Deterrence and the Fragility of Rationality*

Frederick Kroon

I

The 1950s saw the birth of an influential new position in strategic thinking about nuclear weapons. This position made two related claims. On the one hand, it conceded something that had become obvious to many: given the growth of nuclear arsenals, the actual use of nuclear weapons in a full-scale nuclear exchange would be suicidal and hence irrational. But it also contended that despite this—indeed, because of this—the serious threat to use such weapons could well be entirely rational. A threat to use nuclear weapons in massive retaliatory response to their use by others would be rational, so the view held, if it was on balance likely to be both necessary and sufficient for averting this offensive use and hence likely to avert the start of such a suicidal and irrational nuclear exchange. (Most strategic thinkers probably thought that the threat would in that case be moral as well, given the moral importance of the goal of peace and security served by the threat, but they largely shied away from using moral categories to describe their views.) In its most extreme form, this way of thinking came to be known as Assured Destruction or, in its famous symmetric form, as Mutual Assured Destruction (MAD). Because of its history and prominence, I shall call this the "classical" policy of nuclear deterrence, although for much of the time I omit the description 'classical' since we won't be discussing other forms of nuclear defense in this article.

Over the past few decades, philosophers have been among the strongest critics of the morality of the classical policy, often using

* Versions of this article were read at the Universities of Auckland, Canterbury, and Otago, and at the 1995 Australasian Association of Philosophy conference held in Armidale. My thanks to Gillian Brock, Paul Griffiths, David Lewis, and many others for their helpful comments. I owe a special debt to Philip Pettit, Michael Slote, and Christine Swanton, as well as to two anonymous referees and associate editors of *Ethics*.

Ethics 106 (January 1996): 350–377

arguments that focused on the risks and costs associated with pursuing the policy.[1] But what has become the most famous philosophical criticism of the policy is firmly nonconsequentialist and much more direct: even if retaliatory threats deter in the way predicted, so this argument goes, classical nuclear deterrence is still morally wrong since no truly moral agent or agents can be aware of what is involved and yet be of a mind to retaliate in so utterly senseless, destructive, and inhumane a way. At the point of delivery, after all, retaliation doesn't just involve inflicting great harm, but involves inflicting it for no useful purpose since it comes only after deterrence has failed;[2] to that extent, agents who seriously contemplate retaliating in the event of an attack are not guided by the moral implications of what they contemplate and so are immoral (assuming, of course, that they are sufficiently rational to understand these implications).

The other side of the coin is that classical deterrence must then also be an irrational policy, at least if certain rather plausible assumptions are granted. Rational agents can't be of a mind to retaliate in this way if we assume, for example, that retaliation involves choosing overall worse prospects for oneself and one's group (say, because one's chances of survival are less in a world containing more destruction, and one cares above all about survival). And they can't be of a mind to retaliate in this way if a senseless and destructive kind of revenge simply can't form part of the repertoire of ends available to any genuinely rational agents (say, because more might be required for rationality than the proper serving of whatever ends an agent has; we might want more than a merely instrumental conception of rationality).[3]

1. See, e.g., Douglas Lackey, *Moral Principles and Nuclear Weapons* (New York: Rowman & Allenheld, 1984) and *The Ethics of War and Peace* (Englewood Cliffs, N.J.: Prentice Hall, 1989).

2. See, e.g., Michael Dummett, "The Morality of Deterrence," *Canadian Journal of Philosophy* 12, suppl. (1986): 111–27; and Anthony Kenny, *The Logic of Deterrence: A Philosopher Looks at the Arguments for and Against Nuclear Disarmament* (London: Firethorn, 1985). Kenny, in fact, thinks that even being willing to retaliate in this way is already gravely immoral, even if the "agent" is not finally committed to retaliation. (Throughout, I am putting aside the complication that the "agent" doing the threatening is more in the nature of a collusive body than an individual.) Note that for both Kenny and Dummett it is the nature of the retaliation—the targeting of innocents—that makes for the immorality. It is clear, however, that we need not insist on this: even consequentialists will agree that retaliation is immoral when, as in this case, there is no outweighing good.

3. Cf. Hume's statement, "It is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (in David Hume, *A Treatise of Human Nature*, ed. Selby-Bigge [1888; reprint, Oxford: Oxford University Press, 1958], p. 416). Many contemporary commentators reject such a purely instrumentalist conception; see, e.g., Robert Nozick, *The Nature of Rationality* (Princeton, N.J.: Princeton University Press, 1993), and David Schmidtz, "Choosing Ends," *Ethics* 104 (1994): 226–51.

If—as I shall do from now on—we restrict our discussion to situations and conceptions satisfying one or both of these assumptions, we can then argue that those of a mind to retaliate in the manner of the classical policy must somehow fail to see the irrationality of what they contemplate and so must be irrational.[4]

According to this kind of nonconsequentialist critique, then, one reason why nuclear deterrence is not only morally but also rationally problematic is that the agents who implement the policy must be either immoral or irrational (or both). I stress again that this sort of critique is quite different from various other familiar responses—for example, the response that deterrence is immoral and irrational because it is unstable and unlikely in the end to keep the nuclear peace, or that it is immoral and irrational because it is based on a bad misreading of the intentions and/or capabilities of one's opponents and so subjects numerous people to unnecessary risks. To keep the issue as clear-cut as possible, I am going to assume that we are dealing with situations in which classical deterrence is not unstable in this sort of way, that there indeed is a substantial threat to the nuclear peace, that a policy of seriously threatening massive nuclear retaliation is by far the most effective way to ward off this threat while carrying only a minimal risk of the deterrer's actually using her weapons, and that the ends served—in particular, a stable nuclear peace—are of the highest importance from both a prudential and moral point of view. I am going to assume, that is, that the policy of deterrence is being invoked in possible situations in which the policy can't be faulted on moral or prudential grounds of a broadly consequentialist sort, even though acting on the deterrent threat in case deterrence fails means acting irrationally and immorally. I fully acknowledge, of course, that this is a huge and controversial assumption where real-world deterrence is concerned[5]—perhaps real-world deterrence was always riskier, more dangerous, less useful, than many strategic thinkers allowed, and perhaps it never really involved the sort of conditional threats of massive

4. The perception that it is somehow irrational to threaten massive destruction of this kind is not, of course, an unusual criticism of deterrence theory. Usually, however, it is presented as an attack on the believability of the threat rather than as a direct attack on the rationality of deterrence. Robert Foelbar notes one extreme exception to the claim of the irrationality and immorality of retaliation in his "Deterrence and the Moral Use of Nuclear Weapons," in *Nuclear Deterrence and Moral Restraint,* ed. Henry Shue (New York: Cambridge University Press, 1989): one's adversary may be so evil that it should not be left dominating the world.

5. For an excellent account of these matters, see Steve Lee, *Morality, Prudence, and Nuclear Weapons* (New York: Cambridge University Press, 1993). Lee himself offers a complicated conditional moral argument for a form of minimum nuclear deterrence, the condition being that this deterrent stance must "help to bring about conditions that would make the abandonment of nuclear weapons prudentially preferable" (p. 331).

"countervalue" retaliation on which the classical model rests (there is little doubt that modern versions of deterrence have some rather different features). But so long as we remain up-front about the assumption, this can scarcely be a cause for complaint.[6]

Let us now ask, who is right, those who support this form of deterrence under the assumed conditions and claim that the policy is both rational and moral, or their nonconsequentialist philosophical debunkers described above? The writings of Gregory Kavka contain an intriguing and by now famous compromise reply: the supporters of deterrence are right, but so, in a way, are the debunkers; for while the debunkers are wrong to think that the classical policy of deterrence must be either irrational or immoral, they are right in their contention that rational and moral agents could not maintain the kind of intentional mind-set on which the policy rests—the mind-set of those conditionally intent on behavior that in the event would be utterly inhumane and destructive.[7] Put as a kind of paradox (one of his "paradoxes of deterrence"), Kavka's position is that in the relevant circumstances, "It would be right, both morally and prudentially, for agents to perform certain actions [namely, forming the conditional intention to retaliate], it is possible for some agents to perform such actions, but it is impossible for rational and morally good agents to perform them."[8]

6. In an article published in the mid-eighties ("Devil's Bargains and the Real World," in *The Security Gamble: Deterrence Dilemmas in the Nuclear Age,* ed. David MacLean [New York: Rowman & Allenheld, 1984], pp. 141–54), David Lewis objected that taking this assumption—and the ensuing puzzle about morality and rationality—to be a serious and worthwhile object of philosophical reflection was irresponsible to the extent that the assumption represented deterrers as partly corrupt in virtue of their retaliatory intentions, thereby insinuating that real-world deterrers were also corrupt: a "picture that implicitly slanders many decent [American] patriots" (p. 148; to be fair, Lewis thinks this picture portrays deterrers as a mixture of both good and bad). According to Lewis, real-world deterrers never entertained the sort of corrupting retaliatory intentions invoked on the classical policy. I disagree with this charge on two related counts. First, Lewis is wrong, in my view, to think that such retaliatory conditional intentions indicate a (partly) corrupt character, and so is wrong to think that portraying deterrers in the way the assumption does is to denigrate deterrers. (That, in fact, is going to be the burden of the present article.) Second, he is wrong, in my view, to think that real-world deterrers never maintained the classical policy of deterrence, although it may be true that they never maintained it consistently, or only maintained it as the "if all else fails" part of a more flexible policy, or were not sure if they could maintain it in the course of a nuclear attack. Indeed, showing that Lewis and others are wrong about the corrupt character of classical deterrers might help to remove some of the temptation to think that the classical policy is somehow an unthinkable option for a rational, moral state.

7. See Gregory Kavka, "Some Paradoxes of Deterrence," *Journal of Philosophy* 75 (1978): 285–302, "Nuclear Deterrence: Some Moral Perplexities," in MacLean, pp. 123–40, and *Moral Paradoxes of Nuclear Deterrence* (Cambridge: Cambridge University Press, 1987).

8. Kavka, "Some Paradoxes of Deterrence," p. 294.

The only agents able to form and entertain the requisite intentions would thus have to be suffering from a form of rational or moral blindness, a trait that in the circumstances should be encouraged rather than deprecated.

I take it that from the point of view of the early proponents of nuclear deterrence this would not be a concession of any worth. They didn't just think that nuclear deterrers were doing something that happened to be rational (and even moral); they thought that in the specified circumstances nuclear deterrers were acting the part of properly rational agents, that nuclear deterrers were doing what a fully rational agent would be doing if put in the same difficult situation, despite the monstrousness of what was threatened. Call this kind of position an "agent-rationalist" view of nuclear deterrence. More precisely, agent-rationalists about nuclear deterrence are those who think that it is not only the act of threatening retaliation—in the sense of conditionally intending it—that is fully rational in the specified circumstances; the agent who threatens retaliation in these circumstances can also be fully rational, despite the fact that what she threatens to do is irrational. The contrary position held by Kavka I call an "agent-irrationalist" view of nuclear deterrence. On such a view, deterrers must be irrational in some way, perhaps through having undergone a process of corruption that gives them irrational goals or makes them unable to understand the full implications of what they propose.[9] (Although I am mainly interested in nuclear deterrence, the issues, of course, are wider. Thus agent-rationalism and agent-irrationalism can also be understood more broadly as views concerning the rationality of agents who face "Special Deterrent Situations" in roughly Kavka's sense; these situations include our nuclear scenarios but also many other possible situations of conflict between agents. While the argument of this article may be general enough to extend to all such situations, I shall continue to focus on the nuclear case.)[10]

9. Agent-irrationalism may well be the dominant view about the "paradoxical" intentions that underlie deterrence in our (hypothetical) nuclear scenarios. Kavka defends this view, but so, in different ways, do a number of other authors, e.g., Daniel Farrell, "On Some Alleged Paradoxes of Deterrence," *Pacific Philosophical Quarterly* 73 (1992): 114–36. There is also a clear sense in which David Lewis should be considered an agent-irrationalist. For Lewis insists that our (hypothetical) nuclear deterrers are a "strange" mixture of good and evil, and of the rational and the irrational (Lewis, pp. 144–46). By admitting that there is something evil and irrational in deterrers, he depicts himself as both an "agent-immoralist" and agent-irrationalist, since these doctrines claim only that deterrers are not *wholly* virtuous or rational. (In conversation, however, Lewis claims that he is a clear case of an agent-irrationalist only if the standards for full rationality are set very—perhaps inappropriately—high. Despite Lewis's rhetoric, therefore, he may in the end still be a kind of agent-rationalist.)

10. Kavka describes "Special Deterrent Situations" as follows: they are situations where in order to prevent some harmful and unjust offense, an agent must threaten (in the sense of conditionally intend to apply) some harmful sanction should the offense

In the same way, we may talk of "agent-moralism" and "agent-immoralism." Thus agent-immoralism about nuclear deterrence holds that because of the immorality of the retaliatory act, and despite the moral desirability of the threat, no morally good agent can seriously threaten retaliation in the nuclear scenarios described.[11] Any agent able to threaten retaliation must have undergone a process of moral corruption, or be affected in some other way by an element of moral imperfection in her nature. (This is again Kavka's view, but versions of the view are held by many others; David Lewis, for example.)

These various positions are not, of course, exhaustive. Take rationality again. Some theorists think that there can be no situation in which threatening nuclear retaliation is rational.[12] If so, no fully rational agent could be a nuclear deterrer. And in the mid-1980s (but no longer) David Gauthier held that because threatening retaliation is sometimes clearly rational, it would ipso facto be rational in those cases for a deterrer to act on her retaliatory threats should deterrence fail. If so, agent-irrationalist arguments can't get a toehold, and we can no longer deny full rationality to nuclear deterrers. While I reject these various positions, they are not the direct concern of this article.[13]

---

occur, this threat has a high probability of preventing the offense, the amounts of harm involved in the offense and the threatened sanction are both very large, a rational consequentialist calculation would substantially favor having the intention, and the agent would have conclusive moral reasons not to apply the sanction if the offense were to occur (Kavka, "Some Paradoxes of Deterrence," pp. 286–87). If this characterization is to fit the nuclear scenario as I have described it, we should assume that the moral and prudential aspects of the situation coincide (e.g., that it would also be irrational, not just morally wrong, for the agent to apply the sanction).

11. Kavka says only that such an agent can't be both morally good and rational. In fact, I doubt that we would call an agent unequivocally good in this sort of case unless she could rationally grasp the issues facing her and understand the implications of her choices.

12. Alan Gewirth comes close to arguing this in his "Reason and Nuclear Deterrence," *Canadian Journal of Philosophy* 12, suppl. (1986): 129–59.

13. The classic statement of Gauthier's older position is his "Deterrence, Maximization and Rationality," *Ethics* 94 (1984): 474–95 (reprinted with minor changes in MacLean, pp. 101–22). Gauthier there seems to assume that a rational agent, following the dictates of a "maximization of conditional expected utility" decision theory, can come to entertain deterrent intentions, and then argues that the alleged irrationality of retaliation itself (should deterrence fail) is incompatible with the agent's integrity over time: the idea that it is one and the same agent who, simply by following up on certain rationally formed intentions, and without any change inside her, is now declared to be doing something irrational. But if that is the argument, then agent-integrity is better served in other ways; for how can a rational agent, who in other contexts repudiates the sort of vicious, useless action she is conditionally asked to perform in the conditional intention, maintain her integrity and yet form the intention? The present article suggests a way of reconciling the rationality of both the agent and the agent's deterrent stance with a form of agent-integrity. Gauthier himself now rejects his earlier view. In "Assure and Threaten," *Ethics* 104 (1994): 690–721, he defends a rather different view, one more akin to agent-irrationalism: "It would not be rational for a

The debate I am presently interested in is between agent-rationalists and agent-irrationalists, agent-moralists and agent-immoralists: philosophical opponents who all accept that threatening (nuclear) retaliation can be rational and moral where acting on the threats is not.

In this article I am mainly concerned to defend agent-rationalism about nuclear deterrence against its irrationalist critics. That is, my main goal is to show that we can coherently regard both of the following rationality claims as true: not only is the act of forming and maintaining deterrent conditional intentions perfectly rational in the nuclear circumstances envisaged, but in addition forming and maintaining such intentions is something that rational agents are fully capable of, despite their knowing that such intentions conditionally enjoin an irrational act. I thereby take myself to be defending nuclear deterrence against an important and persuasive philosophical attack on the character of those running the policy.

By implication, however, I will also be defending an agent-moralist view of nuclear deterrence and hence defending deterrence against another kind of attack on the character of those running the policy. For the moral case turns out to be similar and in some ways easier. Although there are conclusive reasons of a moral kind against applying a nuclear sanction should deterrence fail, I claim that broadly the same kind of argument can be used to show that a rational and moral agent is nonetheless able to form and have the relevant conditional intention to apply such a sanction. And nothing, as far as I can see, would restrict this conclusion very strongly to certain favored accounts of morality, such as some version of consequentialism. While agent-moralism is not the focus of this article, I hope to say enough to justify these claims.

II

Why suppose for a moment that rational agents cannot form and sustain such deterrent intentions? I can think of five more or less seductive arguments to this effect, some reconstructed from the literature on the topic, others independently plausible. All are based—directly or indirectly—on the content of the conditional intentions contemplated and on the implications for a rational agent who contemplates such intentions. Recall the problem. Because of what any such

---

person to execute an apocalyptic threat, even if she had reasonably expected that her life would go best were she to issue such a threat. Since *a rational agent cannot intend what she believes she will not have reason to do,* there are intentional structures that she is unable to erect, even though she would expect to benefit from erecting them" (ibid., p. 720; the italics are mine). The present article serves as a response to Gauthier's new position by challenging the italicized claim.

intention enjoins, we allegedly have a circumstance where an agent satisfies the following conditions:

> P: P1, the agent is (fully) rational; P2, she conditionally intends to do something E if a certain event C happens; P3, it is clear to her that if C should happen it would be irrational to do E.

This triad of conditions appears inconsistent, however, which suggests that no rational agent can have such a conditional intention in full knowledge of what it involves. But then neither, it seems, can a rational agent *form* such an intention in full knowledge of what it involves; deterrence can't even get started unless the deterring agent first becomes irrational.

Different agent-irrationalist arguments provide different ways of showing how the tension inherent in (P) argues for agent-irrationality. But before I begin my survey of these arguments, let me say a bit more about the idea of agent-rationality itself. The substance of my critique will be that, one way or another, agent-irrationalist arguments variously mislocate or misdescribe aspects of this idea.

What follows is supposed to be uncontentious. To describe an agent as rational is to characterize the agent as epistemically responsible: such an agent responds to evidence in the right sort of way, believing propositions when the evidence supports them (but at any rate not when it is cognitively unsafe to adopt such beliefs) and deciding how to act by taking proper account of her desires and beliefs regarding the likely outcome of actions. This is clearly a dispositional notion, for someone is correctly described as rational to the extent that she is disposed to function in this way, not just that perchance she always does function in this way. But note that the disposition is characterized in terms of a more local rationality: options open to a person have the property of being rational if they are supported by her evidence in the right sort of way or if they reflect her beliefs and desires in the right sort of way.

The proper characterization of this property is, of course, a contentious matter, with different theories defining the property in different ways. Thus, among theories of rational choice we have theories that recommend maximization, whether of evidential expected utility, causal expected utility, or some other agent-value, as well as theories that promote satisficing or some more extreme kind of suboptimizing.[14] In addition we have theories that explicitly allow only for instru-

14. Recent theories of rational choice include Robert Nozick's "maximization of decision-value" account (Nozick, chap. 2), where the decision-value of an act is the summed value of various kinds of expected utilities of the act, each value weighted by the agent's confidence in being guided by that utility. In stark contrast to all such maximizing or "optimizing" theories, Michael Slote presents a radical suboptimizing

mental rationality, as well as theories that allow also for a rational evaluation of agents' goals. (I noted earlier, in fact, that a noninstrumental theory may be just what we want if we count retaliation merely for the purpose of revenge as ipso facto irrational.) For my present purposes, however, there is no need to choose among these theories, so long as we are able to choose theories that declare retaliation, under the circumstances imagined, to be irrational, yet hold the policy of deterrence itself to be rational. What is far more important for my purposes is that all agree that rationality is first and foremost a property of the options available to an agent, a property that applies to an action in virtue of certain independently specifiable features it has or constraints it satisfies. Other notions can then be defined on the basis of option-rationality. Thus a decision might count as rational if (i) the option chosen is rational and perhaps if in addition (ii) the process used by the agent to arrive at her decision is reliably connected to the choice of rational options.[15] Agent-rationality is then understood simply as the disposition to make rational decisions. More precisely, an agent can be said to be (fully) rational if she not only invariably makes rational decisions but also is disposed to make rational decisions. Call this the "basic schema of (agent-) rationality."

I said that this is all supposed to be uncontentious. With such a generous basic schema, about the only resistance will come from those who think that rationality is at bottom a feature of agents rather than of options: an agent-based theory of rationality. But few have tried to develop this approach,[16] and in this article I put the suggestion aside. More, no doubt, might be said in favor of an agent-based theory where the moral case is concerned: consider, in particular, virtue theories of

---

theory of rational choice in his *Beyond Optimizing: A Study of Rational Choice* (Cambridge, Mass.: Harvard University Press, 1989).

15. Thus according to Nozick, "decision theory . . . must refer to the process or procedure by which the action is generated in order to be a theory of rationality" and, more generally, "the rationality of a belief or action is a matter of its responsiveness to the reasons for and against, and of the process by which those reasons are generated" (Nozick, pp. 65, 107). In my view, this is unnecessary: the inclusion of such a reliable process component rests on a confusion between the evaluation of the decision and the evaluation of whatever produced the decision (for a similar point in relation to epistemic rationality, see Richard Foley, *The Theory of Epistemic Rationality* [Cambridge, Mass.: Harvard University Press, 1987], pp. 199 ff.). But even if we grant the importance of a reliable process component in the characterization of rational decision, the manner in which option-rationality is the fundamental notion and agent-rationality is derivative remains unchanged.

16. Roy Sorensen comes close to such a view with his talk of "cognitive vices" ("Rationality as an Absolute Concept," *Philosophy* 66 [1991]: 473–86), but the main feature of his account is not the agent-centeredness of the approach but the fact that rationality gets defined in terms of its foil, irrationality, rather than the other way around.

ethics.[17] But I suspect that agent-based theories in general will be of little comfort where our present problem is concerned, for remember the full dimensions of that problem as we are now conceiving it: we are supposing that in the specified circumstances the act of conditionally intending to do E should C happen is morally as well as rationally right, even though doing E after C has happened is both immoral and irrational, and we are asking how this impacts on the moral and rational status of the agent involved. It is difficult to see how agent-based theories can deal with this problem, given that for agent-based theories the morality and rationality of options is derivative on the morality and rationality of agents.[18]

Once we restrict ourselves to option-based theories of morality, the moral version of problem P is structurally little different from the rational version. Unlike agent-based theories of morality, option-based theories make rightness fundamentally a property of options, whether in terms of an appeal to teleological features of options (maximizing the [expected] satisfaction of people's preferences, say, or some suboptimizing alternative) or in terms of an appeal to independently specifiable right-making features of another kind (their conformity to Rossian prima facie duties, for example).[19] As in the case of rationality, agent-morality will then be derivative. In claiming earlier that my argument in this article is generalizable to the moral case, I meant that claim to apply only to the case of appropriate option-based accounts of morality: option-based accounts that count the policy of deterrence as moral while counting the kind of retaliation contemplated if deterrence fails as immoral. All the agent-immoralists I know advocate some such option-based theory of morality.

III

Back to rationality. I am going to assume the basic schema of rationality for the remainder of this article, and I am going to assume reliance

17. Not all so-called virtue theories of ethics count as agent-based theories. Aristotelian or "eudaemonic" versions of virtue ethics don't, for example, since everything is based on flourishing rather than on traits of the agent. So far as I know, only Michael Slote is an avowed agent-theorist—but so far only about morality, not rationality. See his "Agent-Based Virtue Ethics," *Midwest Studies in Philosophy* 20 (1995), in press.

18. Michael Slote demurs (private communication). He suggests that the intention to retaliate may be moral because it exhibits resolute determination to avoid nuclear war and save human life (this is agent-based) and that the act of retaliation is immoral because it expresses wanton unconcern for human life (this too is agent-based). This looks promising, although what is still far from clear is how moral and rational agents could ever get themselves into such a resolute frame of mind, given the retaliatory means contemplated.

19. See W. D. Ross, *The Right and the Good* (Oxford: Oxford University Press, 1930). Ross's own account of morality is complicated by his admission of an independent notion of goodness that can characterize motives as well as states of affairs.

on a theory of rational choice that declares forming (and entertaining) the conditional threat to do E in the event of C to be rational in the specified circumstances, and the choice of doing E in the event of C irrational.

Let us now turn to the various agent-irrationalist arguments for the incoherence of (P). Our first agent-irrationalist argument is also the most direct. Armed with the basic schema, it imputes agent-irrationality to the agent who satisfies (P) by simply noting the dispositional nature of agent-rationality:

1. Rational agents are, by definition, disposed to act rationally; but if they have the intention to do E should C happen then they are disposed to act irrationally should C obtain. Hence rational agents cannot have such an intention.

But such a direct appeal to our basic schema of rationality is wrong on two counts. In the first place, the connection between intention and disposition is by no means clear-cut. An agent who sincerely intends to do E should condition C arise may well find that actually confronting C makes it hard to follow up on this intention: the real world may import features that engage the agent in ways she did not foresee, and could not reasonably have foreseen. It is not clear that this failure to foresee one's actual reactions need count as a defect of rationality.

But more important, even if we accept that the presence of the conditional intention makes for a certain disposition to act irrationally, the presence of this disposition may have little impact on the question of whether or not the agent is rational. The reason is that although the proper characterization of dispositional properties—fragility, braveness, and so on—requires us to consider situations that may never actually come to pass, it does not require us to consider *all* such situations. A glass may still be fragile even if a physicist can specify highly unusual situations S in which its chemistry would be dramatically altered once we strike or drop the glass, thus preventing it from shattering. The crucial thing is that in relevant circumstances the glass should behave in characteristic ways, and being dropped or struck in situations of kind S is simply not relevant. For precisely the same reason, an agent may be perfectly rational even if in various altered circumstances the same agent would *not* have behaved rationally.

This is not to deny the importance of altered circumstances in the characterization of rationality; clearly, just to choose rational options is not enough to make an agent rational, for sheer chance might allow her to avoid situations where she would have acted irrationally. But if the altered circumstances where she would have behaved irrationally are ones that involve an unplanned and unwanted diminution of her cognitive capacities, a diminution for which she is not responsible and

which she tries hard to avoid (e.g., having Alzheimer's disease), then it seems that her status as a rational agent would not be affected. What makes any of us far from ideally rational agents is not that we *would* behave irrationally if we ever were affected by this kind of cognitive malfunction; the truth of that conditional shows only that we *would* be irrational agents in such a circumstance, not that we actually *are* irrational agents. So the disposition-invoking argument (1) by itself does not provide good grounds for indicting our deterrer's status as a rational agent.[20]

Nor, clearly, do we have good grounds for indicting her status as a rational and moral agent if argument (1) is rephrased in moral terms. An agent who is moral as well as rational is disposed to do what is right—but not just in any counterfactual situation. Like the other arguments to be looked at, the moral version of (1) is no more convincing than the rational version.

The failure of argument (1) seems to rest on its inability to exploit a special feature of an agent who satisfies (P): the fact that the agent, as a rational agent, is aware of the irrationality of doing E even before C ever happens, while nonetheless being committed to doing E in that situation (it is not simply that she *will* be aware should C ever happen). So let us now consider some arguments that take up this reflective aspect, although in rather different ways.

Perhaps the simplest way to capture this reflective aspect is to focus on the sheer deliberateness that marks the agent's being in both states (P2) and (P3). In that way we get something like argument (2):

> 2. Even if a rational agent might act irrationally should a certain condition C obtain, there is no sense in which a rational agent could deliberately arrange to act irrationally should C obtain. Since this is just what having the contemplated intention achieves, a rational agent couldn't have the intention to do E should C happen.

The point is that our deterrer has a deliberate plan to act irrationally should C occur. We are not talking of someone who, through

20. What the agent-irrationalist further needs to show is that the occurrence of C is a relevant circumstance in a way that suffering from Alzheimer's is not. But what would show this? Note that these two altered circumstances have at least this much in common: it is through design—the strategy of deterrence—rather than good luck that a deterrer doesn't end up facing C, and so the altered circumstance in which C happens and she does the irrational E as a result of having the conditional intention is more akin to a circumstance she couldn't help being in and tries everything in her power to avoid—like suffering from Alzheimer's disease— than to a circumstance that she avoids being in through sheer good luck. (For a rather more skeptical view of the idea of "relevant" altered circumstances and the link between dispositions and behavior in such circumstances, see C. B. Martin's "Dispositions and Conditionals," *Philosophical Quarterly*

no design of her own, would find herself acting irrationally if she were to succumb to a certain cognitive disease. But how can an agent with such a plan be fully rational?

This second argument, however, seems as flawed as the first. For consider a person—Smith—who is otherwise perfectly rational. Smith knows that she would lose all powers of rational deliberation and would begin to exhibit pathological behavior if (and only if) a certain undesirable event U were to occur; furthermore, she knows that it is only the foreseen consequences of such behavior that deters those who threaten her with U (imagine that a certain part of her brain would be affected if U were to occur: she would become psychotic, with disastrous consequences for those who imposed U). Suppose also that Smith is aware of techniques that might rid her of these pathological propensities, but that she chooses not to be cured; these propensities are just too valuable to her. To this extent, she is someone who deliberately arranges to act irrationally should certain conditions obtain.

None of this, however, seems a good reason for thinking that Smith is less than fully rational. Perhaps human beings are all biologically constituted to display this sort of pathological behavior should U occur. As in the Alzheimer's example, however, that seems a bad reason for thinking that, as things stand, none of us are fully rational (it is only a reason for accepting the conditional claim that if U *were* to occur we *would* not be fully rational). Now Smith has the further advantage that she knows that she is so constituted, and, applying this knowledge wisely, she then decides not to change this part of her constitution. But this can scarcely make a difference to our assessment of Smith; if anything, her ability to turn this dispositional feature to her own advantage merely confirms her rationality.

Still, in this case we can scarcely say that the agent conditionally intends to adopt the pathological behavior. She sees the behavior as pathological, and so not as a sequence of actions she can imagine herself from the inside as initiating. As such, the case is not a good model of what is supposed to take place in the kind of deliberative strategic reasoning that underlies deterrence. In short, what is missing in argument (2) is sensitivity to the dynamics of rational choice, to the way a rational agent forms and justifies her choices.

That suggests looking more directly at the way in which the rational agent of (P) must have formed the conditional intention to do E if C should happen. One thought is that the triad (P) is inconsistent because there is a tension between the way the agent forms the inten-

---

44 [1994]: 1–8. David Lewis suggests that Martin's skepticism may be relevant to my argument, but I won't explore this issue here.)

tion and the reflective realization that it is irrational to do E if C should happen. And on one conception of what is involved, this tension is blatant and immediate:

> 3. A rational agent's conditionally intending to undertake some action, say X, should C happen must depend on her recognizing that X is the rational option should C happen. Hence, contra (P), a rational agent cannot conditionally intend to do E should C happen, since she sees that doing E should C happen is not the rational option.[21]

Argument (3) posits the tension in (P) as a simple consequence of what is involved in forming and justifying a (conditional) choice. But it thereby begs the question at a crucial point. It assumes that a rational agent can form conditional intentions only by using the following kind of matching deliberative process: form the intention to do X should C happen (if and) only if doing X would be rational in the event of C's happening. But why grant this assumption? The only reason I can think of rests on a certain model of how decision theory is to be applied in ordinary nonconditional cases. On this reading, the assumption that the conditional attractiveness of doing X is to be analyzed in terms of the agent's reflective assessment of X as conditionally rational is just a natural extension of the claim that the *un*conditional attractiveness of doing X is to be analyzed in terms of the agent's reflective assessment of X as *un*conditionally rational.

But if that is what lies behind argument (3), we have every reason to be suspicious. For in its general, unconditional form this gives the wrong picture of rational choice. In general it is not, and it certainly need not be, the case that rational agents choose by determining reflectively that their chosen option fits the demands of some canonical decision theory, where among other things this involves explicitly identifying one's desires as desires: items whose satisfaction counts in a way determined by the theory. All that rational decision theory demands is that the choices an agent makes systematically match the conclusions of whatever account of rationality is chosen as canonical. Rational decision theory need not in addition function as a kind of decision procedure.

So a general defense of strategy (3) fails if it is conducted in terms of the procedures rational agents must follow if they are to conform

21. Kavka seems almost to embrace something like argument (3) when he says that "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" ("Some Paradoxes of Deterrence," p. 292). The words "the agent's *view of reasons* for doing the act" (my italics) sound uncomfortably close to the kind of reflective account I reject in the text. Solution (5) offers a different way of understanding what Kavka says.

to rational decision theory. Perhaps it is possible for rational agents to conform to rational decision theory by reflectively following its dictates in this way, but it surely isn't necessary. (And if it isn't necessary, then it may sometimes not even be possible, in part because it might get in the way of decisions rational agents rationally ought to take, such as forming deterrent intentions.) Absent other defenses, therefore, there seems no reason to grant the grounding assumption of argument (3) that the way to decide on conditional intentions can only be in terms of the deliberative process described. This is a point that will be reinforced below when we look at other ways of understanding how conditional intentions might be formed.

Here is another, rather different, way of confronting the thought that a rational agent recognizes the irrationality of what she considers conditionally intending, thus rendering her incapable of forming or having the intention. Rational agents no doubt think of themselves as rational agents, and rightly so. But how, in that case, can they also think of themselves as agents who act irrationally in certain situations?

4. Agents who are fully rational (over time) are entitled to think of themselves as rational. But if so, they are entitled to think of themselves as agents who would choose the rational option should C ever happen. But as rational agents they also know that to choose rationally in that situation is to choose not-E. Hence they are entitled to believe that they would not do E should C happen. But that means they cannot seriously hold the conditional intention to do E should C happen: that would be a kind of pragmatic inconsistency.

Note that the starting premise—call it "entitlement," for short—is not something which only a straw man would propose. Future planning, for example, often depends on predicting how one would behave in various situations, and that might mean predicting that one would make rational choices in those situations. Thus I might predict that I would act in a rationally appropriate way were I to sit a certain exam, and on that basis I might argue in favor of having a party after the exam. Relying on such an argument is to presuppose entitlement to one's rationality, at least where choices of a certain kind are concerned.

Now it can hardly be denied that ordinary agents are often entitled to assume in this way that they are by and large rational, whether the source of the entitlement is broadly inductive (an agent may recognize her own track record) or broadly a priori (an agent's assumption of her own rationality may be part of a self-applied methodological injunction to interpret people as by and large rational). Any such imputation of entitlement may therefore seem even more soundly based

in the case of fully rational agents, the sort of agent appealed to in argument (4).[22]

But appearances here are deceptive. Whether an agent is entitled to assume that she would be rational in this or that choice-situation is bound to depend on the contingencies of the situation: some things like Alzheimer's disease can make people effectively incapable of rational action, and an agent's awareness that she would act irrationally under such an affliction shouldn't count against her rationality (recall the response to argument [1]). A rational agent might even arrange things so that she knows she would be incapable of rational action should some other event occur (recall the case of Smith in the response to argument [2]). Cases like this seem perfectly good counterexamples to Entitlement. If the occurrence of C is thought to be crucially different, with the rational agent being fully entitled to believe that she would then act rationally, we need to know why the cases are so different. Given the nature of C—destruction of much of the agent's natural and social world—is it so clear that the agent is entitled to think that her rational nature would be preserved intact? If our agent-irrationalist critic continues to insist that this is indeed clear, then we can fairly accuse her of begging the question against the agent-rationalist, since for the agent-rationalist our deterring agent sees herself as liable to act *ir*rationally should C ever occur.[23]

All this is not to say that there can't be otherwise rational agents who see that they won't do E if C should happen; perhaps some agents are secure enough in their knowledge of their motivations to know that they won't ever be able to bring themselves to do E if C should happen. But why describe such agents as fully rational if this knowledge then prevents them from entertaining the sort of deterrent intentions that they might well desperately need, given some of their most basic interests?

IV

Both arguments (3) and (4) aim to establish agent-irrationality by letting the agent reflectively focus on what is involved in rational

22. Note that the appeal of argument (4) to Entitlement is suitably weak: it doesn't state that a rational agent *knows* that she is rational, only that she is entitled to believe that she is. It doesn't even require her to be entitled to assume her rationality over all decisions she may face, but only a selected range of such decisions.

23. Even more problematic would be arguing from a rational agent's being entitled to believe that she would choose rationally should C happen to the claim that this gives her a reason against forming the intention to do E should C happen. That would be a particularly heinous example of a fallacy of circular reasoning, since it presents her as deciding how she should comport herself in the event of C (for clearly, whether to form the intention to do E in the event of C is relevant to that question) by assuming

choice: by allowing the agent to "foreground" the fact that a (conditional) option is or is not rational, or that she, the agent, is herself rational.[24] It is time to look at a natural alternative, one that "backgrounds" any appeal to rationality and lets choices count as rational so long as they accord with the demands of some canonical account of rationality: so long, in particular, as the agent takes account of her beliefs and desires in the right sort of way, without being required to identify this way explicitly as the canonically rational way to deal with such beliefs and desires. It then becomes tempting to adopt something like the following view of conditional intentions. What makes it the case that a rational agent holds the conditional intention to do something X should C happen is that when such an agent considers a scenario in which C does happen, with a view to determining what to do in that (imagined) situation, she chooses option X. But for a rational agent conducting such a thought-experiment, the X in question clearly can't be E: as a rational agent, she will be attracted by the weight of reasons against choosing E.

More precisely: let us say that an agent has deliberative integrity with respect to a possible circumstance S if in deliberations directed at what to do should S obtain the agent identifies with, and argues from, presently held commitments (both desires and beliefs). Then we might say that our triad of conditions (P) is incoherent because in forming the conditional intention to do E in the event of C a rational agent must have deliberative integrity with respect to C, an integrity centering on rationally held commitments. (In this sense, forming and sustaining the conditional intention to do E requires much more than merely believing, knowing, or even bringing it about, that if C should happen one would do E.)

On the present diagnosis, the charge of inconsistency facing (P) can be put as follows:

5. A rational agent can only intend to do E should C happen if in conditionally choosing what to do on the assumption that C does occur she chooses E on the basis of presently held commitments and so exhibits deliberative integrity. It follows that she can't intend to do E should C happen, since to choose E conditionally on the assumption that C occurs would be to choose against the balance of reasons that as a rational agent she identifies with, and so would show a lack of deliberative integrity.

---

that she will comport herself in a particular way. In "Rationality and Epistemic Paradox" (*Synthese* 94 [1993]: 377–408) I offer a general critique of such circular ways of assuming one's rationality.

24. See Philip Pettit and Michael Smith, "Backgrounding Desire," *Philosophical Review* 99 (1990): 565–92.

The idea is simple—beguilingly so. In conditionally choosing, a rational agent goes through some such reasoning as this: "Suppose C has happened. Then it will be of no (further) use to retaliate in kind, for this will just add to the misery, with no compensating benefits for anyone, including me. Hence I won't do E." Here the agent shows in her reasoning that she identifies with certain kinds of reasons that as a rational agent she also identifies with in her *non*conditional choices; after all, a rational agent confronting C surely must reason in the following sort of way: "C has happened. Now it will be of no (further) use to retaliate in kind, for this will just add to the misery, with no compensating benefits for anyone. Hence I won't do E."

Note the difference between this and argument (3): the agent is here going through a bit of (conditional) reasoning about whether to do E; she is not just identifying what, nonconditionally, *would* be the rational option for her to take before deciding what to do conditionally. And note the difference with argument (2): the rational agent of argument (5) is an agent who constructs her conditional intention by planning in terms of presently held commitments and thereby exhibiting deliberative integrity; by contrast, the requirements of argument (2) are already met by someone like our agent Smith, who arranged to be a certain sort of person should U ever occur while being unable to identify with the commitments that person might then have.

Kavka's own argument in "Some Paradoxes of Deterrence" also suggests something like (5). (So do the arguments of various other writers, including those who concentrate on the moral case and then talk about the way deterrers must somehow morally embrace the intended act and what it involves.) According to Kavka, "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."[25] That is what explains the tension in (P), for "suppose . . . that the agent 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, [this is] because the reasons for applying it motivate him more strongly than do the conclusive reasons not to apply it, [and so] . . . he is irrational."[26]

Argument (5) appears to escape all the difficulties in the other arguments; it offers a seemingly attractive account of the internal constraints that face a rational agent when deciding how she should react should C happen, and in doing so it provides what seems by far the most plausible diagnosis of the apparent tension in (P).

I believe, nonetheless, that we should reject the argument. It holds that the agent who conditionally chooses E deliberates in a way that

25. Kavka, "Some Paradoxes of Deterrence," p. 292.
26. Ibid.

a genuinely rational agent cannot identify with: she conditionally chooses for reasons that have no hold on such an agent, that the agent cannot see as attractive once she understands what is involved. But if this is what lack of deliberative integrity comes down to, then the argument operates by excluding some important components of our rational makeup. In particular, it leaves out the ineluctable role that emotion plays in our lives and to that extent works with an impoverished notion of deliberative integrity.

Thus consider the emotion of anger. A person's anger might be behind her decision to perform a certain action, where the action she undertakes is the rational option on independent grounds. So anger is the sort of emotion that can "make" an agent do the rational as well as the irrational. It is only if the agent is "overcome" by anger, and as a result is blind to rational-making features of options, that we can fairly accuse the agent of acting irrationally. A similar point can be made about the case of agent-rationality. The fact that an agent sometimes acts out of anger, or even that she tends to act out of anger in certain types of situations, is not sufficient reason for thinking she is less than fully rational, for the situations that provoke her to act out of anger may well be ones where her actions continue to be rational but where anger is the appropriate response.

There is, of course, a contrary perception which sees the rational agent as inevitably calm and aloof, subject to the coldly calculative exercise of reason, and the angry agent as inevitably irrational because she is subject to quite another, irruptive, sort of motivation; but this contrary perception comes from a tradition that is now generally rejected.[27] Indeed, it is difficult to imagine rational creatures who have no emotional life at all. For on the usual decision theories, rational choice is choice that looks at the satisfaction of an agent's desires in light of her beliefs, whatever counts as an appropriate level of satisfaction and whatever else is involved. But desires impact on our emotions in at least two ways. First, many of our desires can be characterized only in emotion-attributing terms, and so too, therefore, must our tendencies to rational behavior: thus we may act out of love for a person, yet behave rationally to the extent that our action satisfies our desire for our loved one's well-being in light of our beliefs. Second, if a rational agent deems a certain choice of action the appropriate one to undertake, given her most fundamental desires, then she is not likely to take a neutral stance toward a contrary action on the part of another agent that debases these desires. Not only are resentment and

27. One radical criticism of the tradition came from Robert Solomon, whose *The Passions: The Myth and Nature of Human Emotions* (New York: Doubleday, 1976) argued that emotions were just judgments. But as it stands this view is clearly implausible, for one can make judgments without experiencing the corresponding emotion.

anger not irrational in isolation; they may even, in a sense, be required emotions for rational agents if rational agents are to identify in the right sort of way with their desires.

How does all this bear on the idea of deliberative integrity, which looms so large in (5)? My answer is as follows. The idea of deliberative integrity is a device that mediates between merely conditional deliberation and nonconditional deliberation: invoking it seems to allow one to argue that the conditional reasoning a rational agent goes through when she contemplates C must mirror the reasoning such an agent *would* go through if C were to happen. But focusing on this idea underestimates the strategies available to a rational agent by forgetting about her emotional makeup and the knowledge others have of this emotional makeup.

Here is an alternative way of understanding what goes on when an agent entertains deterrent intentions. She entertains the antecedent condition C and (predictably) finds that the thought of C happening—especially the thought of C happening after all she has done to show how seriously she cares about C *not* happening—engages her emotions in a certain way: in the grip of the thought of C happening, she finds that she doesn't particularly care to do what is in her own, and others', best interests, but is emotionally inclined to want to exact revenge. (In the nuclear scenario, the agent is a designated subgroup that has the interests of the nation at heart: it is the [thought of the] virtual destruction of one's nation that then excites the desire to avenge what has happened.) From the conditionally adopted perspective of the state of being thus affected and assuming no other barriers to her acting on the basis of this affective state, the agent now finds it all too easy to decide on doing what hurts her attackers most; she thus decides on E even though she is aware that so to choose were C really to happen would be to choose irrationally.

Let me stress again that we are here talking not of the agent's actually choosing to do E, given that C has happened; for after all, C has not yet happened and, we hope, will never happen. We are talking only of the agent's imaginative preconstruction of her choice, on the assumption that C has happened and given presently held commitments as well as the emotional coloring which her attachment to these commitments brings to her deliberations.[28] If the agent is truly rational and C does actually happen, then the agent will choose against doing

28. Note that I am here talking of an emotional reaction that guides my conditional choice but takes place without my believing that the antecedent condition C has actually occurred; I only imagine that C has occurred. Patricia Greenspan, in *Emotions and Reasons* (New York: Routledge, 1988), pp. 17–20, and Michael Stocker, in "Emotional Thoughts," *American Philosophical Quarterly* 24 (1987): 59–69, discuss some other cases of emotional states that involve entertained thoughts rather than actual beliefs.

E since we have agreed that to choose E is to choose irrationally. But in the conditional choice-situation described C hasn't happened, and the agent only chooses E in the scope of imagining that C has happened.

Of course, to the extent that she believes that her imaginative preconstruction accurately prefigures what she would do were C to happen, the agent finds herself believing that she would act irrationally were C to happen. But it is important to note that she doesn't thereby see this future self as a person whose commitments she can't identify with, a person who is as foreign to her as Smith's psychotic alter ego must seem to Smith in the response to argument (2). On the contrary, in making her conditional choice to do E in the event of C the agent displays deliberative integrity in so far as she makes the choice in the light of her own commitments and certain natural emotional propensities. That is how she is able to see her conditional choice as prefiguring how she would choose were C actually to happen. Equally important, she sees that her opponents realize this too, recognizing as they do how rooted such a conditional choice is in the agent's commitments and propensities.

Now to conclude the story. Talk of letting an agent's imaginative preconstruction be a guide to how she would behave presupposes at the very least that the agent will have the wherewithal to behave in this way. In the case of nuclear deterrence, this means having enough nuclear weapons; in other cases it means having enough of something else. But how can a rational agent take the step of acquiring this wherewithal, or not disposing of it if she already has it? For she accepts, we agreed, that once she has the wherewithal (and maximal opportunity for the use thereof), her imaginative preconstruction becomes a fair guide to what she would do if C should occur, namely E—yet she sees that it becomes pointless, irrational, to do E once C has happened.

The rest of the story is familiar but worth retelling from the vantage point of the present proposal. The agent sees that her having the wherewithal has deterrent value since her opponents are clearly able to see that once she has it the vengeful but all-too-human scenario in which she responds to C by doing E becomes more or less likely, and she knows that her opponents above all do not want E to happen. That is why she acquires the wherewithal for doing E, or doesn't dispose of it if she already has it. And that is why she sets in place a procedure that makes it easy enough to do E (it becomes too difficult if the wherewithal is too difficult to access, for example, or if a decision to do E is subject to extensive review or statutory delays). Given the need for deterrence, she has acted rationally in thus firming up the guiding potential of her imaginative preconstruction, even though she

is aware that her doing so increases the risk of her acting irrationally in the future because of the risk that deterrence will fail.[29]

In short, deterrence is a strategic option available to rational agents to the extent that rational agents are susceptible to human emotions, including anger and the desire for revenge in the face of great insults to individual or group interests or values. It is a susceptibility they recognize in themselves, and they recognize it as something which in effect uses them to their own advantage. For they understand this about deterrence: it works to the extent that others—potential opponents—see that even rational agents are susceptible to the desire for revenge in the case of great insults to their interests and values. In a sense, therefore, rational deterrence rests on both the agent's and her opponent's perception of the fragility of agent-rationality: the fact that rationality is no ironclad disposition but a disposition that even perfectly rational agents can readily imagine losing should circumstances go against them.

For the same reason, deterrence is a strategic option available to morally good agents. Even if such agents recognize that retaliation is immoral, they also recognize that they are naturally susceptible to emotions like anger and the desire for revenge in the face of great insults to interests or values, and they recognize the way this susceptibility is able to use them to their own advantage. Deterrent conditional intentions are mind-sets that reflect this recognition. Agent-moralism is thus as viable a position as agent-rationalism.

V

Now for some objections to this way of understanding agent-rationalism (and, implicitly, agent-moralism).

i) Shouldn't a truly rational agent be credited with the ability somehow to override her strong desire for revenge displayed in her reaction to the thought of C occurring? And so shouldn't the (conditional) deliberations of a truly rational agent be free of the distortions produced by such emotions?

Reply: Even if the answer to the first question is yes (although I am doubtful), the answer to the second should be a clear no. For having the ability doesn't provide the agent with a clear reason to exercise the ability unless it is in the agent's overall interests to do so. To think that it does is again to fall victim to the mistaken view that a truly rational agent can't be guided by emotion. In terms of a more

29. In the case of real-world deterrence, of course, this risk may well be enough to cast doubt on the rational viability of a policy of deterrence (for a discussion of relevant issues, see Lee, chap. 7).

concrete example, recall once again our agent Smith who sees that she would become psychotic if some unwelcome event U were to happen, but who recognizes that she is well off with this possibility since it deters those intent on U. A rational agent would not choose to change the situation even if it meant risking irrational behavior.

ii) Why should the agent's choices in her imaginative preconstruction be a fair guide to what she would really be like in the event of C happening? Only think of the way most of us react when watching movies featuring injustice and violence; the sort of bravado we easily generate in ourselves is likely to be a poor guide to the way we would behave in the real world when confronted by similar situations. Why shouldn't fully rational agents be even more skeptical of the correspondence between imaginative preconstruction and the real world, realizing as they do that actual retaliation would be irrational?

Reply: This is an important point that underscores a real difficulty for effective, credible deterrence. People often have an exaggerated impression of how they will behave in certain circumstances. This is surely also a problem for otherwise rational deterrers. Thus consider the goal of deterring an attack on some friend or ally. Because of a certain distance between one's own interests and the interests of the friend, the sort of vengeful feelings generated in the course of one's imaginative preconstruction of the attack need not be a reliable guide to how one would really react—"in the cool light of reason," as it were. If so, a rational agent would probably be aware of this, as would her opponent, and to this extent the deterrent she tries to exercise would not be fully credible. (That was always the problem with the American attempt to deter a nuclear attack on Western Europe, which involved a move from a policy of continental defense to one of extended deterrence.)

But whatever the truth about this sort of case, note that the problem of an imperfect overlap of interests doesn't exist in the cases we are presently considering, at least not to the same degree. In these cases, the potential attack on the deterrer and her group represents an insult to her deepest interests and values, not to interests that only marginally intersect with hers: hence the depth and quality of the agent's anger in the scope of her imaginative preconstruction. This is what turns her anger and vengefulness into a more or less reliable guide to what she would do in the event of an actual attack, especially in virtue of the way she is also able to motivate the removal of effective barriers to her exercising this desire for revenge (in the interests of turning her vengeful stance into a credible deterrent).[30]

30. I say "a *more or less* reliable guide," for on the strategic policy the agent adopts, we are not warranted in saying more. In particular, the agent is not relying on some

iii) Deterrence might fail: C might after all happen. In that case, the agent's conditional intention to do E should C happen means that the agent will end up doing E, and hence will act irrationally. So the agent will have proved herself to be a less than perfectly rational agent after all, even though motivationally she may be the twin of the successful deterrer whom the above account is prepared to count as perfectly rational: equally clear-sighted, and so on. How can this be?[31]

Reply: First of all, recall the gap between intention and execution. Even if C were to happen despite the deterrent effect of her intention, it does not yet follow that the agent would actually do E, for in the actual world C might have features that even the best-informed imaginative preconstruction simply failed to address (think of the degree to which we still remain almost totally ignorant of the full range of effects of the use of nuclear weapons; it can scarcely be a demand on rationality that we overcome this, for perhaps only the actual use of nuclear weapons would yield the desired information).

This consideration aside, the short answer to objection (iii) is simply that agents may be rationally "unlucky": the world—or at any rate not something of the agent's own making—may conspire to turn an agent who would otherwise be classed as fully rational into someone less than fully rational. That is the fate of the agent for whom deterrence fails. Her twin does not share this fate and so remains fully rational. The contrary thought that rationality must be immune to all forms of luck is just as wrong, I suspect, as the corresponding thought that morality is so immune.[32]

---

inflexible doomsday machine that would remove the deliberative aspect of her strategy. Hence she can't be utterly certain how she would in fact react, especially if retaliation works against her prudential interests. (Recall that on one of our scenarios, retaliation is irrational because it makes her somewhat worse off; if retaliation is only irrational on an ideal, noninstrumental conception of rationality, or if it is immoral without being irrational, her vengefulness may well be a much more reliable guide to how she would in fact react.) Note, however, that this absence of certainty doesn't show that the agent can't then be said to entertain the conditional intention to retaliate, for there is no (pragmatic) inconsistency in the thought that one might entertain a conditional intention, while not being completely certain that one will bring it off—surely a common enough feature of standard intentions, even among otherwise rational agents.

31. There is a trace of this argument in Farrell, pp. 125–26. Farrell asks us to envisage an agent who makes such a commitment, where that commitment persists undiminished up to the time at which the supposed action is to be performed; but then, he comments, the action will be performed, supposing no other changes have occurred. The answer I give in the article is this: so much the worse for the agent. Rational luck is on the side of the rational agent who makes the commitment but doesn't need to perform the irrational action.

32. See the symposium articles by Bernard Williams and Thomas Nagel, "Moral Luck," in *Proceedings of the Aristotelian Society*, suppl. L (1976): 115–35, 137–51. Foley, pp. 199 ff., defends the role of epistemic luck in the case of the rationality of belief. Of

iv) Even if deterrence doesn't actually fail, isn't the mere fact that the agent is of a mind to exact irrational revenge for such an insult to her fundamental interests bound to percolate through elsewhere, and show the agent up as irrational in some more categorical way?

Reply: Here we should ask the objector to provide concrete, convincing evidence. Let me quickly dispose of some attempts along these lines. To begin with, there is no evidence so far that our agent is pathologically inclined to anger or revenge, allowing these emotions to "take over" in some of her nonconditional choices (remember that in talking about the agent's attitude to C we are talking of her attitude to inordinate rather than everyday insults to her interests, insults that she imagines taking place in the face of the clearest possible evidence of her passionate commitment to these interests). More specifically, there is no conclusive evidence to show that the impact of emotions like anger, vengefulness, or even love must make for global rationality but local irrationality, in the kind of way made famous by Robert Frank.[33] That surely depends on how the agent deals with these irruptive motivations in her decision making, just as in a situation where deterrence has failed, the question of whether the deterring agent can continue to be seen as rational depends on whether she does in fact retaliate.

There is also no reason to think that an agent who conditionally chooses E in her imaginative preconstruction must suffer from other kinds of cognitive defects (an inability to do simple math, for example). In arguing that an all-too-human yet rational agent is thus able to choose E we have focused on the nature of both C and response E; we haven't simply argued that our deterrer is someone who can simply choose whatever will deter. Something like the vengeful E is in some ways the natural, all-too-human response. By contrast, an agent who can seriously contemplate adding two plus two and getting five, if this is what it takes to deter, is irrational in a way not touched by this position.

v) But surely the vengeful mind-set of the agent in question is not the mind-set of someone who rationally intends to do E if C should happen; it doesn't seem rooted in the right sort of way in the agent's

---

course, in rejecting immunity to luck I am once again putting aside agent-based theories of rationality which might be expected to make room for such immunity.

33. See Robert Frank, *Passions within Reason: The Strategic Role of the Emotions* (New York: Norton, 1988). I was reminded of Frank's work only after finishing this article and so haven't considered his argument in more detail. One point to note, however, is that Frank's claim that the emotions tend to lead to local irrationality may be true only on certain very narrow accounts of self-interested rationality. What counts as (locally) irrational on such a view may well be entirely rational on a more satisfactory account of rationality.

rationally held goals, which include the desire that E not happen. What we have instead is the intentional state of an agent who, for reasons rooted in rationally held goals, has managed to undergo a process of self-corruption that makes her desire that E happen should C happen. In other words, the position described is essentially just a relabeling of Kavka's position.

Reply: But what then distinguishes rationally intending from other states? Earlier, in connection with argument (5), we saw that for Kavka, "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."[34] But this seems too restricted even for the case of nonconditional intending. Perhaps an agent who only forms intentions when the thing intended follows her "view of reasons" for doing that thing would be too irresolute a character (she might have too few reasons for doing things). Suppose she wants to counteract this trait. One way—perhaps the hard way—is for her to find more reasons for doing things. The other way is for her to form intentions when there are no strong reasons for or against doing the thing intended (thus I might firmly intend to do twenty push-ups a day, for example, just for the resolve this shows and not because my health demands the push-ups). This is a strategy that is surely available to rational agents; it doesn't require self-corruption or the services of a hypnotist.[35]

Conditional intentions, so I have argued, show even more scope for such indirect motivation. In their case there is a further factor—the impact which the contemplation of the antecedent condition has on the agent.[36] If the agent deliberately cultivates her natural emotional reaction to the thought of C occurring, allowing it to use her in the way described for a rationally held goal that she cares about passionately, why shouldn't we say that this is another way in which an agent can rationally intend to do E if C should happen?

34. Kavka, "Some Paradoxes of Deterrence," p. 292.

35. David Copp discusses some intriguing examples of this sort in his "Irrational Deterrence or Rational Retaliation" (unpublished manuscript). Kavka might have thought he had a reply: we need to add something like "*or* the agent's view of reasons for forming the intention thereby comes to justify doing the act," which leaves retaliatory intentions out in the cold. But what warrants this second condition?

36. Another agent-irrationalist argument that (wrongly) forgets about this difference between nonconditional and conditional intentions is due to Daniel Farrell, who writes: "The same feature that would make us say that someone who had actually performed an admittedly irrational act was thereby exhibiting that she was less than rational—namely, her willingness to do what she knew to be irrational—would also be present and enable us to say the same of someone who merely *intended* to do what she knew would be irrational" (Farrell, p. 124). Farrell takes his argument to be a better argument than Kavka's, which he regards as "rather obscure, . . . and much too facile" (p. 123). I suspect that both make the same sort of simplifying mistake.

The second part of the objection claims that this line of reasoning trades on a confusion: although the agent's mind-set rests on certain rationally held goals, this is only true to the extent that these goals support a kind of self-corruption that makes the agent less than fully rational when in the grip of the mind-set. But is this really so? Surely an otherwise rational agent who has undergone a process of self-corruption has some goals or beliefs a truly rational agent can't totally identify with. And our agent is not like that. Her earlier rationally held goals are in no sense compromised by her new mind-set. As we portrayed the agent, she uses the fact that, rationally, retaliation serves no goals of hers (and even subverts some of those goals) to remind her opponent that the vengefulness embodied in her threat to retaliate captures her strong emotional reaction to the thought of her opponents' setting aside her most fundamental interests in the face of clear signs—including her threat—of her commitment to those interests. That is how her threat has credibility; it threatens emotionally intelligible but irrational behavior. But her threat does not show that she now also has a different goal, namely, a desire to do E in the face of events like C. That radically misdescribes the nature and function of the threat, for the agent's stable preference is that her actual behavior should continue to serve her genuine ends even if deterrence fails: she accepts the stupidity of retaliating. Opponents who impose C despite the threat are therefore betting on the agent's keeping these ends firmly in view and remaining rational, a bet that seems ill-advised given the fragile nature of rationality.

VI

I have been concerned in this article to defend the proponents of deterrence against an attack from those—I termed them "agent-irrationalists"—who think that no rational agent could have the required deterrent intentions, even if it is rational to have them because of their deterrent value. In my view, none of the agent-irrationalist arguments canvassed above succeeds in showing this. I have argued, in fact, that the agent-irrationalist position is itself wrong: there is an attractive position according to which deterrent intentions are states that exploit the fragility of rationality without being any less available to rational agents (the same goes for morality).

But suppose I am wrong about this. Suppose instead that some agent-irrationalist argument like (5) proves in the end to be sound, despite the doubts I have expressed. Then agent-irrationalists like Kavka will have been proved right after all. But in the end I doubt that much hinges on this concession, for it doesn't yet show that the proponents of deterrence whose views we began with are wrong. That would follow if the two positions were contraries, but denying that the vengeful mind-set identified above is a case of rational intending

suggests that there are other ways of understanding the deterrent threats on which proponents of deterrence commonly base their doctrine. (Indeed, in my experience proponents of deterrence rarely use the language of conditional deterrent intentions unless they are philosophers; they tend instead to use the language of serious deterrent threats.) Briefly: we can regard such vengeful mind-sets as providing a plausible interpretation of serious deterrent threats even if—as I am now supposing—these mind-sets do not involve rational intentions. Philosophers' talk of conditional deterrent intentions can then be seen as another interpretation, but one that creates trouble for agent-rationalism.

Note that on this alternative nonintentional interpretation, deterrent threats have some of the features of what Schelling calls "the threat that leaves something to chance," that things may just get out of control because of such things as "the role of emotions and misinformation on the leader's decisions or a breakdown in the chain of command."[37] For if C does happen, then on the present picture the agent may well retaliate vengefully without this being something she rationally intends. But note that on my account deterrent threats continue to differ substantially from "the threat that leaves something to chance," for they also involve an important element of planning through their reliance on imaginative preconstructions. While our vengeful mind-sets may not be cases of full-blooded rational intending, they are still cases of a kind of planned commitment whose presence is designed to deter, and so are not just cases of the "threat that leaves something to chance."

In short, even if the vengeful mind-sets described in this article do not involve genuine rational intending, they may well provide a good enough interpretation of deterrent threats to suit the agent-rationalist; good enough because it allows the threats to be both serious and effective, allows the agent a sort of deliberative integrity, and allows the agent her rationality.

37. Thomas Schelling, *The Strategy of Conflict,* new ed. (Cambridge, Mass.: Harvard University Press, 1980), p. 188; and Lee, p. 243. For a discussion of such threats, see Lee, pp. 242 ff.