THE JOURNAL OF PHILOSOPHY

VOLUME LXXV, NO. 6, JUNE 1978

SOME PARADOXES OF DETERRENCE *

DETERRENCE is a parent of paradox. Conflict theorists, notably Thomas Schelling, have pointed out several paradoxes of deterrence: that it may be to the advantage of someone who is trying to deter another to be irrational, to have fewer available options, or to lack relevant information.[1] I shall describe certain new paradoxes that emerge when one attempts to analyze deterrence from a moral rather than a strategic perspective. These paradoxes are presented in the form of statements that appear absurd or incredible on first inspection, but can be supported by quite convincing arguments.

Consider a typical situation involving deterrence. A potential wrongdoer is about to commit an offense that would unjustly harm someone. A defender intends, and threatens, to retaliate should the wrongdoer commit the offense. Carrying out retaliation, if the offense is committed, could well be morally wrong. (The wrongdoer could be insane, or the retaliation could be out of proportion with the offense, or could seriously harm others besides the wrongdoer.) The moral paradoxes of deterrence arise out of the attempt to determine the moral status of the defender's *intention* to retaliate in such cases. If the defender knows retaliation to be wrong, it would appear that this intention is evil. Yet such "evil" intentions may pave the road to heaven, by preventing serious offenses and by doing so without actually harming anyone.

* An earlier version of this paper was presented at Stanford University. I am grateful to several, especially Robert Merrihew Adams, Tyler Burge, Warren Quinn, and Virginia Warren, for helpful comments on previous drafts. My work was supported, in part, by a Regents' Faculty Research Fellowship from the University of California.

[1] *The Strategy of Conflict* (New York: Oxford, 1960), Chaps. 1–2; and *Arms and Influence* (New Haven, Conn.: Yale, 1966), chap. 2.

0022-362X/78/7506/0285$01.80

Scrutiny of such morally ambiguous retaliatory intentions reveals paradoxes that call into question certain significant and widely accepted moral doctrines. These doctrines are what I call *bridge principles*. They attempt to link together the moral evaluation of actions and the moral evaluation of agents (and their states) in certain simple and apparently natural ways. The general acceptance, and intuitive appeal, of such principles, lends credibility to the project of constructing a consistent moral system that accurately reflects our firmest moral beliefs about both agents and actions. By raising doubts about the validity of certain popular bridge principles, the paradoxes presented here pose new difficulties for this important project.

I

In this section, a certain class of situations involving deterrence is characterized, and a plausible normative assumption is presented. In the following three sections, we shall see how application of this assumption to these situations yields paradoxes.

The class of paradox-producing situations is best introduced by means of an example. Consider the balance of nuclear terror as viewed from the perspective of one of its superpower participants, nation *N*. *N* sees the threat of nuclear retaliation as its only reliable means of preventing nuclear attack (or nuclear blackmail leading to world domination) by its superpower rival. *N* is confident such a threat will succeed in deterring its adversary, provided it really intends to carry out that threat. (*N* fears that, if it bluffs, its adversary is likely to learn this through leaks or espionage.) Finally, *N* recognizes it would have conclusive moral reasons *not* to carry out the threatened retaliation, if its opponent were to obliterate *N* with a surprise attack. For although retaliation would punish the leaders who committed this unprecedented crime and would prevent them from dominating the postwar world, *N* knows it would also destroy many millions of innocent civilians in the attacking nation (and in other nations), would set back postwar economic recovery for the world immeasurably, and might add enough fallout to the atmosphere to destroy the human race.

Let us call situations of the sort that nation *N* perceives itself as being in, *Special Deterrent Situations* (SDSs). More precisely, an agent is in an SDS when he reasonably and correctly believes that the following conditions hold. First, it is likely he must intend (conditionally) to apply a harmful sanction to innocent people, if an extremely harmful and unjust offense is to be prevented. Second, such an intention would very likely deter the offense. Third, the

amounts of harm involved in the offense and the threatened sanction are very large and of roughly similar quantity (or the latter amount is smaller than the former). Finally, he would have conclusive moral reasons not to apply the sanction if the offense were to occur.

The first condition in this definition requires some comment. Deterrence depends only on the potential wrongdoer's *beliefs* about the prospects of the sanction being applied. Hence, the first condition will be satisfied only if attempts by the defender to bluff would likely be perceived as such by the wrongdoer. This may be the case if the defender is an unconvincing liar, or is a group with a collective decision procedure, or if the wrongdoer is shrewd and knows the defender quite well. Generally, however, bluffing will be a promising course of action. Hence, although it is surely logically and physically possible for an SDS to occur, there will be few actual SDSs. It may be noted, though, that writers on strategic policy frequently assert that nuclear deterrence will be effective only if the defending nation really intends to retaliate.[2] If this is so, the balance of terror may fit the definition of an SDS, and the paradoxes developed here could have significant practical implications.[3] Further, were there no actual SDSs, these paradoxes would still be of considerable theoretical interest. For they indicate that the validity of some widely accepted moral doctrines rests on the presupposition that certain situations that could arise (i.e., SDSs) will not.

Turning to our normative assumption, we begin by noting that any reasonable system of ethics must have substantial utilitarian elements. The assumption that produces the paradoxes of deterrence concerns the role of utilitarian considerations in determining one's moral duty in a narrowly limited class of situations. Let the *most useful* act in a given choice situation be that with the highest expected utility. Our assumption says that the most useful act should be performed whenever a very great deal of utility is at stake. This means that, if the difference in expected utility between the most useful act and its alternatives is extremely large (e.g., equivalent to the difference between life and death for a very large number of people), other moral considerations are overridden by utilitarian considerations.

This assumption may be substantially weakened by restricting in

[2] See, e.g., Herman Kahn, *On Thermonuclear War*, 2nd ed. (Princeton, N.J.: University Press, 1960), p. 185; and Anthony Kenny, "Counterforce and Countervalue," in Walter Stein, ed., *Nuclear Weapons: A Catholic Response* (London: Merlin Press, 1965), pp. 162–164.

[3] See, e.g., n. 9, below.

various ways its range of application. I restrict the assumption to apply only when (i) a great deal of *negative* utility is at stake, and (ii) people will likely suffer serious injustices if the agent fails to perform the most useful act. This makes the assumption more plausible, since the propriety of doing one person a serious injustice, in order to produce positive benefits for others, is highly questionable. The justifiability of doing the same injustice to prevent a utilitarian disaster which itself involves grave injustices, seems more in accordance with our moral intuitions.

The above restrictions appear to bring our assumption into line with the views of philosophers such as Robert Nozick, Thomas Nagel, and Richard Brandt, who portray moral rules as "absolutely" forbidding certain kinds of acts, but acknowledge that exceptions might have to be allowed in cases in which such acts are necessary to prevent catastrophe.[4] Even with these restrictions, however, the proposed assumption would be rejected by supporters of genuine Absolutism, the doctrine that there are certain acts (such as vicarious punishment and deliberate killing of the innocent) that are always wrong, whatever the consequences of not performing them. (Call such acts *inherently evil.*) We can, though, accommodate the Absolutists. To do so, let us further qualify our assumption by limiting its application to cases in which (iii) performing the most useful act involves, at most, a small *risk* of performing an inherently evil act. With this restriction, the assumption still leads to paradoxes, yet is consistent with Absolutism (unless that doctrine is extended to include absolute prohibitions on something other than doing acts of the sort usually regarded as inherently evil).[5] The triply qualified assumption is quite plausible; so the fact that it produces paradoxes is both interesting and disturbing.

II

The first moral paradox of deterrence is:

(P1) There are cases in which, although it would be wrong for an agent to perform a certain act in a certain situation, it would nonetheless be right for him, knowing this, to form the intention to perform that act in that situation.

[4] Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), pp. 30/1 n; Nagel, "War and Massacre," *Philosophy and Public Affairs*, I, 2 (Winter 1972): 123–144, p. 126; Brandt, "Utilitarianism and the Rules of War," *ibid.*, 145–165, p. 147, especially n. 3.

[5] Extensions of Absolutism that would block some or all of the paradoxes include those which forbid intending to do what is wrong, deliberately making oneself less virtuous, or intentionally risking performing an inherently evil act. (An explanation of the relevant sense of 'risking performing an act' will be offered in section IV.)

At first, this strikes one as absurd. If it is wrong and he is aware that it is wrong, how could it be right for him to form the intention to do it? (P1) is the direct denial of a simple moral thesis, the Wrongful Intentions Principle (WIP): *To intend to do what one knows to be wrong is itself wrong.*[6] WIP seems so obvious that, although philosophers never call it into question, they rarely bother to assert it or argue for it. Nevertheless, it appears that Abelard, Aquinas, Butler, Bentham, Kant, and Sidgwick, as well as recent writers such as Anthony Kenny and Jan Narveson, have accepted the principle, at least implicitly.[7]

Why does WIP seem so obviously true? First, we regard the man who fully intends to perform a wrongful act and is prevented from doing so solely by external circumstances (e.g., a man whose murder plan is interrupted by the victim's fatal heart attack) as being just as bad as the man who performs a like wrongful act. Second, we view the man who intends to do what is wrong, and then changes his mind, as having corrected a moral failing or error. Third, it is convenient, for many purposes, to treat a prior intention to perform an act, as the beginning of the act itself. Hence, we are inclined to view intentions as parts of actions and to ascribe to each intention the moral status ascribed to the act "containing" it.

It is essential to note that WIP appears to apply to conditional intentions in the same manner as it applies to nonconditional ones. Suppose I form the intention to kill my neighbor if he insults me again, and fail to kill him only because, fortuitously, he refrains from doing so. I am as bad, or nearly as bad, as if he had insulted me and I had killed him. My failure to perform the act no more erases the wrongness of my intention, than my neighbor's dropping dead as I load my gun would negate the wrongness of the simple intention to kill him. Thus the same considerations adduced above in support of WIP seem to support the formulation: If it would be wrong to perform an act in certain circumstances, then it is wrong

[6] I assume henceforth that, if it would be wrong to do something, the agent knows this. (The agent, discussed in section IV, who has become corrupt may be an exception.) This keeps the discussion of the paradoxes from getting tangled up with the separate problem of whether an agent's duty is to do what is actually right, or what he believes is right.

[7] See *Peter Abelard's Ethics*, D. E. Luscombe, trans. (New York: Oxford, 1971), pp. 5–37; Thomas Aquinas, *Summa Theologica*, 1a2ae. 18–20; Joseph Butler, "A Dissertation on the Nature of Virtue," in *Five Sermons* (Indianapolis: Bobbs-Merrill, 1950), p. 83; Immanuel Kant, *Foundations of the Metaphysics of Morals*, first section; Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation*, chap. 9, secs. 13–16; Henry Sidgwick, *The Methods of Ethics* (New York: Dover, 1907), pp. 60/1, 201–204; Kenny, pp. 159, 162; and Jan Narveson, *Morality and Utility* (Baltimore: Johns Hopkins, 1967), pp. 106–108.

to intend to perform that act on the condition that those circumstances arise.

Having noted the source of the strong feeling that (P1) should be rejected, we must consider an instantiation of (P1):

(P1′) In an SDS, it would be wrong for the defender to apply the sanction if the wrongdoer were to commit the offense, but it is right for the defender to form the (conditional) intention to apply the sanction if the wrongdoer commits the offense.

The first half of (P1′), the wrongness of applying the sanction, follows directly from the last part of the definition of an SDS, which says that the defender would have conclusive moral reasons not to apply the sanction. The latter half of (P1′), which asserts the rightness of forming the intention to apply the sanction, follows from the definition of an SDS and our normative assumption. According to the definition, the defender's forming this intention is likely necessary, and very likely sufficient, to prevent a seriously harmful and unjust offense. Further, the offense and the sanction would each produce very large and roughly commensurate amounts of negative utility (or the latter would produce a smaller amount). It follows that utilitarian considerations heavily favor forming the intention to apply the sanction, and that doing so involves only a small risk of performing an inherently evil act.[8] Applying our normative assumption yields the conclusion that it is right for the defender to form the intention in question.

This argument, if sound, would establish the truth of (P1′), and hence (P1), in contradiction with WIP. It suggests that WIP should not be applied to *deterrent intentions*, i.e., those conditional intentions whose existence is based on the agent's desire to thereby deter others from actualizing the antecedent condition of the intention. Such intentions are rather strange. They are, by nature, self-stultifying: if a deterrent intention fulfills the agent's purpose, it ensures that the intended (and possibly evil) act is not performed, by preventing the circumstances of performance from arising. The unique nature of such intentions can be further explicated by noting the distinction between intending to do something, and desiring (or

[8] A qualification is necessary. Although having the intention involves only a small risk of applying the threatened sanction to innocent people, it follows, from points made in section IV, that forming the intention might also involve risks of performing *other* inherently evil acts. Hence, what really follows is that forming the intention is right in those SDSs in which the composite risk is small. This limitation in the scope of (P1′) is to be henceforth understood. It does not affect (P1), (P2), or (P3), since each is governed by an existential quantifier.

intending) to intend to do it. Normally, an agent will form the intention to do something because he either desires doing that thing as an end in itself, or as a means to other ends. In such cases, little importance attaches to the distinction between intending and desiring to intend. But, in the case of deterrent intentions, the ground of the desire to form the intention is entirely distinct from any desire to carry it out. Thus, what may be inferred about the agent who seeks to form such an intention is this. He desires *having the intention* as a means of deterrence. Also, he is willing, in order to prevent the offense, to accept a certain *risk* that, in the end, he will apply the sanction. But this is entirely consistent with his having a strong desire not to apply the sanction, and no desire at all to apply it. Thus, while the object of his deterrent intention might be an evil act, it does not follow that, in desiring to adopt that intention, he desires to do evil, either as an end or as a means.

WIP ties the morality of an intention exclusively to the moral qualities of its object (i.e., the intended act). This is not unreasonable since, typically, the only significant effects of intentions are the acts of the agent (and the consequences of these acts) which flow from these intentions. However, in certain cases, intentions may have *autonomous effects* that are independent of the intended act's actually being performed. In particular, intentions to act may influence the conduct of other agents. When an intention has important autonomous effects, these effects must be incorporated into any adequate moral analysis of it. The first paradox arises because the autonomous effects of the relevant deterrent intention are dominant in the moral analysis of an SDS, but the extremely plausible WIP ignores such effects.[9]

III

(P1′) implies that a rational moral agent in an SDS should want to form the conditional intention to apply the sanction if the offense is committed, in order to deter the offense. But will he be able to do so? Paradoxically, he will not be. He is a captive in the prison of his own virtue, able to form the requisite intention only by bending the bars of his cell out of shape. Consider the preliminary for-

[9] In *Nuclear Weapons*, Kenny and others use WIP to argue that nuclear deterrence is immoral because it involves having the conditional intention to kill innocent people. The considerations advanced in this section suggest that this argument, at best, is inconclusive, since it presents only one side of a moral paradox, and, at worst, is mistaken, since it applies WIP in just the sort of situation in which its applicability is most questionable.

mulation of this new paradox:

(P2′) In an SDS, a rational and morally good agent cannot (as a matter of logic) have (or form) the intention to apply the sanction if the offense is committed.[10]

The argument for (P2′) is as follows. An agent in an SDS recognizes that there would be conclusive moral reasons not to apply the sanction if the offense were committed. If he does not regard these admittedly conclusive moral reasons as conclusive reasons for him not to apply the sanction, then he is not moral. Suppose, on the other hand, that he does regard himself as having conclusive reasons not to apply the sanction if the offense is committed. If, nonetheless, he is disposed to apply it, because the reasons for applying it motivate him more strongly than do the conclusive reasons not to apply it, then he is irrational.

But couldn't our rational moral agent recognize, in accordance with (P1′), that he ought to form the intention to apply the sanction? And couldn't he then simply grit his teeth and pledge to himself that he will apply the sanction if the offense is committed? No doubt he could, and this would amount to trying to form the intention to apply the sanction. But the question remains whether he can succeed in forming that intention, by this or any other process, while remaining rational and moral. And it appears he cannot. There are, first of all, psychological difficulties. Being rational, how can he dispose himself to do something that he knows he would have conclusive reasons not to do, when and if the time comes to do it? Perhaps, though, some exceptional people can produce in themselves dispositions to act merely by pledging to act. But even if one could, in an SDS, produce a disposition to apply the sanction in this manner, such a disposition would not count as a *rational intention* to apply the sanction. This is because, as recent writers on intentions have suggested, it is part of the concept of rationally intending to do something, that the disposition to do the intended act be caused (or justified) in an appropriate way by the agent's view of reasons for doing the act.[11] And the disposition in question does not stand in such a relation to the agent's reasons for action.

[10] 'Rational and morally good' in this and later statements of the second and third paradoxes, means rational and moral in the given situation. A person who usually is rational and moral, but fails to be in the situation in question, could, of course, have the intention to apply the sanction. (P2′) is quite similar to a paradox concerning utilitarianism and deterrence developed by D. H. Hodgson in *Consequences of Utilitarianism* (Oxford: Clarendon Press, 1967), chap. 4.

[11] See, e.g., S. Hampshire and H. L. A. Hart, "Decision, Intention and Certainty," *Mind*, LXVII.1, 265 (January 1958): 1–12; and G. E. M. Anscombe, *Intention* (Ithaca, N.Y.: Cornell, 1966).

It might be objected to this that people sometimes intend to do things (and do them) for no reason at all, without being irrational. This is true, and indicates that the connections between the concepts of intending and reasons for action are not so simple as the above formula implies. But it is also true that intending to do something for no reason at all, in the face of recognized significant reasons not to do it, would be irrational. Similarly, a disposition to act in the face of the acknowledged preponderance of reasons, whether called an "intention" or not, could not qualify as rational. It may be claimed that such a disposition, in an SDS, is rational in the sense that the agent knows it would further his aims to form (and have) it. This is not to deny the second paradox, but simply to express one of its paradoxical features. For the point of (P2′) is that the very disposition that *is* rational in the sense just mentioned, is at the same time irrational in an equally important sense. It is a disposition to act in conflict with the agent's own view of the balance of reasons for action.

We can achieve some insight into this by noting that an intention that is deliberately formed, resides at the intersection of two distinguishable actions. It is the beginning of the act that is its object and is the end of the act that is its formation. As such, it may be assessed as rational (or moral) or not, according to whether either of two different acts promotes the agent's (or morality's) ends. Generally, the assessments will agree. But, as Schelling and others have noted, it may sometimes promote one's aims *not* to be disposed to act to promote one's aims should certain contingencies arise. For example, a small country may deter invasion by a larger country if it is disposed to resist any invasion, even when resistance would be suicidal. In such situations, the assessment of the rationality (or morality) of the agent's intentions will depend upon whether these intentions are treated as components of their object-acts or their formation-acts. If treated as both, conflicts can occur. It is usual and proper to assess the practical rationality of an agent, at a given time, according to the degree of correspondence between his intentions and the reasons he has for performing the acts that are the objects of those intentions. As a result, puzzles such as (P2′) emerge when, for purposes of moral analysis, an agent's intentions are viewed partly as components of their formation-acts.

Let us return to the main path of our discussion by briefly summarizing the argument for (P2′). A morally good agent regards conclusive moral reasons for action as conclusive reasons for action *simpliciter*. But the intentions of a rational agent are not out of

line with his assessment of the reasons for and against acting. Consequently, a rational moral agent cannot intend to do something that he recognizes there are conclusive moral reasons not to do. Nor can he intend conditionally to do what he recognizes he would have conclusive reasons not to do were that condition to be fulfilled. Therefore, in an SDS, where one has conclusive moral reasons not to apply the sanction, an originally rational and moral agent cannot have the intention to apply it without ceasing to be fully rational or moral; nor can he form the intention (as this entails having it).

We have observed that forming an intention is a process that may generally be regarded as an action. Thus, the second paradox can be reformulated as:

(P2) There are situations (namely SDSs) in which it would be right for agents to perform certain actions (namely forming the intention to apply the sanction) and in which it is possible for some agents to perform such actions, but impossible for rational and morally good agents to perform them.

(P2), with the exception of the middle clause, is derived from the conjunction of (P1′) and (P2′) by existential generalization. The truth of the middle clause follows from consideration of the vengeful agent, who desires to punish those who commit seriously harmful and unjust offenses, no matter what the cost to others.

(P2) is paradoxical because it says that there are situations in which rationality and virtue preclude the possibility of right action. And this contravenes our usual assumption about the close logical ties between the concepts of right action and agent goodness. Consider the following claim. *Doing something is right if and only if a morally good man would do the same thing in the given situation.* Call this the Right-Good Principle. One suspects that, aside from qualifications concerning the good man's possible imperfections or factual ignorance, most people regard this principle, which directly contradicts (P2), as being virtually analytic. Yet the plight of the good man described in the second paradox does not arise out of an insufficiency of either knowledge or goodness. (P2) says there are conceivable situations in which virtue and knowledge combine with rationality to preclude right action, in which virtue is an obstacle to doing the right thing. If (P2) is true, our views about the close logical connection between right action and agent goodness, as embodied in the Right-Good Principle, require modifications of a sort not previously envisioned.

IV

A rational moral agent in an SDS faces a cruel dilemma. His reasons for intending to apply the sanction if the offense is committed are, according to (P1'), conclusive. But they outrun his reasons for doing it. Wishing to do what is right, he wants to form the intention. However, unless he can substantially alter the basic facts of the situation or his beliefs about those facts, he can do so only by making himself less morally good; that is, by becoming a person who attaches grossly mistaken weights to certain reasons for and against action (e.g., one who prefers retribution to the protection of the vital interests of innocent people).[12] We have arrived at a third paradox:

(P3) In certain situations, it would be morally right for a rational and morally good agent to deliberately (attempt to) corrupt himself.[13]

(P3) may be viewed in light of a point about the credibility of threats which has been made by conflict theorists. Suppose a defender is worried about the credibility of his deterrent threat, because he thinks the wrongdoer (rightly) regards him as unwilling to apply the threatened sanction. He may make the threat more credible by passing control of the sanction to some *retaliation-agent.* Conflict theorists consider two sorts of retaliation-agents: people known to be highly motivated to punish the offense in question, and machines programmed to retaliate automatically if the offense occurs. What I wish to note is that future selves of the defender himself are a third class of retaliation-agents. If the other kinds are unavailable, a defender may have to create an agent of this third sort (i.e., an altered self willing to apply the sanction), in order to deter the offense. In cases in which applying the sanction would be wrong, this could require self-corruption.

How would a rational and moral agent in an SDS, who seeks to have the intention to apply the sanction, go about corrupting himself so that he may have it? He cannot form the intention simply

[12] Alternatively, the agent could undertake to make himself into an *irrational* person whose intentions are quite out of line with his reasons for action. However, trying to become irrational, in these circumstances, is less likely to succeed than trying to change one's moral beliefs, and, furthermore, might itself constitute self-corruption. Hence, this point does not affect the paradox stated below.

[13] As Donald Regan has suggested to me, (P3) can be derived directly from our normative assumption: imagine a villain credibly threatening to kill very many hostages unless a certain good man corrupts himself. I prefer the indirect route to (P3) given in the text, because (P1) and (P2) are interesting in their own right and because viewing the three paradoxes together makes it easier to see what produces them.

by pledging to apply the sanction; for, according to the second paradox, his rationality and morality preclude this. Instead, he must seek to initiate a causal process (e.g., a reeducation program) that he hopes will result in his beliefs, attitudes, and values changing in such a way that he can and will have the intention to apply the sanction should the offense be committed. Initiating such a process involves taking a rather odd, though not uncommon attitude toward oneself: viewing oneself as an object to be molded in certain respects by outside influences rather than by inner choices. This is, for example, the attitude of the lazy but ambitious student who enrolls in a fine college, hoping that some of the habits and values of his highly motivated fellow students will "rub off" on him.

We can now better understand the notion of "risking doing *X*" which was introduced in section I. For convenience, let "*X*" be "killing." Deliberately risking killing is different from risking deliberately killing. One does the former when one rushes an ill person to the hospital in one's car at unsafe speed, having noted the danger of causing a fatal accident. One has deliberately accepted the risk of killing by accident. One (knowingly) risks deliberately killing, on the other hand, when one undertakes a course of action that one knows may, by various causal processes, lead to one's later performing a deliberate killing. The mild-mannered youth who joins a violent street gang is an example. Similarly, the agent in an SDS, who undertakes a plan of self-corruption in order to develop the requisite deterrent intention, knowingly risks deliberately performing the wrongful act of applying the sanction.

The above description of what is required of the rational moral agent in an SDS, leads to a natural objection to the argument that supports (P3). According to this objection, an attempt at self-corruption by a rational moral agent is very likely to fail. Hence, bluffing would surely be a more promising strategy for deterrence than trying to form retaliatory intentions by self-corruption. Three replies may be given to this objection. First, it is certainly *conceivable* that, in a particular SDS, undertaking a process of self-corruption would be more likely to result in effective deterrence than would bluffing. Second, and more important, bluffing and attempting to form retaliatory intentions by self-corruption will generally not be mutually exclusive alternatives. An agent in an SDS may attempt to form the retaliatory intention while bluffing, and plan to continue bluffing as a "fall-back" strategy, should he fail. If the offense to be prevented is disastrous enough, the additional expected utility generated by following such a combined strategy (as opposed to simply

bluffing) will be very large, even if his attempts to form the intention are unlikely to succeed. Hence, (P3) would still follow from our normative assumption. Finally, consider the rational and *partly corrupt* agent in an SDS who already has the intention to retaliate. (The nations participating in the balance of terror may be examples.) The relevant question about him is whether he ought to act to become less corrupt, with the result that he would lose the intention to retaliate. The present objection does not apply in this case, since the agent already has the requisite corrupt features. Yet, essentially the same argument that produces (P3) leads, when this case is considered, to a slightly different, but equally puzzling, version of our third paradox:

(P3*) In certain situations, it would be morally wrong for a rational and partly corrupt agent to (attempt to) reform himself and eliminate his corruption.

A rather different objection to (P3) is the claim that its central notion is incoherent. This claim is made, apparently, by Thomas Nagel, who writes:

> The notion that one might sacrifice one's moral integrity justifiably, in the service of a sufficiently worthy end, is an incoherent notion. For if one were justified in making such a sacrifice (or even morally required to make it), then one would not be sacrificing one's moral integrity by adopting that course: one would be preserving it (132/3).

Now the notion of a justified sacrifice of moral virtue (integrity) would be incoherent, as Nagel suggests, if one could sacrifice one's virtue only by doing something wrong. For the same act cannot be both morally justified and morally wrong. But one may also be said to sacrifice one's virtue when one deliberately initiates a causal process that one expects to result, and does result, in one's later becoming a less virtuous person. And, as the analysis of SDSs embodied in (P1′) and (P2′) implies, one may, in certain cases, be justified in initiating such a process (or even be obligated to initiate it). Hence, it would be a mistake to deny (P3) on the grounds advanced in Nagel's argument.

There is, though, a good reason for *wanting* to reject (P3). It conflicts with some of our firmest beliefs about virtue and duty. We regard the promotion and preservation of one's own virtue as a vital responsibility of each moral agent, and self-corruption as among the vilest of enterprises. Further, we do not view the duty to promote one's virtue as simply one duty among others, to be weighed and balanced against the rest, but rather as a special duty

that encompasses the other moral duties. Thus, we assent to the Virtue Preservation Principle: *It is wrong to deliberately lose (or reduce the degree of) one's moral virtue.* To many, this principle seems fundamental to our very conception of morality.[14] Hence the suggestion that duty could require the abandonment of virtue seems quite unacceptable. The fact that this suggestion can be supported by strong arguments produces a paradox.

This paradox is reflected in the ambivalent attitudes that emerge when we attempt to evaluate three hypothetical agents who respond to the demands of SDSs in various ways. The first agent refuses to try to corrupt himself and allows the disastrous offense to occur. We respect the love of virtue he displays, but are inclined to suspect him of too great a devotion to his own purity relative to his concern for the well-being of others. The second agent does corrupt himself to prevent disaster in an SDS. Though we do not approve of his new corrupt aspects, we admire the person that he *was* for his willingness to sacrifice what he loved—part of his own virtue—in the service of others. At the same time, the fact that he succeeded in corrupting himself may make us wonder whether he was entirely virtuous in the first place. Corruption, we feel, does not come easily to a good man. The third agent reluctantly but sincerely tries his best to corrupt himself to prevent disaster, but fails. He may be admired both for his willingness to make such a sacrifice and for having virtue so deeply engrained in his character that his attempts at self-corruption do not succeed. It is perhaps characteristic of the paradoxical nature of the envisioned situation, that we are inclined to admire most the only one of these three agents who fails in the course of action he undertakes.

V

It is natural to think of the evaluation of agents, and of actions, as being two sides of the same moral coin. The moral paradoxes of deterrence suggest they are more like two separate coins that can be

[14] Its supporters might, of course, allow exceptions to the principle in cases in which only the agent's feelings, and not his acts or dispositions to act, are corrupted. (For example, a doctor "corrupts himself" by suppressing normal sympathy for patients in unavoidable pain, in order to treat them more effectively.) Further, advocates of the doctrine of double-effect might consider self-corruption permissible when it is a "side effect" of action rather than a means to an end. For example, they might approve of a social worker's joining a gang to reform it, even though he expects to assimilate some of the gang's distorted values. Note, however, that neither of these possible exceptions to the Virtue Preservation Principle (brought to my attention by Robert Adams) applies to the agent in an SDS who corrupts his *intentions* as a chosen *means* of preventing an offense.

fused together only by significantly deforming one or the other. In this concluding section, I shall briefly explain this.

Our shared assortment of moral beliefs may be viewed as consisting of three relatively distinct groups: beliefs about the evaluation of actions, beliefs about the evaluation of agents and their states (e.g., motives, intentions, and character traits), and beliefs about the relationship between the two. An important part of this last group of beliefs is represented by the three bridge principles introduced above: the Wrongful Intentions, Right-Good, and Virtue Preservation principles. Given an agreed-upon set of bridge principles, one could go about constructing a moral system meant to express coherently our moral beliefs in either of two ways: by developing principles that express our beliefs about act evaluation and then using the bridge principles to derive principles of agent evaluation—or vice versa. If our bridge principles are sound and our beliefs about agent and act evaluation are mutually consistent, the resulting systems would, in theory, be the same. If, however, there are underlying incompatibilities between the principles we use to evaluate acts and agents, there may be significant differences between moral systems that are *act-oriented* and those which are *agent-oriented.* And these differences may manifest themselves as paradoxes which exert pressure upon the bridge principles that attempt to link the divergent systems, and the divergent aspects of each system, together.

It seems natural to us to evaluate acts at least partly in terms of their consequences. Hence, act-oriented moral systems tend to involve significant utilitarian elements. The principle of act evaluation usually employed in utilitarian systems is: in a given situation, one ought to perform the most useful act, that which will (or is expected to) produce the most utility. What will maximize utility depends upon the facts of the particular situation. Hence, as various philosophers have pointed out, the above principle could conceivably recommend one's (i) acting from nonutilitarian motives, (ii) advocating some nonutilitarian moral theory, or even (iii) becoming a genuine adherent of some nonutilitarian theory.[15] Related quandaries arise when one considers, from an act-utilitarian viewpoint, the deterrent intention of a defender in an SDS. Here is an intention whose object-act is anti-utilitarian and whose formation-act is a utilitarian duty that cannot be performed by a rational utilitarian.

[15] See Hodgson, *Consequences.* Also, Adams, "Motive Utilitarianism," this JOURNAL, LXXIII, 14 (Aug. 12, 1976): 467–81; and Bernard Williams, "A Critique of Utilitarianism," in J. J. C. Smart and Williams, *Utilitarianism: For and Against* (New York: Cambridge, 1973), sec. 6.

A utilitarian might seek relief from these quandaries in either of two ways. First, he could defend some form of rule-utilitarianism. But then he would face a problem. Shall he include, among the rules of his system, our normative assumption that requires the performance of the most useful act, whenever an enormous amount of utility is at stake (and certain other conditions are satisfied)? If he does, the moral paradoxes of deterrence will appear within his system. If he does not, it would seem that his system fails to attach the importance to the consequences of particular momentous acts that any reasonable moral, much less utilitarian, system should. An alternative reaction would be to stick by the utilitarian principle of act evaluation, and simply accept (P1)–(P3), and related oddities, as true. Taking this line would require the abandonment of the plausible and familiar bridge principles that contradict (P1)–(P3). But this need not bother the act-utilitarian, who perceives his task as the modification, as well as codification, of our moral beliefs.

Agent-oriented (as opposed to act-oriented) moral systems rest on the premise that what primarily matters for morality are the internal states of a person: his character traits, his intentions, and the condition of his will. The doctrines about intentions and virtue expressed in our three bridge principles are generally incorporated into such systems. The paradoxes of deterrence may pose serious problems for some agent-oriented systems. It may be, for example, that an adequate analysis of the moral virtues of justice, selflessness, and benevolence, would imply that the truly virtuous man would feel obligated to make whatever personal sacrifice is necessary to prevent a catastrophe. If so, the moral paradoxes of deterrence would arise within agent-oriented systems committed to these virtues.

There are, however, agent-oriented systems that would not be affected by our paradoxes. One such system could be called Extreme Kantianism. According to this view, the only things having moral significance are such features of a person as his character and the state of his will. The Extreme Kantian accepts Kant's dictum that morality requires treating oneself and others as ends rather than means. He interprets this to imply strict duties to preserve one's virtue and not to deliberately impose serious harms or risks on innocent people. Thus, the Extreme Kantian would simply reject (P1)–(P3) without qualm.

Although act-utilitarians and Extreme Kantians can view the paradoxes of deterrence without concern, one doubts that the rest of us can. The adherents of these extreme conceptions of morality are untroubled by the paradoxes because their viewpoints are too one-

sided to represent our moral beliefs accurately. Each of them is closely attentive to certain standard principles of agent *or* act evaluation, but seems too little concerned with traditional principles of the other sort. For a system of morality to reflect our firmest and deepest convictions adequately, it must represent a middle ground between these extremes by seeking to accommodate the valid insights of both act-oriented and agent-oriented perspectives. The normative assumption set out in section I was chosen as a representative principle that might be incorporated into such a system. It treats utilitarian considerations as relevant and potentially decisive, while allowing for the importance of other factors. Though consistent with the absolute prohibition of certain sorts of acts, it treats the distinction between harms and risks as significant and rules out absolute prohibitions on the latter as unreasonable. It is an extremely plausible middle-ground principle; but, disturbingly, it leads to paradoxes.

That these paradoxes reflect conflicts between commonly accepted principles of agent and act evaluation, is further indicated by the following observation. Consider what initially appears a natural way of viewing the evaluation of acts and agents as coordinated parts of a single moral system. According to this view, reasons for action determine the moral status of acts, agents, and intentions. A right act is an act that accords with the preponderance of moral reasons for action. To have the right intention is to be disposed to perform the act supported by the preponderance of such reasons, because of those reasons. The virtuous agent is the rational agent who has the proper substantive values, i.e., the person whose intentions and actions accord with the preponderance of moral reasons for action. Given these considerations, it appears that it should always be possible for an agent to go along intending, and acting, in accordance with the preponderance of moral reasons; thus ensuring both his own virtue and the rightness of his intentions and actions. Unfortunately, this conception of harmonious coordination between virtue, right intention, and right action, is shown to be untenable by the paradoxes of deterrence. For they demonstrate that, in any system that takes consequences plausibly into account, situations can arise in which the rational use of moral principles leads to certain paradoxical recommendations: that the principles used, and part of the agent's virtue, be abandoned, and that wrongful intentions be formed.

One could seek to avoid these paradoxes by moving in the direction of Extreme Kantianism and rejecting our normative assump-

tion. But to do so would be to overlook the plausible core of act-utilitarianism. This is the claim that, in the moral evaluation of acts, how those acts affect human happiness often is important—the more so as more happiness is at stake—and sometimes is decisive. Conversely, one could move toward accommodation with act-utilitarianism. This would involve qualifying, so that they do not apply in SDSs, the traditional moral doctrines that contradict (P1)–(P3). And, in fact, viewed in isolation, the considerations adduced in section II indicate that the Wrongful Intentions Principle ought to be so qualified. However, the claims of (P2) and (P3): that virtue may preclude right action and that morality may require self-corruption, are not so easily accepted. These notions remain unpalatable even when one considers the arguments that support them.

Thus, tinkering with our normative assumption or with traditional moral doctrines would indeed enable us to avoid the paradoxes, at least in their present form. But this would require rejecting certain significant and deeply entrenched beliefs concerning the evaluation either of agents or of actions. Hence, such tinkering would not go far toward solving the fundamental problem of which the paradoxes are symptoms: the apparent incompatability of the moral principles we use to evaluate acts and agents. Perhaps this problem can be solved. Perhaps the coins of agent and act evaluation can be successfully fused. But it is not apparent how this is to be done. And I, for one, do not presently see an entirely satisfactory way out of the perplexities that the paradoxes engender.

GREGORY S. KAVKA

University of California at Los Angeles

A CONCEPTUAL PROBLEM FOR LIBERAL DEMOCRACY *

LIBERAL democratic theory can be viewed as an attempt to articulate, at the same time, both a liberal and a democratic value judgment on political institutions. Roughly, the liberal judgment holds that there are certain aspects of a person's life, including certain of his actions, which are private [1] and against

* I am grateful to Gerald MacCallum for discussion of the questions raised in this note.

[1] This private sphere need not coincide with what is called *private* in ordinary speech and even in the law. In general, the private sphere will include certain public acts, of the kind the U.S. Bill of Rights seeks to protect.

0022-362X/78/7506/0302$00.60