Gregory Kavka, Moral Paradoxes of Nuclear Deterrence, Chapter 1 & 2

1. 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.

Scrutiny of such morally ambiguous retaliatory intentions

An earlier version of this chapter was presented at Stanford University. I am grateful to several persons, especially Robert Merrihew Adams, Tyler Burge, Daniel Farrell, Robert Ladenson, 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.

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 contructing 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. SPECIAL DETERRENT SITUATIONS

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 will see how application of this assumption to these situations yields paradoxical conclusions that conflict with widely accepted bridge principles.

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.[2] (*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 (and sun-blocking ashes and dust) 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 sanctions are very large, and the relevant probabilities and amounts of harm are such that a rational utilitarian evaluation would substantially favor having the intention.[3] 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.[4] 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.[5] 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 us say that *a great deal of utility is at stake* in a given situation if either (1) reliable expected utilities are calculable and the difference in expected

utility between the best act and its alternatives is extremely large, or (2) reliable expected utilities are not calculable and there are extremely large differences in utility between some possible outcomes of different available acts. Our assumption says that the act favored by utilitarian considerations should be performed whenever a great deal of utility is at stake. This means that, if the difference in expected, or possible, utilities of the available acts 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 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 that 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, Richard Brandt, and Michael Walzer, 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.[6] Even with these restrictions, however, the proposed assumption would be rejected by supporters of genuine moral 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 some 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 interpreted to include absolute prohibitions on something other than doing acts of the sort usually regarded as inherently evil[7]). The triply qualified assumption is quite

plausible; so the fact that it produces paradoxes is both interesting and disturbing.

II. PARADOXICAL INTENTIONS

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 that agent, knowing this, to form the intention to perform that act in that situation.

At first, this strikes one as absurd. If it is wrong and the agent is aware that it is wrong, how could it be right for her to form the intention to do it? (P1) is the direct denial of a simple moral thesis, the Wrongful Intentions Principle (WIP): *To form the intention to do what one knows to be wrong is itself wrong.*[8] 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.[9]

Why does WIP seem so obviously true? First, we regard the person who fully intends to perform a wrongful act and is prevented from doing so solely by external circumstances (e.g., a person whose murder plan is interrupted by the victim's fatal heart attack) as being just as bad as the person who performs a like wrongful act. Second, we view the person who intends to do what is wrong, and then has a change of 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 to form the intention 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. It follows that doing so involves only a small risk of performing an inherently evil act.[10] Further, in an SDS, a great deal of utility is at stake, and utilitarian considerations substantially favor forming the intention to apply the sanction. 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,* that is, 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 intending) to intend to do it. Normally, an agent will form the intention to do something because she 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. She desires *having the intention* as a means of deterrence. Also, she is willing, in order to prevent the offense, to accept a certain risk that, in the end, she will apply the sanction. But this is entirely consistent with her having a strong desire not to apply the sanction, and no desire at all to apply it. Thus, while the object of her deterrent intention might be an evil act, it does not follow that, in desiring to adopt that intention, she 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) that 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.[11]

III. THE PRISON OF VIRTUE

(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 formulation 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.[12]

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.[13] And the disposition in question does not stand in such a relation to the agent's reasons for action.

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 it 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 her assessment of the reasons for and against acting. Consequently, a rational moral agent cannot intend to do something that she recognizes there are conclusive moral reasons not to do. Nor can she intend conditionally to do what she recognizes she 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 she 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, if they could, 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 the consideration of the vengeful agent, who desires to punish those who commit serious 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 person would do the same thing in the given situation.* Call this the Right-Good Principle. One suspects that, aside from qualifications concerning the good person'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 person 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. DELIBERATE SELF-CORRUPTION

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).[14] 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.[15]

(P3) may be viewed in light of a point about the credibility of threats that has been made by conflict theorists. Suppose a defender is worried about the credibility of her deterrent threat, because she thinks the wrongdoer (rightly) regards her as unwilling to apply the threatened sanction. She 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 occūrs. What I wish to note is that future selves of the defender herself 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 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 performing an inherently evil act" introduced in Section I. For convenience, let "an inherently evil act" 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 "fallback" strategy, should self-corruption 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 the agent's 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 such an agent is whether she ought to act to become less corrupt, with the result that she 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 herself and eliminate her 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.[16]

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 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.[17] 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 herself to prevent disaster in an SDS. Though we do not approve of her new corrupt aspects, we admire the person that she was for her willingness to sacrifice what she loved – part of her own virtue – in the service of others. At the same time, the fact that she succeeded in corrupting herself may make us wonder whether she was entirely virtuous in the first place. Corruption, we feel, does not come easily to a good person. 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. ACTS AND AGENTS

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 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.[18] Related quandaries arise when one considers, from an act-utilitarian viewpoint, the deterrent intention of a defender of 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.

A utilitarian might seek relief from these quandaries in either of two ways. First, she could defend some form of rule-utilitarianism. But then she would face a problem. Shall she include, among the rules of her 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 she does, the moral paradoxes of deterrence will appear within her system. If she does not, it would seem that her 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 her task as the modification, as well as the 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 – character traits, intentions, and the condition of the will – and these should not be evaluated solely in terms of their consequences. 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 person would feel obligated to do whatever is necessary to prevent a catastrophe, even if this required a sacrifice of personal virtue. 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 character and state of will. The extreme Kantian accepts Kant's dictum that morality requires treating oneself and others as ends rather than means. This is interpreted 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 treated 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, that is, 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 her own virtue and the rightness of her 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 assumption. 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. Conversesely, 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 incompatibility 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 at present see an entirely satisfactory way out of the perplexities that the paradoxes engender.

2. A paradox of deterrence revisited

Since Chapter 1 was originally published in 1978, there have been a number of discussions, by philosophers and others, of the issues treated there. These discussions have focused on the first moral paradox of deterrence, concerning whether it can be permissible to conditionally intend impermissible retaliation, and – in particular – on the application of this paradox to the case of nuclear deterrence. In this chapter, I consider some of the points raised in these discussions and explain my current views concerning this paradox.

The paradox arises from our apparently having good reasons to endorse – as regards Special Deterrent Situations (SDSs) – each of the members of this inconsistent triad of propositions:

(1) It would be wrong to retaliate if the offense were committed.

(2) It is permissible to form the intention to retaliate should the offense be committed, since this is the only reliable way to prevent the offense.

(3) If it would be wrong to do something under certain conditions, then it is wrong to form the intention to do that thing should those conditions arise. (The Wrongful Intentions Principle [WIP])

Unfortunately, the best known and most influential discussion of these issues – that of the U.S. Catholic bishops[1] – appears to endorse all of these propositions (in the case of nuclear deterrence) without acknowledging their mutual inconsistency.[2] That is, the bishops seem to condemn nuclear use and approve nuclear deterrence

I am grateful to Daniel Farrell and David Lewis for helpful comments on an earlier draft of this chapter.

(under specified constraints), while holding fast to WIP. This is quite understandable considering that the bishops' Pastoral Letter is a collectively produced political document operating under the constraints of the popes' limited endorsement of nuclear deterrence, the Catholic tradition's emphasis on the moral importance of intentions, and the need to achieve consensus among diverse opinions within the Church. Nonetheless, the bishops' failure to address the inconsistency of the various principles of nuclear morality that they apparently endorse inhibits our ability to learn clear moral lessons from an otherwise sensible and informative document.

If the bishops failed to notice, or at least acknowledge, the first paradox of deterrence, other writers – operating under fewer constraints – have not. Some, call them Traditionalists, have held fast to propositions (1) and (3), and have concluded that forming retaliatory intentions in SDSs is, after all, morally impermissible. Others, whom we may call Retaliators, have embraced propositions (2) and (3), and have concluded that retaliation would be permissible if deterrence failed in an SDS. Yet others have denied that the paradox applies to nuclear deterrence. In the next three sections of this chapter, I discuss these positions in turn, firmly rejecting the first two and explaining why I do not fully agree with the last.

Before proceeding with these discussions, however, something must be said about the nature of the first paradox of deterrence. As suggested at the beginning of Chapter 1, the first moral paradox of deterrence is analogous to a paradox about rationality noted by strategic theorists: it may be rational (for purposes of deterrence) to form the intention to carry out an irrational act of retaliation. But the analogy is not perfect in all respects. In particular, the moral paradox may apply in situations in which the rational paradox does not. Suppose A must sincerely threaten deadly retaliation against a group containing potential offender B, in order to deter B from committing a horribly destructive offense. If the likelihood of deterrent success is high enough, and the offense is bad enough (relative to the harm contained in the retaliation), the moral paradox arises. For retaliating would wrongly impose deadly harm on the other members of B's group, while intending to retaliate is necessary, and very likely sufficient, to prevent the offense. But suppose A is utterly indifferent to the fate of the members of B's group, but does desire to see serious offenders suffer. Then, given the usual instrumental conception of rationality as choosing effective means

to one's ends, it would be rational for A to retaliate against B's group (to secure revenge on B) if B committed the offense. Since here both forming the intention to retaliate and actually retaliating are deemed rational, there is no paradox of rationality. We have moral but not rational paradox, because morality rules out seriously harming innocent people in the pursuit of vengeance, while rationality, in itself, may not.[3]

Nor is this difference an unimportant one. For some nuclear deterrence situations may fit this pattern and involve us in moral, but not rational, paradox. Consider the example used in Chapter 1 to illustrate the moral paradox: a nation deciding whether to retaliate to a surprise nuclear attack that has left it with virtually nothing to defend. While we argued that such retaliation would be immoral, it would not be irrational if the potential retaliators are totally indifferent to the fate of those outside their nation. That the rational paradox may not arise here – though the moral one clearly does – is revealed by the form in which the rational paradox is often discussed by deterrence theorists. They consider the case of a limited first strike on a nation (or a strike against its allies) that renders retaliation irrational *because it would invite counterretaliation.* They do not seem to feel there is a similar problem about rational retaliation to an all-out first strike on the nation itself.[4] The implicit assumption operating here is that causing the destruction of one's own nation is irrational, while causing the destruction of other nations is not. A parallel claim about the morality of destroying one's own and other nations would not appeal to anyone but the most hardened nationalists. Hence the moral and rational versions of our first paradox apply to different nuclear scenarios.

Still, for purely theoretical purposes, the two paradoxes can be brought back together. We simply stipulate that the potential retaliator is not interested in revenge on the offender for its own sake, and that she knows that the reasons against retaliation will outweigh those for retaliation once the offense is committed. Then we have a paradox concerning the rationality of forming an intention to irrationally retaliate that is precisely analagous to our moral paradox. These two paradoxes should stand or fall together, and should have parallel solutions.

It has been necessary to clarify the relationship between the moral and rational versions of our first paradox because champions of different solutions have tended to focus on different versions. Op-

ponents of forming retaliatory intentions in SDSs have stressed the immorality of forming such intentions, while the main defender of retaliation in SDSs argues the rationality of such retaliation. I will answer each in kind, arguing that the former are wrong about morality and the latter is wrong about rationality.

I. WRONGFUL INTENTIONS

Traditionalists reject proposition (2), the view that, in an SDS, it is permissible to form an intention to perform an immoral act of retaliation.[5] The argument for (2), presented in Chapter 1, depends upon this intention – which I call an *SDS deterrent intention* – having (at least) the following five features:

(A) It is conditional – that is, of the form "If offense O occurs, I will do W."
(B) It is a deterrent intention – that is, one formed to prevent the occurrence of its antecedent condition (O).
(C) The offense O which the intending agent seeks to deter is an unjust and seriously harmful act.
(D) The deterrent intention would very likely prevent O.
(E) Given the magnitudes of O and W, and what is known about the likelihood of the deterrent intention (and alternative courses of action) preventing O, a rational utilitarian balancing of costs and benefits favors forming the intention.[6]

Unfortunately, in their discussions of the first paradox of deterrence, even some of the most sophisticated Traditionalists – such philosophers as Douglas Lackey, James Sterba, and Anthony Kenny – fail to take account of all these features of an SDS deterrent intention. Hence, their criticisms of proposition (2), insofar as they go beyond a mere reiteration of support for WIP, are largely off-target.

Consider first Lackey. He views the defender of deterrence as committed to the following principle: "It is always morally permissible to form an intention to do W if O provided that one has good reason to believe that O will not occur even if W is a wicked action which would be morally wrong to perform if O occurred."[7] And he proposes testing this principle by imagining whether someone who knows that he is unlikely ever to meet a member of a certain

minority group may permissibly form the conditional intention to spit in the face of any member of this group that he does meet. But Lackey's principle, and example, take account only of feature A (conditionality) and part of feature D – its implication that the conditions for carrying out the intention are unlikely to arise. Lackey completely ignores the other crucial features of SDS deterrent intentions. Forming the intention to spit that Lackey describes is, as he suggests, clearly unjustified. But this is partly because that intention does not serve the purpose of deterring a seriously harmful and unjust offense that cannot otherwise be prevented. (Indeed, in Lackey's description, the intention serves no discernible purpose at all but that of expressing the agent's anti-minority feelings.)

To genuinely test proposition (2) by Lackey's case, we must alter that case so that the relevant intention possesses these additional features. This requires imagining fanciful circumstances, but is nevertheless instructive. Suppose you lived in, and could not escape, a community that hated a certain minority group. Members of this minority group are known to sometimes approach members of your influential family, unless these members make clear that such approaches would be firmly rebuffed. (The conventional method of firm rebuff is spitting in the face.) But you are closely watched by members of your family who caution you against having anything to do with minority-group members and credibly warn that they will massacre many members of this group if any of them ever approaches you. At the same time, you have good reason to believe that the minority group has some spies in your community who might well infer your true intentions about how to respond if approached. Further, you know that minority-group members are much less well deterred from making approaches by fear of violence than by fear of rebuff. So you cannot reliably expect to prevent approaches by publicizing your family's warnings, but you can prevent approaches by having – and making known – the intention to firmly rebuff them. In these circumstances, it seems to me that it would be permissible for you (if you could) to form the intention to spit in the face of any minority-group member who approached you – provided that this was done to prevent the killing of innocents that would likely follow such an approach. In any case, one has reason to reject proposition (2) and affirm Traditionalism only if one rejects the permissibility of forming the (conditional) intention to act

immorally *in this sort of case*. Lackey's original version of the spitting case has no bearing on the truth or falsity of this proposition.

Sterba also uses an example to support Traditionalism and undermine proposition (2). The intention to do wrong that he focuses on is that of a gunman who sincerely threatens to shoot you if you do not hand over your money.[8] The gunman is only trying to prevent the occurrence of the circumstances in which he intends to carry out the threat (namely, your not handing over the money), and his threat is likely to succeed and not have to be carried out. Still, his forming the intention to shoot you if you do not surrender your money is wrong. This case, unlike Lackey's, captures the idea that proposition (2) ascribes permissibility to forming *deterrent* intentions that are likely to be successful; that is, it takes account of features A, B, and D.[9] But it ignores features C and E of an SDS deterrent intention – its prevention of an unjust, seriously harmful offense and its utilitarian justification. And, as in Lackey's case, if we add the ignored features, forming the SDS deterrent intention may plausibly be viewed as morally justified. Suppose that you wrongfully stole the money you possess from the gunman, and that taking it back at gunpoint is his only means of securing an emergency operation needed to save his child's life. Given these suppositions, what the gunman seeks to prevent by his intention – your keeping his money – is both unjust and has such serious bad consequences that a utilitarian justification of his forming the threatening intention is possible. But under these circumstances, it is no longer clear that his forming the intention is impermissible. Thus Sterba's example, like Lackey's, fails to come to grips with the case for proposition (2), because it ignores crucial features of an SDS deterrent intention.

Kenny describes a somewhat different argument as being "decisive against those who maintain that it is morally acceptable to have a conditional intention to do something which they agree to be morally unacceptable."[10] The argument seems to be this. If the agent were certain the condition would never arise, then he could not properly be said to have the intention to act (even conditionally). But if the agent lacks such certainty, as generally is the case, forming the intention is wrong (presumably because it can lead to the performance of the wrongful action, e.g., if the condition were to come about).

Depending upon one's interpretation of the slippery notion of intention, the first claim in this argument may or may not be true.

That is, it might be the case that inclinations or dispositions or reasons to act in certain ways in circumstances that one is certain will not arise are too "idle" ever to count as intentions (as opposed to wishes or fantasies). But this is beside the point in evaluating the morality of forming SDS deterrent intentions. For justification here depends only upon a *sufficient likelihood* of success in preventing the circumstances of fulfillment of the threat from coming about, as indicated in features D and E.

The second claim in Kenny's argument denies this; it asserts that anything less than certainty of successful deterrence would render forming the intention evil. But no support is ever offered for this claim, other than Kenny's reiterated endorsement of WIP.[11] Perhaps he is assuming that a morally good person would not – or could not – dispose himself to do wrong, even in circumstances he does not expect to arise. This may be so, as we saw in our discussion of the second moral paradox of deterrence in Chapter 1. But, as emerged in our second and third paradoxes, it does not follow that it is always wrong to form such dispositions. Indeed, one may be obligated to try to do so even at the cost of one's own virtue. This is a genuine (though paradoxical) possibility that Kenny fails to consider, much less argue against.

Lackey defends a weakened version of Kenny's second claim. He says that the acceptable level of risk of deterrence failure depends on the moral gravity of the retaliatory act, and that in the case of nuclear deterrence this implies we must be "nearly certain" of success if forming deterrent intentions is to be justified.[12] I would add that the acceptable risk level depends also on the moral gravity of the offense deterred, in which case something less than near certainty may suffice to justify nuclear deterrence. Lackey goes on to assert that the moral status of forming a deterrent intention to do W if O is the same as that of setting up an automatic retaliator which will do W if O occurs. This may be so, but what does it imply? The retaliator itself – poor machine that it is – has no intentions at all, conditional or otherwise, evil or otherwise. The intention of the agent who sets up the automatic retaliator is to prevent O by so doing. He does not himself intend to do W; at most he intends to risk being the initiator of a physical process that may end in W. *This* intention is wrongful only if that risk is not justified – which would seem to be primarily a matter of weighing costs, benefits, and probabilities. In other words, Lackey's arguments bring us back to a risk

– benefit calculation in determining the moral status of a deterrent intention. But, by feature E (or the definition of an SDS), the results of such a calculation support forming an SDS deterrent intention. There is no reason here to abandon proposition (2) and take the Traditionalist way out of the first paradox of deterrence.

Or perhaps there is. Gerald Dworkin, one Traditionalist who seems to have a very clear grasp of the logic of an SDS, contends that the characteristics of an automatic retaliation device reveal what is wrong with conditional intentions to retaliate immorally.[13] He considers the autoretaliator of our Chapter 4, a hypothetical device that can deflect half of incoming missiles to predetermined targets, with enemy cities chosen as deflection targets for purposes of deterrence. Deploying this device, Dworkin allows, is morally like practicing a policy of deterrence based on the conditional intention to retaliate. But both are to be contrasted with deploying a *bounce-back* device that is capable of deflecting half of incoming missiles to their point of origin, but nowhere else. For Dworkin, bounce-back is morally permissible while neither autoretaliation nor the conditional intention to retaliate is. This is because the bounce-back system, unlike autoretaliation, does not involve the immoral intention to kill civilians, though its existence might predictably lead to the deaths of as many civilians if there were an attack (since the potential attacker's missiles might be based near his cities). That is, bounce-backers do not impose a risk of death on civilians *as a means* of preventing attack, as autoretaliators (and conditional intenders) do.

We can best understand Dworkin's point, I think, by developing his suggestion that the difference between the intentions of the autoretaliator and the bounce-backer are cashable in terms of dispositions to act in certain counterfactual circumstances.[14] What distinguishes the autoretaliator is his willingness to impose risks on civilians for his deterrent ends. This is revealed by his not targeting his deflections (as he could) on oceans or deserts. The bounce-backer, as Dworkin conceives him, is different. His only available means of defense and deterrence – the bounce-back system – does impose risks on civilians.[15] But this risk is not chosen as a means, as is revealed by the fact that the bounce-backer would forgo (or accept less) deterrence rather than reinstate this risk if, for example, the enemy were to move all his missile bases far from cities (while the bounce-backer acquired a retargeting capacity so he could deflect missiles onto cities if he chose to do so). This is the real difference

between the autoretaliator and the bounce-backer as portrayed by Dworkin: in certain nonactual circumstances, the former would place enemy civilians at risk to preserve deterrence, while the latter would not.[16]

Now this difference in action dispositions between bounce-backers and autoretaliators, as characterized by Dworkin, reflects a difference in values between the two and therefore may influence how we evaluate them as moral agents. But it does not follow from this that, in an SDS, only the bounce-backer acts permissibly. Perhaps they both do. For while the autoretaliator is willing to impose a risk on the innocent in an SDS, by definition of this sort of situation, this risk is highly unlikely to eventuate in actual harm and its imposition is favored by utilitarian considerations. It may be that agents willing to suffer harm themselves rather than impose such risks are morally superior, for they sacrifice their interests rather than jeopardize those of others. But those who redistribute risks onto others when an overall utilitarian balancing favors redistribution, and actual harm is highly unlikely, are generally not acting impermissibly. If, for example, one can escape likely serious injury from some dynamite that is about to explode only by tossing it out the window where it could injure passersby, it is permissible to do so. So acting does not reveal a willingness to use others as means in any objectionable sense, though one would be more virtuous if one heroically faced the explosion rather than expose others to risk. But even this conclusion about comparative virtue would be open to question if one's family was also in the room and was endangered by the dynamite. This implies that the case for the permissibility of risk imposition in an SDS is even stronger for collective agents than for individuals. For most members of collectives that practice deterrence favor doing so largely to protect each other, rather than simply to protect themselves.

The upshot of all this is that the differences Dworkin notes between the bounce-back system, on the one hand, and autoretaliation or deterrent intentions, on the other, are not good reasons for thinking that the latter policies would be immoral in an SDS. However, Dworkin does not rest his case entirely on these differences. He offers two other arguments for the Traditionalist position that we must briefly consider. One is that the retaliation threatener must be able to justify her policy to those she places at risk, by showing that this policy benefits them on the whole.[17] But clearly, this is too strict

a requirement for permissible redistribution of risks: it would, for example, seem to imply that it is wrong to inflict substantial punishments on serious law violators. For many of them would be better off not suffering such punishments, even if this entailed an increased risk of being victims of the undeterred (or less deterred) crimes of others. Perhaps Dworkin means to apply this requirement only to risks or harms imposed on the innocent, but even here the requirement seems too strict. To justify quarantines must we be able to show that the contagious victims benefit on the whole from being confined? To justify private ownership and use of automobiles must we show that nondrivers benefit on the whole from such a practice? It is more plausible to suppose that benefits to some groups of a practice justify that practice if they sufficiently outweigh losses to other groups.[18] In at least some SDSs – those in which the utilitarian benefits sufficiently outweigh the costs – this condition will be satisfied.

Dworkin's final Traditionalist argument is that practices (such as deterrent threats) that impose risks of intentional harm are worse than practices that impose risks of accidental harm. He writes, "I do not believe that we would accept an institution which imposed the same risk of injury and death that [automobile] accidents cause, but risks brought about by actions aimed at injury or death."[19] But we do accept at least one such institution: private child rearing and the nuclear family. Every year, private families inflict significant and deliberate violence on thousands of children, and produce thousands of misguided offspring who eventually deliberately injure or kill others. Yet I think we would regard the nuclear family as justified – because of the benefits it provides to most and the central role it plays in their lives – even if we were convinced that some alternative mode of child rearing (e.g., in state institutions) would substantially reduce the rate of violent crime against children and adults. Similarly, if the overall benefits of a practice of deterrence in an SDS are substantial, we may regard that practice as justified even though it risks eventuating in harms done intentionally.

In summary, the Traditionalist arguments we have considered fail to establish that it is always immoral to form a conditional intention to do what is immoral. Many of them fail because they do not address the difficult cases for their position – deterrent intentions in SDSs. By focusing on disanalogous cases and red-herring principles, these arguments avoid rather than respond to the serious challenge

to WIP raised in Chapter 1. Other Traditionalist arguments take account of that challenge, but do not answer it in a persuasive manner.

II. RATIONAL RETALIATION

In Chapter 1, strong arguments were presented for proposition (2), which asserts the permissibility of forming deterrent intentions in SDSs. In the last section it was contended that Traditionalists have not given any good reasons for rejecting these arguments and proposition (2). But this does not yet establish that we must reject WIP or embrace paradox. There is the Retaliator's alternative of rejecting proposition (1), which asserts the wrongness of retaliation if the deterrent threat fails.

The main expositor of this alternative position is David Gauthier.[20] He discussed the rational analogue of our first moral paradox, which may be formulated, for SDSs, in the following propositions:

(1′) It would be irrational to retaliate if the offense were committed, because this would cause harms without producing sufficiently compensating benefits.

(2′) It is rational to form the intention to retaliate should the offense be committed, since this is the only reliable way to prevent the offense.

(3′) If it would be irrational to do something under certain conditions, then it is irrational to form the intention to do that thing should those conditions arise.

Gauthier agrees that (2′) holds true in some SDSs, essentially for the reasons I have given in Chapter 1. He accepts (3′) because, like the traditionalists, he believes that actions, and the intentions from which they flow, must be evaluated together: either both are rational (moral) or neither are. He infers that (1′) is false. He does not deny that once deterrence fails, more bad than good (in the potential retaliator's scheme of values) would be produced by retaliating. Indeed, he stipulates this as a feature of the situations he is most interested in discussing. Rather, he says retaliation must be rational, or else, in view of the truth of (3′), rational agents would not be able to have the deterrent intentions they need to deter offenses in SDSs.

Before turning to consideration of Gauthier's arguments for the rationality of retaliation in SDSs, it will be useful to clear aside a confusion that lends his conclusion more initial plausibility than it deserves. We are used to thinking of deterrence operating in repeatable contexts – like that of criminal punishment – where one's future credibility and ability to deter depends heavily upon one's willingness to carry out one's retaliatory threats once deterrence has failed in the case at hand. Habituated to thinking this way, we may find it easy to suppose retaliatory actions following failed deterrence rational, as Gauthier suggests, in SDSs as well. But the difference between the two cases is crucial. The long-range deterrent effects that may render retaliation rational in a repeatable context are, by definition, either absent or outweighed in an SDS. Hence, in evaluating retaliation in SDSs, we should resist being swayed by intuitions appropriate only for repeatable contexts.

Gauthier, however, does not rely on such misguided intuitions. He offers, as far as I can discern, four arguments for the rationality of retaliation in SDSs – that is, against (1′). Let us consider and respond to these arguments in turn.

Gauthier argues that if (1′) were true, it would be impossible for rational agents to form the rational deterrent intentions that they should have, according to (2′). This is true in one sense and false in another. Rational agents cannot form the intention to retaliate harmfully and pointlessly if they remain rational in all respects.[21] But they can seek to form that intention by exposing themselves to external influences that will render them irrational in the necessary respects. (Indeed, if Gauthier's account of rational retaliation is wrong but often persuasive, a rational agent in an SDS might seek to render himself appropriately irrational by reading Gauthier!) Thus, there are rational paradoxes analogous to the second and third moral paradoxes discussed in Chapter 1, namely:

(R2) There are situations (namely SDSs) in which it would be rational for agents to perform certain actions if they could (namely forming the intention to retaliate), and in which it is possible for some agents to perform such actions, but impossible for fully rational agents to perform them.

(R3) In certain situations (namely certain SDSs), it would be rational for a rational agent to deliberately (attempt to) make himself less rational.

The arguments for these propositions are precisely parallel to the arguments offered in Chapter 1 for their moral analogues. (R2) is true because a fully rational agent cannot intend to act against the balance of reasons, and hence cannot intend to retaliate in an SDS. (R3) is true because the agent's need for deterrence in an SDS may be so great that his ends are best fulfilled overall by making himself partly irrational and thus able to deter. So the answer to Gauthier's first argument is that a rational agent can, in principle, form the necessary intention to retaliate – though only by rendering himself less than fully rational.

This reply leads to Gauthier's second argument against (1′): this proposition precludes the unified assessment of the agent who forms and carries out a deterrent intention in an SDS. We must count him both rational and irrational. This is so, but there is nothing incoherent about it; the agent is simply rational and irrational at different times. In appreciating the case for having the deterrent intention and in setting out to form it, he is rational and acts rationally. Since, however, the intention is an intention to act irrationally (should deterrence fail), he can form this intention only by making himself irrational in certain respects. If he succeeds in doing so he becomes (partly) irrational. And if deterrence fails and he retaliates, he now acts irrationally. We are familiar with the same agents being rational and irrational, and acting rationally and irrationally, at different times. The only oddity about the present situation is that the agent rationally chooses at one time to try to make himself less rational at a later time. This is an oddity called for by the unusual structure of an SDS. It makes assessment of the agent's rationality over time more complex but not, as Gauthier suggests, impossible or incoherent.

Gauthier also claims that the rational agent is one who submits larger, rather than smaller, segments of her activity to rational scrutiny. This agent will assess actions in terms of the rational plans and intentions they flow from rather than from the effects they are likely to bring about. Now there may be something to this "wider segments" view. The general advantages of agents acting according to rules, plans, or policies rather than calculating on a case-by-case basis – for example, lower decision costs, more efficient coordination and cooperation – are well-known. But our normal view of rationality also implies being prepared to change previously formulated plans or intentions when there are significant stakes

involved and relevant new information about outcomes is available. This is precisely the situation that arises when deterrence fails in an SDS. There is much harm to be done by retaliation, and the benefit that motivated formation of the intention to retaliate – prevention of the offense – is now unobtainable. Hence, nonretaliation is now the rational action and is the one our failed deterrer would perform if she somehow regained full rationality after the commission of the offense.

Gauthier's fourth and final argument is that rational agents would be better off – that is, better able to achieve their ends – if it were rational to retaliate. For this would make them more effective deterrers in SDSs and similar situations than they could be if (1′) were true. Let us grant that, given appropriate assumptions about the improbability of deterrence failing and the improbability of rational agents transforming themselves into irrational retaliators, agents would do better on average if they were retaliators than if they were not. This would show retaliation to be rational, as Gauthier claims, only if we assume that rational acts are those flowing from the most beneficial traits. But this assumption is not valid. To see this, let X stand for any trait that we can all agree is irrational (and leads to irrational actions), but not normally so damaging as to make its possessors' lives miserable. If an eccentric billionaire were to heap fortunes upon all and only those having X, this would benefit those possessing the trait, and could make it rational for others to try to acquire the trait, but would hardly make the trait itself (or its possessors or the acts flowing from it) rational. If the environment is structured to reward irrationality, success is no proof of rationality. In an environment studded with enough SDSs (or single-play prisoner's dilemmas[22]), the most rational actors would be unlikely to fare the best. We may regret this, but we cannot really improve things by attempting to redefine "rationality" so as to make it impossible by definition.

In the end, then, none of Gauthier's arguments against (1′) is persuasive. There is, however, an argument against *his* position that is, in my opinion, conclusive. Deterrent intentions in SDSs are a subclass of what I call *problematic intentions.* Problematic intentions are those whose direct effects (i.e., effects of carrying out the intention) are bad, but whose overall expected effects are good because of their good and important *autonomous* effects (i.e., effects of the agent having the intention that are independent of the intention being carried out). A deterrent intention in an SDS is problematic because the bad

effects of carrying out retaliation are outweighed (when probabilities are taken into account) by the good autonomous effect – deterrence of the offense. Gauthier's view about rationality would imply a similar conclusion about problematic intentions in general as about deterrent intentions in SDSs: if it is rational to form and have them, it is rational to carry them out. But this cannot be right, as is shown by the following hypothetical example involving a problematic intention that is not conditional and has a desired autonomous effect other than deterrence.[23]

You are offered a million dollars to be paid tomorrow morning, if at midnight tonight you intend to drink a vial of toxin tomorrow afternoon that will make you very sick for a day. If you believe the offer and believe that the offerers can really tell whether, at midnight, you have the requisite intention, you would clearly have a good reason (in fact, a million good reasons) to form that intention. Suppose that you do so and bank the money the next morning – cashing in the desired autonomous effect of your intention. Would it then be rational for you to carry out your intention and drink the toxin? Surely not. If not, we have a divergence between the rationality of forming a problematic intention and the rationality of carrying it out – (3′) is shattered. Seeing no valid reason to suppose that this principle holds in the special case of problematic deterrent intentions in SDSs, I reject Gauthier's solution to the first paradox of deterrence.

III. NUCLEAR DETERRENCE AND RETALIATORY INTENTIONS

Our first paradox of deterrence has survived the attacks of the Traditionalists and the Retaliators. But does it apply to nuclear deterrence and tell us something about the moral status of that practice? Chapter 1 leaves this question open. It uses one conception of the nuclear balance of terror to illustrate the notion of an SDS (in which the paradox arises), and notes that the balance of terror may actually satisfy the definition of an SDS *if retaliatory intentions are necessary for successful nuclear deterrence.* The caution thus exhibited was appropriate. Nuclear deterrence occurs in an SDS, and thus exemplifies the first paradox, only if (i) its benefits outweigh its costs, (ii) it is actually necessary for defense, and (iii) its success requires possession of an actual intention to retaliate immorally should deterrence fail. In Chapters 3 and 6, it is argued that conditions (i)

and (ii), respectively, may well be satisfied. In this section I will discuss whether intentions to retaliate immorally are necessary for successful nuclear deterrence.

Deterrence works, if it does, by persuading a potential aggressor that the risks of retaliation attached to the contemplated act of aggression outweigh its benefits. If the costs of suffering retaliation are immense, as they clearly are in the case of nuclear retaliation, the probability of that retaliation need not be very high to render aggression for any plausible political gain clearly a bad bargain. So even minimally rational governments will be deterred from engaging in aggressive acts that they believe might lead to their nation suffering nuclear retaliation. This analysis, plus experience with how high government officials actually regard nuclear weapons, has led to the idea that the existence of a nuclear retaliatory capability suffices for deterrence, regardless of a nation's will, intentions, or pronouncements about nuclear weapons use. This basic idea, called "existential deterrence,"[24] has led to various proposals for effective nuclear deterrence without immoral retaliatory intentions – bluffing, deterrence without threatening retaliation, and so on.[25] Here I limit my attention to two of the more interesting proposals, which I call, respectively, *No Intention* and *Scrupulous Retaliation.*

A *No Intention* nuclear retaliation policy is one practiced by a nation having the capability to retaliate if attacked (i.e., survivable nuclear weapons and plans for their possible use), but having no definite intention about whether or not to use this capability. It is not that the nation's leaders intend not to retaliate, they simply put off making up their minds about retaliation unless and until their nation is actually attacked.[26] By contrast, a *Scrupulous Retaliation* policy is one in which a nation intends to retaliate if subjected to nuclear attack, but only in a clearly moral fashion by limited strikes against military and economic assets located far from population centers.[27] Apparently, neither policy involves the conditional intention to retaliate immorally if attacked; hence neither can be shown to be wrong by direct application of WIP. If nuclear deterrence in either form were effective, it seems that we could practice nuclear deterrence (in that form) without being subject to the first moral paradox of deterrence.

There are two key questions to address here. Would these alternative forms of nuclear deterrence be as reliable as deterrence based on a conditional intention to immorally retaliate? Would there

really be significant moral advantages to be gained by practicing one of these policies rather than a policy of deterrence by intention to immorally retaliate (*DITIR*, for short)? Let us consider these questions in turn.

No one really knows whether forms of nuclear deterrence that promise less retaliation (e.g., Scrupulous Retaliation) or retaliation with less certainty (e.g., No Intention) are less effective deterrents than policies threatening more retaliation with greater certainty. If our adversaries were always naive calculators who were prepared to attack us at any time their calculations showed the slightest gain in expected value for them in doing so, it would follow that a threat of greater retaliation with greater certainty would be a more reliable deterrent. If, to take the opposite extreme, our adversaries would always be deterred by the mere possibility of suffering significant nuclear retaliation, a threat of less (but still significant) retaliation with less certainty would be an equally effective deterrent. Doubtless, supporters of existential deterrence are correct that present nuclear adversaries under present circumstances are much closer to the latter extreme than the former – they are strongly disposed to err on the side of caution in deciding whether to use nuclear attack to achieve political-military gains or avoid political-military losses. But, as there is little prospect of achieving nuclear disarmament except over a relatively long period of time, we want our nuclear deterrence policies to be extremely *robust* – that is, effective under the greatest possible variety of circumstances. In particular, we want them to work even against non-risk-averse leaders who may come to power in nuclear-armed countries in the future, and in circumstances in which all the alternatives to using nuclear weapons may seem bleak and undesirable to our adversaries. We also want nuclear deterrence to work without a single instance of failure, including in changed political circumstances that might result from future environmental, population, or resource problems.[28] Now in theory, we might adapt our retaliatory policy to the dangers of the moment and maintain a policy of Scrupulous Retaliation unless and until non-risk-averse nuclear adversaries actually appeared.[29] But in the real political world there would very likely be a substantial time lag before such an adversary's true nature was perceived and appropriate changes in retaliatory policy were put into effect, just as it took a long time for the Western democracies to perceive, appreciate, and respond to the grave threat posed by Hitler. In a nuclear world, the

consequences for humanity of a similar lag in appropriately responding to a non-risk-averse leader (or leaders) of a major power could be catastrophic.

Given the reasonable desire for robustness, the importance of avoiding a single failure, and this time-lag problem, it does not seem irrational to opt for the greater potential credibility provided by a nuclear policy of DITIR. We cannot know that circumstances will ever arise in which having such a policy will be necessary for successful deterrence, nor can we know that they will not arise. But given the momentousness of what is at stake – the avoidance of nuclear war – we should not risk practicing a less effective deterrent policy unless there clearly are overriding moral advantages to be gained by doing so.

Are there such advantages in the case of Scrupulous Retaliation? At first, it seems so, for such a policy appears to spare enemy civilians from the danger of our retaliation. But if Scrupulous Retaliation is a less robust and reliable deterrent this need not be so. In the event of nuclear war, our retaliation practices might (deliberately or accidentally) be much less scrupulous than our prewar intentions. And the environmental effects of nuclear war (e.g., radioactive fallout or nuclear winter) might lead to the death of many of these civilians without our intending it. Thus, if Scrupulous Retaliation raises the probability of nuclear war enough, it may (compared to DITIR) actually *increase* the risks of nuclear destruction for enemy civilians. At the same time, if it is a less effective deterrent, it raises the risks to ourselves and our allies. Thus, if Scrupulous Retaliation does sacrifice robustness, as has been suggested, it possesses no clear moral advantages that would compensate for this sacrifice.

In addition, Scrupulous Retaliation poses the following moral-strategic dilemma. If sincerely proclaimed as official policy and reflected in force design and deployment changes, Scrupulous Retaliation is subject to being made even less effective by countermoves. Adversaries may, for example, try to base their most valuable military and economic assets in or near cities so as to deprive us of meaningful retaliatory targets. Suppose, on the other hand, Scrupulous Retaliation were adopted in secret by high officials. This would leave lower-level officials, missile crews, and ordinary citizens (who would believe that the policy is still one of unrestricted retaliation) in the same state of "nuclear sin" they began in – only the leaders will have improved their moral state.

Before leaving the subject of Scrupulous Retaliation, I should emphasize that I have been discussing it – and questioning its supposed advantages – as a form of deterrent policy. As noted in the Introduction, once a nuclear war started, the morally proper policy to follow then would probably be to carry out only scrupulous retaliatory attacks, if any. To point out this divergence between permissible deterrent intentions and permissible retaliatory actions is, in essence, to restate the first moral paradox of deterrence.

What of the No Intention policy? Does this policy have any decisive moral advantages that might compensate for its potential lack of robustness? To answer this last question, we must consider why the intention to retaliate immorally with nuclear weapons is considered bad. Then we may determine whether, and to what extent, the No Intention policy is itself free from these bad-making characteristics.

One obvious reason for thinking DITIR bad is that it creates an actual risk of death (and other serious harms) for the many innocent people who would suffer attack if the intention were carried out. But would a No Intention policy create a lesser risk for these people? This depends upon a number of indeterminable empirical factors. How likely is it that the top leadership would decide not to retaliate (or to retaliate scrupulously) if their nation suffered a nuclear attack? If they decided not to retaliate, or to retaliate in a scrupulous or otherwise limited way, how likely is it that their decisions would actually be adhered to in the midst of a nuclear war? Does the relative lack of robustness of a No Intention policy increase the risk of nuclear war, and if so, by how much? Depending upon the answers to these questions, the No Intention policy may (or may not) actually *increase* the risks of nuclear destruction undergone by enemy civilians (and the risks of our being complicit in immoral nuclear retaliation).

As we saw in our discussion of the Traditionalists, however, there is a different possible explanation of the badness or evilness of intentions to retaliate immorally. Such intentions entail a willingness (under certain circumstances) to perform a wrong action, and hence reflect a flaw in the agent's values. In this view, the relevant question to ask about the No Intention policy is whether those who pursued it would possess better values than pursuers of DITIR.

Consider first the top leadership who, under the No Intention policy, have not made up their minds whether they would retaliate. Whether their values are better would seem to depend upon what

they would decide if attacked. If they would in fact retaliate, it is hard to see that this reflects better on their values than if they had decided to do so ahead of time (perhaps influenced by considerations of deterrence[30]). We, and perhaps they themselves, do not know what they would decide in the event. Hence, while it is possible that a No Intention policy reflects superior values by top leadership in a given case, it need not, and we would not know whether in a particular case it did (at least until nuclear war had broken out).

What of lower-level officials, soldiers, and ordinary citizens? Does their going along with a No Intention policy reflect superior values to going along with threats of all-out retaliation? Again, it may or may not. If they support or acquiesce in a No Intention policy that could for all they know amount to the same wrongful acts in the event of war as a policy of all-out retaliation, it is hard to see in what way their values are superior. Perhaps if their support is based on the perceived likelihood that there would be no retaliation against civilians, and would be withdrawn if that perception changed, we could infer that the individuals in question had superior values. But if based on other grounds, it would seem that support for a No Intention policy, like support for DITIR, indicates a willingness to risk complicity in mass killing of the innocent. Thus, while a No Intention policy might reflect superior values on the part of some individuals, it need not. Indeed, depending upon the underlying reasoning, support for DITIR might reflect better values – for example, if the individual regards the two policies as equivalent in their effects, and views the No Intention policy as hypocritical and dishonest since "we'd surely retaliate anyway."

We have until this point considered the intentions of individuals. But what of the intentions of collectives such as nations? If WIP applies to them, it is relevant to inquire whether the No Intention policy avoids the *collective* intention to retaliate immorally if attacked. This is not easily determined, however, given that there is no generally accepted theory of what intentions are, even in the individual case. Chapter 1 assumed that rational intentions, at least, are dispositions to act derived from the agent's appreciation of the reasons for and against so acting. But it is not clear whether this account is correct, or even what exactly it means, when applied to the collective case.

Our difficulties in this matter are compounded by the fact that

there is no simple formula for inferring group intentions from the intentions of individuals making up that group. Even all group members sharing the intention to do their part in a joint undertaking is not always sufficient to constitute a group intention. For even if the physical means of carrying out the intention are available, and each fully intends to do his best, it may be apparent that things are not sufficiently organized to get the job done. Thus, for example, there may be enough shelters to protect all, and each may be committed to doing her part to get herself and others into shelters in the event of attack. But if it is obvious that there are not sufficient workable organizational and operational plans to get people into shelters, so that chaos would be the likely actual result of attack, the nation could hardly be said to genuinely intend to protect itself and its citizens by means of shelters, if attacked. So even unanimous individual intentions plus physical capability need not add up to a group intention.

On the other hand, there may conceivably be a group intention to do X even if no individual member of the group intends to do (her part of) X. Suppose each member of a society secretly opposes the official policy of nuclear retaliation, but wrongly believes that all others favor that policy. Each intends not to do her part in retaliating, should the occasion arise. But it is predictable that enough would do their parts, if the occasion arose, because of the pressure of the perceived expectations of their comrades, and the (perhaps true!) belief that if one did not act (e.g., did not press the button firing the missile), someone else surely would.[31] In this case, it would seem appropriate to ascribe the intention to retaliate to the nation, though none of its members at present share that intention.

In light of all this, I am inclined to propose the following as jointly sufficient conditions for a group G *collectively intending* to do X if C occurs:

(a) G has the physical capability to do X if C occurs.
(b) G has plans to use this capability to do X if C should occur.
(c) It is in fact likely that were C to occur, G would put these plans into effect and do X.

Note the rough analogy between the above account of rational intentions and this partial analysis of collective intentions. Conditions (a) and (c) correspond to having a disposition to act, while the

notion of a plan in condition (b) roughly corresponds to the idea of being disposed to act in virtue of reasons for action.

The importance of this analysis, for our present concerns, is this. Suppose that the top leadership has not decided whether to carry out plans to immorally use their nuclear retaliation capacity if their nation suffers a nuclear attack – that is, the nation's policy is one of No Intention. But suppose that the leaders' values are such, or the nature of the command and control system is such, that it is likely that immoral retaliation would in fact take place if there were an attack. In this case, all the conditions in the above analysis are satisfied and the nation possesses a collective intention to retaliate immorally if attacked. In other words, a No Intention policy may, at the collective level, involve the intention to retaliate immorally. This means that if WIP applies to collective intentions, No Intention could have a moral advantage over DITIR only if command and control is highly reliable and top leadership possesses the right moral values (and could be expected to retain and act on those values during a nuclear war). Given the doubts of some experts about the ability of existing command and control systems to maintain nuclear restraint during alerts,[32] much less under nuclear attack, one may rightly wonder whether No Intention is a morally superior policy.

The task of this section was to evaluate the claim that the first moral paradox does not really apply to nuclear deterrence because – in virtue of existential deterrence – reliable nuclear deterrence without immoral retaliatory intentions is possible. Consideration of two representative alternative policies, Scrupulous Retaliation and No Intention, has suggested that only a more ambiguous conclusion than that embodied in the above claim is justified. Because of complex factual and conceptual uncertainties, we simply do not know whether the first paradox applies to nuclear deterrence or not. In the case of Scrupulous Retaliation, we do not know whether the potential loss in robustness it would entail is enough to eliminate it as a viable alternative policy, thus leaving intact the original argument for the permissibility of having immoral retaliatory intentions in the nuclear case. As regards the No Intention policy, there are several key uncertainties. Is it robust enough? Does it, at the individual level, involve the same moral paradox as DITIR, because many of the individuals involved in implementing the policy would likely have the same values that render the latter policy questionable? Does the No Intention policy escape the first paradox of deterrence at the level of collective intentions? Without definite answers of the right sort to these questions, it is hasty – and quite possibly wrong – to conclude that our first paradox does not apply to nuclear deterrence.

IV. CONCLUSION

Traditionalists have given us no persuasive reasons for abandoning proposition (2). Retaliators have given us no persuasive reasons for abandoning proposition (1′) – or, by analogy, proposition (1). Given the strong arguments for (1) and (2) offered in Chapter 1,[33] we are faced with the choice of embracing paradox or rejecting the Wrongful Intentions Principle. The latter is obviously the better choice. Strong intuitions support WIP, but these intuitions are doubtless based on the ubiquity of normal cases in which the intentions in question are nonconditional, nondeterrent, or non-SDS-deterrent. It is highly unlikely that originators of WIP considered the strange case of deterrent intentions in SDSs. The abnormality of this case is highlighted by the fact that some modern philosophical defenders of the WIP have failed to take account of many of its relevant features even when these have been clearly laid out in the literature. In any case, what seems to be the best way out of the first paradox of deterrence is to qualify WIP so that it does not apply to SDS deterrent intentions, or in general to problematic intentions. If this modification of WIP is justified, our paradox has served a constructive purpose by leading us to properly limit the scope of this important moral principle.

It is worth noting in addition that a proper understanding of the argument of Chapter 1 allows us to hold fast to some of the main Traditionalist intuitions underlying WIP, even while rejecting or modifying the principle itself. As the second and third paradoxes of Chapter 1 showed, there is some moral (or rational) defect in the agent who has succeeded in forming the relevant deterrent intention in an SDS. Thus, we may concede to the Traditionalists that there is something morally wrong with the intention itself (namely, it is an intention to act wrongly under certain circumstances) and with the character of the agent who has it (namely, possessing either the wrong values or the willingness to act wrongly under certain circumstances). But, as the arguments of Chapter 1 indicate, it does not

follow from this that it is wrong for the agent to form this SDS deterrent intention. Hence, the unmodified version of WIP may be rejected while the main underlying intuitions are retained.[34]

Appropriately modifying WIP provides a theoretical solution to the first paradox of deterrence. But what does all this imply about the morality of nuclear deterrence? The arguments of the last section indicate that we do not know whether the paradox really applies to the case of nuclear deterrence. But if the proper solution to that paradox involves modifying WIP so it does not apply to SDS deterrent intentions, this does not really matter. If robust and reliable deterrence without immoral retaliatory intentions were possible, we could deter without running afoul of WIP. If not, and nuclear retaliatory intentions occur within an SDS, they fall outside the proper scope of the (modified) WIP. In neither case are they shown to be wrong in virtue of proper employment of the WIP.

This, of course, is not enough to determine that nuclear deterrence in some form is morally permissible. It remains to be shown that such deterrence can reasonably be viewed as having utilitarian benefits that exceed its costs. (Otherwise the nuclear deterrent situation is not a genuine SDS.) And there may be other deontological moral objections to nuclear deterrence that should be answered, besides the charge that it embodies wrongful intentions. Finally, deontological arguments *for* nuclear deterrence must be considered. In examining these matters further, in later chapters, we shall see that the first moral paradox of deterrence is not the only moral puzzle surrounding nuclear deterrence.