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.

constantly starting from scratch – they argue against a strategy of never treating prior plans as settling a practical question. A second answer is that our pursuit of organization and coordination depends on the predictability to us of our actions.[3] Coordinated, organized activity requires that we be able reliably to predict what we will do; and we need to be able to predict this despite both the complexity of the mechanisms underlying our behavior and our cognitive limitations in understanding those mechanisms. In treating prior plans as settling practical questions we make our conduct more predictable to cognitively limited agents like us by simplifying the explanatory structures underlying our actions.

Intelligent planning agents may differ in their desires, cares, commitments, and concerns. They may, in particular, endorse various noninstrumental, substantive ideals of steadfastness, of sticking to one's prior plans in the face of challenge.[4] But we can ask what "instrumental rationality" – rationality in the pursuit of one's desires, cares, commitments, and concerns – requires of planning agents, despite possible differences in those desires, cares, commitments, and concerns.[5] A theory of instrumentally rational planning agency may not exhaust all that is to be said about rational intentions and plans. But it will, if it is successful, characterize important structures of rational planning agency that are, as it is said, neutral with respect to diverging conceptions of the good.

Such a theory needs to be responsive to a fundamental tension. On the one hand, a planning agent settles in advance what to do later. On the other hand, she is an agent who, whatever her prior plans, normally retains rational control over what she does when the time comes. Following through with one's plan is not, after all, like following through with one's tennis swing. We need to do justice to both these aspects of planning agency.[6]

II. A Basic Model

A planning agent, we may suppose, has a background of values, desires, cares, and concerns.[7] These support considered rankings of various kinds of alternatives, in light of relevant beliefs. I will call such rankings *evaluative rankings*. These rankings express the agent's considered ordering at a time, an ordering she sees as a candidate for shaping relevant choices.

A planning agent is in a position to have an evaluative ranking of alternative actions available beginning at a given time. She is also in a position to have an evaluative ranking of alternative plans for acting over time, and of alternative general policies. As a planning agent she will sometimes decide on a future-directed plan or policy. This involves settling on – and so, in an important sense, being committed to – ways of acting in the future. She might, for example, settle on a detailed plan for an anticipated job interview; and she might settle in advance on a general policy about, say, alcohol consumption. In settling on such

plans or policies she comes to have relevant intentions to act in specified ways in specified future circumstances.

By settling now what she will do later a planning agent puts herself in a position to plan appropriate preliminary steps and means, and to filter options that are incompatible with planned action. This will work only if her plans are to some extent stable and she is not constantly reconsidering her prior decisions – not constantly starting from scratch. A theory of instrumentally rational planning agency is in part a theory of intention and plan stability: a theory of when an instrumentally rational planning agent should or should not reconsider and abandon a prior intention.

Many important issues about the rational stability of prior intentions concern appropriate strategies, given limitations of time and attention, for responding to unanticipated information that one's prior planning did not take into account.[8] Perhaps I settled on a plan for an interview on the assumption that Jones would be the interviewer. When I get to the interview, I discover that Smith has taken her place. Should I stop to reconsider and, perhaps, replan? Here issues about our resource limits and the costs of reconsideration and replanning loom large. And here it seems natural to have a broadly pragmatic, two-tier model: we seek general habits and strategies of reconsideration that are, in the long run, effective in the pursuit of what we (rationally) desire. In a particular case we reasonably implement such pragmatically grounded general habits and strategies and, depending on the case, reconsider or refrain from reconsideration. This means that a planning agent may sometimes rationally follow through with a prior plan even though she would have rationally abandoned that plan if she had reconsidered it in light of relevant unanticipated information.

What about, in contrast, cases in which one's circumstances are, in all relevant respects, those for which one has specifically planned? Here it may seem natural simply to say that if one's plan was rational when formed, then surely it would be rational, barring relevant unanticipated information or change in basic desires or values, to execute it in those circumstances for which one specifically planned. But the issues here are complex.

III. Autonomous Benefits: Toxin

Begin with Gregory Kavka's ingenious toxin case.[9] A billionaire has access to a technology that allows her to discern other people's intentions with almost flawless accuracy. She credibly offers to give me a lot of money on Tuesday if I form the intention on Monday to drink a disgusting but nonlethal toxin on Wednesday. I would be more than willing to drink the toxin in order to get the money. However, to get the money I do not need to drink the toxin; I just need to intend on Monday to drink it. But I need to arrive at this intention in a clear-headed way and without exploiting any external mechanisms (e.g., a side bet).

I would love to form this intention, but I have a problem. The benefit of the intention is, to use Kavka's term, "autonomous":[10] It does not depend causally on my actually executing the intention. I know that when Wednesday arrives I will already either have the money or I will not. In either case it seems that on Wednesday I will have no good reason to drink the stuff, and a very good reason not to, in precisely the circumstances in which I would have planned to drink it.[11] So it is not clear that I can rationally form the intention in the first place, despite its autonomous benefits.[12]

There are two ideas in the background here. The first is a principle that links the rationality of a prior intention with the rationality of the later retention and execution of that intention. We may state this as a constraint on rational, deliberation-based intention: If, on the basis of deliberation, an agent rationally settles at t_1 on an intention to A at t_2 if (given that) C, and if she expects that under C at t_2 she will have rational control of whether or not she A's, then she will *not* suppose at t_1 that if C at t_2, then at t_2 she should, rationally, abandon her intention in favor of an intention to perform an alternative to A. Call this statement of a link between rational intention formation and supposed rational intention retention the *linking principle.*[13]

Second, there is the common idea that the instrumental rationality of an action, in the kind of no-unanticipated-information cases of interest here, depends on the agent's evaluative ranking at the time of the action of options available then. Call this the *standard view.* On Wednesday my evaluative ranking will favor nondrinking over drinking. We infer, given the standard view, that even if I had earlier decided to drink toxin, when Wednesday arrives and the money is in the bank, instrumental rationality would require nondrinking. But then, given the linking principle and given that I am aware of these features of the case, I cannot rationally form the intention to drink in the first place. So there is a problem for rationally settling in advance on an intention to drink toxin despite the attractions of the autonomous benefit, a problem traceable to the joint operation of the linking principle and the standard view.

There is a complication. We have been supposing that in my prior deliberation on Monday about whether to drink toxin on Wednesday I take into account the autonomous benefit of an intention to drink. Does this mean that in my deliberation on Monday I am deliberating directly about the plan: intend on Monday to drink toxin on Wednesday, and then on Wednesday drink toxin? The problem with this is that my deliberation on Monday seems instead to be about what to do on Wednesday, not what to intend on Monday, though I know that a decision on Monday about what to do on Wednesday would involve an intention on Monday so to act. Granted, I might deliberate directly about whether on Monday to cause myself to intend to drink on Wednesday – for example, by taking a certain pill, or engaging in self-hypnosis. But that is a different matter.

Suppose, however, that on Monday I am directly deliberating about whether to drink toxin on Wednesday. I am not deliberating directly about the intention, on Monday, to drink. But this does not by itself show that I cannot include in my reasons for deciding to drink the fact that if I do so decide I win the money, whereas if I instead decide not to drink toxin I do not win the money.[14] The barrier to winning the money in the toxin case is not a simple exclusion of the consideration of autonomous benefits in deliberation. If there is a barrier, it is, rather, the combination of the linking principle and the standard view.

IV. Autonomous Benefits: Reciprocation

Consider a second example. You and I are mutually disinterested, instrumentally rational strangers about to get off an airplane. We know we will never see each other – or, indeed, the other passengers – again. We also know that we each have a pair of suitcases, and that each of us would benefit from the help of the other in getting them down from the overhead rack. We each would much prefer mutual aid to mutual nonaid. Given our seating arrangements, however, you would need to help me first, after which I could help you. You will help me only if you are confident that I would, as a result, reciprocate. But we both see that once you help me I will have received the benefit from you that I wanted. My helping you later would, let us suppose, only be a burden for me. Of course, most of us would care about the plight of the other passenger, and/or have concerns about fairness in such a case. But let us here abstract away from such concerns, for our aim here is to determine what is required solely by instrumental rationality. Let us also suppose, again artificially, that my helping you or not would have no differential long-term effects (including reputation effects) that matter to me now. Given these special assumptions, it seems I would not have reason to reciprocate after you have helped me. Seeing that, you do not help me, so we do not gain the benefits of cooperation.[15]

In such a situation I might try to assure you I would reciprocate. But suppose I am not very good at deceit and will only be convincing to you if I really intend to reciprocate if you help me.[16] I would, then, very much like to provide a sincere assurance. Can I?

If an assurance from me issued in a moral obligation to reciprocate, then perhaps I could get a new reason to reciprocate simply by issuing such an assurance. But let us put direct appeal to such moral considerations to one side and see where the instrumental rationality of mutually disinterested agents would by itself lead. To achieve the benefits of a sincere assurance I must intend to reciprocate. But in the very special circumstances described, it seems that I will have, when the time comes, no reason to reciprocate. So it seems I cannot rationally intend to reciprocate, and so cannot gain the benefits of cooperation in such circumstances.

The point is not that there are no relevant considerations of fairness or of

assurance-based obligation. The point is only that we may not get at such reasons, when, in such special cases, we confine our attention solely to instrumentally rational planning agency.

So we have two autonomous benefit cases: toxin and reciprocation. In each case I consider at t_1 whether to act in a certain manner (drink the toxin, help you if you have helped me) at t_2. I know that my so intending at t_1 would or may well have certain benefits prior to t_2 – my becoming richer; my being aided by you. But I also know that these benefits would be autonomous: they would not depend causally on my actually doing at t_2 what it is that at t_1 I would intend to do then. The execution at t_2 of the relevant intention would bring with it only the burden of being sick or of helping you.[17] In each case, however, I prefer throughout the package of autonomous benefit and burden of execution over a package of neither.

Given my considered preference for such a package, can I in such cases rationally settle at t_1 on a plan that involves (conditionally or unconditionally) so acting at t_2? The linking principle tells us that rationally to settle on such a plan I cannot suppose that at t_2 I should, rationally, abandon my intention concerning t_2 – the intention to drink, or to reciprocate. But I know that under the relevant circumstances at t_2 my ranking would favor not drinking, or not helping. The standard view, then, says that at t_2 I should not execute my prior intention. If that is right then, given the linking principle, I cannot rationally and in a clearheaded way decide on the plan in the first place. Instrumental reason is an obstacle to gaining the autonomous benefits in such cases, even though I would gladly drink the toxin, or help you, in order to achieve those benefits.[18]

Is that right?

V. Sophistication and Resolution

An agent who adjusts her prior plans to insure that what she plans to do will be, at the time of action, favored by her then-present evaluative rankings has been called a "sophisticated" planner.[19] Given the conjunction of the linking principle with what I have termed the standard view, an instrumentally rational planning agent will be sophisticated, and so will not be in a position deliberatively to form the intention needed to get the autonomous benefits in our two cases.

The intuition that, to the contrary, instrumental rationality should not always stand in the way of such autonomous benefits suggests an alternative approach, one that retains the linking principle but abandons the standard view. The basic idea is that if it was best in prospect to settle on a prior plan, and if there is no unanticipated information or change in basic values, then it is rational to follow through with that plan in those circumstances for which one specifically planned. Settling on a plan to drink toxin, or to reciprocate if helped, might well be best in prospect, given relevant autonomous benefits. So it may be rational

to follow through with such a plan in planned-for circumstances, even though, at the time of follow-through, one would thereby be acting contrary to one's then-present evaluative ranking of one's then-present options. So, instrumental rationality need not stand in the way of the money or of mutual aid. Borrowing a term from Edward McClennen, we may call this a version of "resolute choice."[20] In anticipation of a later distinction I will call it, more specifically, *strong resolution.*

Call the conjunction of the linking principle and the standard view *sophistication.* Both sophistication and strong resolution accept the linking principle, and both allow one to consider autonomous benefits in deliberation about plans for the future. Where they differ is in their view of rational intention retention and execution. Sophistication accepts the standard view; strong resolution says, instead, that a prior plan settling on which was – because of autonomous benefits – best in prospect, can trump a later, conflicting evaluative ranking concerning planned-for circumstances.

To these two approaches we may add a third, a qualified form of resolution defended by Gauthier.[21] Suppose that settling on a certain prior plan at t_1 is favored by one's evaluative ranking then. Suppose that the attractions of settling on this plan include expected autonomous benefits prior to t_2; and suppose that this plan specifically calls for one to *A* at t_2, given circumstance *C.* Suppose *C* does obtain at t_2; and suppose there is neither unanticipated information about this circumstance nor change in basic values. Gauthier proposes that one should stick with the plan if and only if one thereby does better than one *would have* done if one had not settled on the plan in the first place, at t_1.

This view qualifies strong resolution with a further, counterfactual test on rational follow-through. There are autonomous-benefit cases in which strong resolution would call for follow-through and yet Gauthier would not. These include certain cases of failed threats.[22] In the toxin case and the reciprocation case, however, Gauthier's view matches strong resolution. In drinking toxin or reciprocating one does better than one would have done if one had initially not settled on the plan to drink or to reciprocate. So one may rationally drink the toxin, and one may rationally reciprocate. Let us call Gauthier's view *moderate resolution.*

Both strong and moderate resolution focus on the evaluation of courses of action, as individuated by the agent's intentions and plans.[23] Strong resolution treats the prior evaluation of a course of action as critical, allowing it, in certain cases, to determine the rationality of later follow-through in planned-for circumstances. Gauthier adds a further, counterfactual test on rational follow-through, a test that concerns the comparative evaluation at the time of action of the overall course of action. But both views agree that if one's intentions and plans see one's conduct at t_2 as fitting into a larger course of action that began at t_1 (perhaps only with a decision), then it is the assessment of that larger course

of action, a course of action some of which is already in the past, that is crucial. Strong resolution highlights the assessment at t_1 of that course of action; at t_2 one refers back to that assessment. Gauthier adds a role for a comparison at t_2 of that course of action with its t_1 through t_2 alternatives. But both agree that it is the overall course of action that is one's concern, even at t_2. That is why, for Gauthier, one should follow through and reciprocate if one has been helped; for the course of action that began at t_1 with a sincere assurance that one would reciprocate if helped, and then includes reciprocating after having been helped, is seen at t_2 as superior to alternative courses of action available beginning at t_1.[24]

One evaluates, then, not simply alternatives from now on but courses of action, as individuated by one's intentions and plans. These courses of action can include elements already in the past, elements over which one no longer has causal control. On both views, then, intentional structure can trump temporal and causal location.

This is in tension with a basic fact about our agency. As time goes by we are located differently with respect to our plans. Along with a change in temporal location normally goes a change in the agent's causal powers. What is up to the agent is what to do from now on. So she will normally want to rank alternatives beginning from now on.

Granted, the agent may well rank her alternatives with respect to past events: she may, for example, be grateful for past benefits or want revenge for past harms. The point is not that a rational agent does not care about the past. The point concerns, rather, what is now under the control of the agent. What is now under her control are her alternatives from now on.[25] So it seems she will want to rank those alternatives. Both versions of resolution concern themselves instead with courses of action that typically include elements no longer in the agent's causal control. This seems to me not to do justice to the significance of temporal and causal location to our agency. Strong and moderate resolution, in seeking a strong role for planning in achieving the benefits of coordination over time and across agents, seem not to do justice to the basic fact that as agents we are temporally and causally located.

A reply will be that in giving such priority to intentional structure a resolute agent employs a deliberative procedure that is, in the words of Gauthier, "maximally conducive to one's life going as well as possible."[26] One who employs such a procedure will win money in toxin cases and gain benefits of cooperation. But it is difficult to see why this shows that *at the time of action* one will not reasonably consult one's ranking of options that are at that time in one's control. If one is concerned with what is "maximally conducive to one's life going as well as possible," why wouldn't one be concerned with which action, of those presently in one's control, is "maximally conducive to one's life going as well as possible"? Faced with the toxin on Wednesday, however, the action

presently in one's control that is maximally conducive to that benefit is, we may suppose, not drinking.

VI. Temporary Reversals in Rankings

I am skeptical, then, about strong and moderate resolution.[27] But I also think that sophistication is too simple.

Consider Ann. She enjoys a good read after dinner but also loves fine beer at dinner. However, she knows that if she has more than one beer at dinner she cannot concentrate on her book after dinner. Prior to dinner Ann prefers an evening of one beer plus a good book to an evening with more than one beer but no book. Her problem, though, is that each evening at dinner, having drunk her first Pilsner Urquell, she finds herself tempted by the thought of a second: for a short period of time she prefers a second beer to her after-dinner read.[28] This new preference is not experienced by her as compulsive. If asked, she will say that right now she really prefers to go ahead this one time and have the second drink, despite the impact on her ability to concentrate later, though she will also acknowledge that even now she prefers that she resist similar temptations on future nights. As she knows all along, this change in ranking will be short-lived: after dinner she will return to her preference for a good read.

Prior to dinner on Monday Ann prefers

(1) one beer at dinner on Monday plus a book after dinner

to

(2) more than one beer at dinner on Monday and no book after dinner.

In the middle of dinner, after her first beer, this preference reverses, and she prefers (2) over (1). By the end of dinner, she again prefers (1) over (2), though by then this preference will express itself either in relief or in regret. Throughout dinner, however, Ann continues to prefer

(3) one beer at dinner and a book after, for all nights,

to

(4) more than one beer and no book for all nights.

But it is also true that during dinner on Monday Ann temporarily prefers

(3') more than one beer and no book on Monday, but one beer and a book on all other nights

to (3).[29]

What is Ann to do? We might say to Ann: "You should settle in advance on a policy of having at most one beer at dinner and then stick with that policy in

the face of expected temptations. In that way you will achieve (3) rather than (4), thereby satisfying a preference you will have throughout. Granted, on each night there will be a slightly modified policy you will prefer to (3). On Monday, for example, you will prefer (3') to (3). But this will be only temporary. The preference that will persist throughout is for (3) over (4). By settling on a policy in favor of (3), that is what you can achieve."

Might this be sensible advice? Might Ann rationally settle on such a policy and then rationally stick with it in the face of a diverging preference?

Note that we are not asking whether it is always rational to resist all temptations. Nor are we supposing that if, in a particular case, it would be rational to stick with such a prior policy, that very fact ensures that one does. Note finally that I am understanding evaluative rankings as aspects of the real, explanatory story of action. Although such rankings are susceptible to a broadly functional characterization, they are not merely the reflection of actual choice and action. In this sense of "ranking" it is possible for Ann intentionally to stick with her policy and to act contrary to her present ranking (even though there is also a sense in which, if she so acts, that is what she most wanted).[30] Our question is whether this may be rational.

Sophistication answers in the negative. Despite her prior one-beer policy, at dinner Ann prefers a second beer. So, given the standard view, that is what instrumental rationality requires. A sophisticated Ann cannot even settle on the one-beer policy in the first place.

Such a blanket prohibition on settling on and sticking with such policies in the face of temporary rankings to the contrary seems to me mistaken. It seems to me that instrumentally rational willpower sometimes involves sticking with a sensible prior policy in the face of a diverging temporary preference. Can we make theoretical room for rational willpower of this sort?

We might distinguish here between a reversal of a mere preference and a reversal of an evaluative ranking.[31] Only the latter, we might say, trumps a prior policy. This may work for some cases of temptation, but I do not think that it does justice to all that is at stake. First, some versions of Ann's case may involve temporary changes in evaluative ranking. And second, for reasons that will emerge, there remain important issues about a planning agent's concern with her own future assessments.

Ann's case is in some respects similar to the toxin and reciprocation cases. In all three cases there is a prior plan or policy settling on which is best in prospect. And in all three cases the agent knows that when the occasion for action arrives her rankings of then-present options will argue against following through. But there is also a significant difference between the cases. The underlying desires and values that argue for abandoning a plan to drink toxin, or a plan to reciprocate, are stable. Ann's preference for two beers, in contrast, is temporary. I want to see whether an account of instrumentally rational planning

agency should exploit this difference and, if so, how. But first I need to look at a different kind of case.

VII. Slippery-Slope Intransitivities

Consider Warren Quinn's example of the potential self-torturer who agrees to allow an extremely tiny medical device to be permanently attached to his body. The device generates a constant electric current of varying levels, from 0 (no current) to 1,000 (extremely high and extremely painful current). Each increment, from setting n to setting $n + 1$, is so small that he cannot feel the difference, though he can of course feel the difference between setting 0 and setting 1,000. The device begins at setting 0, and the potential self-torturer is given an initial ten thousand dollars for allowing it to be attached. He is also offered ten thousand dollars for each advance in the setting (something he can choose once each week) from setting n to setting $n + 1$, though he knows that once the device is advanced to a higher setting it cannot be returned to a lower setting.

This poses a problem:

> Since the self-torturer cannot feel any difference in comfort between adjacent settings, he appears to have a clear and repeatable reason to increase the voltage each week. The trouble is that there *are* noticeable differences in comfort between settings that are sufficiently far apart. Indeed, if he keeps advancing, he can see that he will eventually reach settings that will be so painful that he would then gladly relinquish his fortune and return to 0.[32]

This potential self-torturer has intransitive preferences.[33] He prefers setting 1 to setting 0, setting 2 to setting 1, and so on. But he prefers setting 0 to setting 1,000. Further, these intransitive preferences are there all along. This is not a case of preference change over time, though, once the process gets going, different preferences are engaged at different times.

What is the potential self-torturer to do? Quinn suggests he should decide in advance on a "*reasonable* stopping point" and then stick with it when he gets there.[34] In that way he gets more than enough money to compensate for the discomfort but does not find himself in unacceptably extreme suffering.[35]

This is good advice. But to follow this advice the agent will need to stick with his prior decision in the face of a stable preference to go on. Suppose that he prefers 15 to 0, and 0 to 16: 15 is, so to speak, the *switch point relative to 0.* Suppose that for this reason the agent decides in advance to stop at this switch point – to stop at 15.[36] When he gets to 15 he will prefer to move to 16, and that is a preference that was there all along. In order to stick with his prior decision he must act contrary to his ranking of 16 over 15, and that would violate the standard view and so sophistication.

I have rejected strong and moderate resolution as applied to our autonomous-

benefit cases. For these cases sophistication is a superior response to the fact that our agency is located temporally and causally. But as a view about cases of temptation, and of slippery-slope intransitivities, sophistication seems overly simple. We need to steer a path between resolution and sophistication.

VIII. Planning Agency and Future Regret

Ann prefers, at the time of action, to drink a second beer; I prefer at the time of action not to drink toxin. Given, in each case, a prior intention to the contrary, why should Ann's drinking a second beer be a potential candidate for rational criticism whereas my refraining from drinking toxin is not? Why should rational intention stability distinguish in this manner between toxin and temptation?

Suppose that you are an adviser to Ann, or to the potential self-torturer. You might well say: "Stick with your plan or policy. If you do, you will be glad you did. And if you do not, you will wish you had." We can spell this out as an argument offered at the time of action:

(a) If you stick with your prior intention, you will be glad you did.
(b) If you do not stick with your prior intention, you will wish you had.

So, other things being equal,

(c) Though you now prefer to abandon your prior intention, you should nevertheless stick with it.

Statement (a) says, roughly, that the agent would not regret sticking with her prior intention; (b) says, roughly, that she would regret not sticking with her intention. Let us say that when (a) and (b), suitably interpreted, are true, following through with one's prior intention satisfies the *no-regret condition.*[37] Sophistication (since it accepts the standard view) holds that in our no-unanticipated-information cases it is instrumentally rational to follow through with one's prior intention to A at t only if one's evaluative ranking at t favors A over one's other options at t. But consideration of the no-regret condition suggests an alternative view: in the kind of no-unanticipated-information cases we are considering, the agent's reasonable anticipation at the time of action that follow-through would satisfy the no-regret condition can sometimes make follow-through rational even in the face of a present ranking to the contrary.[38]

The agent, then, is to ask at the time of action, t_2, about her attitude at some appropriate later time, t_3, concerning options still available at t_2. I will say more shortly about what counts as an appropriate later time. Note, though, that the anticipated attitude at t_3 that is at issue concerns options still in one's power at t_2. We are not considering one's anticipation at t_2 of an assessment at t_3 of overall courses of action beginning *earlier* than t_2. We want the options being assessed to be options still available to the agent at the time of plan follow-through, t_2, for we are trying to be responsive to the fact that agency is located temporally and causally.

Such a view would continue to subscribe to the linking principle, but it would reject the standard view for some cases like that of Ann or of Quinn's potential self-torturer. I want to spell out how such a view would work.

Begin with Ann. She knows that she will be glad after dinner if she has stuck with her policy and had only one beer; and she knows that after a second beer, faced with the later part of the evening, she will wish that she had stuck with her one-beer policy. So she knows that in following through with a one-beer policy she would satisfy the no-regret condition.

The case of the potential self-torturer is more complicated. We need first to ask how far into the future he is to look. After all, very shortly after moving from 15 to 16, he may still be glad he gave up on a prior intention to stop at 15. However, at the time of his choice between 15 and 16, he can ask: "If I abandon my prior decision to stop at 15, what will then transpire?" And it seems he may reasonably answer: "I would then follow the slippery slope all the way to 1,000." His prior decision to stop at 15 was his best shot at playing the game without going all the way; if he does not stick with that decision, there is little reason to think he would stick with any other decision short of the bottom of the slippery slope.[39] Further, he can anticipate that were he to slide all the way to 1,000 he would then wish that he had instead stopped at 15: he would then wish he had earlier followed through with his yet-earlier decision and stopped at 15 rather than abandoning that decision and sliding all the way to 1,000. This line of reasoning can reasonably lead him to accept versions of (a) and (b), appropriately interpreted, concerning his following through with his plan to stop at 15: he would be glad later if he stuck with his plan and would regret it if he did not. So he can conclude at the time of action that his following through with his intention to stop at 15 satisfies the no-regret condition.

Now consider the toxin case. Suppose on Wednesday you try saying to me: "I know you prefer not to drink toxin, despite your prior intention to drink. But you will later be glad if you did drink it, and if you do not drink it you will later wish you had." I think I would surely object. On Wednesday I already either have the money or not. If I have the money and yet abandon an intention to drink, I will be *glad* I abandoned that intention and so avoided the pains of the toxin!

It might be replied that even after Wednesday I still prefer money and drink to no money and no drink. So perhaps I would later be glad I had stuck to my intention and drunk the toxin, given my preference for the package of toxin plus money. But recall that we are considering my reflections at the time of action, on Wednesday. By this time the first part of the package – whether I have the money or not – is already fixed. My choice at that time – what remains under my control then – concerns the second part of the package: to drink or not to

drink. What I want on Wednesday to know is how I will later assess *these* options. And it seems that I will reasonably conclude on Wednesday that at the end of the week, and holding fixed the past prior to my Wednesday decision, I would regret following through and drinking the toxin. Granted, if I did follow through I might later be glad that I am that kind of guy – the kind of guy who wins the money in such cases. But that is not later to favor the option on Wednesday of drinking over the option on Wednesday of not drinking, given that the money is, by Wednesday, already in the bank. So following through with an intention to drink toxin would not satisfy the no-regret condition, properly interpreted.

A similar point can be made for the case of reciprocation, as we have understood it. Suppose I intended to reciprocate, you have helped me with my luggage, and I am now considering reciprocating. Given the special assumptions we are making about the case, I will see that, after all is done, I would later favor not following through, for then I would have thereby gotten the benefit without the burden. As in the toxin case, if I did follow through I might be glad I am that kind of guy, but that seems a different matter.

The no-regret condition, then, seems to divide the cases in the manner we anticipated: It is reasonably believed by the agent at the time of action to be satisfied by follow-through in some cases of temptation and of slippery-slope intransitivity, but not to be satisfied by follow-through in our cases of toxin and reciprocation. In our temptation case, the agent can anticipate that looking back later she will be glad of earlier follow-through. In the toxin case the agent can anticipate that looking back later he would regret earlier follow-through.[40]

To deepen our discussion we need to reflect further on regret. Regret should be grounded in some appropriate evaluative ranking. In particular, the agent's regret at t_3 concerning abandoning her prior intention at t_2 is, we may suppose, grounded in some appropriate evaluative ranking. What ranking?

In our temptation case, the answer is clear: Ann's later regret that earlier she had a second beer is grounded in her later ranking of one beer over two – a ranking she did not have at the time of drinking the second beer. Matters are more complex, however, for the potential self-torturer. He can see that if he abandons his intention and opts for 16 there is good reason to expect that he will continue all the way to 1,000. And he can see that when he gets to 1,000 he will wish he had stuck at 15: he will regret having abandoned his intention to stop at 15. But this regret is not grounded in a ranking of 15 over 16: there is no reason to think that he has reversed his ranking of 16 over 15. In what ranking, then, is the regret grounded?

Well, 15 is the switch point relative to 0. Is the relevant ranking his ranking of 0 over 16?[41] But we want the regret to concern what is still available to the agent at the time of choice between 15 and 16; for we want to respect the way in which agency is temporally and causally located. And 0 is no longer available at the time of choice between 15 and 16. This suggests that the ranking that

is critical is, instead, his ranking of 15 over where he ends up, 1,000,[42] for both of those remain available at the relevant time. Perhaps, then, we can understand the relevant anticipated regret as grounded in *that* ranking: if he opts for 16, he will end up at 1,000; if he sticks with his intention, he will stay at 15; and he will regret his failure to stick with his intention to stop at 15 because he ranks 15 over 1,000.

But if that is the ranking that grounds the relevant regret, we have a puzzle about this case that does not arise in the temptation case. The ranking that grounds the later regret relevant to Ann's case is not a ranking that Ann has when she is faced with the temptation. But if the potential self-torturer ranks 15 over 1,000, that is a ranking that is there all along. In particular, it is there at the time of the choice between 15 and 16. If this ranking is relevant to the rationality of that choice, why isn't it relevant in a straightforward way, at the time of the choice itself? Why is there a need to look to later regret?

There is a good reason why this ranking, at the time of action, of 15 over 1,000 would not by itself support the choice of 15 over 16. The choice of 16 is *evidence* that one will go all the way to 1,000: it is, we are supposing, evidence that one's underlying psychology is such that one will likely go all the way. But the choice of 16 does not itself *cause* one's going all the way to 1,000; it is, rather, itself an effect of the mechanisms that will cause one's going all the way. There are large issues here, issues associated with "Newcombe's problem."[43] For present purposes let me just say that it seems to me that it is normally not a reason in favor of a choice that it is merely evidence of, and does not contribute to, something that is valued. At the time of the choice between 15 and 16 the agent could reasonably appeal to the ranking of 15 over 1,000 if he thought that the choice of 16 would cause his going all the way to 1,000. But that is not what the potential self-torturer thinks: he only thinks that a choice of 16 would be evidence that he will go all the way.

That explains why we cannot appeal to the ranking, at the time of action, of 15 over 1,000 to explain why it might be rational to stick with the intention to stop at 15. But if we cannot appeal to that ranking at the time of action, how can we appeal to it at the end, when the agent is in the throes of pain experienced at setting 1,000?

The answer seems to be that there is a kind of regret that is grounded in a ranking of what would have resulted from certain past conduct as compared with what has actually transpired.[44] At the end of the day the self-torturer sees that, indeed, after choosing 16 he did go on all the way to 1,000, and he sees that that would not have happened if he had stuck with his intention to stop at 15. If he had stopped at 15, he would, as a result, not have ended up at 1,000. He therefore regrets not having stuck with his intention to stop at 15. This regret seems to be grounded in his ranking of 15 over 1,000, even though he does not see his choice, instead, of 16 as causing his ending up at 1,000. Given his

ranking of 15 over 1,000, it is enough to support this regret that he believes that if he had stuck with his intention to stop at 15 he would (as a result) not have ended up at 1,000 (which is in fact where he did end up). It is the potential self-torturer's anticipation of such later regret that supports the argument that it may be rational for him to stick with his intention to stop at 15.[45]

IX. Why Future Regret Can Matter

Why should anticipated satisfaction of the no-regret condition matter to an instrumentally rational planning agent? Let us reflect on the very idea of a planning agent. Planning is future oriented. In being engaged in planning agency, one seems to be committed to taking seriously how one will see matters in the relevant future. One seems, in particular, to be committed to taking seriously how one will see matters at the conclusion of one's plan, or at appropriate stages along the way, in the case of plans or policies that are ongoing.[46] This gives anticipated future regret or nonregret on relevant future occasions a special significance to an agent engaged in settling on and following through with plans. That is a major reason why the anticipated satisfaction of the no-regret condition matters to an instrumentally rational planning agent. This also helps somewhat to clarify how far into the future the agent is to look. Implicit in one's planning is, normally, a rough conception of what counts as – as we might say – plan's end.[47]

The idea is not simply that anticipation of future regret or nonregret can change one's present evaluative ranking, though no doubt it can. The idea, rather, is that anticipation of future regret or nonregret can be relevant to the stability of a prior intention of a planning agent; it can be relevant to the question of when it is reasonable to reconsider and abandon a prior intention, and when not.[48] Our concern with stability, recall, is a concern with when it is rational to stick with a prior intention, given that one already has it; it is not simply a concern about the formation of a new intention from scratch.

This clarification helps defuse a possible objection. I have argued that anticipated future regret or nonregret can have a special relevance to a planning agent. But, faced with temptation, why couldn't an agent simply abandon any relevant planning and thereby escape the rational pressures of such anticipations?[49] The answer is that the agent comes to the temptation with relevant prior intentions, and so there is already an issue about whether she may rationally simply give them up. This is the issue of rational intention stability that we have been addressing. And so long as she has these intentions she is a planning agent in a manner that makes salient relevant, anticipated future regret.

In some cases, granted, one will reasonably side with one's present ranking and abandon one's prior plan, while recognizing that one will later regret it. Perhaps one now sees one's anticipated later regret as deeply misguided, or perhaps one anticipates that one's later regret will itself not be stable. The inference from

(a) and (b) to (c) in the earlier argument is defeasible. Indeed, at the level of generality at which we have been proceeding there may be no simple principle that sorts out those cases in which this inference goes through from those in which it does not. But this inference can still have force in certain cases for a planning agent: that is what sophistication fails to see, and that is the key to our explanation of how rational intention stability distinguishes between toxin and temptation.

My claim is not that the no-regret condition has force simply because, in the words of Thomas Nagel, one sees "oneself as a temporally extended being for whom the future is no less real than the present."[50] Nagel argued that such a conception of oneself as temporally extended supports a concern with one's future desires. But that is not my argument. The force of the no-regret condition is not grounded simply in the recognition that one is a "temporally extended being." It is grounded, further, in one's actual engagement in relevant planning agency, and in the resulting significance to one of how one will see matters specifically at plan's end.

I am appealing to certain later attitudes toward now-available options, later attitudes one now anticipates that one will actually have if one proceeds with one's plan and completes it in a certain manner (or, alternatively, if one abandons one's plan). My appeal is not merely to some ranking one would have if one were to step back from pressures of present choice, nor is my appeal to regret or nonregret at the time of plan follow-through concerning one's earlier decisions, nor is it to a ranking made from some detached perspective on the whole of one's life.[51] Finally, the relevant, anticipated later attitudes concern courses of action that are still available to the agent at the time of the anticipation, at the time of plan follow-through. They are not rankings of general traits of character or of general procedures of deliberation.[52]

Earlier I indicated my endorsement of a broadly pragmatic, two-tier approach to plan stability and rational reconsideration in the face of resource limits and unanticipated new information. My main concern here, however, has been with perplexities about certain no-unanticipated-information cases in which one's ranking at the time of action argues against follow-through, and in which issues of resource limits are not germane. For some of these cases I have rejected the standard view, and so sophistication. But in doing this I have not appealed to a pragmatic, two-tier theory of plan stability; for I have argued that such an appeal in these kinds of cases would not do justice to the fact that our agency is temporally and causally located. Instead I have appealed to a planning agent's concern with how she will see her present decision at plan's end. It is this concern, not an appeal to a two-tier pragmatic structure, that supports a distinctive kind of intention stability in certain no-unanticipated-information cases, and thereby a path between resolution and sophistication.[53]

Resolution does not do full justice to the way in which our agency is located

temporally and causally. Sophistication does not do full justice to the way in which our engagement in planning agency normally bestows a special significance on how we will see our now-present action at plan's end. By avoiding both extremes we arrive at a view of instrumentally rational planning agency that does justice both to the fact that we are planners and to the fact that we are temporally and causally located agents. Instrumental rationality does limit access to certain kinds of autonomous benefits, even for a planning agent. Nevertheless, there are no-unanticipated-information cases in which an instrumentally rational planning agent can reasonably commit herself in advance to a plan or policy and then reasonably follow through, rather than simply conform to her rankings at the time of action.

Notes

1. This is a major theme in my *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987).
2. This is in the spirit of work by Herbert Simon. See, e.g., his *Reason in Human Affairs* (Stanford: Stanford University Press, 1983).
3. A similar point is made by J. David Velleman in his *Practical Reflection* (Princeton: Princeton University Press, 1989), pp. 225–6.
4. Jordon Howard Sobel emphasizes the possibility that an agent may "put a premium on steadfastness"; see his "Useful Intentions" in his *Taking Chances: Essays on Rational Choice* (Cambridge: Cambridge University Press, 1994), pp. 237–54, esp. p. 249. Wlodek Rabinowicz, in a complex and subtle discussion, also emphasizes that a rational agent may "assign value to resoluteness and to commitment to previously chosen plans." "To Have One's Cake and Eat It Too: Sequential Choice and Expected-Utility Violations," *Journal of Philosophy* 92 (1995), 586–620, at p. 611. Matters here are delicate: Such valuations may lead to odd forms of bootstrapping, as I argued in my *Intention, Plans, and Practical Reason,* ch. 2. But here I want simply to put such views to one side, for my interest is in an account of instrumentally rational planning agency that does not begin by presupposing such intrinsic valuations.
5. "Instrumental" is here understood broadly: it is not limited solely to causal means to an end. For example, my going to a concert tonight might be promoted by my going to hear the Alma Trio, though my going to hear the Alma Trio is not a causal means to my going to a concert. (See Bernard Williams, "Internal and External Reasons," in his *Moral Luck* [Cambridge: Cambridge University Press, 1981], p. 104.) The crucial point is that I am trying to discuss structures of planning agency in a way that appeals to the nature of such agency and to demands of instrumental reason but does not depend on arguing that practical reason, by itself, mandates certain ends.
6. This problem is similar to the problem posed by the trilemma I discuss in *Intention, Plans, and Practical Reason,* p. 5. See also Paisley Livingston, "Le dilemme de Bratman: Problèmes de la rationalité dynamique," *Philosophiques* 20 (1993), 47–67.
7. Much of the model to be described is discussed in my *Intention, Plans, and Practical Reason,* which also provides other details.
8. These were a primary concern in my discussions of stability in my *Intention, Plans, and Practical Reason,* esp. chs. 5–6. See also my "Planning and the Stability of Intention," *Minds and Machines* 2 (1992), 1–16.
9. Gregory S. Kavka, "The Toxin Puzzle," *Analysis* 43 (1983), 33–6.
10. Gregory S. Kavka, *Moral Paradoxes of Nuclear Deterrence* (Cambridge: Cambridge University Press, 1987), p. 21.
11. This assumes that on Monday my mere intention, even in the special, science-fiction circumstances of the toxin case, does not by itself amount to an assurance to the billionaire of a sort that induces an obligation to drink the toxin. It also assumes that my intention to drink the toxin is not itself an intrinsic desire to drink toxin, an intrinsic desire of a sort that would give me an instrumental reason for drinking. Both assumptions are implicit in standard discussions of the toxin puzzle, and here I follow suit.

    In his comments at the March 1995 Pacific Division meeting of the American Philosophical Association, Gilbert Harman challenged the second of these assumptions. (This challenge is also presented in his chapter "The Toxin Puzzle" in this volume, an essay that derives from his comments at the meeting.) Harman argues that the intention to drink toxin would be an intrinsic desire adopted for instrumental reasons. In this respect, he suggests, it would be like a new intrinsic desire to win a game, adopted because it is more fun to play when you care about winning.

    I agree that if one does somehow come to have such a new intrinsic desire to drink, that may make it instrumentally rational to drink. But I do not see that an intention to drink toxin would generally be like this. After all, in coming to have the intrinsic desire to win, you come to care about winning – winning is now something that matters to you, if only temporarily. In intending to drink toxin in the kind of case we are discussing you would not in the same way care about drinking it.
12. Kavka writes that "you cannot intend to act as you have no reason to act, at least when you have substantial reasons not to act" ("The Toxin Puzzle," p. 35). My remark in the text is in the same spirit, though it is offered as a remark about rational intention, rather than about intention *simpliciter.*
13. I refer to versions of this principle also in my "Planning and Temptation," in *Mind and Morals,* ed. Larry May, Marilyn Friedman, and Andy Clark (Cambridge, MA: Bradford/MIT, 1995), pp. 293–310, and in "Following through with One's Plans: Reply to David Gauthier," in *Modeling Rational and Moral Agents,* ed. Peter Danielson (Oxford: Oxford University Press, 1998), pp. 54–65. Both Brian Skyrms, in remarks at the conference held in honor of Gregory Kavka, and Gilbert Harman, in his comments on my paper at the 1995 Pacific APA meeting, have suggested that the linking principle is challenged by cases of rational irrationality. (See Derek Parfit, *Reasons and Persons* [Oxford University Press, 1984], p. 13, and Thomas Schelling, *The Strategy of Conflict* [Cambridge, MA: Harvard University Press, 1980], p. 18.) These are cases in which it seems to be rational to cause oneself to have an intention to do something one knows it would be irrational to do. Now, it does not follow from the fact that it would be rational to cause oneself to intend to *A* if *C* that it would be rational so to intend. But in any case the formulation I have offered here of the linking principle is intended to circumvent these worries by limiting the cases to occasions on which the agent expects to retain rational control. These are, after all, the cases that are central here. (On this point I am in agreement with David Gauthier's remarks about rational irrationality in his "Commitment and Choice: An Essay on the Rationality of Plans," in *Ethics, Rationality, and Economic Behavior,* ed. Francesco Farina, Frank Hahn, and Stefano Vannucci [Oxford: Oxford University Press, 1996], pp. 217–43, at pp. 239–40.) My formulation also aims at forestalling

complexities raised by Alfred Mele's case of Ted in his "Intentions, Reasons, and Beliefs: Morals of the Toxin Puzzle," *Philosophical Studies* 68 (1992), 171–94.

Both Skyrms and Harman indicated a preference for a principle that instead links a rational intention to *A* with a belief that one will *A*. I agree that a full story will include some appropriate belief condition (see my *Intention, Plans, and Practical Reason*, pp. 37–8), but I do not see this as precluding the linking principle formulated here.

14. In *Intention, Plans, and Practical Reason* I wrote that "in . . . deliberation about the future the desire-belief reasons we are to consider are reasons for various ways we might *act* later" (p. 103). This precluded appeal to autonomous benefits in deliberation. I have changed my mind about this in response to criticisms from David Gauthier in his "Intention and Deliberation," in Danielson, *Modeling Rational and Moral Agents*, pp. 40–53. The linking principle I formulate here aims to retain a tight connection between rational intention and supposed rational execution of that intention, without precluding appeal to autonomous benefits in deliberation. T. L. M. Pink has offered a different criticism of my cited remark. See his "Purposive Intending," *Mind* 100 (1991), 343–59, and in more detail in his *Psychology of Freedom* (Cambridge: Cambridge University Press, 1996). Pink supposes that my remark disallows an appeal in deliberation to certain kinds of coordination benefits of a prior intention. Suppose, for example, that if I intend to go running tomorrow my intention will insure that I get a new pair of running shoes before then, thereby making my running more attractive. Pink thinks that my remark precludes appeal to this fact in my deliberation now about whether to run tomorrow. I am not sure that my remark has this implication. (In the example, note, the benefit of running with new shoes is a benefit of the act of running.) In any case I agree with Pink that we should allow appeal in deliberation to such coordination benefits.
15. Examples along these lines have figured prominently in the work of David Gauthier. See for example his "Assure and Threaten," *Ethics* 104 (1994), 690–721.
16. We could make this more realistic by assuming only that I know that a sincere assurance is considerably more likely to be successful than an insincere one.
17. Recall that I am assuming in both of these cases that simply by forming the intention concerning t_2 I do not newly come to have a reason-giving intrinsic desire so to act.
18. These last two paragraphs draw (with changes) from my "Following through with One's Plans: Reply to David Gauthier." For a trenchant discussion leading to a similar conclusion about the toxin case, see Daniel Farrell, "Intention, Reason, and Action," *American Philosophical Quarterly* 26 (1989), 283–95.

    There is a possible complication concerning the reciprocation case (pointed out to me in different ways by Meir Dan-Cohen and David Gauthier). Suppose I do not follow through and so do not reciprocate even though you have helped me. I get the benefit of your help without the burden of my helping you. But I also get evidence about myself – evidence that I tend not to follow through in such cases. This evidence may make me in the future more skeptical than I would have been if I had followed through that I would follow through with such intentions in the yet farther future, and thereby make me less likely in the future to form intentions to reciprocate. I would, then, be in the future less likely to achieve associated autonomous benefits. So perhaps in those cases in which I expect to be in an indeterminate number of future situations of potential cooperation (even with different potential partners) I do have reason now to follow through and reciprocate.

    This argument cites what we might call a *reflexive* reputation effect. I tried to abstract away from reputation effects in my characterization of our reciprocation case. But it might be objected that if we preclude even such reflexive reputation effects we are imposing an overly severe limitation on our discussion.

    My response is, first, that if this argument succeeds, we should just grant that it is only a limited range of cases of potential reciprocation that are our concern here, namely, those cases that really do have the structure of the toxin case, a structure in which the primary consideration in favor of follow-through derives from the autonomous benefit of the prior plan rather than from the future effects of follow-through. But, second, I am skeptical that the argument succeeds. The argument depends in part on the claim that if I do follow through and reciprocate this time, I will as a result have reason to be more confident that I would follow through in the future and so will, as a result, achieve such self-confidence. But we are assuming that I am, and know I am, generally an instrumentally rational agent. So, for my present follow-through to support a rational belief in my own future follow-through, it needs to support the belief that such future follow-through would be instrumentally rational; otherwise I will tend to infer that my present follow-through is not a good predictor of my future conduct. But it is not clear how the appeal to the reflexive reputation effects of my present follow-through supports a claim about the rationality of future follow-through. (Perhaps what is crucial is not the reflexive reputation effects of my present follow-through but rather that my present follow-through gives me evidence that later follow-through would itself have certain reflexive reputation effects that would tend to make that later follow-through rational. But if my present follow-through only gives me evidence of that – if the rationality of later follow-through is not itself an effect of my present follow-through – it is not clear how this helps the argument.) A related concern is that the linking principle says that I can in the future rationally intend to reciprocate only if (roughly) I judge then that it would be *rational* in the farther future to follow through. It will, again, not be enough for me just to expect that I would (perhaps not rationally) follow through. And it is not clear how appeal to reflexive reputation effects of present follow-through can show that such follow-through in the farther future would be rational.
19. I learned this terminology, and much else, from Edward F. McClennen, *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge: Cambridge University Press, 1990).
20. Ibid. See also Laura DeHelian and Edward F. McClennen, "Planning and the Stability of Intention: A Comment," *Minds and Machines* 3 (1993), 319–33. I do not try to do justice to the complexity and subtlety of McClennen's detailed views here. In particular, his defense of his version of resolute choice is limited in important respects. My broad characterization of strong resolution will, I think, suffice for the purposes of the present discussion.
21. See his "Assure and Threaten," "Commitment and Choice," and "Intention and Deliberation."
22. This is a change from Gauthier's earlier views about deterrence. See his "Deterrence, Maximization, and Rationality," *Ethics* 94 (1984), 474–95. Gauthier's views about following through with a failed threat are complicated and involve consideration of general policies of issuing and carrying out certain kinds of threats. For a probing discussion and criticism see Joe Mintoff, "Rational Cooperation, Intention and Reconsideration," *Ethics* 107 (1997), 612–43.
23. Gauthier writes: "in deliberating rationally, one considers whether one's course of action is best . . . where a course of action is distinguished and demarcated by its intentional structure." "Assure and Threaten," p. 717.

24. What about the sequence: Sincerely assure at t_1; do not reciprocate at t_2? This is not a sequence that one could decide on at t_1, since the intention not to reciprocate would mean that the assurance is not sincere. So it is not a "course of action" available beginning at t_1.
25. Compare Bernard Williams's remark that "The correct perspective on one's life is *from now.*" *Moral Luck* (Cambridge: Cambridge University Press 1981), p. 13.
26. Gauthier, "Assure and Threaten," p. 701.
27. In a recent essay David Velleman tries to anchor a Gauthier-like view about reciprocation and assurance in a fundamentally different line of argument, one that appeals to the idea that action has a constitutive aim. I do not try to assess Velleman's alternative strategy here. See David Velleman, "Deciding How to Decide," in *Ethics and Practical Reason*, ed. Garrett Cullity and Berys Gaut (Oxford: Clarendon Press, 1997), pp. 29–52.
28. I assume that there really is a preference shift, that she is not merely confused about what her preferences are. My discussion of the case of Ann owes much to George Ainslie, *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person* (Cambridge: Cambridge University Press, 1992).
29. Ainslie, in *Picoeconomics*, tries to show that temporary preference reversals like Ann's would occur in agents who have certain – as he believes, extremely common – highly bowed temporal discount functions. But we do not need to discuss here Ainslie's diagnosis of such cases to agree that they are common. Our preferences for certain goods – be they beer, mystery novels, chocolates, or others you can cite from your own experience – do seem susceptible to this kind of temporary shift.

    I consider Ainslie's views in "Planning and Temptation," where I discuss a wine-drinking pianist whose problem is similar to Ann's, except that whereas Ann's preference reversal is triggered by her drinking the first beer, the pianist's preference reversal is triggered by the arrival of dinnertime. Given this difference, Ann's case may not cohere with Ainslie's claim that the primary mechanism underlying such preference changes is generally one of temporal discounting.
30. Here I agree with similar remarks of Gauthier's in "Commitment and Choice," pp. 238–9. For a different approach to preference see Sarah Buss, "Autonomy Reconsidered," *Midwest Studies in Philosophy* 19 (1994), 95–121.
31. Compare Gary Watson, "Free Agency," *Journal of Philosophy* 72 (1975), 205–20. This appeal to Watson's distinction was a suggestion of J. L. A. Garcia, in conversation.
32. The example is from Warren Quinn, "The Puzzle of the Self-Torturer," in his *Morality and Action* (Cambridge: Cambridge University Press, 1993), pp. 198–209, at p. 198. Quinn provides references to relevant literature. Thanks to Liam Murphy for bringing Quinn's essay to my attention.
33. Quinn says that such intransitivities bar the potential self-torturer from saying that each setting is better than the preceding one, for *better than* is, Quinn says, a transitive relation. But Quinn also says that the preferential ranking may be thoughtful and informed, and so an appropriate candidate for shaping choice (ibid., p. 199). So we may allow it to provide evaluative rankings in the sense relevant here.
34. Ibid., p. 206.
35. In his discussion Quinn seems to endorse "the principle that a reasonable strategy that correctly anticipated all later facts (including facts about preferences) still binds" (ibid., p. 207). We need to be careful, however, not to interpret this principle in a way that would justify sticking with a plan to drink toxin. (In an earlier essay Quinn had indicated that he would not welcome such a result. See "The Right to Threaten and the Right to Punish," in his *Morality and Action* [Cambridge: Cambridge University Press, 1993], pp. 52–100, at p. 98.)

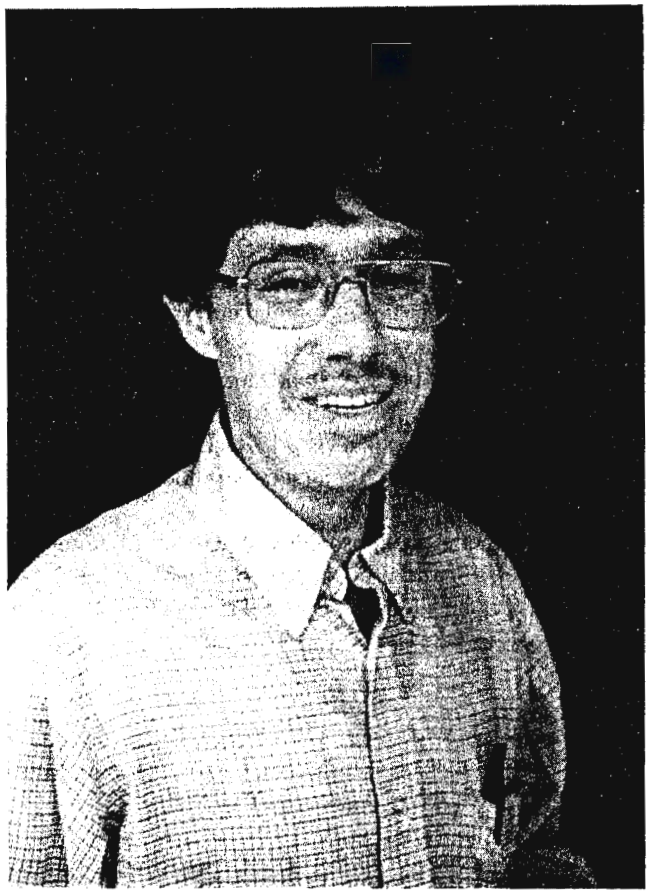
36. I owe to David Gauthier the suggestion that such a switch point relative to 0 is a reasonable point at which to settle in advance on stopping. (In the absence of some such argument the agent might be in a Buridan situation: he might know that there is reason to decide in advance on a stopping point, but there might be no single point such that there is reason to decide to stop there rather than at some competitor.) Note that in reaching a decision to stop at the switch point relative to 0 the agent may know that there is also a later switch point *relative to 15* – a later point that is preferred to 15 but whose successor is dispreferred to 15.
37. Note that this no-regret condition includes both the absence of regret at having followed through and the presence of regret if one did not follow through.
38. Versions of the idea that anticipated future regret, or its absence, can matter to the rationality of present conduct appear in a number of studies. See, e.g., Graham Loomes and Robert Sugden, "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty," *Economic Journal* 92 (1982), 805–24. (Note, though, that the regret that is central to the Loomes and Sugden theory is the result of new information that was not available at the time of the (regretted) action; my focus, in contrast, is on anticipated later regret that does not depend on such new information.) See also John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), pp. 421–3.
39. He need not think that his descending the slippery slope all the way to 1,000 would be *caused* by his failure to stop at 15. It is enough that he believe that if he does not stick with his plan to stop at 15 he will go all the way to 1,000. I return to related matters at the end of this section.
40. Consider the much-discussed case of Ulysses and the Sirens. (See, in particular, Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality*, rev. ed. [Cambridge: Cambridge University Press, 1984].) Suppose Ulysses decides in advance to sail by the Sirens, but when he hears them his ranking changes in just the way he had anticipated. Ulysses is like Ann in one respect: he knows that if he sticks with his prior decision to sail by he will be glad he did. But on some versions of the Ulysses case, and unlike the case of Ann, it is also true that if he does not stick with his prior decision he will be glad he did not! So follow-through would satisfy one but not *both* parts of the no-regret condition. So there will be important cases of temptation and the like that are similar in certain respects to the one I have discussed but will need a different treatment.
41. A suggestion of Gideon Yaffe's.
42. I am assuming that this is indeed a ranking of our agent.
43. See, for starters, Robert Nozick, "Newcombe's Problem and Two Principles of Choice," in *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher et al. (Dordrecht: Reidel, 1969).
44. I do not say: "a ranking of what certain past conduct would have been evidence for (but not a cause of) as compared with what has actually transpired."
45. Gideon Yaffe has wondered whether there is an instability here. The potential self-torturer, let us suppose, sticks with his prior intention to stop at 15 in part because he believes that if he instead goes on to 16 he will (likely) go on to 1,000. But if he does stop at 15 he can, perhaps, reasonably believe that if he instead intended to stop at the (later) switch point *relative to 15* (supposing there is one) he might well pull that off. (After all, we have given reason to think it would then be rational to do so.) Suppose the agent had earlier decided on 15, the switch point relative to 0. Faced

with the choice of 15 or 16, he wonders whether to stick with his prior decision. He sees that if he *does* stick with it, he *would* (probably) stick with a decision to stop at the (later) switch point relative to 15. So why not go on to that later switch point? The answer seems to be that if he does go on past 15 he will not have this evidence that he will stop at the later switch point. That seems sufficient to support his stopping at 15. Having stopped at 15, it may seem that one has available an argument for going on to the next switch point, but that argument would be undermined by one's going on and so seems not to have practical force.

46. This qualification concerning ongoing plans or policies should be understood throughout the discussion that follows.
47. These last two sentences benefited greatly from conversation with Elijah Millgram. For some suggestive remarks broadly in the spirit of this paragraph, see Gerald J. Postema, "Morality in the First-Person Plural," *Law and Philosophy* 14 (1995), 35–64, at pp. 56–7. See also Thomas E. Hill, Jr., "Pains and Projects," in his *Autonomy and Self-Respect* (Cambridge: Cambridge University Press, 1991). Hill writes that "the commitment to make my choices justifiable to myself later seems implicit in any project of deep deliberation" (p. 186). I am suggesting that a somewhat analogous commitment is implicit in planning agency more generally.
48. Of course, it may be that anticipated regret can play other roles in practical reasoning as well. See, for example, Robert Nozick, *The Nature of Rationality* (Princeton: Princeton University Press, 1993), p. 185, n. 21.
49. Tim Schroeder suggested an objection along such lines.
50. Thomas Nagel, *The Possibility of Altruism* (Oxford: Oxford University Press, 1971), p. 69. See also Rawls's remarks about seeing oneself "as one continuing being over time." *A Theory of Justice*, p. 422.
51. David Velleman emphasizes the significance of those evaluations of a person that "are relative to the perspective of his life as a whole" in his "Well-Being and Time," *Pacific Philosophical Quarterly* 72 (1991), 48–77, at p. 67.
52. I am not saying that at the time of follow-through one might not anticipate such later assessments – including forms of regret – concerning general character traits, or courses of action that began well before the time of follow-through. I am only saying that it is not one's anticipation, at the time of follow-through, of those later attitudes that is critical to plan stability, for those later attitudes are not focused on what is now, at the time of follow-through, in one's control.
53. I do not claim this is the only source of the cited intention stability. My concern is only to identify a major source of such stability, one that responds differently to toxin and to temptation, and one that is not grounded in a two-tier pragmatic structure. In my discussion of the toxin case in *Intention, Plans, and Practical Reason*, ch. 6, I argued, as I do here, that rationality stands in the way of follow-through. But my argument there assumed that our approach to stability in such no-new-information cases should stay roughly within the two-tier framework I had developed primarily for new-information cases in which our resource limits play a central role and in which the crucial issue is whether or not to reconsider one's prior intention. I no longer accept that assumption.

    Remarks in this and the preceding paragraph in the text are intended to indicate briefly what seem to me to be some significant differences between my view here and ideas about temptation and related cases in Gauthier's "Resolute Choice and Rational Deliberation: A Critique and a Defence," *Noûs* 31 (1997), 1–25. Concerning cases of temptation, Gauthier contrasts preferences that the agent has at the time of action with "(temporal) *vanishing point* preferences that he acknowledges when

choice is not imminent" (p. 20). The notion of a vanishing-point preference is in some respects similar to, but is not the same as, my notion of one's attitude at plan's end: vanishing-point preferences are preferences, either earlier or later, when "choice is not imminent," not specifically at plan's end. Gauthier's view about cases like Ann's is "based on a comparison of the effects on an agent's overall prospects of different modes of choice" (p. 23). I have here, in contrast, eschewed such a two-tier pragmatic approach to such cases and appealed instead to the significance to a planning agent of anticipated regret or nonregret at plan's end. Finally, Gauthier appeals (p. 24) to possible regret, at the time of action, concerning an earlier decision. The regret I appeal to is regret one anticipates at the time of action that one will have later, at plan's end.



Gregory S. Kavka (Photo courtesy of Terence Parsons.)

Rational Commitment and Social Justice

Essays for Gregory Kavka

Edited by

JULES L. COLEMAN
CHRISTOPHER W. MORRIS

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

First published 1998

Printed in the United States of America

Typeface Times Roman 10/12 pt. *System* Quark XPress™ [AG]

*A catalog record for this book is available from
the British Library.*

Library of Congress Cataloging in Publication Data

Rational commitment and social justice : essays for Gregory Kavka /
edited by Jules L. Coleman, Christopher W. Morris.

p. cm.

ISBN 0-521-63179-3 (hardbound)

1. Social justice. 2. Commitment (Psychology) I. Coleman, Jules L.
II. Morris, Christopher W. III. Kavka, Gregory S., 1947–1994.
JC578.R365 1998
303.3'72–dc 21 98-20681
CIP

ISBN 0 521 63179 3 hardback

Contents