The Justification of Deterrent Violence*

Daniel M. Farrell

I shall use the phrase 'deterrent violence' to refer to violence that is aimed at inducing someone not to do something she might otherwise be inclined to do. And I shall suppose that violence can be so aimed, and hence properly called "deterrent" violence, independently of whether or not the person whose behavior one is trying to control is the same as the person on whom the relevant violence is inflicted. Thus, I might harm someone as a way of attempting to induce *him* not to try to harm me again, or I might harm someone as a way of inducing other, potential attackers not to do me harm. And, of course, I might harm someone, once she has wrongfully harmed me, in hopes of deterring both her and others from acting against me in the relevant ways in the future.

So conceived, deterrent violence is not easy to defend. For it has—especially when its aim is the deterring of persons other than the person currently being harmed—the appearance of being a way of "using" one person as a means of protecting ourselves from others.[1] Of course, it is sometimes said that we can avoid this difficulty by resorting to deterrent violence only after having *warned* potential recipients that they are liable to be treated in the relevant ways if they do the sorts of things we are trying to deter. And in fact I shall suggest a defense of deterrent violence along just these lines in the body of the present paper. Notice here, though, that this appeal to the use of antecedent warnings—or *threats*, as I should prefer to say—raises difficulties of its own. For it is not at all obvious how warning someone that we will harm her in some way, if she does something we want her not to do, is connected with the justifiability of actually harming her in that way if and when our threat is ignored.

* An earlier version of this paper was presented at the 1987 Pacific Division meetings of the American Philosophical Association in San Francisco, and I am indebted to Warren Quinn, who was my commentator on that occasion, for a number of extremely helpful comments on that and on a subsequent draft. In addition, I am very much indebted to James Bogen and to Don Hubin for their invaluable help with that earlier paper, as I am also indebted to the very useful comments suggested by two anonymous referees for *Ethics*.

1. The classic statement of this objection is Kant's, in *The Metaphysical Elements of Justice*, trans. John Ladd (Indianapolis: Bobbs-Merrill, 1965), pp. 100–101.

Ethics 100 (January 1990): 301–317

My aim in what follows is to lay out at least a rough sketch of a general solution to the problems just described. What I shall try to show is that all the deterrent violence we are intuitively inclined to think is morally justifiable can be justified by principles that are remarkably close to the principles that justify violence with a much less controversial aim—namely, violence undertaken in direct self-defense. I shall begin by showing that in certain cases deterrent violence can be justified by exactly the same principle that justifies violence aimed at direct self-defense. I shall then explore the question of what we have to add to our account of self-defense if we are to be able to justify some of the other cases of deterrent violence that we are normally inclined to think are morally justifiable. Of particular interest here will be the question of how it is that having threatened deterrent violence bears on the justifiability of actually inflicting that violence if one's threat is ignored. I shall attempt to answer this question in two stages: first, by examining a type of account that has received a certain amount of attention lately but that I think fails, and then by presenting a positive account of my own.

I

When we conceive of deterrent violence in the manner suggested above, we can distinguish violence of this sort from violence perpetrated in what I shall call "direct self-defense" as follows: violence perpetrated in direct self-defense is violence that is aimed at stopping an on-going attack; deterrent violence, by contrast, is violence that is aimed at preventing, through intimidation, the initiation of the sort of attack that direct self-defense is aimed at stopping once an attack is under way. Given this distinction, we can see why violence aimed at direct self-defense might be of special interest to those who are interested in the justification of deterrent violence. For most of us are inclined to think that under the right circumstances, violence aimed at direct self-defense is a paradigm of whatever violence might be morally justifiable. Hence if we could show that deterrent violence is, under the right circumstances, itself a kind of self-defense, or that it is close enough to self-defense as to be clearly justified by whatever justifies the latter, we would have made at least some progress toward explaining when and why deterrent violence is itself morally justifiable.

Now I have argued elsewhere that we can justify at least a certain amount of deterrent violence on exactly the same grounds that we can justify violence used in self-defense. To see this, we need to note, first, that the right to self-defense can itself be seen as arising out of the rights attributed to us by the following principle of distributive justice:

> P1: When someone knowingly brings it about, through his own wrongful conduct, that someone else must choose either to harm him or to be harmed herself, justice allows the latter to choose that the former shall be harmed, rather than that she shall be harmed, at least if the harm inflicted on the former

> is roughly proportional to the harm that would otherwise be inflicted on the latter.[2]

It will perhaps be obvious how this principle grounds the right to direct self-defense. Less obvious, however, but equally important, is the fact that if we accept this principle, we can also justify at least a certain amount of *deterrent* violence as conceived above. What's more, we can do this without invoking the notions of threat or warning that we suggested above might be necessary for a complete account of the justification of the latter form of violence. To see this, we need to note that in certain situations an innocent party will have been made more vulnerable to subsequent attack, as a result of an already accomplished attack, if the latter attack is left unpunished. Suppose, for example, I am situated in a Lockean state of nature and someone wrongfully harms me. Clearly, someone else, observing this attack, and observing as well that I do not retaliate against it, might be inspired by what he sees to attack me himself. And he might be inspired to do this even though he would not have dared to attack me if my original attacker either had not attacked me in the first place or had been punished by me once he in fact attacked. But, then, in such circumstances we can say that my original attacker has put me, by attacking, into exactly the sort of situation P1 describes: a situation in which I must choose, as a result of another's wrongful action, either to harm him or to be harmed myself.[3]

So P1 can be used to ground at least a certain amount of deterrent violence. Unfortunately, I think it is clear that not all the violence we are inclined to think we are justified in inflicting, as a way of securing either special or general deterrence, can be justified by a principle of this sort. To see this, we simply need to note that P1 applies, and justifies the relevant sort of violence, only if we assume that the wrongdoer in

2. Here I am indebted to Phillip Montague, "Punishment and Societal Defense," *Criminal Justice Ethics* 2 (1983): 31–36. See too, for an elaboration of the remarks of the next few paragraphs, my paper "The Justification of General Deterrence," *Philosophical Review* 94 (1985): 367–94, and also my "Punishment without the State," *Nous* 22 (1988): 437–53. It will be obvious, I hope, from what follows, that I do not intend anything I say in the present paper as a defense of P1 or of any of the other principles I articulate below. The point is simply to show how far we can go, in the justification of deterrent violence, if these principles are granted.

3. I am assuming, of course, that to be subjected to a higher probability of being wrongly harmed is itself a way of being harmed. For a defense of this assumption, and a discussion of some other, related difficulties with the argument from P1, see my "Punishment without the State." (One should also note, in this connection, that the argument above would obviously not be available to us if P1 restricted a potential victim's admissible options to protective actions that are necessary if she is to avoid being harmed *by her current attacker*. A principle that limited a potential victim's options in this way, however, would itself be a mere corollary of P1. And while space prohibits attempting to prove this here, I think it can be shown that the very considerations that would incline us to accept the narrower principle could also be used, without additional moral assumptions, to support the more general principle. For more on this, see Farrell, "The Justification of General Deterrence"; and Montague.)

question is both causally and morally responsible for our increased vulnerability to others' wrongdoing. Suppose that in a certain situation this is not the case. Suppose, for example, that while my vulnerability will not be *heightened* if I do not retaliate against a given attack, I have good reason to believe that if I do retaliate, I will increase my overall security. Or suppose that while my vulnerability will indeed be heightened if I do not retaliate against a given attack, I know that I can enhance my previous level of security, vis-à-vis others, by doing even more to my attacker than I need to do in order to bring my security level back to where it was prior to his attack. It will be tempting, of course, to retaliate, in the first sort of case, and, in the second, to do more than I have to do in order to get back to where I was. If I do retaliate, however, in the one case, or do in fact do more, in the other, than our principle of distributive justice allows, and if I wish to say that I am justified in doing so, I obviously cannot base my actions on P1. For by hypothesis the wrongdoer in question is not responsible for my being in a position where I have to decide either to harm him (at least to the degree to which I am tempted to harm him) or to let myself be harmed by others. If we assume, therefore, that I may justifiably harm him as a means of reducing the probability that I will be harmed by others, it must be because we are assuming the validity of justificatory principles other than P1.

II

As we have already seen, it is tempting to think that a plausible alternative account can be constructed via the notion of antecedent warnings or threats. The idea, very roughly, is to show that we are certainly justified in *issuing* threats as a way of deterring aggressive violence, and then to show that, at least in the right circumstances, we are also justified in enforcing such threats when they have been justifiably made and then ignored. Proceeding in this way, we might suppose that our account would ultimately rest on some such principle as the following:

> P2: When my situation is such that either I make and then enforce certain threats as a way of protecting myself from unjust aggression, or I do not make and enforce such threats, thereby leaving myself more vulnerable to aggression than I would otherwise have been, I am entitled to choose the former alternative over the latter, at least if what I threaten is proportional to what I am trying to prevent by issuing these threats in the first place.

It is not entirely clear, of course, exactly how an argument from a principle like this would enable us to get around the Kantian objection sketched above; this is something we shall have to explore at greater length below. Still, the direction such an argument would take is clear enough: we defend making the relevant threats by appealing to an appropriate principle of self-defense, and then we defend enforcing them by appealing to P1

or some near relative of P1 (i.e., by showing that in ignoring our threat, our attacker has put us into a position where we must either harm him or be harmed ourselves because of a possible loss in credibility). Before pursuing this line of argument, however, I want to take note of an argument that has been advanced by a number of other writers and that, while aimed at establishing a principle very much like P2, aims at establishing that principle in a way that turns out to be much simpler and much more straightforward than the way I shall adopt below. Very briefly, this latter line of argument suggests that the best way to understand the nature and justification of the sort of strategy that is suggested by a principle like P2—I shall call it an "ordinary threat strategy" or an "OTS" for short—is in terms of an antecedent understanding of another strategy of a rather different sort: what I shall call an "automated-retaliation strategy" or an "ARS" for short.[4]

Central to the notion of an ARS is the notion of a certain kind of device: what I shall call an "automated retaliation-device" (an "ARD" for short). An ARD is a device very much like the celebrated "doomsday devices" of nineteen-fifties popular fiction: a device that can be programmed to retaliate against aggressive action and that, once in place, cannot be "de-activated" for some specific period of time. Indeed, the only difference—and it is, of course, an important one—between "doomsday devices" and ARDs is that the latter are to be presumed to be programmable for all sorts of different retaliatory tasks.[5]

Suppose it were possible to construct ARDs and to construct them so that they could be programmed to do, to any potential wrongdoer, whatever we want a potential wrongdoer to expect will happen to her in the event she does the relevant wrong. It would then be possible, it seems, to do with ARDs exactly the sorts of things we might want to do with ordinary threats in the interests of protecting ourselves against unjust aggression. For we could simply program the relevant devices to do, for each kind of offense, whatever we would otherwise wish to threaten (and then do, if our threat were ignored) in relation to that kind of offense, publicly informing everyone who will be subject to the relevant retaliation that this is what we are doing.

So much for the general idea behind the utilization of an ARS. Let us now ask why anyone would even dream of saying that if we were capable of constructing devices of the relevant sort, we would, at least in the right circumstances, be justified in using them as a way of protecting

4. See, especially, Lawrence Alexander, "The Doomsday Machine: Proportionality, Punishment and Prevention," *Monist* 63 (1980): 199–227; and Warren Quinn, "The Right to Threaten and the Right to Punish," *Philosophy and Public Affairs* 14 (1985): 327–73.

5. Here I follow Quinn as opposed to Alexander, who holds the very odd view that we would be justified in protecting ourselves with ARDs, if we have them, by programming them to enforce the same penalty—death—against every act of aggression that anyone might ever be tempted to undertake (provided, of course, that potential offenders have been warned of what will happen to them if they harm us). See Alexander, pp. 74 ff.

ourselves from possible instances of unjust aggression. Obviously, if we make a number of important assumptions, we might say that someone who believes in the permissibility of resorting to such devices implicitly commits himself to some such principle as the following:

> P3: When my situation is such that either (*i*) I resort to an ARS as a way of reducing the risks of unjust harm to myself or (*ii*) I remain, by virtue of not having done this, at greater risk of unjust harm myself, I may rightly choose (*i*) over (*ii*), at least if what I program the relevant device to do remains within certain limits.[6]

But why would anyone ever suppose that this principle is sound?

A complete answer to this question would require more space than is available to us here. I think it is clear, though, that one likely defense of P3 will have to do with the following feature of the situations to which it applies: in such situations, it is by hypothesis possible to lower the absolute probability of harm to certain innocents without actually doing harm to anyone, innocent or otherwise. Of course, to effect this reduction in the absolute probability of harm to the innocent, we have to raise the conditional probability that those who do wrong to the innocent will themselves be harmed. However, this latter feature of an ARS, which we can think of as the "price" of adopting such a strategy, is presumably a price that the advocates of such a strategy would say we have a right to impose. For is it not better, they would say, that the innocent be spared a certain probability of absolute harm than that their potential attackers be spared an increase in the conditional probability that they will be harmed if in fact they wrongfully attack the innocent? After all, the latter have no right to do the relevant wrongs in the first place, and hence by increasing the relevant conditional probabilities, we are not depriving them of the right to do anything they would otherwise have a right to do. And if, by imposing those increased (conditional) probabilities, we can protect the innocent, without thereby doing any direct harm to others, why should we not do it, especially since others will be harmed only if they ignore the existence of the ARDs and do wrong themselves?

There are difficulties with this line of argument which we shall not be able to pursue here. Difficulties to one side, though, it will be clear why one might be tempted to think that resort to an ARS would be justifiable under the relevant circumstances and why one might be tempted to think that the emplacement of the relevant ARDs might plausibly be

6. I am assuming for the sake of simplicity that in the circumstances to which P3 is designed to apply, an ordinary threat strategy is for some reason unavailable to us (e.g., because we know it won't work). In addition, I am skirting, because of limitations of time and space, the question of just what limitations we would have to observe in programming an ARD for retaliation against a given (kind of) offense. Obviously, consideration of this latter question would be crucial if we were interested in anything more than the in-principle justifiability of the use of ARDs.

said to be a kind of self-defense. For all we are doing in putting them in place is trying to reduce the probability that we will be unjustly harmed. And since we are doing this not by harming any other innocents, but simply by ensuring that those who harm us will be harmed themselves, it is hard to see how anyone could object to it on the grounds that in resorting to such a strategy we are somehow doing wrong to others.

III

Suppose we accept this admittedly very sketchy account of the justification of automated-retaliation strategies as basically sound. What is the relevance of this account to the problem that actually interests us—that is, the problem of justifying the making and enforcing of ordinary threats as a way of attempting to deter unjust aggression against us? One very tempting view is what I shall call the "direct entailment view" or the "direct view" for short. According to this view, what we are doing when we attempt to control aggression by making and enforcing ordinary threats is really no different, in principle, from what we would be doing if we were to attempt to control aggression by constructing and then publicly activating a suitable set of ARDs. In particular, this view holds that (*i*) the principle that justifies us in making and enforcing such threats is essentially the same as the principle that would justfiy us in utilizing an ARS in exactly analogous circumstances, and hence that (*ii*) any circumstances in which we would be justified in using an ARS to protect ourselves from another's aggression are *eo ipso* circumstances in which we would be justified in making an ordinary threat to achieve this same end and then in enforcing this threat in the event it is ignored.[7]

I shall explain in a moment why I believe the direct view is false. Notice first, though, how convenient it would be if it could be sustained. For one thing, it suggests a very straightforward answer to our question about the normative basis of deterrent activities, when these latter take the form of making and enforcing the relevant sorts of threats. For it tells us that the basis for these activities is a principle that is essentially the same as the principle we have suggested for the justification of ARSs: namely, the principle P3. Second, though, and quite important, notice that if we could provide an account of the limits that P3, and hence the use of ARDs, requires us to honor, the direct view would provide us with at least a rough account of what we may justifiably threaten and then do (if our threats are ignored) by way of attempting to diminish the probability of unjustified attacks against us. For if the direct view is right, what we may justifiably threaten, and then do, if our threats are ignored, is equivalent to what we could justifiably program an ARD to do in an analogous situation.

7. Alexander; and Quinn. The latter is by far the more sophisticated statement (and defense) of the direct view, and it is this version that I have principally in mind in what follows.

Unfortunately, I think it can be shown that the direct view is overly simple and hence, at least as we have stated it, unacceptable. To see this, we must begin by noting one important difference in what one is doing when one installs an ARD as a way of deterring potential aggressors and what one is doing when one makes an ordinary threat in order to achieve the same end. In resorting to an ARD, one is deliberately taking the issue of whether or not retaliation will occur out of one's own hands. One is acting to *ensure* that retaliation will occur, that is, in the event one's threat is ignored. In the case of making an ordinary threat, by contrast, one is not—in making the threat—taking the issue of whether or not retaliation will occur out of one's own hands. For regardless of how sincere I am in what I say—that is, regardless of whether or not I really mean to do what I say I will do—making or issuing that threat does not *ensure* that retaliation will occur. On the contrary, I think we can say it is a *threat,* in the ordinary sense of the term, precisely because it is a declaration (sincere or not) of what I intend *to do* if the considerations stated are ignored.

Now this point has the following consequence. In the case of an ARS, there is just one thing—I mean just one action—to be justified: installing the ARD. In the case of an OTS, by contrast—when we suppose such a strategy is being used to achieve the same end as the corresponding ARS—there are, at least potentially, two things to be justified: the act of issuing the threat and, if it is ignored, the act of enforcing it (supposing it is enforced). It is possible, of course, that whenever the relevant threat would be justified, enforcement would be justified as well. This remains to be seen. The point to be noted here is simply that with a threat-and-enforcement strategy, there are in fact two separate actions that will sometimes have to be justified.

We may now return to the direct entailment view. As we have seen, this view holds that attempting to deter aggression by making and enforcing ordinary threats will be morally justifiable whenever an ARS would be justified for the same purpose. And it holds, as well, that the principle that justifies making and enforcing ordinary threats is essentially the same as the principle that justifies the use of ARDs. What, then, is that principle? One possibility is the following:

> P4: When my situation is such that I must either threaten another with conditional retaliation as a way of reducing the risks of unjust harm to myself, or refrain from threatening him with retaliation, thereby leaving myself at greater risk of harm, I may justifiably issue the relevant threat—provided that what I threaten remains within certain limits—and I may justifiably *enforce* that threat, if I am able to do so, in the event that it is ignored.

Principle P4 is indeed analogous, at least superficially, to P3: each suggests that when my situation is such that I must either suffer an increased

probability of harm or else subject my potential antagonist to an increased probability of harm, I may choose the latter, at least if (*i*) my threat remains within certain limits, (*ii*) the question of whether my antagonist will really be harmed depends entirely on whether he chooses to harm me, and (*iii*) my antagonist is aware of the fact that whether he will be harmed depends entirely on whether he chooses to harm me. What's more, if we suppose that P4 is sound, what I shall call the direct view's "extensionality component" would appear to hold as well: P4 does appear to entail, that is, that whenever I would be justified in attempting to deter aggression by activating an ARD, I would also be justified in attempting to deter that aggression by making an ordinary threat and then enforcing that threat if it is ignored. There is, however, at least one very serious problem with P4: it is not clear why we should suppose that its final clause, according to which a threat that is justifiably made may justifiably be enforced, is anything but ad hoc. To be sure, the proponent of the direct view needs this clause in order to ensure that what is justifiable, on his view, is not just the *making* of threats but their enforcement as well. And he needs this, of course, because his claim is (very roughly) that a strategy of making and enforcing deterrent threats is morally justifiable whenever an ARS would be morally justifiable in exactly similar circumstances. Still, it is not clear why we should suppose that the advocate of the direct view is entitled to assert the final clause of P4, even if we suppose that P3 is sound and hence that the advocate is entitled to assert, in an analogous principle, whatever the thrust of P3 will warrant. For the point of P3 is that in certain circumstances we may take certain steps to see to it that we are better protected against unjust aggression than we would be if we did not take those steps. And it is not at all clear that this supports anything more than making threats, when self-protection requires this, and then enforcing them when this too is required for self-protection.

We shall consider in a moment the possibility that the defenders of the direct view can get around the problem that has just been raised by constructing an argument to show that the clause that interests us is not in fact ad hoc. First, though, let us consider the following possibility. Let us suppose the direct view's defender abandons P4 in favor of a principle like the following:

> P5: When my situation is such that either I make *and then enforce* a threat of conditional retaliation, thereby lowering the risk of unjust harm to myself, or I do not make and enforce such a threat, thereby leaving myself open to a higher risk of unjust harm, I may make and enforce the relevant threat, provided that what I threaten and then do, if my threat is ignored, remains within certain limits.

We shall ask later exactly how this principle would be applied and why it might be thought to be more plausible than P4. Here I simply want

to observe that the defender of the direct entailment view cannot in fact abandon P4 in favor of a principle like P5 and still maintain his view: namely, that attempting to deter aggression by making and (when they are ignored) enforcing ordinary threats is functionally, foundationally, and extensionally equivalent to attempting to deter aggression via the construction and public installation of an ARD. For if we continue to hold P3, the substitution of P5 for P4 entails that there are cases where I would be justified in attempting to deter aggression with an ARD but where I would not be justified, in exactly the same sort of situation, in attempting to deter aggression by making and then, if it is ignored, *enforcing* an ordinary threat. To see this, imagine that I am situated on a desert island with just one other person, and suppose that my only way of deterring this person from an expected act of homicidal aggression is to install an ARD which is programmed to kill him if he attempts to kill me. If our account above is right, I would be justified in installing the ARD in such a case, announcing that I am doing so, as a way of deterring the relevant crime. But now let us suppose that in an exactly analogous situation I attempt to deter the relevant aggression not with an ARD but with an ordinary threat. And let us suppose my threat is ignored: I am attacked, but the attack fails and I am not killed. In a moment we shall examine at length the question of when I would be justified in enforcing the threat of death that I had previously made, and why, and when I would not. Notice at once, though, that there is one kind of situation in which it is arguable that I would not be justified in enforcing it, and in which, in any case, I would certainly not be justified in enforcing it on the basis of P5. This is the case in which I discover, after my threat has been ignored, that there is no chance whatsoever that my solitary adversary will ever attempt to harm me again. He has a complete change of heart, let us suppose, and at the same time I come to have conclusive evidence that he neither would nor could ever attempt to attack me or anyone else again.

P5 does not support enforcement in a case like this, of course, because P5 makes the justifiability of enforcement contingent on the need for enforcement as a way of protecting oneself from avoidable aggression. And this is exactly what P4 does not do: according to P4, if the threat is justified, then enforcement is justified as well. And this, as we have seen, is what is problematical about P4: it is simply not clear why we should suppose that enforcement is justifiable simply because making the threat was justifiable.

IV

It is no accident, of course, that P4 makes the justifiability of enforcing a given threat independent of the question of whether or not there is some forward-looking reason for enforcing it when the time comes to enforce or not enforce it. For the central point of the direct entailment view, which P4 was introduced to serve, is that making and enforcing a

threat is morally justifiable whenever the installation of an ARD would be justifiable in similar circumstances. And, clearly, the justifiability of installing an ARD is not contingent, on our account, on the need for "enforcement" once the "threat" that is constituted by the existence of the ARD is ignored. An ARD is justified, at least on our account, not by considerations that have to do with whether there will be some forward-looking reason for harming a given wrongdoer once she has done wrong, but by considerations that have to do with whether a given wrongdoer is likely to be deterred by being informed of the existence of the relevant device. Thus, regardless of how we specify alternatives to P4, any principle that is going to do the job that P4 was introduced to do will have to have the feature that now concerns us: it will have to entail that enforcement will be morally justifiable in a case like the one described above.

Now I believe that this feature of the principle to which the so-called direct view implicitly appeals constitutes a good reason for rejecting that view. And I believe this because I believe that in a case like the one described above, one could not justify—on strictly nonretributive grounds—enforcing the threat one had previously—and justifiably—made. No doubt, one could justify enforcement in such cases by bringing in various retributive and quasi-retributive assumptions. Our interest here, however, is in a theory of deterrent violence that rests not on retributive assumptions but on assumptions that are no more controversial than the assumptions on which we based our earlier accounts of the right to self-defense and of the right to protect oneself with ARDs. And this, I believe, is something we cannot have, unless we assume that the justification of deterrent violence is in at least one respect importantly different from the justification of self-protection via the use of ARDs. I now want to bring out what I think these differences are, therefore, and at the same time begin the development of an account of the justification of deterrent violence that is sensitive to the constraints we want our theory to respect.[8]

V

We may begin by noting that it is the *enforcement* of ordinary threats that raises the problem we have claimed the direct view cannot handle; the actual making of the relevant threats does seem to be something that can be justified by a principle very much like P4. To see this, we simply need to reflect on what we are trying to do when we make a conditional threat of retaliation in the sorts of cases that interest us: by saying that we will do X if our antagonist does Y, we are trying to induce him not to do Y. And this would seem to be something we have a perfect right

8. There is a great deal more that might be said, of course, in defense of the direct view (see, especially, Quinn, pp. 359–73). Here, however, I am more interested in what an alternative view would look like than in attempting to do justice to the subtleties of the direct view. I attempt a far more detailed critique of the latter in "On Threats and Punishments," *Social Theory and Practice* 15 (1989): 125–54.

to do in the sorts of cases that interest us. To be sure, to do so is to introduce an element of intimidation into our relations with this other person. But this seems perfectly justifiable provided that certain conditions are met: for example, our threats are threats to harm those who wrong us, not those who are innocent of wrongdoing but whose welfare might be of interest to potential wrongdoers; we observe certain as yet unstated proportionality limits in deciding what to threaten for any given potential offense; our threats are made only as a way of preventing people from doing wrong to us, not as a way of keeping them from doing things they have a right to do; and so on.

The principle that underlies such threats, then, would seem to be something like the following:

> P6: When my situation is such that either (*i*) I threaten someone with retaliation as a way of deterring him from wronging me in some way, or (*ii*) I do not so threaten him, thereby leaving myself more vulnerable than I would otherwise be, I may rightly choose (*i*) over (*ii*), at least if what I threaten remains within certain limits and is directed only at potential wrongdoers for the wrongs they might otherwise do.

Our problem, of course, is to determine when threats that have been made in accordance with this principle may justifiably be enforced.

We can begin by considering the following variation on the case imagined above. Suppose I am on our imaginary desert island with just one other person, and suppose I have reason to believe that unless I threaten that person in a certain way, there is a good likelihood he will attempt to harm me in some (specific) way. I issue the requisite threat, therefore, being careful to see to it that the threatened penalty is within whatever limits are appropriate for a case of the relevant sort (e.g., being careful not to threaten more than I would be entitled to do in order to prevent the relevant offense in a case of direct self-defense). Finally, suppose that, as in our earlier case, this other person ignores the threat and attempts to do the relevant wrong. Suppose as well, though, that in this new case, there are good reasons, once the threat is ignored, for enforcing it: I have reason to believe, let us suppose, that if it is not enforced, I will be unjustly harmed again.

In a case like this, it seems to me, most of us would be inclined to say that I would be justified in enforcing my threat. Those of us who accept the argument of Section III above, however, would be inclined to say this not simply because we are supposing that the threat had been justifiably made, but because we are supposing both that it had been justifiably made and that, now that it has been ignored, there exists a forward-looking reason for enforcing it. Obviously, this latter feature of the new case would be crucial for those of us who do not accept the view that threats can justifiably be enforced whenever they have been justifiably

made. Hence it is to this latter feature of the new case that we must look if we are to develop an account of the sort we want.

We must proceed carefully here, however. There are, as we have seen in Section I, cases in which I would be justified in harming another, for deterrent purposes, even if I have not previously warned this other party that I would harm him if he did me wrong: cases, for example, where, if I do not harm him, I will be worse off than I would have been had I not been attacked by him in the first place. Obviously, to focus on cases like these would not help us to see how threats are relevant to the justifiability of subsequent deterrent actions. What we need is a case where the relevant violence would not have been justifiable in the absence of an antecedent threat and yet where the threat alone does not seem sufficient to justify that violence either—that is, where we need, in addition to the threat, some forward-looking reason for enforcing it once it is ignored.

Consider, then, the following elaboration of the preceding case. Suppose that the amount of harm that would be justified by P1 in this case is less than the amount of harm that I have previously threatened and, moreover, is less than what would be required to significantly reduce the probability of a second attack. Suppose, for example, that while P1 would justify me in imprisoning my antagonist for six months, I have in fact threatened to imprison him for two years, this latter penalty being, as it happens, roughly what I would have to do to him in order to be likely to deter a second attack. Suppose, though, at the same time, that the harm I have threatened, in my effort to deter the relevant wrong, is within whatever limits are required by the proportionality demands of P6 once we have worked these out.

In such a case, it seems to me, we are intuitively inclined to say that I would have a right to impose the additional harm as a way of deterring the relevant violence, given that I have warned the relevant wrongdoer that I would do so if he attempted the initial attack. But why? What is it about the fact that I have warned him that makes violence that otherwise would not be clearly justifiable justifiable nonetheless?

An obvious answer is this: once my initial threat has been ignored in a case like this, I am in a situation where it is reasonable to believe that I will be more vulnerable to subsequent harm if I do not enforce my threat. What's more, I am in this situation because my antagonist has chosen to do me wrong, knowing that this would put me in the position of having to enforce my threat, in order to protect my credibility, or else allowing my credibility to be undermined, at least to a degree, as a result of not enforcing it. Hence, at least on this view of things, I am entitled to enforce my threat because my antagonist has put me in a situation where I have to enforce it if I am to keep myself from being made even more worse off, as a result of his depredation, than I have already been made.

I believe this line of thought is on the right track. There is, however, at least one rather serious difficulty with it, as we can see by imagining my antagonist replying to this argument as follows. In a case of this sort, he might observe, I (the threatener) am in the position just described—that is, a position where I must either enforce my threat or lose (some of) my credibility if I do not—only because I made that threat in the first place. And this, my antagonist might say, is something I *chose* to do, not something that I was forced to do by anything that he did. To be sure, he might say, I am entitled, by P1, to retaliate in the present case to whatever degree is warranted by that principle: that is, to whatever degree is required to restore myself to that level of security that I would have enjoyed had I neither threatened him nor been attacked by him in the first place. However, we are supposing at present that I have threatened to do more than I would be entitled to do by P1, the question being, What entitles me to enforce that threat? And the answer to this question, according to my antagonist, cannot be that I am entitled to enforce it because he has put me in a position where I must either harm him, beyond what P1 allows, or suffer some signficant loss in my credibility (thereby becoming more vulnerable to harm myself). For he has not put me in this position, at least all by himself. Rather, I have put myself in this position, he might say, by virtue of having made the relevant threat in the first place. If I had not done that, I would not be in the position where I have to decide whether to enforce the threat, in order to maintain my credibility, or not to enforce it, thereby possibly undermining that credibility.

What this objection shows, of course, is that one very simple defense of the position that interests us will not work: we cannot say that we are justified in enforcing our threats in the relevant sorts of cases *simply* because in those cases we are in the position, as a result of another's wrongful choice, of having to decide whether to harm him or to be harmed ourselves. For, as our critic above points out, in the cases that interest us we can ourselves be said to have played an active part in making it the case that, if he wrongs us, we will be in the sort of position we are in.

Despite the fact that this very simple defense is not available to us, I think it is clear that the brief defense sketched above can nonetheless be made good. To see this, suppose we take very seriously for a moment the possibility that in the sorts of cases that interest us we might be justified in making the relevant sort of threat, as a way of lowering the probability that we will be wrongfully attacked, but not justified in enforcing it once it is ignored. Obviously, this would leave anyone who was willing to do only what she was morally justified in doing in an extremely vulnerable position, strategically, so far as deterring unjust aggression is concerned. For once any one of her initial threats was ignored, she would be unable to enforce it, with the result that the rest of her threats would not be very likely to be believed.

Once the situation of the upright individual is presented in this way, we can see how an effective defense of the position that interests us might be made out. Recall, to begin with, the situation of the individual whose only way of deterring unjust aggression is to adopt an ARS: here, we said, it seems reasonable to suppose that under certain circumstances one could justifiably adopt an ARS, provided, among other things, the penalties one programmed one's ARDs to impose were somehow proportionate to the harms one was thereby seeking to prevent. The intuitive idea was that if one must choose between a higher (absolute) probability of unjust harm to the innocent and a higher (conditional) probability of harm to those who would unjustly harm the innocent, we are entitled to choose the latter over the former, provided potential wrongdoers are aware of what will happen to them if they do wrong *and* the existence of the relevant autoretaliation scheme can plausibly be said to be necessary to reduce the risk of unjust harm to the innocent.

Turn now to the situation of the individual who is faced not with the options of adopting an ARS or remaining at higher risk of unjust harm if she does not, but with the options of making and then, when necessary, enforcing deterrent threats or not making and then enforcing such threats (i.e., either not making them or making them but not enforcing them). As in the case of the choice of the ARS, it seems to me the individual in this second sort of situation can reason as follows: either I make and then, when necessary, enforce threats of retaliation against acts of wrongful aggression, or I do not make and enforce such threats (i.e., in the latter case, either I don't make them or I make them but don't enforce them when they are ignored). If I choose the former, the probability of unjust aggression against me will be lowered, provided I actually enforce the threats I make, while the probability of harm to those who wrong me will be higher than it would otherwise be. (Assume that in the absence of threats, I will do less harm to those who wrong me than I would otherwise do.) If, on the other hand, I choose the latter, the probability of unjust harm to me will be higher, while the probability that I will harm a given wrongdoer will be lower. (Same assumption.) In such circumstances, I am entitled to choose the former option over the latter. For as in the case of the ARS, it seems reasonable to suppose that if one must choose between a higher probability of harm to the innocent and a higher probability of harm to those who would unjustly harm the innocent, one is entitled to choose the latter over the former, at least if potential wrongdoers are aware of what we will do to them if they do wrong and our doing this to them can plausibly be said to be necessary to reduce the risk of unjust harm to the innocent.

The crucial move here, of course, is in seeing enforcement of otherwise justifiable threats as a critical part of a larger strategy that is designed to lower the absolute probability of unjust harm to the innocent by raising the conditional probability that those who unjustly harm the innocent will themselves be harmed. If we suppose, in light of our brief remarks

about the related phenonmenon of ARSs, that some such strategy is indeed morally justifiable, and if we suppose, as well, that in the case of OTSs it is essential to the effectuation of such a strategy that we actually enforce our threats once they have been made and then ignored, it follows that we will after all be able to make out a case for enforcing them, the objection presented above notwithstanding.

VI

The principle defended in the preceding section can be stated as follows:

> P7: When my situation is such that either (*i*) I enforce a conditional threat of retaliation that I have previously and justifiably made, thereby protecting myself from a decrease in my credibility and hence from an increase in my vulnerability or (*ii*) I do not enforce the relevant threat, thereby jeopardizing my credibility and hence increasing my vulnerability to aggression I might otherwise have deterred, I am entitled to choose (*i*) over (*ii*), provided that the penalties thus threatened and imposed are within certain limits and are directed only at offenders for offenses.[9]

It will perhaps be obvious that once the defense of both P6 and P7 has been secured, we can restate them as a single principle and that that principle will be equivalent to P5 above, which in turn was equivalent to P2. Less obvious, perhaps, but equally important, is the fact that all three versions of the principle that interests us are designed to bring out its affinity to P1, on which, we claimed, our right to direct self-defense can itself be based. Given the time and space, we would, of course, want to proceed at this point to a discussion of whether this successor principle really is as plausible, intuitively, as P1, and then, of course, to a discussion of whether the intuitive plausibility of P1 can be upheld in any more rigorous (nonintuitive) way.

Unfortunately, we have neither time nor space to pursue these matters here. Nor do we have time, or space, to pursue the all-important question of just what the limits are that our discussion above has presupposed: the limits, that is, within which we would allegedly have to stay in programming an ARD to deal with any particular kind of crime and hence

9. The final qualification here is of course crucial: P7 is meant to apply only to justifiable threats to harm those who do an innocent person wrong (this is what makes them threats of *retaliation*). In a fuller treatment of these matters, we would need to discuss the basis for this qualification and also the very interesting question of when, if ever, one is justified in imposing harms on admittedly innocent individuals in an effort to deter other individuals from doing wrongful harm. Here it must suffice to note that while it is conceivable that I might find myself in a situation where by threatening harm to an innocent party I can possibly keep some potential wrongdoer from doing harm to some other innocent party, and while it is also conceivable that under certain circumstances I would in fact be justified not just in making such a threat but also in enforcing it if it is ignored, such a justification is not provided by either P6 or P7.

within which we would have to stay in making and then enforcing the threats that our successor principle says we have a right to make and then enforce. It will have to suffice, for now, to say that these limits are, in my view, exactly the same as the limits we would say we must honor in direct self-defense: for any given kind of offense with which we might be faced, there is just so much that we are entitled to do in order to prevent someone from perpetrating that offense against us, the seriousness of what we are entitled to do being a function of the seriousness of what we are trying to keep the relevant offender from doing to us.