The Philosophical Quarterly, Vol. 53, No. 212 July 2003
ISSN 0031–8094

A DETERRENCE THEORY OF PUNISHMENT

By Anthony Ellis

I start from the presupposition that the use of force against another is justified only in self-defence or in defence of others against aggression. If so, the main work of justifying punishment must rely on its deterrent effect, since most punishments have no other significant self-defensive effect. It has often been objected to the deterrent justification of punishment that it commits us to using offenders unacceptably, and that it is unable to deliver acceptable limits on punishment. I describe a sort of deterrent theory which can avoid both of these objections.

I

If punishment is to be justified, it must be predominantly by reference to deterrence. I hold this for two reasons, neither of which I can argue for here.

First, each of the so-called retributive theories is either internally incoherent or has implausible moral implications.

Secondly, even if there were an otherwise acceptable retributive theory of punishment, we should have to reject it, for I believe that violence is justified only in self-defence against aggression, that is to say, against behaviour violating a constraint which one has a right to uphold in self-interest (as is usual, I include in this the defence of others' self-interest). This claim, of course, does not itself restrict us to a deterrent theory, for it is consistent with other possible justificatory aims: direct prevention, for instance, or reform. But most punishments have no significant preventative effect, and are not intended to have any; and their beneficial reformative effect is scarcely more pronounced.

The simplest version of the deterrence theory of punishment, that we may punish offenders to deter other potential offenders, has usually been rejected, for two reasons.

First, the theory commits us to accepting that in punishing one person in order to deter others we are unacceptably *using* him. There have been many ways of trying to avoid this. For now, I shall simply register my agreement

with Tony Honoré that none of them will work.[1] Broadly, attempts to avoid the problem hold that punishing offenders to deter others does not unacceptably use them (because, for instance, they have implicitly assented in advance to this treatment, or because they deserve their punishment), or that they have forfeited the right not to be used. But I think that the objection is correct, and the only response a proponent of the theory could make would be to accept this, and simply tough it out.

Secondly, the theory lays down no acceptable limits on punishment. Presumably if we are concerned simply with deterring offenders, then the appropriate level of punishment should be whatever is necessary to deter. And this might be, in particular cases, either too lenient or too severe. Of course, we can solve the problem by laying down independent moral constraints upon the acceptable levels of punishment.[2] But this may seem *ad hoc*; one might have hoped for a more unified theory, one in which the point of the institution itself generates the constraints.[3]

In this paper I shall sketch a deterrent theory which avoids these objections.

II

The theory understands punishment as a form of self-defence. I shall assume that we have the right to use force in legitimate self-defence. I shall also assume that groups may use force in self-defence, and that a society or a nation may do so. Do we, in addition, have the right to threaten retaliation against potential aggressors in order to deter them? It seems plausible to think that at least sometimes we do; all else being equal, it must surely be better to try to prevent aggression rather than to have to deal with it forcefully when it occurs. Of course, all else may not be equal. Greater force may have to be used to deter aggression than would be required to deflect it once it has started. Or the methods used to deter could be intrinsically unacceptable. Or the threat may have little chance of succeeding, or may even constitute a perverse incentive to potential aggressors. But in the absence of such factors, threats of retaliation are surely acceptable. However, this raises a familiar question: if it is indeed permissible to *threaten* some level of retaliation to deter aggression, is it also permissible to *carry out* that retaliation if the threat fails to work? The justification for threatening is that it will prevent aggression (and any harm involved in a self-defensive

[1] Tony Honoré, *Responsibility and Fault* (London: Hart, 1999), p. 19.

[2] This is the position of H.L.A. Hart (or one of his positions): see *Punishment and Responsibility* (Oxford: Clarendon Press, 1968), ch. 1.

[3] Cf. P. Montague, *Punishment as Societal-Defense* (Lanham: Rowman & Littlefield, 1995), p. 91.

response). If the threat does not work, however, that justification will not carry over into a justification for carrying out the threat, for carrying out the threat will, by hypothesis, do nothing to prevent these harms. Retaliation, then, seems unjustified (at least by the principles of self-defence).

But what if in addition to merely issuing the threat I somehow *bound myself* to carry it out?[4] I mean by this not just that I pledged myself to carry it out, but that I somehow made it the case that if the threat were ignored, retaliation would be inevitable. We could suppose, for instance, that I constructed a booby-trap surrounding my domain which, once it was constructed, I could not dismantle, and whose operation was automatic as soon as anyone crossed my border; I might then, with the intention of deterring anyone from crossing the border, announce that I had done this. If the threat of retaliation failed, the actual retaliation would be automatic.

When retaliation against an offender is carried out by such an automatic system (I shall just say 'punished' from now on), can he complain that he is being used? Initially, at least, it hardly seems so. That charge was based on the claim that offenders were being made to suffer in order to modify the behaviour of others, and that seems no longer true. After all, it may be known that the offender's punishment will have no such effect; how then can it be said that he is being punished in order to deter others? His punishment is simply a direct response to his own action, carrying no thought of how it may affect others.

However, the core of the worry may remain. Any particular offender's punishment may be known to have no deterrent effect, but the fact that punishments are actually administered must play some role, otherwise it would not be justifiable to go beyond a system in which the threat is merely a bluff. The obvious role is making the threat credible to others. It may then seem that the actual offender is punished because the system requires that offenders in general must be punished in order to maintain the general effectiveness of the system. And this may seem essentially no different from a system in which we pick a number of people and punish them as a deterrent demonstration to others, knowing that, for whatever reason, some of these punishments would not in fact serve that purpose: here each offender could correctly say that he was being used to deter others, even though, in some cases, the particular punishment would not have that effect.

[4] Warren Quinn also used this idea in a discussion of punishment: see 'The Right to Threaten and the Right to Punish', *Philosophy and Public Affairs*, 14 (1985), pp. 327–73. Daniel Farrell has shown that Quinn's argument rests upon a premise which is false: see 'On Threats and Punishments', *Social Theory and Practice*, 15 (1989), pp. 125–54. The idea that punishment can be justified as a form of self-defence has also been defended by Montague (see *Punishment as Societal-Defense*), and by Farrell: see, for instance, 'Deterrence and the Just Distribution of Harm', *Social Philosophy and Policy*, 12 (1995), pp. 220–40, and 'On Threats and Punishments'.

But maintaining the credibility of the threat is not the only reason we could have for making it a genuine one. We could imagine that, for various reasons, a bluff might actually work – sufficiently so at least to maintain the system's effectiveness. In such circumstances, there might still be reasons to make the threat a real one. It may simply be more convenient. Or again, a system of bluffs would require dishonesty, and we might object to that. There is, of course, an indefinite number of further possible reasons, having nothing to do with the deterrent effectiveness of the system. So long as one of these is our reason, no one is punished because the system requires that offenders must be punished in order to maintain the deterrent effectiveness of the system. This should remove the last vestige of the worry that offenders are being used. It would in no sense be true that they are being punished in an attempt to modify the behaviour of others. The threat is addressed to each individually, and each is punished because he, individually, chose to ignore the threat; the others, and their potential behaviour, are irrelevant.

There are, of course, other possible problems with such retaliatory systems. One is that such systems may seem to involve the intention to cause harm at a point in time when this harm may be known to have no beneficial effects. But that is a tendentious description. Less tendentiously, the system requires that at time t_1, with the intention of preventing an evil, we set in motion an irreversible process, which, if our intention is not realized, will cause harm at t_2. Now it is certainly impermissible to intend, *tout court*, to cause pointless harm; but it does not follow from this that it is impermissible to intend to prevent harm by setting in motion a train of events which will, if one's intention is not realized, cause harm. That will depend upon the totality of the circumstances, and in particular whether the agent has observed certain conditions which I shall outline later.[5]

A different sort of problem we may call the problem of scatter: unless the system's method of detection were infallible, the threat would be triggered even by innocent people whom the system mistakenly took to be aggressors. (Who is 'innocent' is a problem to which I shall return.) But this does not by itself make such systems impermissible, for there are clearly circumstances in which it is permissible knowingly to put innocent people at risk as long as this is not our aim. Indeed, we do this constantly in our criminal justice systems, requiring only that benefits and risks be appropriately balanced. And so long as there is the appropriate balance, there is, as yet, no objection to such deterrent threats.

[5] There is, of course, an immense and ancient literature to which I cannot here do justice on the logic of the intentions embedded in threats. For a relevant discussion, see Farrell, 'A New Paradox of Deterrence', in J.L. Coleman and C.W. Morris (eds), *Rational Commitment and Social Justice: Essays for Gregory Kavka* (Cambridge UP, 1998), pp. 22–46.

One way of making it easier to achieve that balance would be by making the system slightly less automatic, allowing it to be stopped when there is reason to think that an apparent aggressor is in fact innocent. I might do this by dividing the operation into different functions and placing each in the hands of a different person. So one agent might be authorized to apprehend suspected aggressors, another might be authorized to decide whether they really were aggressors, and another might be authorized to administer punishment if appropriate. But no one would have authority to deactivate his part of the system except in special circumstances. From my point of view, this system would be substantially similar to the simpler model I sketched a moment ago. Once I have issued my threat, I have nothing further to do in order for an aggressor to be punished: from my point of view it is now automatic. It can now properly be said that the aim of the system is deterrence, but that no one is punished in order to deter others.

III

It may seem that this model is too remote from reality to tell us anything about punishment in the real world. However, I shall suggest that it is a useful model of criminal justice systems in the USA and UK.

A criminal statute, we may say, lays down a threat, which we can think of as a threat of retaliation against anyone in violation of its prohibition. And it is, in the important sense, like the threat in the model I have suggested: once someone has transgressed, the procedure goes forward fairly automatically. Legal systems differ, of course. In some, such as the German and Austrian systems, prosecution of all offences is, in principle, mandatory where there is adequate evidence to sustain a conviction. In others, such as the UK and US systems, there is broad discretion at most levels. It is enough, of course, for my present purposes that we can imagine a realistic system in which the normal course, though not necessarily mandatory, is prosecution. And, I shall suggest, this is the best understanding of the US and UK systems.

The procedure is automatic in the sense in which the system just imagined is automatic – a system which, of course, was meant to reflect the separation of powers that characterizes modern democracies. The legislature issues threats, but it has no authority to decide whether those threats are acted on in particular cases. Prosecutors and police have that authority; but they in turn have no authority to decide whether a convicted person is punished. Actually, in some jurisdictions the police can deal with fines for minor traffic offences. And in some systems prosecutors are empowered to dismiss cases if the accused is prepared to pay some compensation to the

victim, or even make a payment to a charity nominated by the prosecutor. Judges (and jurors) decide upon sentences, but only in accordance with previous law. Punishment is then in the hands of the executive, which, in general, has no authority not to administer due punishment. It could be said that we, through the legislature, issue a threat; 'we', however, cannot revoke it. From 'our' point of view, punishment is automatic.

This may seem to ignore the discretion exercised by each branch of the penal system. The police are granted some authority not to pursue a complaint, or merely to issue a warning to an offender; a prosecutor may choose not to prosecute a case; a judge typically has some discretion in sentencing. There are also exceptions to the general picture I have sketched, such as the power of executive clemency. And in so far as this discretion extended to permit the authority to take account of the deterrent effects of particular punishments it would, of course, reintroduce the original worry.

That discretion does so extend in our own legal systems seems clear. But we are not bound to this. Though there is much discretion in the operation of the law, none of it is arbitrary. When an offender has been apprehended, the default presumption is that he will be prosecuted and punished, and at every level, if the law is not to take its normal course, this should be because allowing it to take its normal course would be against 'the public interest'. I need not now give a full account of what is meant by that phrase, but it need not be, and is not, construed in a narrowly utilitarian way: it already encompasses notions of what is just or fair, for instance. It would not, then, be an unrealistically radical step to embody in this notion the value of individuals, which lays a moral constraint upon merely using them. So we can plausibly say that if, at any stage in a particular case, the question is raised of whether the law should take its normal course, that question should not be answered by reference to whether the normal course will deter others. (The UK Criminal Justice Act 1991 signalled a clear move away from deterrent considerations in sentencing: individual sentences were to be based primarily upon the 'seriousness' of the offence, with the possibility of incapacitative sentences for some offenders who posed a great risk to the community.)

We have, then, a system in which the legislature issues a threat whose execution is, from its point of view, automatic: it need, and can, do nothing further to ensure that it is carried out. Of course those who are involved in carrying it out are granted some discretion in this. But so long as that discretion does not extend to deciding individual cases on the basis of deterrent effectiveness, it can correctly be said that the point of the penal system is to deter, but that the offender is not punished in order to deter others.

IV

I turn now to the limits on punishment. If deterrent threats are made in self-defence, then they are subject to the principles of self-defence, a central one of which is that it is permissible to use in self-defence only such force as is 'reasonably necessary'. What does that mean?

Imagine that I have only three possible responses to an act of aggression already under way: doing nothing, rational persuasion, and counter-force. What I may permissibly do will be a matter of the likely costs and benefits, broadly construed.

For each course of action, the primary benefit aimed at will be that the threatened harm is avoided or minimized. So we need to know both how great the threatened harm is, and how likely it is to occur given each course of action. If the threatened harm is slight and merely speculative, then normally the correct course of action will be to do nothing, though in such circumstances there will usually be alternatives besides the three just mentioned. At the other extreme, a high probability of serious harm if I do not use force will often justify it. But not always: in particular, if it has in any case no chance of warding off the threat, then it will be pointless violence, and cannot be justified under the principles of self-defence.

As to what count as costs, there will be some dispute. Some will consider the harm to the aggressor a cost, but others will not. Most would think that the degree of the aggressor's responsibility would need to be taken account of, but not necessarily as limiting the victim's rights; it may be unkind to use force against an 'innocent threat', but it is not necessarily unjust. Some may consider it a (moral) cost to oneself to indulge in violence, even if it is justifiable to do so, whereas others will not. But there are some costs that are clear: a forceful response to aggression may simply trigger a forceful counter-response on the part of my aggressor, for instance; or it may cause harm to innocent third parties. And on any plausible view, forceful action will in general be more costly than non-forceful action.

These considerations, which I shall refer to as 'the restraining considerations', determine when it is reasonable to use force, and how much force it is reasonable to use. Dispute about their precise content can be settled only by appeal to their source. About that too, of course, there will be dispute, and I can do little more here than state baldly what I believe it is. It stems, I believe, from the general social necessity to minimize the use of violence. It is clear that we need disincentives to violence, and recognizing the right to self-defence is one of the most important of them. But forceful self-defence

itself needs restrictions, for it too poses dangers: it may spill over into, or be confused with, vengeance, with all of the attendant dangers; it may be used as a cloak for violence desired for other ends; it may provoke further violence from the aggressor, either in self-defence or in revenge; and it poses dangers to those who are uninvolved. On the other hand, an absolute prohibition on the use of force would obviously not be desirable, for this would remove a major disincentive to aggression. And it would not be generally adhered to; most would think that, in its encouragement of aggression, the prohibition made no sense, and they would not respect it. In any case, self-interest would ensure that people would not routinely forego the use of force when this entailed a significant sacrifice to themselves.

So a compromise has to be found between an absolute prohibition and no restriction at all. It is impossible to say with precision just where it will be located. But the guiding principle will be that the use of force should be confined to limits we can sensibly expect people to abide by, and force which is reasonably necessary is simply force within those limits. There will, obviously, be some disagreement about what those limits should be; but there is also considerable agreement. Virtually everyone agrees, for instance, that the victim of aggression should be required to forego self-defensive force when there is a significant disparity between the harm threatened and the amount of force required to thwart it: this is a restriction that normal people will see the point of, and will generally abide by. And when it is reasonably necessary to use force, the restriction on the amount of force that may be used has of course the same point. Its effect is to reduce forceful countermeasures to the level that we can sensibly demand. We require, then, that victims should sometimes be prepared to sustain some loss when responding forcefully to an aggressor by choosing a less forceful measure than they might prefer. There is, of course, no way in practice of saying precisely how great a loss it is reasonable in general to require; it is no surprise, therefore, that judgements vary over this matter.

To turn now to *deterrent threats*, again imagine that I have three possible responses to a threat of aggression: I may do nothing, I may try to persuade potential aggressors not to aggress, or I may threaten them with retaliation in order to deter them.

The restraining considerations operate much as before: we need to know the scale and likelihood of the threatened harm, the likelihood of success of each course of action, and the costs of each course. But different factors will now become salient. (They, or counterparts, are of course present in the case of direct self-defence too.) For instance, since we are dealing with merely potential aggression, we shall need to think of how tempting is the aggression to potential aggressors: the more tempting it is, the greater

the harm it will be justified to threaten (other things being equal). There is also the possibility that the threat may act as a perverse incentive, or that retaliation will lead some to seek revenge. And if we are talking of semi-automatic systems of retaliation, the possible harm to innocent third parties will loom much larger.

It may seem natural to think that it would be permissible to threaten in deterrence all and only that which it would be permissible to do in direct self-defence. But this is not, of course, correct. The restraining considerations mentioned above generate crucial divergences. First, we may take account of the probability that the threat itself may work, and the retaliation thus may not be triggered. This will often relax the limits. Indeed, other things being equal, there would perhaps be no limit at all on what might be threatened if we could be absolutely sure that the threat would be effective.

On the other side, there is the crucial question of immediacy. Potential aggressors can often be deterred by the threat of much less force than would be required to prevent their aggression once it has started. (It may be necessary for you to kill someone who is trying to kill you; but he might have been deterred if you had let it be known that if someone kills you, your friends will administer him a beating.) More generally, all sorts of resources that might have been available in advance will not be available in the heat of the moment.

The issue of immediacy also raises the question of scatter. If I booby-trap my house, and post a notice to this effect, the trap may still be triggered by innocent intruders who have not seen the notice. This problem is not confined to self-defensive threats, of course: it applies also to direct self-defence. But it is presumably a more serious problem in the case of automatic and semi-automatic systems of threatened retaliation than in most cases of direct self-defence, and requires a greater stringency in the restrictions governing their use.

Further complications arise when we move from *individual* self-defensive deterrence to *collective* self-defensive deterrence on the part of society against those who aggress against its members; there will then be special empirical facts, and, arguably, special normative considerations to take account of. But the goal of reducing violence will still be appropriate. A threat of retaliation may be a successful disincentive against aggression. But even, or perhaps especially, when wielded by the government, it must be hedged about with restrictions. For one thing, we are rightly reluctant to give government more coercive power than is necessary, for its agents are ordinary people, as likely to misuse power as others. And we are reluctant to allow government to resort to force too easily, for this is a lapse from the ideal relationship

between government and citizens. And even when serious punishment is completely justified, it is likely to foster alienation from the government on the part of at least some citizens (the family and friends of the offender, for instance). There is also the question of likely harm to innocents. As well as the wrongly convicted, there are others, such as the innocent families of properly convicted offenders. And the more serious the punishment we threaten, the greater the costs we shall feel constrained to pay in the attempt to avoid miscarriages of justice. The government is required, then, not to use the threat of punishment to prevent crime if there is a reasonable, less violent alternative; and it may threaten no more than is reasonably required for deterrence.

So though the principles are the same, what it is permissible for an individual to do in direct self-defence will often diverge from what it is permissible for society to threaten in self-defensive deterrence. Overlooking this may make it seem that punishment and self-defence are morally separate phenomena:

> Proportionality in punishment ... is more rigorous than proportionality in self-defence. Using the death penalty for rape, for example, violates the principle of proportional punishment.... Yet if a woman is threatened by rape, she may legally resist by killing the aggressor. Even legal systems that have abolished the death penalty permit the use of deadly force in the defence of vital interests. While proportionality in punishment requires that the sentence fit the crime, clearly more is permitted in self-defence.[6]

Fletcher speaks here of a distinction between what is permissible in self-defence and what is permissible in punishment. But we could equally well think of it as a distinction between what is permissible in direct, individual self-defence and what is permissible in collective, self-defensive deterrence. A woman faced with imminent rape faces an immediate, serious threat, which there may be no way of repelling, short of killing her attacker. A legislature deciding upon the sentence for rape faces a different problem. Its problem is to deter potential rapists, and a threat of death is not reasonably necessary for this end; for the most part, the threat of lesser punishment achieves it. So an individual woman may kill in order to resist rape, whereas the state may not punish rapists by killing them.

But if the threat of imprisonment works only 'for the most part', why should we not say that a threat of death, which may have greater deterrent force, is reasonably necessary?

First, threats of death would be unlikely to achieve the desired result. Most of those who are not deterred by the threat of relatively heavy prison

[6] George P. Fletcher, *A Crime of Self-Defense: Bernhard Goetz and the Law on Trial* (New York: Free Press, 1988), p. 29.

sentences would not be significantly more effectively deterred by the threat of capital punishment. In addition, juries would generally be unwilling to send rapists to their deaths, and this would decrease its deterrent value. This problem might be alleviated by restricting capital punishment to only a few types of rape; but this would increase the uncertainty attaching to the consequences of the offence and would correspondingly reduce its deterrent value again. Secondly, even if it did decrease the number of rapes, we should have to set against that the likely costs. They include the destruction of human life. Many people would think this justifiable to prevent an otherwise certain rape, but not justified as part of a somewhat speculative process designed to deter possible rapes. The likelihood of wrongful convictions would also strike most people as an unacceptable cost.[7] The point is illustrated in most legal systems. A woman immediately threatened with rape may kill to protect herself. But in most civilized jurisdictions she may not, in order to deter potential rapists (or even potential murderers), display a deadly weapon with intent to use it; and even those jurisdictions which allow the open carrying of firearms would baulk at a woman booby-trapping her body with a bomb which would explode if she were attacked; these measures would pose an unreasonable threat to innocent people. The government, in adopting measures to deter rapists, is in the same position.

In order to explain the phenomenon remarked by Fletcher, then, we do not need principles of punishment intrinsically more restrictive than the principles of self-defence. On the other hand, if we derive the principles of punishment from the principles of self-defence, then anything that is permissible in direct self-defence will indeed be permissible, in some imaginable circumstances, in deterrent threats. Imagine, for instance, a desert island inhabited by two people, one of whom is a determined rapist who can be deterred only by the genuine threat of death. Would it in those circumstances be permissible to set up a deterrent threat of automatic deadly retaliation? In these unlikely circumstances, it might be. It would depend in part upon the amount of violence that this threat, and its realization, might themselves trigger, and in part upon the relative evaluation of sexual autonomy and human life. About that latter consideration the protective-deterrent theory has itself, like other theories of punishment, nothing to say. All that can be said is that *if* it is permissible to kill to avoid an otherwise certain rape, *then* there will be some conceivable circumstance in which it would be permissible to mount a credible threat of death to deter a potential rapist. But the legal system in the real world never finds itself in circumstances like these.

[7] Since 1973 'At least 96 people have been exonerated and freed from death rows in 22 states' (*The New York Times*, August 24th, 2001).

V

What would a legal system look like that followed these principles?

The range of offences that could be punished would be those alone which could plausibly be brought under the umbrella of self-defence. It would be impermissible to punish behaviour merely on the ground that it was immoral. And it would be impermissible to punish behaviour which harmed only the agent. The most obviously punishable behaviour would be paradigm acts of aggression, such as rape and murder. But regulatory offences, such as traffic offences, would also be covered, since the relevant regulations, if legitimate, are constraints that we are justified in protecting in self-interest.

It would have a robust and plausible requirement that the punishment should fit the crime: it would restrict punishment to what it is reasonable to threaten in order to deter the offence. Deterrent considerations, however, would play no further part in the treatment of particular cases.

A particular aspect of the idea that the punishment should fit the crime is the idea that the innocent should not be punished, and the deterrence theory can explain this. Punishment is retaliation against those who have ignored a warning. The system is set up so as to be triggered by those who ignore that warning, not by others; and no acceptable self-defensive intention would be achieved by setting it up in any other way. 'Punishing' those who are innocent, then, will have no justification, for they have not failed to comply with the warning. Unless we were, unacceptably, using them in a deterrent display, there would thus be no way of bringing their punishment under the umbrella of self-defence.

We could, of course, set up the system so that innocents might suffer the 'retaliation'. We could issue threats to punish the innocent relatives of potential lawbreakers, for instance. Or we could threaten to harm entire groups when only certain members of the group transgress. But it is clear that punishment in these circumstances would be *using* people. Those who are punished would, either directly or indirectly, be made to suffer as part of an attempt to modify the behaviour of others; the deterrence theory can quite consistently object to that.

Questions of guilt and innocence, of course, refer to more than whether the defendant committed the prohibited act. First, punishment is hedged around with the requirements of *mens rea*, or responsibility. One is not punished at all unless one acted in an appropriate state of mind. And the severity of one's punishment may depend upon one's state of mind when

one committed the offence, whether one did it deliberately or negligently, for instance. Secondly, the law recognizes a number of excuses, such as necessity or duress, which can lead to an acquittal or mitigation of sentence. All of this may seem to reflect nothing of self-defence, for our right of self-defence against one who threatens us does not turn upon his level of responsibility, or whether he is coerced into harming us. Why, then, are such questions so important in the law?

Taking responsibility first, a requirement of responsibility enters into the limits of self-defence as soon as we move from direct self-defence to deterrent threats. The reason is simple: there is no point in issuing threats to those who cannot heed them, and any violence occasioned by such threats would therefore be without justification. Anyone who suffered such violence would have been done an injustice. The law, then, may not justifiably threaten with punishment, for instance, the severely mentally ill, and so may not punish them. (It may, of course, detain them for the safety of themselves and others.) The threat of punishment is not addressed to them.

But some of the mentally ill, kleptomaniacs, for example, are capable to some degree of being deterred, and threats, though they may eventually be ignored, are not without effect. Is it then permissible to punish such people? If it would not be pointless to threaten them with punishment, then punishment, within the restraining considerations, would not be unjust. So they may be punished without injustice. It is a misfortune for them that they suffer this obsession; but others are not required to bear the burden of their misfortune, and may use whatever force is reasonably necessary to protect themselves against its consequences. Whether carrying out punishment in such circumstances would be stupid, or callous, are, of course, different questions.

To turn now to negligence, we seem to have the same right of self-defence against a negligent threat as against a deliberate one. But, typically, the law treats deliberate misconduct more severely than negligent misconduct. How can that be, if punishment is a form of self-defence?

We do indeed have the same rights of self-defence against negligent and deliberate threats, but when we think about self-defensive deterrence, as opposed to direct self-defence, we should be more concerned about the prospect of deliberate misconduct than the prospect of negligent misconduct. Deliberate misconduct is in general more likely to cause harm than negligent misconduct: we have more reason to be concerned about someone who intends to kill than about someone who simply drives his car recklessly, for, in general, the former will pose a much more serious threat. Additionally, deliberate misconduct threatens the fabric of the social order in a way negligence does not. Further, it is generally easier to deter negligent

behaviour, which requires only more care, than deliberate misconduct, which requires abandoning a positively desired course of action. All of this, other things being equal, justifies a correspondingly less forceful measure.

To turn now to excuses, such as coercion, the actions of one who is severely coerced are guided by reasons, and so this is not like the case of the mentally ill. On the other hand, we can be quite sure that those who are severely coerced will not respond to threats of punishment (those who coerce them will usually find it easy to ensure this); so threats would again be pointless, and any violence they involved would be unjustified. So legal threats are not addressed to those who are severely coerced; coercion is an excuse. (Typically, only severe coercion is an excuse. It does not usually excuse in murder cases. But if threats of punishment are no more likely to deter potential murderers than others, coercion should be an excuse here too, as some legal theorists have argued.)

Does this mean that it is unjustifiable to issue threats of punishment whenever we are sure that this will not deter? And does that in turn mean that it is unjustifiable to threaten punishment to hard-core, recidivist criminals? The answer to the first question is 'Yes'; to the second question 'No'. 'Hard-core recidivist criminals' may offend whatever we do, but they would commit far more offences if they had *carte blanche* to do so. Deterrent effectiveness is always a matter of degree, and one who eventually offends may none the less have been deterred *to some degree*, in the sense that the range of circumstances in which he would offend is restricted. The threat is thus not pointless, and the violence it involves may be justifiable.

I shall deal with one last issue, the law of attempts. An unsuccessful attempt to commit a crime may itself be a separate crime: one who tries to murder but fails may be convicted of attempted murder. This may seem odd from the point of view of the deterrence theory:

> those who set about crime intend to succeed and the law's threat has all the deterrent force it can have if it is attached to the crime; no additional effect is given to it if unsuccessful attempts are also punished.[8]

But the deterrence theory can explain why we have separate offences of attempt. If we were setting up a system of semi-automatic retaliation, we should not set it up so that the retaliation would be triggered only by the completed offence. Our first thought might be that we should set it up, if possible, so that it would be triggered *before* the offence was complete, indeed as soon as the offender had fully committed himself to his course of action. But the issue is more complicated. Given the goal of reducing violence, a rational strategy would not only be to give potential aggressors an incentive

[8] Hart, *Punishment and Responsibility*, p. 128. (Hart thinks that this is a 'fallacy'.)

not to aggress in the first place; we should also want to give actual aggressors an incentive to desist from their actions even when they had embarked on them, and to give them such an incentive for as long as reasonably possible. However, if we set the trigger point very late, we might encourage potential offenders to embark on offences, and to continue with them at inconvenience to others, knowing that they could later withdraw if it seemed prudent. If, on the other hand, we set the trigger point very early – as soon as they had started planning the act, for instance – then this would deprive offenders of what might have been an effective incentive to desist once the act was under way; in addition, enforcing this would require enormous resources, and considerable general deprivation of liberty. The challenge, then, is to find the point at which one's self-defensive strategy would be optimized within the restraining considerations mentioned earlier.[9]

This goal in turn makes intelligible why unsuccessful attempts are usually punished less severely than completed attempts. The emphasis on moral desert that characterizes many theories of punishment makes this puzzling.[10] But from the point of view of the protective-deterrence theory, moral desert will of course not be the issue. All that will be relevant are the considerations mentioned earlier: roughly, what will be the likely costs and benefits of threatening various levels of retaliation against, say, unsuccessful attempts? We have reason to want offenders to desist from their offences even when they have already embarked upon them; given this, it would be perverse to threaten them with the full punishment as soon as they embark on the offence, for then they would have no incentive to desist as soon as the likelihood of apprehension is as great as when the offence is accomplished. A natural thought would be a sliding scale of retaliation: roughly, and other things being equal, the further along the course of his action the offender had progressed, the greater would be the retaliation. This would not, of course, be practical in the criminal law. Our best approximation is to fix a point at which an attempt really has been made, as opposed to mere preparations, and a point at which the attempt has been completed; between those two points we punish, but less severely than for the completed attempt.

Virginia Commonwealth University

[9] And we find in jurisprudential thought just what we should expect, given this aim. But for a different approach, see Duff, *Criminal Attempts* (Oxford: Clarendon Press, 1996).

[10] Duff has tried to explain the puzzle in *Criminal Attempts*. I have criticized Duff's argument in my 'Criminal Attempts', *Journal of Applied Philosophy*, 15 (1998), pp. 207–12.