least very difficult to separate judgments of wrongdoing from judgments of blameworthiness. Nevertheless, there are compelling reasons for regarding wrongness and blameworthiness as distinct concepts.

One such reason is that wrongness is a property of actions while blameworthiness if a property of persons: actions can be wrong, but people cannot be; and, whereas people can be blameworthy, actions cannot be. Secondly, if wrongness is not distinguished from blameworthiness, then there would seem to be no way adequately to accommodate the difference between justification and excuse. Or, to put the point more precisely, if wrongness were indistinguishable from blameworthiness, then the considerations relevant to whether some person has acted wrongly would be the same as those relevant to whether the person is blameworthy for having acted. Yet these two sets of considerations differ from each other in at least one significant respect: whereas those relevant to blameworthiness include excuses, excuses are irrelevant to whether wrongdoings have occurred.[7]

[7] The considerations respectively relevant to each of the two concepts might differ from each other in additional ways. For example, the wrongness of actions might be an entirely "objective" matter, while blameworthiness might depend at least partly on "subjective" considerations – that is, on facts about the mental lives of agents. Or wrongness and blameworthiness might both depend at least partly on subjective considerations, but on distinct sets of such considerations. We need not determine precisely how wrongness differs from blameworthiness in order to recognize that there are much better reasons in favor of distinguishing the concepts from each other than there are for equating them.

With the preceding remarks in mind, consider again the hypothetical system of punishment referred to above. Recall that it is a Mill-system that grades punishments according to seriousness of wrongdoing. When wrongness and blameworthiness were assumed to be indistinguishable from each other, the system automatically graded punishments according to degrees of blameworthiness, and therefore served as a counterexample to the gradation objection's second premise. Now that wrongness is assumed to be independent of blameworthiness, however, there is no guarantee that the system grades punishments as the gradation objection requires.

So the gradation objection presents problems for the justifiability of Mill-systems of punishment only if wrongness and blameworthiness are distinct concepts. And, for the reasons presented above, I will assume that the two concepts are indeed distinct.[8] Let us now consider Hart's and Feinberg's replies to the gradation objection.

II

Hart presents his reply in this passage:

> . . . those who concede that we should attempt to adjust the severity of punishment to the moral gravity of offences are not thereby committed to the view that punishment merely for immorality is justified. For they can in perfect consistency insist on the one hand that the only justification for having a *system* of punishment is to prevent harm and only harmful conduct should be punished, and, on the other, agree that when the question of the *quantum* of punishment for such conduct is raised, we should defer to principles which make relative moral wickedness of different offenders a partial determinant of the severity of punishment.[9]

At the beginning of this passage, Hart implies that legal moralism is not entailed by the proposition that punishments must be graded according to degrees of blameworthiness. If the argument he offers in support of this proposition were sound, then a system of punishment could restrict criminal offenses to harmful acts (and hence be a Mill-system), while also allowing the requisite correlation

[8] We will also assume that wrongness and blameworthiness aren't necessarily connected in certain ways – that, for example, it isn't the case that people are necessarily blameworthy whenever they engage in wrongdoing.

[9] L. A. Hart, *Law, Liberty, and Morality* (Stanford, California: Stanford University Press), p. 37.

of punishments with blameworthiness (and therefore be a Devlin-system). Hart would therefore have refuted the gradation objection. In fact, however, Hart's argument is irrelevant to this objection.

In presenting his position, Hart relies heavily on the distinction between justifying systems of punishment on the one hand, and justifying certain rules of such systems on the other. In the present context, the relevant rules concern how punishments should be graded, but Hart also employs his distinction in addressing the question of whether utilitarian theories of punishment can accommodate prohibitions against punishing innocents. He argues for an affirmative answer in this passage:

> It is perfectly consistent to assert *both* that the General Justifying Aim of the practice of punishment is its beneficial consequences *and* that the pursuit of this General Aim should be qualified or restricted out of deference to principles of Distribution which require that punishment should be only of an offender for an offence.[10]

---

[10] Hart, *Punishment and Responsibility* (Oxford: Oxford University Press, 1968), p. 9. I am interpreting Hart's references to justifying "the practice of punishment" in terms of its "beneficial consequences" as references to justifying the establishment of systems of punishment in terms of their beneficial consequences. I do so because Hart's mode of expression doesn't distinguish between justifying the punishment of individuals and justifying something else, and the only reasonable candidate for this "something else" is the justification of systems of punishment. Since Hart's references to justifying the practice of punishment are clearly not meant by him to concern the justifiability of punishing individuals, he can reasonably be construed as concerned with the justifiability of establishing systems of punishment.

That this is a reasonable way in which to interpret Hart is further supported by the following passage:

> I shall assume that Retribution ... may figure among the conceivable justifying aims of a system of punishment. Here I shall merely insist that it is one thing to use the word Retribution *at this point* in an account of the principle of punishment in order to designate the General Justifying Aim of the system, and quite another to use it to secure that to the question "To whom may punishment be applied?" (Hart, *Punishment and Responsibility*, p. 9.)

It is noteworthy too that Feinberg also refers to the justification of systems of punishment in his discussion of punishment's general justifying aim (Feinberg, *Harmless Wrongdoing*, p. 146f).

Combining these remarks with those quoted earlier, we can reasonably regard Hart's overall view as centering on this proposition: that punishment's General Justifying Aim can provide a basis on which to determine whether systems of punishment are justified without thereby constituting a justifying condition for rules of those systems. More specifically, systems of punishment can be justified on grounds of harm-prevention and yet contain rules that prohibit punishing innocents and – more importantly for this discussion – rules that grade punishments according to degrees of blameworthiness. Let us refer to this as Hart's "compatibility thesis."

It is important to bear in mind that Hart's appeal to this thesis is a major component of his reply to the gradation objection. As we have been interpreting this objection, it states that only Devlin-systems of punishment are justified, and denies that Mill-systems can be Devlin-systems. While the compatibility thesis might well be true, however, it has nothing directly to do with Mill-systems of punishment. Rather, it concerns what I will call "Hart-systems." These are systems of punishment whose implementation conforms to punishment's justifying aim (as stated by Hart) of harm-prevention. To be sure, both Mill-systems and Hart-systems are centrally concerned with the concept of harm. But it is one thing for a system of punishment to prevent harm when implemented, and quite another for the system to restrict punishable offenses to harmful wrongdoings. Hart seems not to recognize this distinction, as is evidenced by his statement that "the only justification for having a *system* of punishment is to prevent harm and only harmful conduct should be punished." That this last remark lumps together two distinct ideas becomes especially clear on recognizing that a system of punishment might be justified on grounds of harm-prevention even though it is not a Mill-system – even though it classifies some harmless acts as punishable.[11]

[11] For example, punishing people for negligent or reckless acts that turn out not to harm anyone could deter others from performing negligent or reckless acts that would be harmful.; and this could result in less overall harm than that which would result if no harmless recklessness or negligence were punished. Similar remarks would apply to the punishment of failed attempts at doing harm. To be sure, Hart might maintain that attempts at doing harm or reckless or negligent acts that are not harmful in the strict sense, are actually harmful in some loose sense. Hart does not suggest this, however; and to do so would require him not only to explain the

So regardless of how successful Hart might be in establishing his compatibility thesis (which is about the relation of Hart-systems to Devlin-systems), his arguments have no bearing at all on the gradation objection (which concerns whether Mill-systems can be Devlin-systems). Feinberg's reply to the gradation objection resembles Hart's in significant respects and is similarly flawed. Feinberg states his position in the following passage:

> [a defender of Mill's doctrine] would say that the justifying aim of the whole system [of punishment] is to prevent ... harms, while insisting that the rules governing the system's operations at every level must be *fair*. Fairness to the accused requires gradation of punishments in accordance with two distinct sets of considerations: the wrongdoer's degree of *responsibility* for his deed and degree of *blameworthiness* as determined by his motive and circumstances.[12]

One respect in which Feinberg's position coincides with Hart's is quite obvious: both purport to demonstrate that systems of punishment can both restrict punishable offenses to harmful wrongdoings and grade punishments according to degrees of blameworthiness. In fact, however, they argue that systems of punishment can do the latter while conforming to the justifying aim of harm-prevention.[13]

Hart and Feinberg not only wish to show that systems of punishment can accommodate Mill's doctrine while also grading punish-

---

nature of this loose sense of harm and argue for its existence, but also to revise his entire discussion of punishment in significant respects. (In this connection, Hart might attempt to rely on Feinberg's notion of harms as "invasions of interests," but the latter is itself problematic in important respects.)

Hart's mistaken assumption about the relation of Mill's doctrine to forward-looking theories of punishment has a parallel, namely, the assumption that legal moralism is essentially related to backward-looking (especially retributive) theories of punishment. This latter assumption is easy to explain: proponents of such theories seldom (if ever) restrict punishable offenses to wrongdoings that are harmful. But backward-looking theories must surely exclude some wrongdoings from the class of criminal offenses; and nothing in the nature of these theories prevents them form excluding all harmless acts.

[12] Feinberg, *Harmless Wrongdoing*, p. 148.

[13] Devlin too confuses the idea of justifying punishment in terms of harm-prevention with the idea of punishing only harmful acts. As noted earlier, he says that "if what justifies the making of law is simply the prevention of harm, the offender must be punished accordingly." When taken in context, this remark clearly means that, if what justifies punishment is harm-prevention, then punishments must be graded according to amounts of harm rather than degrees of blameworthiness.

ments according to degrees of blameworthiness, but they also argue that systems of punishment ought to grade punishments in this way. The reasons they offer in support of this latter claim are worth examining for various reasons, not the least of which is that doing so points to an interpretation of the gradation objection very different from the one we have been relying on so far.

As the previously quoted remarks indicate, Feinberg maintains that grading punishments according to degrees of blameworthiness is a matter of fairness. He goes on to argue that relevantly dissimilar cases ought to be treated dissimilarly, and that – in applying this principle to punishment – "the degree of moral blameworthiness ... [is] a 'relevant' characteristic."[14] Hart presents much the same argument in this passage:

> There are many reasons why we might wish the legal gradation of the serious of crimes, expressed in its scale of punishments, not to conflict with common estimates of their comparative wickedness. One reason is that ... principles of justice or fairness between different offenders require morally distinguishable offences to be treated differently and morally similar offences to be treated alike.[15]

According to Hart and Feinberg, then, justice (or fairness) requires that systems of punishment grade punishment according to degrees of blameworthiness. This claim conflicts, however, with the idea that punishment's General Justifying Aim is harm-prevention. This problem become evident on looking more closely at the relation between justifying systems of punishment on the one hand, and justifying rules of such systems on the other.

This relation is actually closer than Hart's and Feinberg's remarks might suggest, since (as was noted earlier) systems of punishment are at least partly composed of sets of rules of certain sorts. Systems of punishment can therefore differ from each other by virtue of differences in the rules they contain. Moreover, such systems do in fact vary from jurisdiction to jurisdiction, sometimes slightly and sometimes significantly. Hence, determining whether a system of punishment is justified is not simply a choice between punishing or not. Rather, it consists in choosing among systems that can differ from each other in numerous ways in virtue of differences in the rules that they contain.

---

[14] Feinberg, *Harmless Wrongdoing*, p. 149.

[15] Hart, *Law, Liberty, and Morality*, pp. 36–37.

Now consider two systems of punishment, S and S′, that differ from each other in only one respect: whereas S assigns the death penalty to certain types of homicide, S′ assigns life imprisonment to acts of those types. If we faced a choice between these two systems and no others, then the positions of Hart and Feinberg imply that selecting one of them over the other should be guided by punishment's General Justifying Aim of harm-prevention. What precisely this means, however, is extremely unclear.

Sometimes the proposition that harm-prevention is punishment's General Justifying Aim seems to be equated with the proposition that a system of punishment is justified if and only if it prevents (or minimizes) harm. At other times, however, the former proposition is treated as equivalent to something like "A system of punishment is justified if it prevents harm *and other things are equal*" – implying that a system's capacity for preventing harm is but one among several considerations that are relevant to whether the system is justified. Unfortunately, however, the idea that harm-prevention is one of several considerations relevant to whether systems of punishment are justified is not kept separate from the idea that considerations other than harm-prevention are relevant to whether rules of systems of punishment are justified.

Thus, contrast the following two propositions: (a) whether a system of punishment is justified depends on whether it prevents harm and also on whether it conforms to principles of justice; (b) whether a system of punishment is justified depends entirely on whether it prevents harm, but whether the rules of a system of punishment are justified depends on whether they are just. Bearing in mind that the notion of a General Justifying Aim is introduced by Hart and Feinberg in order to distinguish between considerations relevant to justifying systems of punishment on the one hand, and considerations relevant to justifying rules of systems on the other, interpretation (b) seems definitely preferable to (a). I will therefore assume here that, according to Hart and Feinberg, a system of punishment is justified if and only if it conforms to punishment's General Justifying Aim of harm-prevention.

Recall now our two systems of punishment, S and S′, where S assigns the death penalty to certain types of homicide, and S′ assigns life imprisonment to acts of those types. Following Hart and

Feinberg, we should choose S over S′ if and only if implementing S would result in less harm than would result from implementing S′. These same conclusions apply *mutatis mutandis*, moreover, to systems of punishment that differ only in the respect that one allows the punishment of innocents while the other does not, or in the respect that one – but not the other – grades punishments according to degrees of blameworthiness. Hence, while Hart and Feinberg are surely correct in maintaining that the rules of a system of punishment must be just or fair, and while they might be correct in claiming that justice requires punishments to be graded according to degrees of blameworthiness, systems of punishment can conform to the justifying aim of harm-prevention without taking blameworthiness into account, and without being just or fair in any respect.

In and of itself, preventing (or minimizing) harm is essentially unrelated to justice in general, or to grading punishments justly in particular. And if the justifying aim of systems of punishment is harm-prevention, then the justice or injustice of a system is – in and of itself – irrelevant to its justifiability. To be sure, systems of punishment that prevent harm can happen coincidentally to be just, and in certain circumstances the goal of harm-prevention might be achievable only by establishing a system with just rules. To guarantee just rules, however, justice must somehow be built into punishment's General Justifying Aim. The same is true of other features that systems must have if they are to be justified – for example, prohibitions against punishing innocents. For the justifiability of systems of punishment to depend necessarily on their containing such prohibitions, punishment's justifying aim must be appropriately specified; and the justifying aim of harm-prevention clearly fails to satisfy this condition.

The problems here are, of course, more familiarly associated with rule utilitarian theories of punishment. These theories are commonly assumed by their proponents to avoid certain fatal objections to act utilitarian theories, and yet analogous objections to the "rule" theories are just as fatal. Neither act nor rule utilitarian theories can accommodate the *necessary* relevance of considerations of guilt and innocence to whether punishment is justified. Furthermore, since considerations of justice possess this same necessary relevance, and since utilitarian theories – whether "act" or "rule" – can accord them

only contingent relevance, we have an additional basis on which to reject such theories.[16]

Lying at the heart of these criticisms of utilitarian theories of punishment is a general constraint on theories of punishment which can be put as follows:

> A theory of punishment is true only if it is compatible with the proposition that, necessarily, only just systems of punishment are justified.

And if we say (taking our cue from Hart and Feinberg) that, necessarily, systems of punishment are just only if they grade punishments according to degrees of blameworthiness, then we can infer that

> A theory of punishment is true only if it is compatible with the proposition that, necessarily, systems of punishment are justified only if they grade punishments according to degrees of blameworthiness.[17]

This proposition provides the basis for a version of the gradation objection that applies to theories of punishment rather than to systems of punishment as is the case with the original version. When the proposition is applied to theories that equate punishment's general justifying aim with harm-prevention, it implies that such

---

[16] According to Hart, "Retribution in General Aim entails retribution in Distribution" (Hart, *Punishment and Responsibility*, p. 9). That is, if the justifying aim of punishment is interpreted as retributivist in character, then systems of punishment must restrict punishment to offenders. For some reason, however, Hart denies (at least implicitly) that a utilitarian general aim would imply that considerations of utility determine who should be punished. The suggestion being made here is that specifying punishment's general justifying aim determines criteria for establishing rules contained in systems of punishment. As was noted above, Hart might wish to claim that considerations besides harm-prevention are relevant to whether systems of punishment are justified – considerations of justice in particular. But this sort of claim would seriously undermine Hart's view of the importance of separating the justification of systems of punishment from the justification of rules contained in such systems.

[17] Nothing in the discussion to this point depends on its being true that systems of punishment must grade punishments according to degrees of blameworthiness. And, as will become clear below, this remains the case for the rest of the discussion.

theories are false; and this result squares with our earlier remarks about Hart's and Feinberg's favored theories.

III

As originally formulated, the gradation objection centers on a justification-requirement for systems of punishment, and is directed against Mill-systems. Our recently formulated truth-requirement for theories of punishment can be used as the basis of a revised gradation objection. In this revised form, the objection is directed against theories of punishment that justify Mill-systems. This revised gradation objection results from conjoining the truth-requirement for theories of punishment with the following proposition:

> No theory of punishment that justifies systems of punishment incorporating Mill's doctrine is compatible with the proposition that, necessarily, systems of punishment are justified only if they grade punishments according to degrees of blameworthiness.

This proposition is vulnerable to counterexamples, however.

Thus, consider versions of retributivism according to which punishment's General Justifying Aim is insuring that *certain* wrongdoers – namely, blameworthy harmdoers – receive their just deserts. This theory implies that systems of punishment are justified only if they restrict punishable offenses to harmful acts for the performance of which their agents are blameworthy. With blameworthiness related in this way to the justifiability of systems of punishment, there is no reason at all to doubt that the theory in question can justify systems of punishment that grade punishments according to degrees of blameworthiness.

Partly because they are entirely backward-looking, however, retributivist theories are seriously problematic in various respects.[18] What follows is the sketch of a theory that avoids these problems, and also lacks the defects present in entirely forward-looking theories (for example, those that equate punishment's justifying

[18] For a discussion of these problems, see Phillip Montague, *Punishment as Societal Defence* (Lanham, Maryland: Rowman & Littlefield, 1995), 11–23; 80–90.

aim with harm-prevention or harm-minimization). The difficulties associated with theories of both these types are avoidable by interpreting punishment's General Justifying Aim as the just distribution of harm.[19]

The central task of a theory of punishment is to identify considerations that are capable of defeating the moral presumption against punishment arising from its essentially harmful nature. One way in which to approach this task is by considering whether there are principles that justify harming others outside of punishment contexts, and that are also applicable within such contexts. Perhaps the most obvious contexts in which to begin looking for such principles are those involving self-defense against culpable aggressors. At least under certain conditions, people do justifiably harm others in these contexts; and the question therefore arises of whether the principles that justify harming others in self-defense might also justify harming people through punishment.

This approach has two rather obvious drawbacks. One is that identifying the principles that justify self-defense is itself a notoriously difficult problem. And the second is that, whatever these principles are, they justify harm done by innocent people as the only means of preventing themselves from being harmed, whereas punishment is imposed after the fact – after innocent people have already been harmed. The second of these problems is less serious than it appears to be at first glance, however. Indeed, it is avoidable by relying on the distinction between justifying individual punishment on the one hand, and justifying systems of punishment on the other. The idea is that, although the considerations that justify individual self-defense against culpable aggression are inapplicable to the justification of individual punishment, they do apply to the justification of systems of punishment. In order for this approach to succeed, however, the problem of justifying individual self-defense must be solved.

---

[19] Although I have followed Hart and Feinberg in referring more often to harm-prevention than to harm-minimization, the latter presents a far more realistic goal than the former. In this same vein, justly distributing harm is more realistic than preventing harm as a justifying aim of punishment.

A theory that interprets punishment's General Justifying Aim as the just distribution of harm is developed and defended in Montague, *Punishment as Societal Defense*.

For simplicity's sake, we will focus on situations in which innocent people can avoid being killed by culpable aggressors if and only if they kills the aggressors. In such a situation, someone will be harmed regardless of what the intended victim does: if he does nothing, he will be killed, whereas if he acts in his own defense, the aggressor is killed. In other words, harm is unavoidable from the intended victim's standpoint, although he can determine who is harmed. What we have, then, is a problem of justice in the distribution of burdens- in the form of harm – in situations possessing this special feature: it is the fault of some of the potential recipients of the harm that there is harm to be distributed. In such situations, a just distribution is one in which the harm is done to those who are to blame for the existence of the situations.

This result is based on the following principle:

(J) If, in a given situation, there are unavoidable burdens to be distributed; and if, in that situation, it is the fault of some but not all of the potential recipients of the burdens that there are burdens to be distributed; then (other things being equal) imposing the burdens to these latter individuals is justified as a matter of distributive justice.

J can be used to justify both self- and other-defense against culpable aggressors, as well as choices to harm certain individuals rather than others in situations having nothing to do with defense against aggression. In addition to justifying such direct distributions of burdens to individuals, J can also be used to justify the establishment of institutions or practices whose implementation results in harm to individuals. As J applies to punishment, it directly justifies establishing systems of punishment under certain conditions, and only indirectly justifies the punishment of individuals.

Thus, imagine a society certain of whose members will inflict wrongful harm on innocent members if nothing is done to prevent or deter them from doing so. Imagine further that, if the society establishes a system that incorporates real and credible threats to punish those who do harm innocent people, then some of the former will be deterred from harming the latter. Imagine finally that some members of the society will in fact ignore the threats of punishment and harm

innocent people, and that at least some of these wrongdoers will be apprehended and punished.

Under these conditions, our imagined society faces a choice: do nothing, in which case innocent people will be wrongfully harmed; or establish a system of punishment, in which case fewer innocent people will be harmed, but some wrongdoers will be harmed. This situation is therefore one in which harm is unavoidable from the society's standpoint. Moreover the situation is brought about by those who will harm innocent people if not deterred, and they are blameworthy for bringing it about.[20] Given the choice between allowing innocent people to be harmed, and establishing a system of punishment which will result in harm to those whose fault it is that there is harm to be distributed, justice requires (*ceteris paribus*) the society to choose in favor of the former rather than the latter.

So a theory of punishment based on J (call it "punishment as societal-defense") justifies systems of punishment that reserve punishments for blameworthy harmdoers. These systems therefore incorporate Mill's doctrine – which implies that only Mill-systems are justified by the theory. According to punishment as societal-defense, moreover, punishment's justifying aim is the just distribution of harm – and hence the theory justifies only those systems of punishment whose rules are just. If justice requires punishments to be graded according to degrees of blameworthiness, then punishment as societal-defense justifies only those systems that grade punishments in this way – that is, only Devlin-systems. The theory therefore satisfies the truth-requirement for theories of punishment that is contained in our most recent version of the gradation objection.

One final point is worth making.

Nowhere in the discussion have questions been raised about whether systems of punishment must indeed grade punishments according to degrees of blameworthiness. The reason for not raising such questions when sorting out the issues separating Stephen and

[20] The imagined situation can be regarded as brought about by members of the society prior to their actually harming anyone or even engaging in overtly threatening behavior. This is because the situation exists in virtue of facts about certain members of the society – facts that concern matters over which they have control.

Devlin from Hart and Feinberg are clear enough. It should now be equally clear why the issue of whether punishments should be graded according to degrees of blameworthiness can remain unresolved.[21] Punishment as societal-defense requires that harm be distributed justly; what counts as a just distribution of harm depends on principles of justice that are independent of the theory itself.

Western Washington University
Bellingham, WA 98225-5996
USA

[21] As can the issue of how the notion of grading punishments should be interpreted (see note 5 above).