"In the Neighborhood of the Newcomb-Predictor (Reflections on Rational Choice)," *Proceedings of the Aristotelian Society* 89 (1988–9), 179–94, and Edward McClennen, *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge: Cambridge University Press, 1990). I discuss these views critically in "Utility-maximizing Intentions and the Theory of Rational Choice," pp. 60–77.

33. The argument that follows was first sketched, in a somewhat protean form, in my "Strategic Planning and Moral Norms," pp. 66–73. I did not there, though, appreciate the fact that my conclusion can be said to be genuinely paradoxical in the sense I now believe Kavka to have had in mind when he used this term.
34. It is not clear to me that it is even correct to say that the agent in a case like this intends harm to the innocent in the event the device fails, but that is not an issue I can pursue here. The point is that, even if we can say this, her situation is crucially different from the situation of someone who does intend to harm the innocent, needlessly, if her threat fails to deter the first strike. The latter individual is committed to performing not just an immoral action but an irrational action, by our earlier argument, and that is what makes it impossible for a rational agent to make the commitment she has made (or to adopt the intention she has adopted). The former individual, by contrast, even if we say that she is committed to the suffering of the innocent in the event her strategy should fail, and hence "intends" the suffering that will ensue if her strategy fails, is not committed to performing what we can so far say is either an immoral or an irrational act.
35. Notice that our argument for P_2 applies to utilitarians and nonutilitarians alike. The point is that no ideally rational and moral individual could adopt the relevant intention, given that we are assuming, for whatever reason, that it is an intention to do what that individual grants it would be morally wrong to do.
36. Farrell, "On Threats and Punishments"; "The Justification of Deterrent Violence," *Ethics* 100 (1990), 301–17; and "Deterrence and the Just Distribution of Harm," *Philosophy and Social Policy* 12 (1995), 220–40, reprinted in *The Just Society*, ed. Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul (Cambridge: Cambridge University Press, 1995).
37. Ibid.
38. Kant does not require that a prior threat have been made in order for retributive punishment to be appropriate. I should note that, contrary to a widespread misconception, the theory of punishment that Kant defends in *The Metaphysics of Morals* is a nonutilitarian deterrence-based theory, not a retributive theory in any ordinary sense of the term. See especially *The Metaphysical Elements of Justice*, tr. John Ladd (Indianapolis: Bobbs-Merrill, 1965). For a relevant and extremely insightful analysis of Kant's theory of punishment, see Sarah Holtman, "Toward Social Reform: Kant's Penal Theory Reinterpreted," *Utilitas* 9 (1997), 3–21.
39. See especially Warren Quinn, "The Right to Threaten and the Right to Punish," *Philosophy and Public Affairs* 14 (1985), 327–73.
40. My argument here is, of necessity, very brief. For a much more careful treatment, including a critical analysis of Quinn's view, see my essay "On Threats and Punishments."
41. For a defense of this claim, see, again, "On Threats and Punishments."

Rethinking the Toxin Puzzle

DAVID GAUTHIER

I

"As our beliefs are constrained by our evidence, so our intentions are constrained by our reasons for action."[1] With Gregory Kavka's conclusion to "The Toxin Puzzle" I have no quarrel. As a rational person, I can intend only what I expect to have reason to do. What follows from this? Kavka notes that "we are inclined to evaluate the rationality of the intention both in terms of its consequences and in terms of the rationality of the intended action" (p. 36). Combining his conclusion with his claim about evaluation, we should infer that an intention is rational if and only if it is directed at an action that would be rational and no alternative intention directed at an action that would be rational has more favorable consequences. And with this I have no quarrel. But we could easily be misled by the way in which I have expressed this inference. For we could suppose that whether an intended action is rational can be determined independently of and prior to considering whether the intention to perform that action has best consequences. And this I deny.

Consider the toxin puzzle. I shall be paid "one million dollars tomorrow morning if, at midnight tonight, [I] *intend* to drink" a vial of "toxin tomorrow afternoon" that "will make [me] painfully ill for a day, but will not threaten [my] life or have any lasting effects" (p. 33).[2] The only problematic feature of this account that need detain us concerns how my intention is to be established. Kavka postulates a "'mind-reading' brain scanner and computing device" that I am to believe "will correctly detect the presence or absence of the relevant intention" (p. 34). But since I doubt that such a machine is possible, I shall fall back on the claim that I am well acquainted with the person who must decide whether to make the payment and am convinced from both firsthand experience and the testimony of others that she is an extraordinarily astute judge of the real intentions of her fellows, so that I should be foolish indeed to think that at midnight tonight I could deceive her about whether I intend to drink the toxin.

Kavka thinks that I have good reason to intend to drink the toxin (since so intending will almost certainly gain me $1 million). He also thinks that I have no reason to drink it (since drinking it will gain me nothing and make me ill for a day). If, as he says, "our intentions are constrained by our reasons for action" (p. 36), then it seems that we must conclude that I cannot (rationally) intend to

drink the toxin. Even though the intention would have best consequences, it is not directed at an action that it would be rational to perform.

But I disagree. I grant, of course, that drinking the toxin does not have best, or even good, consequences, so that I have no *outcome*-oriented reason to drink it. But drinking the toxin is part of the *best course of action* – in terms of its consequences – that I can embrace as a whole. For I do better to intend to drink the toxin, even at the cost of actually drinking it, than not to intend to drink the toxin. And although I should do better still to intend to drink the toxin but not drink it, I cannot embrace this as a single course of action.[3] To be sure, I am "perfectly free to change [my] mind after receiving the money and not drink the toxin" (p. 34). I know this at the outset. And I know that I should like to change my mind. But if I am rational, and understand my situation, this knowledge is of no use to me. Either I suppose that I shall have no reason to drink the toxin, in which case insofar as I am rational I cannot have the mind to do so, or I suppose that I shall have reason to drink it, in which case I can have the mind to do so but no good reason to change it. Changing my mind is not part of a course of action that I can embrace.

Intending to drink the toxin is part of my best course of action. And come tomorrow afternoon, I can and shall still recognize this. Tomorrow afternoon I shall have no ground for doubting that intending to drink the toxin is part of my best course of action, and so I shall not then have good reason to change my course of action. Intending to drink the toxin, I shall drink it. My reason for drinking it will be that drinking it is part of the best course of action that I could embrace as a whole – best not only prospectively, but still best at the time of drinking.

As a rational agent, I can intend only what I expect to have reason to do.[4] I can intend to drink the toxin, because I expect to have reason to drink it – not on account of its own consequences, but because it is part of the course of action with best consequences. The intention to drink the toxin is rational, because it is directed at an action that is rational and has best consequences among intentions so directed. Note in this connection that the intention not to drink the toxin is also directed at an action that is rational – since if I do not intend to drink the toxin I have no reason whatsoever to drink it. But of course, intending not to drink the toxin has consequences that are not as favorable as the consequences of intending to drink it, and so is not rational.

II

Let me spell out the crucial steps in my argument. (A) It is rational for me to form an intention if (i) were I to form it, I should expect to have adequate reason to execute it, and (ii) among alternative intentions satisfying condition (i) it has best consequences. (B) I should expect to have adequate reason to execute an intention if I should expect that, were I to execute it, I should be doing better than had I not formed it.

Consider now the intention to drink the toxin. Were I to form it, I should expect to have adequate reason to execute it, since I should expect to be doing better were I to execute it than had I not formed it. So condition (i) is satisfied. And compared with the alternatives – intending not to drink the toxin or not intending anything (either of which also satisfies condition (i)) – it has best consequences. So it is rational for me to form it.

I claim that this is the correct resolution of the toxin puzzle. On any account, the situation is puzzling – either because one supposes that I cannot form the enriching intention, or because one supposes that I have reason to perform an action that has only undesirable consequences. But the puzzle arises because intending here has consequences independent of those of the intended action but of greater overall significance. We do not easily accommodate our intuitions about what is rational to such situations. We need to reflect on the role that deliberation plays in enabling a person to realize his overall concerns, and to appeal to this role in assessing his reasons for intending and acting. Someone who took himself to have no reason at all to drink the toxin, because he had no *outcome-oriented* reason to drink it, would be deliberating ineffectively in a situation in which what mattered most to the realization of his concerns was not what he did but what he intended.

A person deliberates in order to decide on and realize his concerns. In the toxin puzzle we assume that the only relevant concerns arise from the money that may be gained and the illness that may be incurred, and that the former outweighs the latter. The considerations that weigh with a person in determining what to do in order to realize his concerns are what he takes to be his reasons for acting. But he may be mistaken. His real reasons for acting are those considerations that would weigh with him in deliberation *directed effectively* at the realization of his concerns. These of course include, but are not restricted to, outcome-oriented considerations.

To guard against misunderstanding my account of deliberation, it is essential to emphasize that deliberative reasons relate to effective direction. They are not simply whatever considerations would need to weigh with someone if he is to realize his concerns.[5] If I were subject to a being whose power enables her to control what happens to me and whose astuteness enables her to judge accurately my deliberative procedures, then I might be unable to engage in rational deliberation, directed effectively at the realization of my concerns. For she might see to it that, on the one hand, if I take her directives as reasons for acting in themselves, independently of how they relate to my concerns, then I should do well, whereas on the other hand, if I consider how best to realize my concerns, I should do badly. Deliberation *directed* at the realization of my concerns would then be ineffective, whereas the *effective* realization of my concerns

would depend on deliberation that ignores them. Subjected to such a being, I should do best were I to believe that her directives in themselves afforded me reasons for acting. My belief would of course be false. Her directives would be relevant to my deliberation not in themselves, but only in virtue of her power to relate how well I do to how I deliberate. But were I to believe this, I should do badly.

The pragmatic standard for rational deliberation and reasons for acting that I embrace does not lead to the absurd view that rationality is simply a matter of what in fact pays. A being of sufficient power and astuteness could frustrate any attempt to deliberate rationally and could make a particular form of deliberation pay, even though it was bad or irrational. I shall consider presently whether such a being manifests herself in the toxin puzzle.

III

It is rational, I claim, to form the intention to drink the toxin, and to drink it. More generally, it is rational to form an intention, if one reasonably expects at the time that forming and executing it will better realize one's objectives than not forming it; and it is rational to execute an intention, if one reasonably expects at that time that one's objectives will be better realized after executing it than they would have been had one not formed it. So, if you will confer some benefit on me if you expect that I shall reciprocate, and I do better to receive the benefit and reciprocate than not to receive it, and I believe that offering a sincere assurance that I shall reciprocate is likely to be both necessary and sufficient to give you the expectation that I shall reciprocate, then it is rational for me to form the intention to reciprocate that a sincere assurance requires. And if you do confer the benefit on me, and when the time comes to reciprocate I still judge that reciprocating leaves me better off than I could have expected to be had I not sincerely assured you of my intention to reciprocate, then it is rational for me to reciprocate, even if some other action would then better realize my objectives.

There is an important difference between reciprocation, as I have characterized it, and the situation envisaged in the toxin puzzle. In a situation calling for reciprocation, your concern is not with what I intend but with what I do. Of course you must act before I act, so you cannot make what you do depend on what I do, but you take my assurance of my intention as justifying your belief about what I shall do. You have no concern with my intention as such, but only with its evidential value for my prospective action. But in the toxin puzzle, the concern of the person who offers the $1 million is strictly with my intention, and not at all with my action. As Kavka insists, I am "perfectly free to change [my] mind after receiving the money and not drink the toxin" (p. 34).

Intention plays only a secondary role in reciprocation. Indeed, in many contexts we may dispense with any reference to it. In situations in which one person will benefit another in the expectation of an appropriate return, a rational deliberator will normally take herself to have adequate reason to reciprocate where its cost to her is less than the benefit she receives. For if she characteristically deliberates in this way, then others who know her will expect an appropriate return and will not need to seek any assurance from her of her intention. And when she is asked for an assurance, she can offer it simply because she takes herself to have adequate reason to reciprocate. Rather than forming an intention to reciprocate which then gives her reason to do so, she recognizes a reason to reciprocate on which she can base her intention to do so.

Mutual benefit through reciprocation is a familiar and readily intelligible form of interaction. A person can understand the rationale of engaging in such interaction as a reliable reciprocator, even without valuing the interpersonal relationships that are based, or partially based, in reciprocation. She can see the benefit of her acts of reciprocation, even though it is not *her* benefit. The toxin puzzle does not represent any common mode of interaction. We are not typically willing to reward others for their intentions when we have no interest in or expect no benefit from the performance of the intended actions. Indeed, we may see an attempt to exercise a form of thought control in the offer of such a reward, especially for an intention to act in a way that is harmful to the agent.

But is there such an attempt? Consider an alternative puzzle, which does seem more directly to involve an attempt at thought control. Suppose that what is required of me, if I am to receive $1 million tomorrow morning, is not that at midnight tonight I intend to drink a vial of toxin tomorrow afternoon, but rather that at midnight tonight I believe that I shall have good or sufficient reason to drink a vial of toxin tomorrow afternoon. If, as I have assumed, I can intend only what I expect to have good reason to do, then this revision may seem to make no difference. For if having the belief is necessary for forming the intention, I can be rewarded for having the intention only if I also have the belief. Or rather, this holds insofar as my intention is rational. For I may simply unthinkingly intend to do something that, were I to reflect, I should recognize that I did not expect to have reason to do. But since our concern is with rational agency, we may put this qualification aside.

And now there does seem to be a problem. Considering whether to form the intention to drink the toxin, I reflect on the benefits of intending even at the cost of performing, in relation to not intending, and I make up my mind to drink – or so I claim. But considering whether to believe that I shall have good reason to drink the toxin, what do I do? Do I reflect on the benefits of so believing, even at the cost of performing, in relation to not believing? This does not lead me to adopt the belief. For believing is believing *true,* and reflecting on the benefits of believing that I have reason to drink the toxin seems quite irrelevant to determining whether the belief is true. It may seem plausible to claim that if I would benefit from forming an intention, despite the cost of executing it, then

I have reason to form and (if all turns out as I expect) carry out the intention. But it does not seem plausible to claim that if I would benefit from adopting a belief, despite the cost of acting in accordance with it, then I have reason to adopt and (if all turns out as I expect) act on the belief.

It is not valid to argue: *p*, because it would pay me to believe that *p*. Substituting "I have a reason to drink the toxin" for *p* does not improve the argument. I do not have a reason to drink the toxin because it would pay me to believe that I do. And no other reason to drink the toxin seems in the offing. So if what is required for me to receive $1 million is that I believe myself to have reason to drink the toxin, then it would seem that I am unable, as a rational person, to acquire the belief and gain the $1 million.[6] But this argument moves too quickly. For it ignores the possibility that I can give myself a reason to drink the toxin, and so come rationally to believe that I have such a reason. In some cases *p* may be such that, if it would pay me to believe that *p*, I can bring about *p*, and thereby come rationally to believe that *p*. Is my having a reason to drink the toxin such a case?

Can I give myself a reason to drink the toxin? It may seem that I can do so by forming the intention to drink it. Although in most situations I would form an intention by considering the outcome-oriented reasons for performing the intended act, here I recognize that the intention has consequences of its own, and so I consider the entire course of action – intention and execution. And, as in the original version of the toxin puzzle, it may seem that the best course of action that I can adopt as a whole, whether I consider the choice prospectively or at the time of performance, is to intend to drink the toxin and to drink it. But alas, this is not so. For recall that it is rational for me to intend to drink the toxin only because I expect to have adequate reason to drink it, and that I expect to have adequate reason to drink it because (and, in the circumstances, only because) I expect that I should do better were I to drink it than had I not formed the intention to drink it. The likely effect of my forming the intention is that I gain $1 million. In the revised puzzle, forming the intention does not have this effect. What would gain me $1 million would be believing that I had reason to drink the toxin.

In the situation of the original puzzle, I believe reasonably that, were I to form the intention to drink the toxin, I should have reason to drink it. And having formed the intention, I believe reasonably that I shall have reason to drink the toxin tomorrow afternoon. Forming the intention gives me reason to execute it. But it does so because forming the intention enables me to gain $1 million. My reason for executing the intention is explained by the beneficial effect of forming it. In the revised puzzle this is not so. Forming the intention has no beneficial effect, and so gives me no reason to execute it. I have no basis for believing that I have reason to drink the toxin.

So the earlier analysis is confirmed: in the revised puzzle, I cannot give myself reason to drink the toxin, and so am unable, as a rational person, to acquire the belief that I have such reason. Although neither intentions nor beliefs can be produced to order, there are significant differences in the conditions that must be met for them to be rationally formed, and these are revealed by comparing the original and revised puzzles.

The comparison should help allay the concern that there is any attempt at thought control in the original puzzle. Rewarding someone for the formation of an intention is not rewarding him for the formation of a belief. The deliberation that leads to forming – and executing – the intention does not involve the acquisition of a belief on grounds of utility rather than truth. And it has a parallel in deliberation about reciprocating benefits, even though intention plays a different and lesser role in reciprocation.

IV

But, as I have insisted, the parallel between the toxin puzzle and benefit reciprocation is by no means a complete one. And it is not only the role of intention that differs between the two. There is a deep structural difference, which emerges if we try to conceptualize the toxin puzzle as involving an exchange of benefits. For if we do this, then the benefit that I confer, in return for the $1 million, can be only the formation of the intention to drink the toxin. Drinking the toxin is explicitly dismissed as of no concern or relevance. But if formation of the intention is the "benefit," then it does not directly involve reciprocation. In a situation calling for reciprocation, you are prepared to confer a benefit on me if you expect me to return it. I am the second performer, and the problem lies in establishing my reason for benefiting you. But in the toxin puzzle, the person offering the $1 million is prepared to confer a benefit on me if I have already "benefited" her by forming the intention to drink the toxin. I am the first performer. The problem lies in the peculiar nature of my "performance" – in the fact that I can rationally perform only if I take myself to have reason to carry out a further act that is costly to me and confers no benefit on the other party.

We may contrast the toxin puzzle with a Newcomb problem. In a standard Newcomb problem, a person whose astuteness in judging her fellows makes her an excellent predictor of their choices offers me the opportunity to take only an opaque box, or an opaque box and a second transparent box containing $1,000. If she has predicted that I will take only the opaque box, she has put the familiar $1 million in it; if she has predicted that I will take both boxes, she has put nothing in the opaque box. Here the parallel with reciprocation is much closer. The person offering the choice is prepared to benefit me if she expects that I shall then "benefit" her by taking only the opaque box. Now, the problem is not usually presented in these terms. A person who thought of a Newcomb problem as involving reciprocation, and who correctly understood the rationale for being a reciprocator, would, I think, be disposed to take only the opaque box.

But of course this way of conceptualizing the problem can be undermined if the predictor makes clear that her interest in the situation is simply in the experimental study of choice behavior – that I should not think of myself as benefiting her by taking only the opaque box. (After all, if she thinks that I shall choose only the opaque box, it will cost her $1 million.) Nevertheless, the problem shares the structure of reciprocation situations. Just as it is rational for me to form the intention to make a return, so that you, recognizing my intention, will have good reason to benefit me, so it is rational for me to form the intention to take only the opaque box, so that the experimenter, recognizing my intention, will have good reason to put $1 million in it.

The toxin puzzle is not a Newcomb problem. Consider, then, a different comparison. Sometimes one person acts to benefit another with the expectation of an appropriate return, but also with the recognition that he is incurring a subsequent cost that in itself is unnecessary to the exchange of benefits. You are strapped for cash; I advance you some money, with the explicit expectation that you will make yourself available to house-sit for me sometime next summer. But in advancing you the cash, I knowingly run myself short, so that I have to pay late fees and interest charges on some of my bills, Even with these charges I consider our exchange worthwhile, but I should of course avoid them if I could. Now this situation has a structure similar to that of the toxin puzzle, except of course that when my bills come due I cannot choose but accept the fees and charges, whereas tomorrow afternoon I can choose not to drink the toxin.

Of course, if I were to find myself after all able to pay my bills on time, I would welcome the opportunity to do so. With this in mind, one may be tempted to think that the toxin puzzle provides a similarly welcome opportunity. Just as I expect to incur late fees and interest charges, so I expect to incur a day's illness from the toxin. But when the time to drink comes, I realize with relief that I can avoid the cost. Why should I not take advantage of my good fortune? Why should I not look upon drinking the toxin as an unwelcome aftereffect that, happily, I can avoid?

But there is a crucial difference between the toxin puzzle and the usual situations with unwelcome aftereffects. In our example, I make you an advance in the expectation that I shall unfortunately run myself short, but making the advance is in no way affected by whether I have such an expectation. In the toxin puzzle, I form the intention to drink the toxin in the expectation that I shall drink it, and here forming the intention requires that I have this expectation. If, realizing in advance that drinking the toxin is unnecessary to gaining the $1 million, I think that I therefore have no reason to drink, then I no longer have the expectation that I shall drink, and I cannot rationally form the intention to drink.

An unreflective person, faced with the toxin puzzle, might not think that he will actually be in a position to choose whether to drink the toxin after the decision whether to put $1 million in his bank account has been made and might, then, simply form the intention to drink because it seemed obviously advantageous to do so. Such a person might come to realize that he would have a choice, with nothing to gain by choosing to drink, only after midnight had passed, and he would have every reason then to abandon his intention and so not to drink the toxin. He would have had the intention at the time that mattered and would have no further use for it. But in forming his intention he would not have deliberated in a fully rational way about his situation. Rationally forming an intention requires looking ahead to its execution and considering whether one may expect to have reason to carry it out.[7] We need not suppose that the person who offers to reward the intention to drink the toxin limits her offers to rational deliberators. But the thought that some persons could gain the $1 million without drinking the toxin by misrepresenting the situation as one with an unwelcome aftereffect that at first seems unavoidable but later proves not to be is of no use to those whose correct understanding of the situation prevents this misrepresentation.

It may seem puzzling that someone who has considered her reasons for and against drinking the toxin, in forming the intention to drink it, will have good reason to drink it when the time comes, whereas someone who considers his reasons for and against drinking the toxin only when the time comes will have good reason not to drink it. Surely one's reasons for and against drinking the toxin do not change. And indeed, construed narrowly they do not. What does change is the context within which these reasons are weighed. The person who, in forming her intention, considers her reasons for and against drinking the toxin assesses her course of action – intention and execution – as a whole. She asks herself whether she has good reason to drink the toxin as part of her course of action, and, in concluding that she does, she recognizes that her reasons for forming the intention to drink outweigh her reasons for not drinking. And once she undertakes a course of action as a whole, she must rationally continue to assess her particular actions as part of that whole, unless she comes to have reason to abandon her course of action. On the other hand, the person who considers his reasons for and against drinking the toxin only when it comes time to decide whether or not to drink has not assessed and chosen his course of action as a whole. If he has formed the intention to drink the toxin, he has done so without deliberating about executing his intention. For him, the choice of whether or not to drink is not a choice within a course of action that he has undertaken. And so he considers only his reasons for and against drinking the toxin considered in itself, concluding, of course, that he has good reason not to drink.

V

The nature of the situation exemplified by the toxin puzzle, as it emerges from our discussion, is this: if I benefit the other party, then she will benefit me in

return; the benefit I confer on her leads to an aftereffect for me that is unwelcome (though worth the benefit I receive), which I am aware of, which I know I must choose to bring about or avoid, but which, if I am to confer the benefit, I must intend to bring about. In the toxin puzzle, the benefit is of course the intention to bring about the unwelcome aftereffect, but what is essential, I propose, is only that the benefit requires that I have this intention, not that the benefit be the intention itself.

When we generalize the toxin puzzle in this way, we can use it to illustrate the role of future-directed intentions in rational deliberation. Usually, no doubt, a person adopts such an intention on the basis of what she expects to be her reasons for performing the intended act. She deliberates, as Michael Bratman says, "about what to *do* then, not what to intend now, though of course a decision about what to do later leads to an intention now so to act later."[8] But "usually" is not "always," and in situations such as that of the toxin puzzle a person may adopt an intention on the basis of her present reasons for performing an act that requires it. Now, of course she may not rationally ignore what she expects to be her reasons for – and against – performing the intended act. Nor may she simply treat her present reasons for performing the act requiring the intention as overriding those reasons. She must consider, in her deliberations, both the intention-requiring act and the intention-executing act – and it is important in treating the general case to recognize that the required intention may be conditional, so that she will have to choose whether or not to execute it only if some condition is satisfied.

Deliberating from her present standpoint, she may first suppose that she has good reason to perform the intention-requiring act if, taking for granted that she would perform the intention-executing act should the question of performing it arise, she expects to do better than if she performs any alternative act. But to deliberate in this manner is to assume that, should the question of performing the intention-executing act arise, she will have adequate reason to perform it simply because she *expected* to do better at the outset by performing the intention-requiring act. ***But this expectation may have been falsified.*** It may be that she expected to do better because she expected that the question of performing the intention-executing act would not arise.

Consider this variant of the toxin puzzle. Suppose that persons with an extremely rare genetic configuration would be permanently disabled by drinking the toxin. I have no reason to believe that I have this rare configuration – the odds against it are 10 million to 1 – and I think the minuscule risk of being disabled by the toxin worth running in order to gain $1 million. But tomorrow noon, before I actually decide whether to drink the toxin, a doctor will examine me to determine whether I have this configuration. If I were found to have it, then I should be stark bonkers to go ahead and drink the toxin, whether I have $1 million in my bank account or not. I should be far worse off were I to drink

than if I had not formed the intention to do so. Even if it would be rational for me to drink the toxin, knowing that I had a 1 in 10 million chance of being disabled by it, it would not be rational, knowing that I was the 1 in 10 million who would be disabled by it. And realizing all this now, I cannot rationally form an intention to drink that would extend to the case in which I were found to have the adverse genetic configuration.

If the expectation that I shall do better to perform the intention-requiring act than any alternative is falsified at the time that I must decide whether to perform the intention-executing act, so that I should not only do better not to perform it, but have done better never to have performed the intention-requiring act, then it is not rational for me to perform the intention-executing act. But I can form an intention rationally only if I expect to have reason to execute it, and so I cannot form an intention rationally if I am aware that it would apply to circumstances in which I should do worse executing it than had I not adopted it. In the toxin puzzle, I can rationally intend to drink only in circumstances in which I should expect to do better to drink than had I not formed the intention to do so.

Rational deliberation concerning future-directed intentions thus must consider both the formation and the execution of the intention. At the time of formation, may one rationally expect to do better overall by performing the intention-requiring act than by performing any alternative? At the time of execution, may one rationally expect to do better by performing the intention-executing act than one would have done had one not performed the intention-requiring act? Deliberatively, the second question must be resolved prior to the first. Only intentions that one expects one would do better to execute than one would have done had one not formed them are eligible for adoption.

The toxin puzzle may still occasion unease. Is it – can it be – really rational to drink the toxin, when all that one accomplishes by drinking it is to make oneself ill for a day? Yes, indeed it can, in the quite unusual circumstances in which the question whether to drink arises. I have tried to show that deliberation about the formation of the intention to drink the toxin, and about the subsequent execution of the intention, may be accommodated in a more general account of deliberation about future-directed intentions. This more general account has a pragmatic rationale in the role that deliberation plays in directing persons to act in ways that best fulfill their overall concerns. Good deliberators should drink up!

Notes

1. Gregory S. Kavka, "The Toxin Puzzle," *Analysis* 43 (1983), 33–6, at 36. Cited hereafter in parentheses in the text, with page number.
2. I have substituted the first for the second person.
3. More precisely, I cannot do this straightforwardly. I might of course be able to arrange to be hypnotized so that I would intend to drink the toxin, and then to be released

from the hypnosis before actually drinking it. And if such hypnosis were available at a cost less than that of a day's illness, no doubt I should do well to avail myself of it. But it need not be – and for the purposes of the present argument we may assume that it is not – available.

4. I take this to express a conceptual truth about intention: an agent rationally intends to do only what she expects that it will be rational for her to do. For present purposes I must leave this as an assumption of my argument.
5. Thus, what I said in another essay – "deliberative procedures are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible" – needs emendation. As a first approximation, we might say that deliberative procedures are rational if and only if they are effectively directed to making one's life go as well as possible. David Gauthier, "Assure and Threaten," *Ethics* 104 (1994), pp. 620–721, 701.
6. Recall that I am assuming that unorthodox methods of belief acquisition, unrelated to the truth of the belief acquired, such as hypnosis, are unavailable.
7. An agent who formed her intentions without looking ahead to their execution and considering whether she might expect to have reason to carry them out would not, in general, be forming them in a way effectively directed to realize her concerns. To be sure, she would do better in the unusual circumstances of the toxin puzzle. One might, then, think that a truly rational agent would normally form her intentions while looking ahead to their execution but would refrain from doing this if faced with a situation such as the toxin puzzle. But alas, she could realize the benefits of refraining only after she had looked ahead. And, as rational, she could intend only what she would expect to have reason to do.
8. Michael E. Bratman, *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987), p. 103.

Toxin, Temptation, and the Stability of Intention

MICHAEL E. BRATMAN

I. Instrumentally Rational Planning Agency

We frequently settle in advance on prior, partial plans for future action, fill them in as time goes by, and execute them when the time comes. Such planning plays a basic role in our efforts to organize our own activities over time and to coordinate our own activities with those of others. These forms of organization are central to the lives we want to live.[1]

Not all purposive agents are planning agents. Nonhuman animals who pursue their needs and desires in the light of their representations of their world may still not be planning agents. But it is important that we are planning agents. Our capacities for planning are an all-purpose means, basic to our abilities to pursue complex projects, both individual and social.

Why do we need to settle on prior plans in the pursuit of organized activity? A first answer is that there are significant limits on the time and attention we have available for reasoning.[2] Such resource limits argue against a strategy of

An earlier version of this essay was presented at the conference held in honor of Gregory Kavka ("Rationality, Commitment, and Community," Feb. 10–12, 1995, University of California, Irvine). A revised version was presented at the March 1995 Pacific Division meeting of the American Philosophical Association, and parts of that version were presented in my 1995 Potter Lecture at Washington State University. The present essay is a substantially revised version of the APA paper. A number of the ideas in this essay were also presented, and usefully criticized, in yet-earlier papers given at Yale University, the University of North Carolina at Chapel Hill, NYU, Rutgers University, Johns Hopkins University, the University of Maryland, and the University of Arizona. The paper, in very roughly its present form, was presented and usefully criticized in March 1996 at Davidson College and Duke University, and at the University of California at Berkeley School of Law Workshop on Rationality and Society in November 1996. I have greatly benefited from the comments and criticisms of many people, including Bruce Ackerman, Nomy Arpaly, Lawrence Beyer, John Broome, Daniel Farrell, Claire Finkelstein, Gerald Gaus, Olav Gjelsvik, Jean Hampton, Gilbert Harman, John Heil, Thomas Hill, Frances Kamm, Keith Lehrer, Edward McClennen, Alfred Mele, Elijah Millgram, Christopher Morris, Michael Pendlebury, John Pollock, Samuel Scheffler, Tim Schroeder, David Velleman, and Gideon Yaffe. I have learned a lot from a series of exchanges – formal and informal – with David Gauthier. Special thanks go to Geoffrey Sayre-McCord for a long and extremely helpful discussion. Final work on this essay was completed while the author was a Fellow at the Center for Advanced Study in the Behavioral Sciences. I am grateful for financial support provided by the Andrew W. Mellon Foundation.