PHILOSOPHICAL TOPICS
VOL. 21 NO. 1, SPRING 1993

Utility-maximizing Intentions and the Theory of Rational Choice

Daniel M. Farrell
Ohio State University

Imagine yourself in the situation described in Gregory Kavka's famous Toxin Puzzle: an eccentric billionaire has guaranteed you a million dollars, no strings attached, if, at midnight tonight, you intend to drink, at noon tomorrow, a mildly toxic fluid that will make you quite ill for a day or two but that is certain to have no worse effects than a couple of days of misery. He (the billionaire) emphasizes that you will get the million dollars for having the relevant intention at midnight tonight, not for actually acting on it tomorrow, and hence he guarantees that, once you've gotten the million dollars (if you get it), it will be yours to keep, regardless of whether or not you actually drink the toxin at noon tomorrow. Finally, he makes it clear that the deal is off if you try to find a way to ensure, between now and midnight tonight, that it will in fact be in your interest to drink the toxin at noon tomorrow (betting an acquaintance ten thousand dollars, for example, that you will actually drink it). The point is that he will give you a million dollars if you really do intend, at midnight tonight, to act, at noon tomorrow, in a way that you believe, at midnight, it will not be in your interest to act tomorrow (at least so far as drinking or not drinking the toxin is concerned).[1]

I have argued elsewhere that, given just one further constraint—namely, that she is not allowed artificially to manipulate her beliefs, desires, and intentions in certain ways—it can be shown that a reflective, basically rational agent would be unable to get the million dollars in the case just described,

because she would be unable to adopt the intention she needs to have in order to get the money. My argument begins with the assumption that it would be irrational actually to *drink* the toxin, supposing one got the million dollars for intending to drink it, and with the further assumption that a reflective agent would see that this is so. I then show that since i) a rational agent cannot, logically, intend to do what she grants it will be irrational to do, it follows that ii) a rational agent could not intend to drink the toxin in a case like the one Kavka has described. Saying just this, of course, leaves open the possibility that a rational agent could nonetheless get the million dollars in Kavka's case, since the argument just glossed does not preclude the possibility that a rational agent could, by adopting the requisite intention, make herself a millionaire while at the same time bringing it about that she is, by virtue of having that intention, less than fully rational. I attempt to block this possibility, though, by showing that iii) on a plausible analysis of what a future-directed intention is, there are compelling reasons for thinking that in fact a reflective and basically rational agent could not just *adopt,* in the way in which one ordinarily adopts one's intentions, an intention that, because it is an intention to do what she grants it will be irrational for her to do, would make her less than fully rational once she has adopted it. Hence, I conclude that, paradoxically, a reflective and basically rational individual would be at a disadvantage, relative to other, unreflective or less rational individuals, in cases like the one Kavka has described—cases, that is, where it is clearly in an agent's interest to intend to do something he correctly believes it will not be in his interest *to do* once the time comes to do it.[2]

Not surprisingly, many people, including a number of philosophers and decision theorists, find it hard to believe that the argument just glossed can be sound. After all, if we suppose, as I think we must, that a rational agent could, consistent with her rationality, *drink* the toxin, if that were what was necessary to get the million dollars, how can it be that she cannot adopt an *intention* to drink it, if that is what is required to get the million dollars?

I continue to be persuaded that the argument sketched a moment ago is sound. However, since I now believe that showing this is considerably more difficult than I had previously thought, I want in what follows to make another attempt to make good on my earlier claims. I shall begin, in section I, by laying out as carefully as I can what I take to be the positive argument for the view I wish to defend. Then, in sections II and III, I shall consider what I take to be the two most important ways in which my argument might be resisted—one of them suggested by David Gauthier in a recent paper, another suggested by Edward McClennen in his recent book. Our discussion in these later sections will require us to raise the general question of how individual actions that are part of overarching plans should be evaluated with respect to their rationality, and to raise as well the more narrow question of whether something like McClennen's notion of "resolute choice" can afford us a way around the implications of a view like my own. It is the discussion

of these latter matters, I try to show, that enables us to see most clearly why Kavka's puzzle ought to be of interest to those whose primary professional concern is the theory of rational choice.

I

Suppose we begin by assuming that, whatever else we say about your situation in the circumstances that interest us, there is no question it would be irrational for you to actually *drink* the toxin at noon tomorrow, regardless of whether or not you have become a millionaire by then as a result of having successfully formed the intention, the night before, to do so. After all, at noon tomorrow you will either have the million dollars or you won't. Refraining from drinking the toxin will not alter this fact, and since drinking it will make you quite ill, what reason could you possibly have to drink it? (Recall that we are ruling out side-bets, and so on, that would make it reasonable for you to drink the toxin despite the fact that you know that doing so will make you ill.)

This assumption—that it will be irrational for you to drink the toxin tomorrow, because doing so will not maximize your expected utility given your other option—turns out to be controversial, for reasons I shall discuss in sections II and III below. For now, however, I want to ignore this controversy in order to show why I believe that those of us who think it *would* be irrational to drink the toxin must also believe that someone who agrees with us would not be able to adopt an intention to drink it and hence would not be able to secure the million dollars, despite the fact that he would want the million dollars as much as anyone else and would certainly be willing to actually drink the toxin if that would get him the money. Having established this claim, at least to my own satisfaction, I shall then return to the question of whether it really is reasonable to assume that it would be irrational to drink the toxin just because, by hypothesis, doing so will not be utility-maximizing when the time comes to drink or not drink it.

We can begin with the following question: given that we are assuming it would be irrational to drink the toxin at noon tomorrow, regardless of whether or not one had intended, the night before, to drink it tomorrow, what can we say about someone who drinks it nonetheless—someone who claims to know perfectly well that there's no reason whatsoever to drink it, for example, and good reason not to drink it, but who picks up the cup, shrugs, and drinks the toxin anyway? Initially, of course, we might be inclined to suppose that, despite what he has said, this person *does* believe it's reasonable to drink the toxin—maybe he's secretly quite superstitious, we might conjecture, though embarrassed to admit it, and quite fearful, because of his superstition, about what will happen to him if he doesn't stick to his earlier

resolve. Or perhaps he's secretly pleased by the effect he knows his bravado will have on us and values the momentary but very gratifying pleasure he expects to feel while exhibiting that bravado far more than he disvalues the misery he will suffer as a consequence of drinking the toxin.

Now, some writers would say there *has* to be a reading of the agent's behavior in a case like this that renders what he has done rational, despite what he has said, at least as long as we suppose that he did what he did both knowingly and freely. I'm not one of these, however, and I ask the reader to assume, with me, that we needn't be one of them. Obviously, there has to be some *explanation* of why the agent did what he did in such cases, and, given the assumption that he did it freely, this explanation will no doubt appeal, inter alia, to his current desires. But to say that there has to be an explanation of why he did what he did is consistent with saying that in some cases what he did was clearly irrational, because it was not the alternative among his options that was likeliest to maximize the satisfaction of his desires, and is consistent with saying, as well, that what he did was irrational by his own lights. Or so, at any rate, I want to assume.

Return now to the question of what we are to say of someone who drinks the toxin in a case like the one with which we are concerned, despite the fact that he admits that doing so is irrational. One thing we would say, it seems to me, is that not only was what he *did* irrational, *he* showed a kind of irrationality, in himself, in doing it. As in the case of someone who freely and knowingly acts *immorally*, that is to say, we would make in the present case a judgment about the *agent* as well as about his act. And, roughly speaking, the judgment would be that, given that his action was fully informed and free, he can't be counted a fully rational individual, and he can't be so counted *regardless* of precisely what we think being a fully rational individual involves. For, surely, whatever one's views are about what must be true of a person if she is to count as an ideally rational agent, someone who knowingly and freely performs what she grants is an irrational act *cannot,* logically, count as such an agent.

What about someone who *intends* to perform what she admits will be an irrational action, though, but who has so far actually *done* nothing whatsoever that would incline us to judge her as anything less than a perfectly rational individual? Is it possible that just intending to act irrationally is also enough to render a person less than what a fully rational individual has to be? Or must we say that the fact that a person intends to do what she grants it will be irrational for her to do is not in itself sufficient to tell us anything at all about her relative degree of rationality?

Answering this question, of course, requires us to raise and answer yet another question first—namely, the question of exactly what is involved in intending, at one point in time, to perform some action or series of actions at some later point in time. Consider, then, the following answer to this latter question: to intend, at one point in time, to perform some action (or series of

actions) at some later point in time, is to be *committed,* at the one point in time, to performing that action (or those actions) at the later time. Obviously, if we find this answer at least prima facie plausible, our next question must be what exactly is involved in the "commitment" that supposedly constitutes future-directed intending. For if we could say what this commitment involves, and if it appeared that a rational agent cannot, *qua* rational agent, be thus committed to performing what she grants will be a clearly irrational act, we will have answered the question with which we were concerned a moment ago—namely, the question of whether merely *intending* to perform an admittedly irrational act, like actually performing such an act, is enough in itself to count against the claim that the intender is a fully rational individual.

So, what is this "commitment" that we are supposing constitutes future-directed intending (or "future-directed intentions," as I shall sometimes say)? A number of different answers to this question have been defended in recent years, but the best answer, it seems to me, is that provided by Michael Bratman in a recent book devoted almost entirely to exactly this question. To be *committed,* in the relevant sense, at one point in time, to the (conditional or unconditional) performance of some action *A* at another (later) point in time, Bratman argues, is to be such as to be disposed i) to *perform* that action if the relevant future time arrives and the conditions are as one imagined they would be, and ii) to *reason* in certain ways in the meantime (specifically, to refrain from reconsidering one's intention to perform that act, in the absence of new information about the desirability of performing it, and to adopt or not adopt certain other intentions depending upon whether having them would support or subvert one's carrying out one's intention to do *A*).[3]

Suppose, as in fact I believe is the case, this account of the nature of future-directing intending is sound. Does it suggest any reason to think that someone who intends to perform what she admits will be an irrational act is, like the person who freely and knowingly performs such an act, *eo ipso* less than an ideally rational individual?

It seems to me it does. To see why, we simply need to ask, of the case in which we imagine this same person actually performing the relevant act, why we would say that in performing it she thereby shows herself to be less than fully rational. Is it because she has actually *performed* that (admittedly irrational) act, or is it because in performing it she has shown us something about herself—something that might very well have been true of her even if she hadn't performed it?

A fully adequate answer to this last question is one I cannot provide here. I think it's quite clear, however, that it is not the actual performance of the admittedly irrational act that is important in our (negative) judgment about the relative degree of rationality of the person who performs it. After all, we would make exactly the same judgment, it seems to me, if she had been just about to perform that act, freely and with full knowledge that it would be irrational for her to perform it, but was prevented from performing it by

circumstances outside of her control—if she had been just about to drink the toxin, for example, quite irrationally, but was prevented from doing so by some entirely fortuitous event. (Notice, by the way, that the analogy to moral appraisals, to which we alluded briefly above, holds here as well, and in fact supports the point we are now making: someone who is fully prepared to betray, for the sake of her own advancement, a close and trusting friend, doesn't have to have actually betrayed that friend in order to be properly subjected to negative moral appraisal for being disposed to treat him in this way.[4])

But now recall what we are assuming about the nature of a future-directed intention: a person who intends, at one point in time, we said, to perform a certain action, *A*, at some later point in time, is a person who is disposed, inter alia, to actually *perform* that action when the relevant time and anticipated circumstances arrive. This is what Bratman calls the "volitional" component of a future-directed intention: one doesn't really intend to perform an act if it's not the case that one would in fact perform it if the relevant time and anticipated circumstances were to arrive and one hadn't for some reason ceased to be committed to performing it. But now, in the sorts of cases we are imagining, not only is it true of the relevant agent that she is thus disposed to perform a clearly irrational action, it is true as well that she is so disposed while at the same time admitting that it will be irrational for her to perform it. But, then, how could such an agent not but fail to be all a fully rational agent has to be? After all, she differs from an otherwise identical version of herself that has just performed the relevant act only in that, because the time for so acting has not yet arrived, she hasn't yet had the opportunity to do what that other person has just done. So far as all of her relevant beliefs, desires, and dispositions are concerned, she is no more different from *that* person than she is from the person she would be if the time arrived, the circumstances were as anticipated, and she was prevented from performing the admittedly irrational act only by the occurrence of some entirely fortuitous event that happened to make it impossible for her to perform that act.

Suppose this line of argument is basically sound: a fully rational individual *cannot*, logically, intend to do what she grants it will be irrational for her to do. Why should we suppose it follows from this that, in the sorts of circumstances that interest us, a fully rational individual could not adopt (and retain, at least for a few moments) the intention she would need to adopt in order to get the million dollars—i.e., the intention to drink the toxin the next day, despite the fact that she knows, and admits, that drinking it will be irrational? After all, even if the argument above is sound, what is to prevent us from saying that in such circumstances a rational individual would gladly sacrifice her claim to (complete) rationality by adopting an intention the having of which will render her less than fully rational?

The best answer to this last question, it seems to me, is this: intention-adoption, for a fully rational individual, is a function not of the desirability of having the relevant intention, but of the reasons the agent foresees himself as having (or not having) for performing the action he would be adopting an intention to perform if in fact he adopted that intention. As in the case of adopting his beliefs, in other words, so too in the case of adopting his intentions, the rational person looks not to what he will get, or likely get, from being in the relevant state—believing that *p;* intending to do *A*—but to something quite different: in the case of beliefs, whatever evidence there is for the proposition *'p';* in the case of intentions, whatever reasons there are for performing the action *A*.

Unfortunately, I am not prepared to defend this answer here—too much would need to be said about exactly what the (intentional) objects of intentions are, about exactly what it is to adopt an intention to act in a certain way, and about exactly what norms it is reasonable to suppose must govern the intention-adoptions of a rational agent. Instead, I shall give an answer to our previous question that, while not as deep and far-ranging as the answer just sketched, is a perfectly compelling answer, it seems to me, nonetheless: a reflective, basically rational agent could not intentionally commit herself to performing an action she grants it will be irrational for her to perform because such a commitment would necessarily be *unstable,* in the case of an otherwise rational individual, and would be unstable in a way that would be inconsistent with its being the sort of commitment we are supposing is necessary for a future-directed intention to exist. To see why this is so, let us suppose that a fully rational individual *could* adopt, at least momentarily, a commitment to act irrationally (in the future) of just the sort that would, if it persisted, clearly constitute an intention, conditional or otherwise, to perform the relevant action. Could such a commitment be maintained, for more than a moment, by an otherwise rational individual (i.e., by an individual who would by hypothesis be *fully* rational except for the fact that she is currently at least momentarily committed to the performance of what she admits will be an irrational action)? It seems to me it could not. For suppose it could—i.e., suppose that reflection on the fact that it was a commitment to act irrationally would not necessarily undermine it. It would then be possible for an otherwise rational individual to make such a commitment and for that commitment to persist, undiminished, up to the time at which the intended action was to be performed. But then, the action will, at that time, be performed, supposing no other changes have occurred, since we are supposing that the commitment in question is such that, were it present at the time of action, the action would be performed. This, however, is absurd, since, as we have already seen, we cannot suppose that a rational individual could knowingly perform an action that she grants it is irrational for her to perform. We must suppose, therefore, that reflection on the irrationality of what one intends to

do *will* inevitably undermine one's intention, at least if one is otherwise a basically rational individual, since its not being undermined would show one to be even more deeply irrational than one's supposed intention already shows one to be. And this means that one's commitment is not, after all, a commitment of the sort that is required for the existence of a genuine intention to perform the relevant action. For such an intention, we have supposed, can only be constituted by a commitment with the dispositional and reasoning-centered components sketched above. And how can it be said of someone whose commitment to action is unstable, in the sense just sketched, that she is disposed not to think about whether she really means to perform the relevant action when the time comes to perform or not perform it?[5]

Where does all this leave us? We have seen, it seems to me, that, paradoxically, a reflective, basically rational individual would, because of her rationality and reflectiveness, not be able to secure the million dollars in circumstances like those hypothesized in Kavka's Toxin Puzzle, even though a less rational or deeply unreflective person might well be able to do so. Or, rather, we have seen that this appears to be so if we make the assumptions required by our argument thus far, including the assumption that, whatever else we say about your situation, if you find yourself in the circumstances hypothesized in Kavka's Toxin Puzzle, it will be irrational for you to actually drink the toxin, when the time comes to drink or not drink it, and it will be irrational for you to drink it regardless of whether or not you have succeeded, the night before, in getting yourself to intend to drink it the next day. I now want to look at this last assumption a bit more carefully. For, quite apart from what can be said for or against our other assumptions, if it can be shown that under certain circumstances, a rational agent *could* drink the toxin in a case like the one described by Kavka, without thereby sacrificing her claim to rationality, the argument developed above would be undone.

II

The case against the rationality of drinking the toxin is obvious enough: one would have no obligation to drink it—moral, legal, or otherwise—and one would by hypothesis *gain* nothing by drinking it except for a few days of misery (*unwanted* misery, we may suppose). Why, then, would anyone maintain that it might nonetheless be rational to drink it?

One answer, defended recently by David Gauthier,[6] goes roughly as follows. Assume that the *rational* choice in any given choice-situation is, by definition, the choice a rational agent would make in that situation. Then ask yourself the following question: how does a rational agent make his choices? On the basis of what principles or other sorts of considerations, that is to say, are a rational agent's choices made?

On one view—the "orthodox" view among contemporary decision-theorists, Gauthier would say—a rational agent will choose, from among his options in any given situation, that option that promises him at least as much expected utility as any of the other options that are available to him. On this view, as we have seen, a rational agent will choose not to drink the toxin, and so, on the definition proposed a moment ago, drinking the toxin would be irrational for him, with potential consequences for his financial future that we have already explored.

Gauthier, however, believes that a proper understanding of the concept of practical rationality shows that the orthodox view is incorrect. In some situations, he argues, a rational agent will *refrain* from choosing the option he correctly believes will maximize his expected utility and will choose, instead, to act in accordance with the requirements of a utility-maximizing *plan* that he has previously and rationally adopted. And he will so choose, at least in certain circumstances, even if the choice in question is not only suboptimal but is, in addition, the final step in the relevant plan.

Gauthier illustrates what he has in mind here not by discussing the implications of his view for Kavka's Toxin Puzzle, but by discussing the implications of his view for a variant of Newcomb's Problem. I now want to take up this variant of the Newcomb Problem, therefore, both because of its intrinsic interest and because of its relevance to the issue before us. It will quickly become obvious, I think, how Gauthier's analysis of his version of the Newcomb Problem is relevant to the problem with which we have been grappling so far.

As in the standard version of the Newcomb Problem, one is offered just two choices in Gauthier's version: one can take just Box A, as Gauthier calls it, or one can take both Box A and Box B. Both boxes, however, are transparent, in Gauthier's version of the problem, with B clearly containing ten thousand dollars and A clearly containing either a million dollars or nothing. The contents of A have been determined, moreover, in advance, as follows: if the predictor has predicted that the chooser will take just Box A if he sees a million dollars in it, the billionaire has put a million dollars into A; if the predictor has predicted that the chooser will take both boxes if he sees a million dollars in A, the billionaire has put nothing in A.

Imagine yourself presented, without advance warning, with the boxes and options described in Gauthier's version of the problem. You are free to take just Box A, or both A and B; you see a million dollars in Box A and, of course, ten thousand dollars in Box B; and you're told the story just recounted about how the million dollars got into Box A. Money has positive utility for you, and neither choice has any consequences, positive or negative, beyond the fact that you will acquire a certain amount of money either way. Which choice would you make?

Surely you will choose both boxes, Gauthier observes, and you will choose both boxes regardless of whether, *vis-à-vis* the standard Newcomb

Problem, you are a "one-boxer" or a "two" (i.e., regardless of whether you favor an "evidential" version of orthodox decision-theory or a "causal" version). After all, how could you do otherwise? You've been presented with a choice of securing either $1,010,000 or $1,000,000, nothing matters but the money, etc., so the choice is obvious.[7]

But now consider the following change in the scenario just described: instead of being presented with the relevant choice without advance notice, you're told, in advance, that you are one of a small group of people some one of whom will soon be selected for evaluation by the billionaire's predictor and then given the choice just described. You are not allowed to take steps to see to it that, if you're chosen, you will have a good, independent reason to take just Box A—an eleven-thousand-dollar side-bet with a friend, for example, that you'll take just that box—and in fact you're allowed to keep the million dollars, if you get it, only if in choosing it you actually chose, in taking it, to forego exactly that amount of utility that you would have gained had you taken both boxes, and hence the extra ten thousand dollars, instead.

What do you do? Suppose it occurs to you that if you could successfully commit yourself, right now, to taking just Box A if you are the person chosen—i.e., if you could adopt a sincere intention (now) to take just box A if you get the choice—and if you could continue to be so committed (or to so intend) as time goes by, you would very likely see a million dollars in Box A if you were the one selected to make the relevant choice, since your having this intention would very likely lead the predictor to predict that you will take just A if you see a million dollars in it. And suppose *that* sort of "precommitment"—adopting and retaining the relevant intention, if you can—is allowed. Would you do it? *Could* you?

Notice, before we try to answer these questions, that when we imagine a person situated in the sorts of circumstances just described, we are imagining them in a situation that is quite different from that proposed in a standard Newcomb Problem: after all, we are supposing, in this version, that you have, at least in theory, the ability to influence, causally, the predictor's prediction. And notice, as well, that a person who finds himself in this sort of situation has in fact found himself in a situation that is formally analogous to the situation he would be in if he were presented with the offer described in Kavka's Toxin Puzzle: he is in a position to secure a huge increase in his expected utility if he can bring himself to intend to do something that, on the orthodox view, it will clearly be irrational for him to do.

We can now return to the question raised at the end of the previous paragraph but one: would you, confronted with the circumstances described in Gauthier's version of the Newcomb Problem, be able to commit yourself to taking just the one box, and then be able to retain this commitment for a reasonable length of time? Our argument in section I suggests that, if you are a reflective, basically rational individual, you would not be able to do this—

because you *could* not. This, of course, is because if we continue to suppose, as we did above, that actually *taking* just the one box, when you see the million dollars in it and the ten thousand in the other, would be irrational, then, by our argument in section I, you will be unable to adopt an intention to take just the one box, much less adopt *and retain* such an intention for any length of time, and you will be unable to do this despite the fact that you would very much want to do it because it would be very much in your interest to do it.[8]

It is at exactly this point, however, that Gauthier's attack on what he calls the "orthodox" view of rational choice is relevant. For Gauthier believes you *would* be able to adopt and retain the intention you need to adopt and retain in order to get the predictor to put the million dollars into Box A, and that you would be able to do this because you would be able to commit yourself to a plan for getting the million dollars into that box that includes adopting the requisite intention as one of its parts. What's more, he believes you would be able to do this, if you really are a basically rational individual, because, being such, you will see that taking just Box A, given your plan, is the rational thing to do.

Now, Gauthier does not say much, at least in general terms, about what he thinks a plan is, nor about what it is in his view to adopt or commit oneself to following a given plan. Suppose for now, though, that answering these questions is relatively unproblematical. A plan, let us suppose, is a kind of blueprint for action—an "action-schedule," as I shall sometimes say—for achieving some end. And suppose that to adopt a plan, or commit oneself to following it, is simply to commit oneself to acting as the plan requires, and to so commit oneself for the sake of achieving the end one believes following the plan will enable one to achieve.

What is the plan Gauthier thinks a rational agent would adopt and follow, in the version of the Newcomb case he has described, and what is his argument for the claim that a rational agent would, and hence could, adopt and follow that plan in order to secure the million dollars that would otherwise be beyond his reach? The plan that Gauthier favors is simple enough, at least superficially, as is his argument for the claim that this is the plan a rational agent would adopt if situated in the sorts of circumstances that interest us. His proposed plan, in fact, has just two elements: first, one is to adopt, in advance of being selected, the intention to take just Box A in the event one sees a million dollars in A after one has been selected, evaluated by the predictor, and given the opportunity to choose one or both boxes; and, secondly, one is to *take* just Box A, when the moment for taking just that box or both boxes in fact arrives. As for why one ought to adopt this plan, according to Gauthier, rather than some other plan or no plan at all, just consider the alternatives to doing so, he says. If one adopts no plan at all, then, as in the case above, where we imagined ourselves presented with the two boxes without prior warning, it seems clear that a rational person, at the time of

choice, will see no reason to take just A and good reason to take both A and B. But then, the predictor, given his record, will no doubt have predicted this, and hence there will be nothing *in* Box A when one makes one's choice.

What about alternative plans? Recall that one is not allowed to resort to *standard* precommitment strategies, like side-bets, pills, or behavioral therapies that change what would otherwise have been one's subsequent preferences, and so on. In fact, all one can do is either adopt no plan at all, or adopt a plan whose first element is either an intention to choose just Box A or an intention to choose both A and B, and whose second element is either choosing just Box A or choosing both A and B. And this means that there are really only four plans one can even think about adopting:

1. Adopt an intention to choose only box A; choose both boxes.
2. Adopt an intention to choose only box A; choose only box A.
3. Adopt an intention to choose both boxes; choose both boxes.
4. Adopt an intention to choose both boxes; choose only box A.

Now, clearly, Gauthier argues, Plan 4 is, from any standpoint, a nonstarter. What's more, Plan 1, we must suppose, would be unavailable to a rational person in the relevant circumstances *as a plan*—among other reasons, because of arguments like the one developed in section I above. But then, there are really only two choices that are tenable for an agent who is at least superficially disposed to think about adopting a plan and who is trying to decide which one he might adopt: namely, Plans 2 and 3. And, of course, it's obvious enough, Gauthier concludes, which of these he ought to adopt (and then implement, if he is selected to make the all-important choice), since following one of them will in all likelihood make him a millionaire, while following the other will leave him a relatively poor man. Hence, a rational agent will choose Plan 2, Gauthier argues, and will stick to it if he is selected and actually gets to choose.[9]

One might concede at this point that a rational agent would adopt the plan Gauthier favors *if he could,* but then go on to claim that it's hard to see how a rational agent could in fact adopt that plan. After all, the final step in the plan Gauthier has proposed is one we may suppose the agent knows he will have better reasons not to perform than to perform, when the time comes to perform or not perform it. But then, since, as we have already seen, a rational agent cannot adopt an intention to perform an action he correctly believes it will be irrational for him to perform, it follows that a rational agent will be unable to adopt the plan Gauthier has shown he would *want* to be able to adopt.

This objection, though, implicitly assumes the truth of the orthodox view—namely, that, in the end, a rational agent will choose whether to take the one box or both by determining which choice will maximize his expected utility. Suppose, instead, Gauthier says, we assume that the orthodox view is wrong and that a rational agent's choices will sometimes be determined

not by her preferences for the outcomes of particular choices, but by her preferences for the outcomes of plans, and that her choices will be so determined, in certain cases, even when the relevant (non-maximizing) choice is the last step in the relevant plan. Then, supposing you are the agent in the case that interests us and you accept the view just described, the objection mooted a moment ago will have no force, Gauthier argues. For taking just the one box is not irrational on your view of how a rational agent will choose in the relevant circumstances, and hence the objection in question cannot get off the ground.

Those who are familiar with Gauthier's work on the foundations of ethics will not be surprised by the structure of his argument for the view that in certain circumstances a rational agent's choices will be determined by her preferences for the outcomes of plans, rather than by her preferences for the outcomes of particular choices, and that they will be so determined even when the choice in question is the last step in the relevant plan. The argument goes as follows. Assume, first, as Gauthier believes his version of the Newcomb Problem shows we must, that agents who sometimes take their reasons for choosing from their preferences for the outcomes of plans, rather than from their preferences for the outcomes of particular choices, and who do this even when making the choice in question is the last step in the relevant plan, do better, in maximizing their overall expected utility, than agents who do not, even in the relevant circumstances, identify their reasons for choosing in this way. Next, assume, with those who hold the orthodox view, that utility-maximization is what rational choice is all about. It then follows, Gauthier says, that we must agree that the orthodox view is defective in at least one important respect—namely, in instructing us always to make our choices in accordance with our preferences for the outcomes of those choices rather than, in some cases at any rate, in accordance with our preferences for the outcomes of plans. For if utility-maximization is what rational choice is all about, and if an agent will do better if she sometimes makes her choices in accordance with utility-maximizing plans, rather than in accordance with her preferences for the outcomes of the final choice of any particular plan, it seems clear that a rational agent will sometimes make her choices in the former rather than the latter way.[10]

Now, one problem with this argument is that it's not clear why we should suppose that its final, key premise is true—namely, that if an agent's taking certain considerations to be reason-providing would make her better off than *not* taking those considerations to be reason-providing would make her, she will, if she is rational, see such considerations as reason-providing. This, of course, is also a key premise in Gauthier's work in the foundations of ethics, and important and seemingly compelling critical appraisals abound. Here, however, I want to skirt this controversy and simply grant Gauthier that a rational agent who had rationally adopted a plan of the sort he favors, and who had then gone on to perform its first step, would see herself as rationally

bound to perform the second step as well, when the time comes—i.e., to take just the one box, given that that's the final step in the plan she has (rationally) adopted. Even if we grant Gauthier this much, he needs yet another concession if his argument is to go through. For he needs, in addition, the assumption that a rational agent could indeed adopt, consistent with her rationality, a plan with the relevant features, and then rationally perform the first step in that plan by adopting the intention to perform the action that, once the relevant intention has been adopted, she will later rationally perform.

Why should we think there is a problem here? Recall, to begin with, what we are supposing it is to adopt or commit oneself to a given plan: to adopt a plan, we said, or commit oneself to following it, is to commit oneself to performing the actions the plan requires one to perform, and to so commit oneself for the sake of achieving the end the plan is designed to enable one to achieve. What is it, though, to "commit" oneself to performing the actions that are listed on the "action-schedule" that constitutes a plan? A lot could be said in reply to this question that I cannot say here. At bottom, though, I think it is clear that the commitment that is involved in adopting a plan is exactly the sort of commitment that constitutes a future-directed intention: it is a commitment, that is to say, with exactly those "volitional" and "reasoning-centered" dimensions that, following Bratman, we are taking to be constitutive of future-directed intentions. But then, if this is right, to commit oneself to the plan Gauthier claims a rational agent would commit herself to following, in the circumstances hypothesized in his version of the Newcomb Problem, is simply to adopt an intention to perform the "actions" that constitute that plan—the actions, that is, of forming an intention to take just the one box if one sees a million dollars in it, and then taking just the one box when the time for choice arrives. And this means that, just to commit herself to the favored plan, the agent in question has to adopt an intention to adopt an intention to *take* just the one box and, at the same time, adopt an intention to take just the one box when the moment for choice arrives.

Assume for the sake of argument that there's no problem with the notion of adopting an intention to adopt an intention to perform a certain act. (In fact, I think this notion is deeply and importantly problematical, but that is an issue we cannot pursue here.) And assume, as well, that there is no problem with the assumption, almost explicit here, and certainly clearly implied, that adopting an intention is (at least in certain circumstances) itself an intentional action—something one can intentionally choose to do, that is to say. (Again, as I have hinted above, and as I have argued at length elsewhere, I think this assumption is in fact extremely implausible and, in any case, in need of careful qualification if we insist on making it. But this too is an issue I cannot pursue here.) Still, even granting all of this, we must suppose, I shall assume, that just as there are (normative) constraints on what "first-order" intentions a rational agent can adopt—this was the point of section I above and is, as

we have seen, a point that Gauthier himself accepts as well—so too there are (normative) constraints on what "second-order" intentions a rational agent can adopt (on what intentions to adopt first-order intentions he can rationally adopt, that is to say). What, then, might these latter constraints be? One of them, surely, is a constraint that is analogous to the constraint on first-order intentions that holds that a rational agent cannot adopt an intention to perform an action he believes it will be irrational for him to perform—i.e., that he believes he will not be able to perform *qua* rational agent. As applied to second-order intentions, I shall suppose, this constraint will hold, by analogy, that a rational agent cannot adopt second-order intentions to adopt first-order intentions he believes a rational agent would not be able to adopt.

Return now to the case at hand. A rational agent will be able to commit herself to the plan Gauthier favors only if she will be able, *qua* rational agent, to commit herself to performing the "actions" that make up that plan. And she will be able so to commit herself, we are supposing, only if the relevant "actions" are actions she believes a rational agent will be able to perform. Now, the first of these "actions" is the act of adopting an intention to take just the one box even though in taking it (rather than both) she will be leaving behind a not insignificant amount of money (ten thousand dollars, to be precise). Is this an act to the performance of which a rational agent could commit herself, consistent with the norm for such actions we elicited a moment ago? Prima facie, it is not. For, as we have already seen, a rational agent will see adopting the intention to take just one box as (rationally) problematical. Of course, Gauthier wants to show that a *truly* rational agent will see her way around (or through) this difficulty, since, he claims, such an agent will see the taking of just the one box, and the adopting of the intention to take just that box, as parts of a (rational) plan the adoption and following of which will make her rich. In order to establish the rationality of following such a plan, however, Gauthier must first establish the rationality of adopting it—i.e., of committing oneself to it, and its component parts, as one's plan. And the latter is what it now appears Gauthier hasn't done—and in fact *cannot* do, without begging at least one of the questions that are at issue here. To see this, we simply need to recall that Gauthier is granting, *arguendo,* that it will initially appear, to a reflective, rational individual, that she will be unable to adopt the intention to take just the one box (in order to influence the predictor in the desired way), because she will see that, as a rational individual, she will be unable to *take* just that box when the time comes. Gauthier's strategy is to convince her, and us, that these first appearances are deceptive and that in fact she *can* adopt the relevant intention, once she sees doing so as part of a plan for getting the million dollars. Unfortunately, though, this otherwise irrational "act" can rightly be thought of by her as part of a (rational) plan, *of hers,* only if she can make it part of such a plan by committing herself to performing it, along with the other parts of the plan, as a way of achieving the relevant end. And how can she do this without

presupposing the rationality of adopting the very intention that her commitment to the plan is supposed to make it rationally possible for her to adopt?

What I am saying, of course, is that Plan 2 is no more an option for a rational agent situated in the circumstances Gauthier has imagined than is Plan 1—and for exactly similar reasons. Plan 1 is ruled out, as Gauthier observes, because it would require the agent to adopt an intention to perform an action he believes it will be irrational for him to perform. Plan 2 is ruled out, however, because it too would require the agent to adopt an intention to perform an "action" he believes a rational agent could not perform—namely, the "action" of adopting a certain intention under circumstances where adopting that intention is rationally problematical because of what it is an intention to do. The fact that this is so—that Plan 2 is unavailable to a rational agent, that is to say, who starts out with the assumptions with which Gauthier asks us to start out—is obscured, in Gauthier's account, by the fact that he talks so very loosely about the possibility of "committing" oneself to a given plan, without ever asking either what a plan *is* or exactly what it is to "commit" oneself to a plan. Once we are clear on what these actually rather subtle notions involve, it emerges that Gauthier's argument involves an objectionable form of normative "bootstrapping": he needs to *assume* the rationality of adopting an intention to take just the one box, as part of his argument for the rationality of adopting that very intention and then actually taking just the one box when the time for choice arrives.

III

We asked, at the beginning of section II, why anyone would be inclined to hold that under certain circumstances it would be rational to drink the toxin, in Kavka's toxin case, supposing one had already gotten the million dollars for intending to drink it, and supposing, as well, that actually drinking it promises one nothing more than a day or two of unwanted misery. It will be obvious, I hope, how Gauthier would defend the rationality of drinking the toxin, given the appropriate antecedents, and obvious, as well, why I think his answer would be mistaken. Gauthier's, however, is not the only recent attempt to ground the possibility of securing the relevant gains, in cases like Kavka's, on a series of assumptions about the capacity of a rational agent to make a non-maximizing choice in light of the adoption of a utility-maximizing plan. Edward McClennen, in an extremely provocative study of a well-known set of problems in the foundations of the theory of rational choice, also claims to have solved, in exactly this way, the problem that Kavka's puzzle appears to pose.[11] I now want to turn to McClennen's account, therefore, first explicating his proposed solution to Kavka's puzzle, and then indicating why I believe that it too fails to do what it purports to do.

McClennen's remarks about the Toxin Puzzle presuppose an understanding of his more general views about the rationality of what he calls "resolute choice," and these, in turn, presuppose an understanding of his views about the constraints it is reasonable to suppose a rational agent has to honor in situations involving "dynamic" or sequential choice. Obviously, it will be impossible, here, to do justice to these more general views. Still, I think we can fairly quickly get a good enough hold on McClennen's overall position, to be able to appreciate why he thinks he has found a compelling solution to Kavka's puzzle. What I shall try to show is that even if we suppose McClennen's general views are sound, there are serious problems—insuperable problems, I believe—with his attempt to use these views to solve the puzzle with which we have been concerned. What's more, these are problems, I shall suggest, that are almost identical to those that undermine what would appear to be Gauthier's solution to this same puzzle.

McClennen is interested, among other things, in the plausibility of the so-called *weak-ordering* and *independence* axioms in modern, Bayesian theories of rational choice. Indeed, he begins his book with a series of reflections on a well-known sequential decision problem in which the agent's preferences violate the independence axiom. It will be useful to begin by describing this problem, since doing so will greatly simplify the task of summarizing McClennen's theory of resolute choice.

Consider first of all, then, the following prospects, where [$2,400, 1] is to be read as "The agent will get $2,400 with probability 1," and [$X, p; $Y, 1-p] is to be read as "The agent will get $X with probability p, and $Y with probability 1-p":

g_1 = [$2,400, 1]
g_2 = [$2,500, 33/34; $0, 1/34]
g_3 = [$2,400, 34/100; $0, 66/100]
g_{3+} = [$2,401, 34/100; $1, 66/100]
g_4 = [$2,500, 33/100; $0, 67/100]

Now imagine an individual who, while he prefers the prospect g_1 to g_2, also prefers g_4 to g_{3+}, and g_{3+} to g_3. "In the presence of certain other seemingly uncontroversial assumptions," McClennen observes, "such a preference pattern can be shown to violate the independence principle."[12] Assume this is true. Then, finally, imagine that the individual in question is exposed to these prospects by virtue of being confronted with the sequential decision problem illustrated in figure 1 (with squares designating choice points and circles designating chance happenings).[13]

How should we suppose a rational decision-maker would proceed when confronted with this problem? A "myopic" chooser, McClennen observes, will see himself as facing, initially, a choice between the prospect g_{3+}, supposing he heads downwards, and g_4, supposing he heads upwards, and so will choose to reject g_{3+} in favor of his preferred alternative. (He will

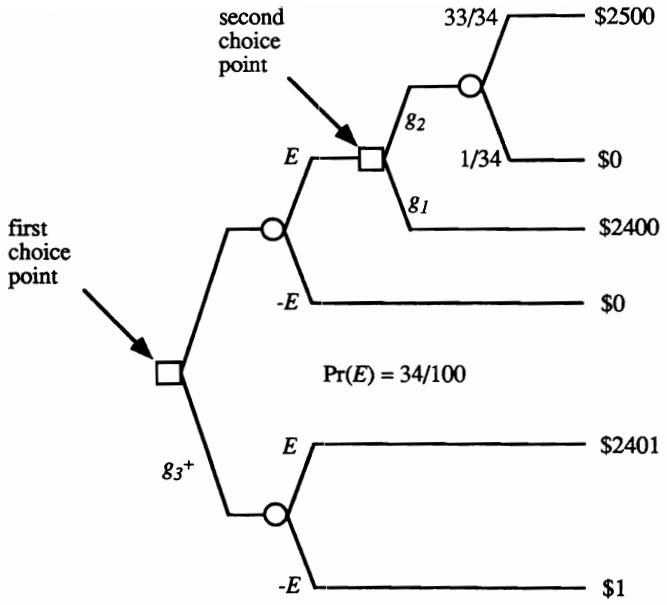

FIGURE 1

see the alternative to g_3+ as offering him the prospect of g_4, of course, only if he supposes that, given the chance, he will subsequently choose g_2 over g_1. Assume that this is what he imagines he will do, given the chance.) Unfortunately, while he sees himself as accepting the prospect of g_4 when he makes this first move, the myopic chooser inevitably undermines exactly this view of things when, circumstances permitting it, he gets to the second choice point, where he must actually *choose* between g_1 and g_2. For at that point he will actually choose g_1, which we are supposing he prefers. And this, of course, depending on one's view of these matters, shows *either* that he should not have been acting, in the first place, on a set of preferences that violated the independence axiom, *or* that, given those preferences, he should have chosen g_3+ right from the start. After all (the argument for the latter claim goes), given exactly the good fortune that got him to the second choice point on his chosen route—namely, the occurrence of E—he would have secured $2,401 instead of $2,400, had he taken the other route, while, supposing he hadn't had that good fortune, he would have gotten $1, by going the other way, instead of nothing.

This much is familiar enough. Also familiar, McClennen notes, is the response of the "sophisticated" chooser to the choice problem illustrated above. This is the agent who, foreseeing that he will choose g_1 over g_2, should a choice of the upward route eventuate in that option rather than a payoff of $0, chooses to reject the upward route from the start and, instead,

immediately heads downwards to g_3+. After all, knowing that he will choose g_1 over g_2, should he get the chance, and that he will get nothing otherwise (supposing he has headed upwards from the start), this agent sees that his initial choice is not between g_3+ and g_4 but, rather, between g_3+ and g_3. Since his preference, given this choice, is for g_3+, he will, if he is rational, head in the latter direction.

Now, one way of characterizing what happens to the myopic chooser, who starts out planning to choose g_2 over g_1, should he get the choice, but who in fact chooses g_1 over g_2 when he actually gets that choice, is to say that over time his choices show an important kind of inconsistency: his views, at the outset, about how he will (or ought to) choose if he gets the chance to choose between g_1 and g_2 are inconsistent with what he actually ends up choosing when he gets the opportunity to make this choice. In McClennen's terminology, his sequence of choices violates the principle of *dynamic consistency.* The sophisticated chooser, by contrast, avoids precisely this inconsistency by acknowledging at the outset how he must choose, should he get the chance to choose between g_1 and g_2, and, given this, plus his preference for g_3+ over g_3, choosing to move downward right from the start.

The sophisticated chooser pays a price, however, at least as McClennen sees things, for what his own conception of rationality forces him to do. Because he knows he will be unable to choose g_2 over g_1, should he get the choice, he is unable, McClennen observes, to expose himself to a prospect he would in fact prefer to face: namely, the prospect represented by g_4. After all, if, for some reason, he could correctly assume that he *would* (rationally) choose g_2 over g_1, should he get the choice, then, like the myopic chooser, he would face, at the outset, a choice between g_4 and g_3+ rather than between g_3+ and g_3. What's more, if, unlike the myopic chooser, he could in fact *choose* g_2 over g_1, consistent with all other requirements of rationality, he would be able to avoid the charge of inconsistency.

It is at this point that McClennen introduces the concept of the "resolute" chooser. This is an agent who, appreciating what the sophisticated chooser loses by virtue of having to see the upward route as a choice of g_3 rather than g_4, and appreciating as well why the sophisticated chooser is bound to see his situation in this way, resolves to avoid the problems of *both* myopic and sophisticated choice by adopting and acting on a certain *plan.* The elements of that plan are exactly those required if he is to create for himself, on the one hand, the prospect the sophisticated chooser is forced to lose, while avoiding, on the other hand, the subsequent choice that leads the myopic chooser into dynamic inconsistency: he will head upwards, rather than downwards, at the outset, thus exposing himself to the prospect of g_4; and he will actually choose g_2 over g_1, subsequent to this first choice, if and when he gets to choose between them.

Why, though, should we suppose the resolute chooser will be able to do what the myopic agent, if he is rational, will be unable to do—namely, actually make the choice of g_2 over g_1, if he gets the opportunity to do so? After all, we are supposing the resolute chooser has the same preferences as his myopic counterpart, and this means that, like the latter, he prefers g_1 to g_2.

Answering this question is, of course, one of the central aims of McClennen's book, and the details of his answer are not something we can hope to present here. We can, though, get a sense of the general thrust of that answer by quickly considering a number of its key elements. First, he says, notice that a rational agent must believe that it will be impossible for him to choose g_2 over g_1, given his initial preferences, only if we suppose that a rational agent is constrained to choose, at any point on a decision tree, as though the only factors relevant to his decision are the consequences of that decision from that point on. Suppose we don't make this supposition. Instead, suppose we assume that under certain circumstances a rational agent will make a certain choice not because of what will follow *from that particular choice,* but because that choice is required by a utility-maximizing plan to which he has previously (and rationally) committed himself. Then, if it can be shown, in any given case, that a seemingly irrational choice is in fact irrational only if we look at that choice in isolation from some plan that requires it, and if, in addition, it can be shown that that plan was one it was rational (because preferable) for the relevant agent to adopt, it will arguably be rational for her to choose in accordance with her plan rather than in accordance with what appear to be her preferences for the (probable) outcomes of that particular choice.

Stated as baldly as we have stated it here, McClennen's view looks almost indistinguishable from Gauthier's view, as summarized above. However, despite important similarities, I think it would be a mistake to assimilate the two views. For one thing, McClennen's view is developed only after a careful and very thorough examination of the grounds various writers have given for the claim that the weak-ordering and independence axioms are indeed plausible constraints on rational preference and choice. His strategy is to show, first, that none of these arguments succeed, and then, secondly, that in fact four other axioms, or principles, are, at least in dynamic contexts, individually necessary and jointly sufficient for deriving the weak-ordering and independence principles: the *simple reduction principle* (SR); the *normal-form/extensive-form coincidence principle* (NEC); the *principle of dynamic consistency* (DC); and the *separability principle* (SEP). The last of these, of course, is precisely the principle that McClennen's resolute chooser violates, because he refuses to accept it as a legitimate constraint on rational choice; and in fact resolute choice is possible, for an otherwise rational individual, only if we suppose that separability is not a legitimate constraint on rational choice.[14]

One difference, then, between McClennen's and Gauthier's views has to do with the fact that the former's view is embedded in a much deeper and much more widely ranging theory about the foundations of rational choice. Another difference is that McClennen's view is meant to show how agents whose preferences violate either the weak-ordering or the independence axioms can manage to avoid the pragmatic pitfalls traditionally thought to be inevitable for them, at least according to many supporters of these axioms as legitimate constraints on rational preference and choice. Given the rejection of separability, it is precisely the resolute agent's capacity for resolute choice that enables him to avoid these pragmatic difficulties and, at the same time, create prospects for himself that the sophisticated agent, with his own way of avoiding those same difficulties, cannot enjoy.

Perhaps the most significant difference between their views, however, lies in the theory of the *self* that McClennen develops in his attempt to explain and defend the *feasibility* of resolute choice.[15] The idea, very roughly, is to see the problem illustrated by decisions like the one discussed above as essentially a problem calling for intrapersonal cooperation between different temporal instantiations of the same self. And key to this (admittedly rather bold) metaphysics, of course, is the notion of one such instantiation—or *later self,* as McClennen sometimes puts it—remaining true to, and thus resolute in light of, the previous resolutions of an earlier instantiation (an *earlier self,* as McClennen sometimes says). Hence the aptness of the labels "resolute chooser" and "resolute choice," as McClennen uses them: for the sake of the gains to be had from doing so, an agent resolves, at one point in time, to act in various ways at determinate later points in time, and then sticks to this resolution at those later times, despite the fact that, had the relevant resolutions not been made, this later self would have been required, given his current preferences, to choose in a very different way.

All of this, of course, is rather heady stuff, and in a full and leisurely analysis of McClennen's work, much of it would rightly be the object of rather serious critical concern. Suppose, though, that as applied to cases like the one we have so far been considering, we agree to assume that McClennen's view is right: faced with a decision problem like that illustrated and discussed above, a rational agent would indeed be capable of resolving to choose g_2 over g_1, given the choice, and of actually choosing g_2 over g_1, in light of his decision to go upwards rather than down, if in fact he gets that choice. How is this supposed to help a rational agent situated in the circumstances described by Kavka's Toxin Puzzle?

McClennen appears to think the implications of his general view for the toxin case are quite straightforward. All that someone who accepts the former view would have to do, it appears, if she found herself in a case like Kavka's—or, we might add, like Gauthier's, in the revised Newcomb case—is "decide on a[n] [appropriate] plan and then follow through on it." So far

so good. But what exactly is the plan upon which she would decide, according to McClennen, and how is it that accepting McClennen's theory of resolute choice would make adopting and acting on this plan possible for a rational, reflective individual?

McClennen doesn't say, explicitly, exactly what the elements of the plan he favors would be—perhaps because he thinks it's obvious. On any plausible reading I can think of, though, there's a deep problem with the claim that a reflective, rational individual would be able to resolve to adopt and implement, in the circumstances we are imagining, a plan that would get her the million dollars. To see why, suppose we assume the relevant plan would be analogous to the plan we imagined a resolute chooser adopting in our discussion of the sequential decision problem discussed above. Suppose we assume, that is to say, it would be a plan with two parts, each comprised of an "action" (or *choice*) that the plan tells the agent to make at each of two separate points in time. Then, it seems natural to think of the plan McClennen has in mind as exactly analogous to the plan Gauthier proposes for his version of the Newcomb Problem: first, one is to adopt the intention to drink the toxin, then one is actually to drink the toxin when the time comes. Obviously, since we are granting McClennen that a rational agent will adhere to a plan that she has rationally adopted, we must suppose that if a rational agent could adopt this plan, she *would* adopt it and would resolutely adhere to it as time goes by.

But how could a rational agent, even a resolute agent, adopt such a plan in the circumstances we are imagining? Adopting a plan, as we saw earlier, or committing oneself to it, is equivalent to adopting an intention to act as the plan requires one to act, for the sake of achieving the goal the plan is intended to achieve. On the present interpretation of McClennen's proposed plan, as in the case of the plan proposed by Gauthier, this means adopting an intention to adopt an intention that we must suppose a rational agent could not adopt. After all, as in the case of Gauthier's plan, the intention the agent must adopt an intention to adopt, on this version of McClennen's plan, is an intention to drink the toxin when the time comes. Since drinking the toxin would be irrational, even on McClennen's view, unless doing so were part of a rational plan to which the agent had previously committed herself, we must suppose that *adopting an intention* to drink the toxin would *also* be impossible for an agent who wasn't already committed to a (rational) plan that required her to adopt that intention. But then, if we continue to suppose, as we did above, that a rational agent cannot adopt second-order intentions to adopt first-order intentions that she believes a rational agent could not adopt, it follows that, on this reading of McClennen's proposed plan, a rational agent would not be able to adopt the requisite plan. She would not be able to adopt it, moreover, for exactly the same reason Gauthier's imaginary agent would not be able to adopt the analogous plan in the revised Newcomb case: in order to commit herself to the first stage of that plan, she has to

presuppose the rationality of adopting the very intention the plan is supposed to make it possible for her rationally to adopt.

Perhaps, though, the plan McClennen has in mind is somewhat different from the plan we've been discussing. What, then, might it be? A possible clue, it seems to me, lies in a plausible, alternative reading of the resolute agent's plan, and planning behavior, in the sequential decision case discussed above: instead of imagining him committing himself, in advance of his first choice, to a plan that has two parts—namely, making the first choice, and then making the second if and when he gets the chance—suppose we imagine him simply *making* the first choice, without any previous commitments, while at the same time resolving to make the appropriate second choice (g_2 over g_1) if and when the appropriate time arrives. This, it seems to me, is just as plausible a reading of McClennen's remarks about that first case as the reading we have just been considering, and, moreover, it seems to me to reflect, quite accurately, what actually goes on in many cases in which a person, confronted with a series of possible choices, commits herself to following a certain plan as time goes by: she thinks things over, she identifies what she believes is the best plan (without thereby, as yet, committing herself to it), and then she takes the first step in that plan, resolving, at the same time, to take the further steps without which her choice of the first step would be irrational.

Notice, before we apply this reading to the toxin case, that, whatever problems it faces, this view of what it is to adopt a plan, and then resolutely stick to it, is not at all uncongenial to, or obviously subversive of, McClennen's overall view. On the contrary; it strikes me as a reading of his *general* view that puts the latter in what is, at least pre-analytically, a quite favorable light. If, for example, one puts oneself, while contemplating the sequential decision problem discussed above, in the frame of mind of someone who, having started out as a sophisticated chooser, has just become convinced of the rationality of resolute choice, the most natural way to think of what one will then do, it seems to me, is the way suggested by this new reading: one doesn't first lay out the steps one now proposes to follow, then somehow formally commit oneself to performing them, and then take the first (and later, possibly, the second) step; rather, one sees what one has to do, one takes the first step, resolving at the same time to take the appropriate second step if and when one gets the chance, and then, ideally, one actually takes that second step, given the opportunity, when the time comes. For this, and a host of other conceivable cases, it seems to me, this is a quite plausible conceptualization of what McClennen would have us do, and it is, moreover, a way of conceiving of his view that makes it—again, at least superficially—quite attractive. (One big problem, of course, is in understanding how a rational agent could resolve, while taking the first step, to take the second, given what the latter involves. But this is precisely the problem McClennen is anxious to raise and show that he can answer.)

Return now to the problem of how McClennen's view is supposed to apply to the Toxin Puzzle. We saw above that, on one natural reading of what he is proposing, McClennen's "solution" to the Toxin Puzzle runs into exactly the same difficulty Gauthier's solution ran into, according to our remarks in section II. How, then, does McClennen fare if we read him as making a proposal like the one just sketched, according to which, rather than committing herself *in advance* to both elements of the imagined plan, the agent simply takes the first step—or makes the first choice—while at the same time resolving to take the appropriate second step if and when she gets the chance?

Well, what would the first step be, on this way of understanding McClennen's view? Evidently, it would be choosing to adopt the intention to drink the toxin. After all, that's what's required to get the million dollars, and it's hard to think of any other choice that could function as the first step. Suppose this is right. What's the second step, then, which the agent must, as she takes the first step, resolve to perform when she gets the chance? Obviously, the second step is actually drinking the toxin when the time comes; for the whole point of McClennen's discussion of the toxin case is to show that a rational agent *could* get the million dollars, contrary to what we have argued in section I, but only at the price of actually drinking the toxin when the time comes. But then, on this reading of McClennen's view, the *resolution* that must accompany the first step in the agent's plan—the resolution that *makes* it a plan—is indistinguishable from the first step in that very plan. For what could it be to resolve, at one point in time, to do something at some later point in time, but to *adopt the intention* to perform that act at that later time?

Is this a problem for McClennen, interpreted as we are now interpreting him? Surely it is. To see why, recall why we need to imagine the agent making a resolution to make the *later* choice, on this way of understanding McClennen's view, as she makes the *first* choice. The point is that, without the appropriate resolution, plus the capacity actually to carry out that resolution, the first choice is one a rational agent could not make. In the sequential decision problem discussed above, for example, we saw that the price a sophisticated chooser has to pay to avoid the dynamic inconsistency that undermines the myopic chooser's initial choice is to forgo the prospect the myopic chooser sees himself as facing when he makes that choice. The resolute chooser, by contrast, is able to give himself the latter prospect, which he prefers to its alternatives, only because he is able to make and keep, given the chance, a resolution to make the subsequent choice if and when he gets the chance to do so.

But now, if, as in the toxin case, making the first choice and adopting the requisite resolution are *one and the same act,* we face an obvious problem. For we are supposing a rational agent won't be able to make the first choice—i.e., won't be able to adopt the intention to drink the toxin—unless

she adopts (or makes) the relevant resolution at the same time. If, however, as we are now supposing, that resolution just *is* the act of adopting the intention to drink the toxin, adopting *it* will *also* be impossible for a rational agent, given our other assumptions. And this means that even if, on this way of understanding it, McClennen's general view is sound, it is not a view we can use to solve the Toxin Puzzle—i.e., it's not a view that a rational resolute chooser could use to make herself rich, should she be "lucky" enough to find herself in the circumstances Kavka has imagined.

IV

Where does all this leave us? It's obvious, I think, that anyone interested in a general theory of rationality must take an interest in situations like that exemplified by Kavka's Toxin Puzzle. Less obvious, though, is the fact that reflection on such situations has the potential to affect, quite dramatically, our views about the nature and basis of *rational choice.* What I have been concerned to show above is that, while such reflection does indeed have this potential, neither of two very intriguing recent accounts succeeds in showing that a correct view of such situations supports anything other than the very puzzling conclusions defended in section I: a reflective, basically rational individual really would be at a disadvantage, relative to other, less reflective or less rational individuals, in situations like the one Kavka has described, and, so far as I can see, no amount of further reflection, thoughtful planning, or robust resolution will be able to change the fact that this is so.[16]

NOTES

1. Gregory Kavka, "The Toxin Puzzle," *Analysis* 43 (1983): 33–36.
2. See especially my "Intention, Reason, and Action," *American Philosophical Quarterly* 26 (1989): 283–95. I discuss some of the implications of the view defended here in "Strategic Planning and Moral Norms: The Case of Deterrent Nuclear Threats," *Public Affairs Quarterly* 1 (1987): 61–77, and "On Threats and Punishments," *Social Theory and Practice* 15 (1989): 125–54. For an analysis of Kavka's treatment of the Toxin Puzzle, as well as of some related matters, see also my "On Some Alleged Paradoxes of Deterrence," *Pacific Philosophical Quarterly* 73 (1992): 114–36.
3. Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge, Mass.: Harvard University Press, 1987), *passim,* but especially 15–18. My formulation of Bratman's view here follows the formulation in Farrell, "Intention, Reason, and Action," 289.
4. For a fuller discussion of this point, see my "Immoral Intentions," *Ethics* 102 (1992): 268–86.
5. Note that this argument, which is reworked from the previously published essays mentioned above, depends on the assumption that momentarily adopting a commitment that would, if it lasted, be an intention to perform an irrational act, is not the sort of thing that

is likely, in and of itself, to bring it about that one is insufficiently rational to appreciate the incongruity between that commitment and the action towards which it is directed. (I am indebted to David Velleman for pointing this out to me.) I can think of no reason why we should suppose that this seemingly very plausible assumption is false, especially for cases where the individual in question is, apart from his momentary commitment to acting irrationally, otherwise a basically rational individual; but, obviously, a fuller account of these matters would have to say more on this score. Notice, also, that my argument here is meant to apply only to intentions adopted in what I have elsewhere called "the normal way" ("Intention, Reason, and Action"). Obviously, an agent could avoid the implications of the argument sketched above if he could find a way to make himself unaware of the irrationality of his intended action or sufficiently irrational as not to care about the irrationality of that action. Here I am supposing that these are not things one could do without special help—special drugs, for example, behavioral conditioning, and the like—of the sort we may suppose is ruled out for purposes of the present discussion.

6. "In the Neighborhood of the Newcomb-Predictor (Reflections on Rationality)," *Proceedings of the Aristotelian Society* 89 (1988–89): 179–94.
7. Note that Gauthier makes both of his boxes transparent precisely because he wants it to be clear that the issue he wishes to raise is independent of the controversy between proponents of so-called "causal" and "evidential" theories of rational choice.
8. Recall that I am assuming that one is not allowed artificially to manipulate one's psychological states in certain ways—with appropriate drugs, for example, behavioral conditioning, and so forth. Our claim, above, was that a rational agent would not be able to adopt the relevant sort of intention *"in the normal way"*—i.e., in whatever way it is we adopt our intentions in everyday life, when we have no recourse to special methods for manipulating our beliefs, desires, or intentions.
9. The implications of this argument for Kavka's Toxin Puzzle are obvious, I shall suppose: confronted by the billionaire's offer, one will, if Gauthier is right, commit oneself to a plan whose component parts are (a) adopting the intention to drink the toxin at noon tomorrow, and (b) actually drinking it at noon tomorrow.
10. Gauthier, op. cit., esp. 184 and 189–93.
11. *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge: Cambridge University Press, 1990). McClennen discusses the Toxin Puzzle on 226–31.
12. Ibid., 7.
13. Ibid. The discussion that follows omits McClennen's arguments for the assumption that the upper half of the following tree can be seen as offering *either* the prospect of g_4 or the prospect of g_3, depending on the agent's assumptions. I think it will be obvious why he thinks this is so, but the reader with doubts should consult McClennen, op. cit., 7–13. (Note that I am granting McClennen's claims in this regard only for the sake of argument. My own view, formed as a result of correspondence with Jim Joyce, is that, in the problem illustrated here, *neither* g_3 nor g_4 correctly describes the prospect offered by the upper half of the tree. But this is an issue we cannot pursue here.)
14. Intuitively, we can say that the separability principle directs an agent, at any given point in time, to consider only those consequences of his possible choices that are realizable *at that point in time*. (Cf. McClennen, 12.) For the formal characterization of this principle, see ibid., 120–22.
15. Ibid., 216–18.
16. I am indebted to Don Hubin and Jim Joyce for helpful comments on earlier versions of this paper and to Robert Batterman for invaluable technical assistance. I am especially indebted, though, to Rebekah Kaufman, for a long series of extremely helpful conversations on the subject of this paper, as well as for her written comments and much-needed encouragement as it moved through its various drafts.