



Assure and Threaten

Author(s): David Gauthier

Reviewed work(s):

Source: *Ethics*, Vol. 104, No. 4 (Jul., 1994), pp. 690-721

Published by: [The University of Chicago Press](#)

Stable URL: <http://www.jstor.org/stable/2382214>

Accessed: 01/11/2012 12:21

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.



The University of Chicago Press is collaborating with JSTOR to digitize, preserve and extend access to *Ethics*.

Assure and Threaten*

David Gauthier

I

What is the relation between my aim and my reasons for acting or deciding or choosing? Simply to have a convenient specification I shall let my aim be that my life go as well as possible.¹ This aim in itself will not be particularly helpful to me in deciding what to do. I shall have to fill in more specifically what it is for my life to go well. Here the various particular concerns that I have, some perhaps lasting over the course of my life, others less enduring, are relevant. But for present purposes I can let them be as they may. And indeed, I should insist that nothing in my present argument requires that my concerns be

* Versions of this essay have been read at several universities in the United States and Australia, and the resulting discussions have led to significant modifications, as have written comments from Annette Baier, John Broome, Richmond Campbell, Gregory Kavka, Christopher Morris, Howard Sobel, and two anonymous readers. Revisions to the essay were written while a Visiting Fellow in the History of Ideas program, Research School of Social Sciences, Australian National University, and also while resident at the Bellagio Study and Conference Center of the Rockefeller Foundation and Fellow of the John Simon Guggenheim Memorial Foundation; I am grateful for this support.

1. The idea of starting a discussion of deliberative rationality by specifying an aim comes from Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), p. 3. But in what follows I largely ignore Parfit's discussion of self-defeating theories in chap. 1 of his book, even when I consider accounts of deliberation that would represent it as self-defeating in relation to the aim. I explore some of my differences with Parfit's views in "Rationality and the Rational Aim," in *Reading Parfit*, ed. Jonathan Dancy (Oxford: Blackwell, in press). I should also note that I make no attempt in this essay to discuss or relate my arguments to the most recent major study of intention, deliberation, and rationality, Michael E. Bratman, *Intention, Plans, and Practical Reason* (Cambridge, Mass.: Harvard University Press, 1987), or to the groundbreaking study of modes of sequential choice, Edward F. McClennen, *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge: Cambridge University Press, 1990). For some points of contact with both Bratman and McClennen, see my "Commitment and Choice: An Essay on the Rationality of Plans," in *Ethics, Rationality, Economic Behaviour* (tentative title), ed. Francesco Farina, Stefano Vannucci, and Frank Hahn (Oxford: Oxford University Press, in press).

Ethics 104 (July 1994): 690–721

© 1994 by The University of Chicago. All rights reserved. 0014-1704/94/0404-0003\$01.00

self-directed; I could take as my aim that the world, viewed from my perspective, go as well as possible.²

And now I want to consider, not my specific reasons for some particular action, but the kind or kinds of reasons I have, given my aim. My reasons take their character from this aim. It is natural to assume that they do so directly, so that in considering what to do or to choose, I simply consider the ways in which my various possible actions would affect how my life goes, for better or worse. Since my actions cannot affect how my life has gone prior to their performance, I consider the ways in which my possible actions would affect how my life would go from the time of their performance. I form expectations about the consequences for my life of my possible actions. Such expectations constitute my reasons,³ and the action or actions best supported by them is the one, or are the ones, that of those possible for me, would at the time of performance be part of or lead to a life that would go best for me.

But as I have argued before, this type of account will not do.⁴ It is often the case that the action best supported by considerations about the consequences of my actions is the one that would best serve my

2. Although my concerns need not be self-directed, they must be self-based; thus it would not do to take my aim as simply that the world go as well as possible. For it is plausible to suppose that if agents had this aim, and were fully informed, they would agree on their ranking of states of affairs in relation to their aim. And I am concerned primarily with deliberation in situations in which agents disagree—in which the outcome that would best satisfy my aim is not the outcome that would best satisfy yours. Those who think that this sort of disagreement shows some imperfection or fault in at least one of the parties to the disagreement will find little to hold their attention in what follows.

3. And insofar as they are reasonable expectations, they constitute good reasons. Unless I say or imply otherwise, I shall assume that expectations are reasonable. But reasonable expectations may be mistaken, and this helps give rise to some—not all—of the problems about deliberative rationality that I shall examine.

4. I have been arguing this, explicitly or implicitly, ever since my paper “Reason and Maximization” (*Canadian Journal of Philosophy* 4 [1975]: 411–33). In addition to the discussion in chap. 6 of *Morals by Agreement* (Oxford: Clarendon Press, 1986), especially pp. 182–87, my argument here builds on (and sometimes departs significantly from) what I have said in “Deterrence, Maximization, and Rationality,” *Ethics* 94 (1984): 474–95, “The Unity of Reason: A Subversive Reinterpretation of Kant,” *Ethics* 96 (1985): 74–88, “Reason to be Moral?” *Synthese* 72 (1987): 5–27, “Hobbes’s Social Contract,” in *Perspectives on Thomas Hobbes*, ed. G. A. J. Rogers and Alan Ryan (Oxford: Clarendon Press, 1988), pp. 125–52, “War and Nuclear Deterrence” in *Problems of International Justice*, ed. Steven Luper-Foy (Boulder and London: Westview, 1988), pp. 205–21, “In the Neighbourhood of the Newcomb-Predictor (Reflections on Rationality),” *Proceedings of the Aristotelian Society* 89 (1989): 179–94, and “Economic Man and the Rational Reasoner,” in *From Political Economy to Economics—and Back?* ed. James H. Nichols, Jr., and Colin Wright (San Francisco: ICS Press, 1990), pp. 105–31. It also relates to the two papers referred to in n. 1 above. I shall make specific references to my earlier papers only to note explicit differences with my present position.

aim—that would be part of a life that goes best for me. But often is not always, and an account of how my aim determines my reasons aspires to generality. I shall not achieve full generality in this discussion, but I do hope to indicate some of the complications that a satisfactory account must accommodate.

II

I shall begin by retelling a story, adapted originally from Hume, which I frequently employ.⁵ My crops will be ready for harvesting next week, yours a fortnight hence. Each of us will do better if we harvest together than if we harvest alone. You will help me next week if you expect that in return I shall help you in a fortnight. Suppose you do help me. Consider my decision about helping you. I have gained what I wanted—your assistance. Absent other not directly relevant factors, helping you is now a pure cost to me. To be sure, if I were to help you I should still be better off than had I harvested alone and not helped you, but I should be better off still if having received your help, I did not return it. This calculation may appear shortsighted. What about next year? And what about my reputation? If I do not help you, then surely I shall harvest alone in future years, and I shall be shunned by our neighbors. But as it happens I am selling my farm when the harvest is in and retiring to Florida, where I am unlikely to cross paths with anyone from our community. I might of course have some positive feeling for you, so that I should not want to take advantage of you, but suppose I do not—suppose I am simply indifferent to you. If my reasons for deciding what to do are taken directly from my aim, then since my life will go better for me if I do not help you, that is what I have most reason to do.

Being rational persons, we both know this. The scenario I have sketched is one each of us can sketch—and each of us knows it to be true. It would be pointless for me to pretend otherwise. So you know that I would not return your help, and being no sucker, will therefore leave me to harvest my crops alone. Neither of us will assist the other, and so each of us will do worse than need be. We shall fail to gain the potential benefits of cooperation.

I have diagnosed this type of situation before—my diagnosis being that in the scenario as I have sketched it, I have got my reasons wrong. If my aim is that my life go as well as possible, then I should not take all of my reasons for acting directly from that aim, considering only which action will have best consequences for my life. For if I always deliberate in this way, then my life will not go best for me.

5. See David Hume, *A Treatise of Human Nature* (Oxford: Clarendon Press, 1888), pp. 520–21.

Let us see why. Suppose that I begin deliberating by considering our upcoming harvests and realize that it is worth my while to get you to help me with my harvesting, even if I end up helping you in return, given that my alternative is harvesting alone. And I realize that you will help me if and only if you believe that by helping me you will gain my assistance, whereas by not helping me you will forfeit my assistance. I want you to believe that by helping me you will gain my assistance. Now it is entirely possible that you will believe this if I offer you a sincere assurance that I will help you and that you will not believe this if I offer you an assurance that is not sincere or no assurance at all.⁶ For we may suppose that you are a good judge of sincerity, and I am a poor deceiver. Thus I judge that it is worth my while to offer you a sincere assurance that I will return your help, even if and though in consequence I end up actually assisting you. I must then offer you a sincere assurance if I am rationally to expect my life to go as well as possible.⁷

But I can't offer you this assurance if I take my reasons for acting directly from my aim, and if I know this, and also know or believe myself to be rational. If I take my reasons directly from my aim, then I shall not have, and I know that I shall not have, sufficient reason to return your help. And I can't sincerely assure you that I shall do something that I know I shall have sufficient reason not to do, and believe that as a rational person, I therefore shall not do. Or at least, I can't give you this assurance unless I have some means of making myself not only irrational at the time of decision, but irrational in just the right way—bringing it about that I shall choose what I now know that I shall have sufficient reason not to choose. I shall return to the idea of making oneself irrational shortly, but putting it now to one side, if I take all of my reasons for acting directly from my aim, then I must resign myself to harvesting alone, and my life will not go as well as possible, contrary to my aim.

6. Of course sometimes an insincere assurance will be effective, and sometimes I may expect my life to go better if I am insincere rather than sincere. But not always. In this essay I am concerned only with situations in which an agent reasonably expects sincerity to further her aim better than deception. I take such situations to occur with sufficient frequency that they are worth attention. I also suppose, although I cannot argue the matter here, that the need to be sincere is no sign of weakness, imperfection, incapacity, or irrationality on the part of an agent whose aim is that her life goes as well as possible. Furthermore, the moral status of sincerity and deception, and the light, if any, that their moral status sheds on the relation between morality and rationality, are matters altogether beyond my present concern.

7. Why assurance? Why not, e.g., promise? I want to avoid both the relative specificity and the moral connotations that philosophic attention has conferred on promising. To offer an assurance, as I am using the phrase, is to do more than to express a (mere) intention, but to do less than to invoke the moral considerations that attach to a promise. The present essay is about rationality, not morality.

But is this so? One might object that my life does go as well as possible given that I am a rational person. Someone else might be able to do better because she could give an assurance that I cannot give. But that does not mean that I can do what would make my life go better. The best I can do, and so the best my life can go, must be relative to the sort of person I am. Harvesting alone is the best I can do if I am a rational person with nothing to gain from helping you harvest your crops. The situation I have characterized is one in which persons who have relevant long-term interests at stake, or who place sufficient value on keeping their word, or who care for their fellows, or who are talented at deceiving their fellows will do better than persons like me, of whom none of these things are true. Why should that be surprising or troubling?

I think it should be. Being a rational person, in the sense of someone who acts in accordance with her best or strongest reasons, is not a relevant determinant of what is possible for one to do, or of the best one can do, in the way in which one's various capacities and character traits may be. A rational person is not one for whom only the action best supported by her reasons is possible but, rather, one who selects on the basis of her reasons among her possible actions. An act is possible for a rational person, if she would (or at least might) perform it should she have sufficient reason to choose or decide on it.⁸

If a person's reasons take their character from her aim, then it is surprising and troubling if acting successfully in accordance with her reasons, she must sometimes expect to do less well in relation to her aim than she might. If my aim is that my life go as well as possible, and I act successfully in accordance with the reasons determined by that aim, then should I not expect my life go as well as possible? If the orthodox account of the connection between aim and reasons were correct, then sometimes I should not expect success in acting on my reasons to lead to my life going as well as possible. And so I propose to rethink the connection. I shall be able to characterize rational deliberation in a way that provides a stronger link between acting on one's reasons and fulfilling one's aim than if one supposes simply that one's reasons pick out an action that at the time of performance would lead to one's life going as well as possible. In the end, alas, I shall have to conclude that the changing temporal perspectives from which one deliberates make it impossible to relate reasons and aim so that one may always expect that if one successfully acts on one's reasons, one's life will go as well as possible. But we are now far from that end; my

8. The parenthesis is needed since a person might fail to perform her chosen act, without it being impossible for her to perform it. Complications evidently lurk here, which (thankfully) fall outside the scope of this essay.

present task is to consider how I must deliberate if I am to be able to give sincere assurances.

III

In the situation we are considering, I do best to give you a sincere assurance that, if you help me harvest, I shall in return help you. Giving you this assurance is part of a life that goes best for me. But to give you this assurance, I must have, or at least suppose myself to have, reason to carry it out. I must suppose that, should you help me, I should then have sufficient reason to help you. My reason cannot be that helping you will be part of a life that thenceforth will go best, for that is not the case. But it is not therefore unrelated to my aim. For of the courses of actions that I can choose, taken as wholes, giving you a sincere assurance that I shall return your help, and then, should you help me, honoring my assurance, is part of a life that goes best for me.

To be sure, my life would go better were I to give you the sincere assurance and then not honor it. But although I can do these things—I can be perfectly sincere in assuring you that I shall return your help and yet not return it—I cannot choose to do them taken together. I cannot simultaneously choose both to give you a sincere assurance to return your help and not to honor my assurance when the time comes. Furthermore, if I choose to give you a sincere assurance, then in so choosing I must intend to honor it and believe that I shall honor it.⁹

My reason for helping you, it may therefore be proposed, is that helping you is part of the best course of action that I can choose to follow—part of the course of action that makes my life go as well as possible. Suppose that someone, whom I shall identify as our objector, agrees that this is indeed a reason for helping you. And so he claims that I can offer you a sincere assurance. But, he urges, it is not a good enough reason actually to act on. For there is also a reason against helping you—that when it comes time to help you, then no matter what has happened, not helping you will then make my life go as well as possible. And this he claims is a better reason, since if I act on it my life will go better. But his line of argument is mistaken; he is wrong

9. Neither my intention nor my belief rules out the possibility that I may change my mind and that I may do so with reason. Sometimes reasonable change of mind results from my coming to believe that the situation in which I offered the assurance was not as I took it to be; sometimes it results from my finding that the situation in which I must honor or violate the assurance is significantly different from what I had reasonably anticipated. Sometimes also reasonable change of mind results from alterations in my values and priorities. These and related possibilities do not concern me in the present essay, but nothing that I say is intended to dismiss them. Indeed the account of deliberation that I shall present offers a natural way of assessing change of mind, although I shall defer pursuing this to another occasion.

to suppose that I can offer you a sincere assurance to return your help. For if my reason not to help you is a better reason, or a sufficient reason, then as a rational person I should be aware of this. I shall believe that when the time comes I shall not help you, and so I cannot sincerely assure you that I shall. And so once again I shall harvest alone, and my life will not go as well as possible. It is not enough for me to accept as a reason, that helping you is part of the best course of action I can choose; I must accept it as a better reason than that not helping you would make my life go as well as possible.

Our objector therefore shifts ground. The force of a reason, he claims, need not be fixed over time. Initially, when I wish to assure you that I shall return your assistance, I think in terms of the various courses of action that I can choose, and I correctly weigh most heavily the fact that returning your help would be part of the best course of action. At that time it is my best reason for acting, and so I can sincerely assure you that I shall return your help. Later, when I must choose whether or not to assist you, the situation is different, and I correctly weigh most heavily the fact that returning your help would not be the best action. But our objector's argument is again mistaken. No doubt the force of a reason need not be fixed over time, but it may vary only in such a way that its force at a given time is not undermined by awareness at that time, of its force at some other time. If I am aware that what would weight most heavily at the moment of choosing—call it time t —is that not assisting you would then make my life go as well as possible, then I can act now in a way that gives that consideration lesser weight, but I cannot act now in a way that requires my giving that consideration lesser weight at t . But this is precisely what would be required, for me now to assure you sincerely that I shall assist you at t .

It is clearly futile for our objector to advance any proposal that takes the fact that assisting you will not be part of a life that goes as well as possible for me, as sufficient reason against my actually assisting you, and takes me to be rational and to be always aware of my rationality, and yet includes the supposition that I may sincerely assure you that if you help me, I shall help you in return. If our objector continues to insist that I always have most reason to perform the action that at the time of performance is part of my life going best for me, and yet does not want to condemn me to harvesting my crops alone, what opening is left for him? We may rule out as a dodge, any attempt to change the situation so that assisting you will have best consequences for how my life goes. In practice it may of course be possible to do this. But making such a change will almost certainly prove more costly to me than giving you a simple, sincere assurance. And it fails to face the real issue—that taking my reasons for acting directly from my aim is in certain situations counterproductive and, indeed, self-defeating in relation to that aim.

Before considering our objector's next move, I should pause to note that reflection on the possibility of giving assurances reveals a complication in my account of the connection between possible actions or choices and rational actions or choices. I argued that rationality is not a relevant determinant of possibility, so that an action may be possible for a rational agent even though as rational she cannot perform it because performing it would be contrary to her reasons for acting. But although the rationality of an action x is not a relevant determinant of its own possibility, it may be a relevant determinant of the possibility of some other action y . For if performing y requires that an agent intend to perform, or believe that she will perform, x , then if the agent believes that she is rational and that performing x would be irrational, she cannot perform y . This is exactly what has been involved in my discussion of giving assurances. Giving a sincere assurance requires that one intend to perform, and believe that one will perform, the assured action. One's beliefs about one's rationality and one's possible reasons for performing the assured action thus affect the possibility of giving the assurance.

IV

Let us return from this digression to our objector, who does have another move that has yet to be closed to him—the proposal that it is rational to make oneself irrational. He claims that, although it is irrational for me to help you, it is rational for me to convince myself otherwise.¹⁰ If I believe that there is sufficient reason for me to return your help, then I can sincerely assure you that I shall do so. And since my life goes best if I give you this assurance, it is rational for me to make myself able to give it, and then to give it. Of course, when the time comes to decide whether or not actually to help you, it would be nice, the objector remarks, if I could then do what is really rational. But this, he concedes, could occur only as a matter of happy chance. I cannot arrange to hold one view now of what is rational and another view later—and if I could, then presumably you could realize this, and my present assurance, however sincere given my carefully selected current view of what is rational, would avail me nothing.

So our objector now maintains that success in achieving one's aim that one's life go as well as possible, depends, or at least may depend, on believing that one sometimes has sufficient reason to perform an action other than the one that at the time of performance would be part of one's life going as well as possible. It is altogether reasonable

10. This objector might be Derek Parfit; see *Reasons and Persons*, esp. pp. 9–13. But I shall not consider how far Parfit's arguments correspond to those I ascribe to the objector.

to hold this belief; one has sufficient reason to hold those beliefs about one's reasons for acting that lead to one's life going as well as possible. But one's reason for holding this belief about one's reasons for acting has nothing to do with its truth, and it is, the objector continues to maintain, quite false.

Rationality, the objector supposes (and with this I of course agree), takes its determinate character from one's aim. Since my aim is that my life go as well as possible, it is, he claims, rational for me on each occasion of action or choice, to do or to choose what, judged on that occasion, best promotes my life going as well as possible. This is the objector's key thesis, the ultimate object of my critique. Unfortunately, he notes, being rational is not maximally conducive to my aim. Now of course, it does not follow that being irrational is maximally conducive to my aim, if irrationality is left quite unspecified. What would be maximally conducive to my aim would be a very particular form of irrationality, one that may best be thought of as an alternative mode of deliberation. A person deliberates rationally, the objector supposes, insofar as she considers in each situation, what action or choice among those possible for her will be part of a life that thenceforth goes as well for her as possible. Now she should not abandon deliberation. Rather, she should deliberate by considering what action or choice among those possible for her is required by the overall course of action—among those courses of action from which she is able to choose—that is part of a life going as well for her as possible.¹¹ She may reasonably expect her life to go better if she deliberates in this way than if she deliberates rationally.

It is not always rational to deliberate rationally. (Or more precisely—provided one is able to come to deliberate in a way that will make one's life go better, then it is not rational to deliberate rationally. Henceforth I shall take this proviso as understood.) How are we to understand this claim? First, we need to disambiguate the idea of rational deliberation. We may say that deliberation is rational in reference to its outcome, or to its manner. Deliberation is rational in reference to its outcome insofar as it yields a rational action or choice. Deliberation is rational in reference to its manner insofar as it proceeds by a consideration of those factors that make its end rational. These may diverge; deliberating on the basis of good rules of thumb may be more efficient in yielding a rational choice than deliberating in terms of the actual factors that make a choice rational. An agent is primarily concerned with the outcome and not the manner of her

11. I intend this only as a rough statement of the proposed mode of deliberation. In effect what the objector should advocate, as the irrational mode of deliberation that it is rational to adopt, is what I shall defend in later sections of this paper as the rational mode of deliberation.

deliberation; it may therefore seem that deliberating in a rational manner is itself rational only insofar as it is maximally conducive to yielding a rational outcome. And in this sense it is clearly not always rational to deliberate rationally. Indeed, in some situations it may not be rational to deliberate at all.

But it is not this sense that our objector has in mind, when he claims that it is sometimes not rational to deliberate rationally. His claim is that it is not rational to deliberate in a way that yields a rational outcome, an action or choice best supported by one's real reasons. If I believe that, when the time comes to decide whether or not to assist you, I shall deliberate in a way that yields what is (on his view) in fact the rational outcome, then I cannot sincerely assure you that I shall reciprocate your assistance. So, the objector claims, it is rational for me to convince myself that reciprocally assisting you is rational, and that I should believe that my reasons for acting are considerations that support assisting you, and that it is rational to deliberate in whatever way will yield the choice to assist you as its outcome. Of course none of this is true. But if I convince myself that it is, then I can assure you that I shall reciprocate your assistance, and I shall achieve my aim that my life go as well as possible, at least insofar as it is affected by this situation. And so if I can convince myself, I should.

When we acknowledge that it is not always rational to deliberate in a rational manner, we are recognizing the imperfection of much actual deliberation. An ideal deliberator could not do better in arriving at the rational outcome than to deliberate in a rational manner—that is, in terms of the actual factors that make her choice rational. Real persons, not being ideal deliberators, may do better to deliberate in other ways. But the claim of the objector, that it is not always rational to deliberate to a rational outcome, does not reflect the imperfection of actual deliberation. The objector is not urging us to guard against our weakness as deliberators but, rather, against the failure of those who deliberate rationally and correctly to bring about lives that go as well for them as possible. He is urging us not to be tempted to emulate ideal deliberators, in circumstances in which such emulation might be possible.

What makes our objector's position seemingly attractive is, I think, an assumed parallel between action and belief. To believe is to believe true. Truth sets a standard for belief, and this standard (on my realist view) is independent of epistemic procedures. Since epistemic procedures are fallible in relation to truth, they are imperfect, in Rawls's terminology.¹² Rational belief is characterized in relation to these im-

12. See John Rawls, *A Theory of Justice* (Cambridge, Mass.: Harvard University Press, 1971), pp. 85–86, for a discussion of perfect, imperfect, and pure procedures in the context of justice.

perfect epistemic procedures; a belief is rational if and only if it is adequately supported by appropriate epistemic procedures.¹³ Given this, it makes perfect sense to suppose that in some situations it is not rational for an agent to form rational beliefs. For the rationality of forming a belief, considered as a possible action, is related to how well the agent's life goes, and his life may go better if he forms a belief that is not well supported by procedures directed at truth. Someone then might warn us, not against our weaknesses as believers—which would include such matters as wishful thinking, unwarranted trust in authority, and naive credulity—but against the failure of those who form their beliefs rationally and correctly to bring about lives that go as well for them as possible. To be sure, we may be unable to adopt direct devices for preventing this failure; because believing is believing true, I cannot simply set myself to form a belief that is not well supported by procedures directed at truth. But I can in some cases do this indirectly—a matter discussed by Jon Elster (following Pascal).¹⁴

There are different ways in which one might propose to draw a parallel between action and belief, but here we need consider only one that takes the agent's aim as the standard for his actions. I shall label this "success"; an action is successful if and only if at the time of performance it is part of a life that goes as well as possible for the agent. Success is independent of the agent's deliberative procedures, and these procedures are fallible in relation to it; hence they are imperfect. Rational action is characterized in relation to these imperfect deliberative procedures; an action is rational if and only if it is adequately supported by appropriate deliberative procedures. Now at this point a strict parallel with belief breaks down. A person's life may go better if he forms a belief that is not well supported by procedures directed at truth, and he may sometimes be in a position to recognize this. Although his life also may go better if he performs an action that is not well supported by procedures directed at success, he cannot be in a position to recognize this at the time of performance and so cannot suppose it rational to eschew such procedures on that account. However, the ground for a partial parallel has been established by the recognition, acknowledged by our objector, that in some situations, one may reasonably expect one's life to go better if others believe that one eschews procedures directed exclusively at success, and their belief may itself depend on one being sincerely willing to eschew such proce-

13. I do not suppose either that in any context an appropriate procedure must yield at most one belief or that in any context there is at most one appropriate epistemic procedure. The logic of belief is another of the subjects that I am thankful to avoid in this essay.

14. Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge University Press, 1979), pp. 47–54.

dures. For then the objector may say that it is rational for an agent to be willing to use deliberative procedures that are inappropriate in relation to success, and to perform actions supported by these procedures, even though these actions may be irrational. There is a standard for actions, success, corresponding to the standard for belief, truth; there are imperfect procedures directed at that standard; and in some situations there are reasons for eschewing these procedures.

But why should we follow the objector in accepting this parallel between action and belief? Instead of relating the rationality of deliberative procedures to the success of the actions they determine, why not relate their rationality to the contribution made by their exercise to attaining the agent's aim? Why not determine their rationality by taking into account the effect on how well one's life goes, not only of the actions they determine, but also of those they make possible? Deliberative procedures that make it possible to offer sincere assurances contribute to one's life going better by enabling an agent to choose courses of action that would otherwise be unavailable to her; surely this effect is relevant to the rationality of the procedures.

I have said that an agent is primarily concerned with the outcome and not the manner of her deliberation, but we should now recognize that this statement is ambiguous. Is the agent's primary concern with the overall effect of employing certain deliberative procedures, or with the particular outcomes that her procedures recommend? Surely the former, for it is the effect of employing certain procedures that bears most directly on her aim that her life go as well as possible. And so I conclude that 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, where this effect includes, not only the actions they determine, but also the actions they make possible. We may now agree with our objector that an action is rational if and only if it is adequately supported by rational deliberative procedures. But since the direct link between rational deliberation and particular outcomes has been severed, an action may be rational even though at the time of performance it is not, and is not believed to be, part of a life that goes best for the agent.

If I may expect my life to go better if I am able to offer and honor an assurance, and I then do so, than if I choose any other course of action, then my life goes best if, in deciding whether to honor an assurance, I employ deliberative procedures that are not straightforwardly directed at successful actions, that is, to those actions that at the time of performance would be part of a life that goes as well as possible for me. So these procedures are rational, and the actions they determine are rational, even though I may not expect them to satisfy the standard of success.

The alleged parallel that treats success in action as playing a role comparable to truth in belief is mistaken. There is no reason to sup-

pose that there is a standard for particular actions that provides the appropriate focus for deliberative procedures in the way that truth provides the appropriate focus for epistemic procedures. This is not to deny that there is a truth of the matter about whether actions are rational or not. But this truth, on my view, is settled by relating actions to deliberation, and the truth about the rationality of deliberative procedures is settled by determining which ones will prove most conducive to the agent's aim.

Our objector must suppose that, although an agent's deliberation takes its rationality from her overall aim, it does so in such a way that rational deliberation may prove self-defeating in relation to that aim. Even if a rational agent is able and indeed rationally required to turn herself into an appropriately irrational deliberator, and so to avoid actually engaging in self-defeating activity, the objector cannot deny the self-defeating character of rational deliberation itself but can only claim that its self-defeating character does not entail that an agent's attempt to realize her aim must be defeated in practice. I do not claim that there is any inconsistency or incoherence of a strictly formal nature in this position. The objector is able to give an account of both the way in which deliberation takes its rationality from the aim and the way in which it is then self-defeating. I claim only that one can give an alternative account of the way in which deliberation takes its rationality from the agent's aim, such that deliberation is not self-defeating.¹⁵ On such an account, an agent need not resort to acquiring an appropriate form of irrationality in order to avoid being defeated in practice by her deliberations. And it is surely mistaken to treat rational deliberation as self-defeating, if a non-self-defeating account is available.

V

May we at last dismiss our objector? I have claimed to have an alternative to his account of the relation between my aim and my reasons for acting; I must now make that claim good. How should we characterize rational deliberation? Consider this proposal. Let us say that an action x is intentionally compatible with the prior actions of an agent, if each of those actions could have been performed intentionally had the agent consciously intended to perform x .¹⁶ If I sincerely assure you that I shall reciprocate your assistance, then not reciprocating is not

15. Or perhaps I should claim only that one can give an alternative account that minimizes the self-defeating effect of changing temporal perspectives; see Sec. X below.

16. Could have been performed, not would have been performed. I buy a ticket to fly to Boston with the intention of using it; I would not have bought the ticket absent this intention. But not using it is intentionally compatible with having bought it; I could have bought the ticket with no intention to use it.

intentionally compatible with my prior actions; I could not have sincerely assured you that I should reciprocate had I intended not to do so. Then deliberative procedures are rational if and only if they select an action that is intentionally compatible with the agent's prior actions and part of a life that she expects to go at least as well as any in which her action is intentionally compatible with her prior actions. And an action is rational if and only if it is or would be selected by such a deliberative procedure. Thus only actions intentionally compatible with one's prior actions can be rational.

But as it stands, this proposal is clearly unsatisfactory. For suppose that foolishly but sincerely, I promise you that I shall meet you tomorrow come what may, quite forgetting that I need the time to prepare for an interview of great importance to my future. Assuming that our meeting has no great significance, so that it will be no more than a minor inconvenience to you if I fail to show, then it is surely clear that it is rational for me to break my promise. But breaking my promise is intentionally incompatible with my prior actions, since one of these was sincerely making the unconditional promise, and so it is irrational according to my proposal.

One way to meet the problem of foolish assurances and their like would be to suppose that rational deliberation should select an action that is intentionally compatible with the agent's prior rational actions, and part of a life that she expects to go at least as well as any in which her action is intentionally compatible with her prior rational actions. If I foolishly promise to meet you, then breaking my promise does not thereby become intentionally incompatible with any of my prior rational actions. And since breaking my promise is part of the life that goes best for me, it is rational for me to do it.

The original proposal fails because it rejects perfectly rational acts as irrational. But the revision accepts irrational acts as rational. For suppose that, although I have good reason to offer you some assurance about my future behavior, the actual assurance I give you is one that I should realize is less beneficial or more costly to me than necessary. For example, I offer you a sum of money, say \$100, in return for some object of yours when I should know that you would be happy to part with it for half the amount. But if you give me the object, it may then occur to me that I should be better off not paying you; the circumstances may be such that you can neither make me pay nor blacken my reputation if I do not. To be sure, not paying you would be intentionally incompatible with my offering you \$100, but I realize that I did not act rationally in offering you so much, since I was in a position to know that you would have accepted \$50. So not paying you would not be intentionally incompatible with any of my prior rational acts, and it makes my life go better than any other act now open to me. Hence the revised

proposal treats not paying you as the rational thing for me to do. But this is surely wrongheaded.¹⁷

You are willing to sell me some object, provided you expect me to pay you at least \$50. And I am willing to buy that object, even if I pay you \$100—for this is what I offer you. But if you think that, whatever I may offer, I shall actually pay you only if at the time set for payment I consider that my offer was the best I could have made, then you would be foolish to expect payment simply because I make you an offer, and so foolish to deal with me. In order to obtain the desired object, I want to induce in you the expectation that I shall pay you, but I am unlikely to succeed if you know that I consider it rational to renege on any offer that I come to consider excessive. And so despite your willingness to sell and mine to buy, we fail to exchange. My life goes less well than had I been prepared (and had you then believed that I had been prepared) to carry out my actual offer, even though I should have made (and my life would have gone even better had I made) a different offer.

How may we distinguish assurances that it is rational to honor from those that it is not? Let us begin with the idea that an assurance is foolish if the agent could expect to do better (though not necessarily best) were he to offer no assurance at all. Promising to meet you come what may is a foolish assurance, since I could expect to do better by offering no assurance at all—although perhaps I could do best by giving a less categorical assurance about some alternative way for us to meet. But offering \$100 for what you would sell for \$50 is not a foolish assurance, even though I could do better by offering less, since I could not expect to do better by making no offer at all if indeed the object is worth \$100 to me. Then it is rational to honor an assurance that at the time of performance one does not expect to lead to one's life going as well as possible, if but only if it does not prove foolish, that is, if but only if at that time one expects honoring it to lead to one's life going better than had one given no assurance at all.

Generalizing this suggestion to provide a suitable characterization of rational deliberation is no easy matter. The key idea is that deliberative procedures should in some cases require the intentional compatibility of one's chosen action with one's previous actions; the problem is to determine the scope of this requirement. As a first approximation to a solution, consider the following. Let us say that an act is potentially intentionally restrictive if and only if the agent might come to face a choice among possible acts some of which would be intentionally incompatible with it. And let us say that an act proves to be actually

17. And it will no doubt strike the reader as patently immoral. But this essay is about rationality, not morality. After we understand rational deliberation we can consider how it relates to morality.

intentionally restrictive if and only if the agent comes to be faced with a choice among possible acts some of which are intentionally incompatible with it. Now in deliberating, if one identifies no act in one's past behavior that is actually intentionally restrictive with respect to the choice at hand, so that none of the acts one might choose is intentionally incompatible with any of one's past acts, then of course one should choose an act that one expects would at the time of performance be part of one's life going as well as possible.

However, if one identifies an actually intentionally restrictive act, then one should proceed to compare two expectations. The first expectation is determined by supposing that one's choice is restricted to those acts intentionally compatible with one's prior acts; the second expectation is determined by supposing that one had performed no potentially intentionally restrictive act.¹⁸ One then constructs a best scenario for each supposition. One considers what act would make one's life go best, among those intentionally compatible with one's past acts, and one forms an expectation about how well one's life would then go. And one considers what would have happened, and what acts would have made or would make one's life go best, had one performed no intentionally restrictive act, and one forms an expectation about how well one's life would have gone. One then compares these two expectations. Either the first is at least as great as the second, or the first is smaller than the second. Consider each case in turn.

If the first expectation, based on a choice among acts intentionally compatible with one's prior acts, is at least as great as the second, based on whatever choices one would have had in the absence of any prior potentially intentionally restrictive act, then one expects one's life to go better accepting the intentional restriction, than had one never restricted oneself. The course of action that includes making and abiding by one's actual intentional restriction is part of a life that (one expects) goes better than the course of action that omits any comparable intentional restriction. It is rational to follow the better course of action, and so one should make the choice that is associated with the first expectation. That is, one should choose an act that one believes will make one's life go best, among those possible acts that are intentionally compatible with one's past behavior.

But now suppose that the first expectation is smaller than the second. Then one expects one's life to go less well accepting the intentional restriction than one expects one's life would have gone had one never restricted oneself. Of course one has restricted oneself. One

18. That is, one supposes not only that one had not performed the act that has proved actually intentionally restrictive but also that one had not performed any potentially intentionally restrictive alternative to it.

cannot make whatever choice or choices one would have made, had one performed no intentionally restrictive act. But that act has proved foolish, and so it is not rational simply to abide by it. One should choose an act that one believes will make one's life go best, among all of the acts actually possible for one.

Note that the procedure I have just characterized is not one in which one compares two expectations and chooses the act associated with the greater. The second expectation, since it depends on having acted differently in the past, need not be associated with any choice that is now open. And even if in some sense the choice is open, so that one may choose from the same possible acts, the second expectation is based on choosing among these acts not in one's actual circumstances, but in the circumstances that one would have been in had one performed no potentially intentionally restrictive act. If the second expectation is greater, one chooses an act that is part of one's life now going as well as possible, but this of course need not be the act that one would have chosen, had one not performed any potentially intentionally restrictive act.

Indeed, the best act open to one may happen to be intentionally compatible with one's past acts despite the folly of those acts. Although it is not rational to abide by a foolish assurance just because it is an assurance, it may still prove rational to abide by it. Suppose as before that I offer you \$100 for some object that you are eager to part with, but I do so on a whim. Almost immediately I regret it, but in this case even though I could renege, I should damage my reputation in ways even more costly to me than paying out \$100. And so I choose to honor my offer, not because doing so is the best act intentionally compatible with having made the offer, but because doing so in the circumstances is simply the best act open to me. However, were renegeing not damaging to my reputation, then doing so would be the best act open to me and I should choose it. Intentional compatibility with past acts drops out of rational deliberation when the intentional restrictions imposed by those acts lead to one's life going less well than one would have expected had one never introduced them.

An agent who accepts this account of rational deliberation will find herself unable wittingly to perform certain potentially intentionally restrictive acts. Let x be an act that is potentially intentionally restrictive with respect to y ; in other words, an agent performing x must recognize the possibility of a future choice in which y is an option and is intentionally incompatible with x . So the agent cannot perform x if she consciously intends, if actually faced with that choice, to perform or choose y . But let us suppose, not that she intends straightforwardly to choose y , but simply that she would include it among the options in her deliberations. For her to do this, she must think it possible for her to perform y , and she must intend to choose y should her deliberations

lead her to consider it best. She may of course be quite sure that she will not come to think it best; nevertheless its presence in her deliberations commits her to the conditional intention to choose it should it prove best. And it seems to me that just as she cannot perform an act potentially intentionally restrictive of y if she intends to choose y , so she cannot perform such an act if she intends to include y in her deliberations, and so to choose y should it prove best.

I shall not try to say more here about the scope of intentional compatibility in rational deliberative procedures, although I do not doubt that further refinements on the account I have given will prove necessary.¹⁹ But here I want to emphasize that, although an adequate account may prove complex, its underlying rationale is very simple. Given that my aim is that my life go as well as possible, then normally, I deliberate rationally by choosing that action that at the time of performance I expect to be part of or to lead to my life going as well as possible. But sometimes my life will go better if I am able to commit myself to an action even though, when or if I perform it, I expect that my life will not thenceforth go as well as it would were I to perform some alternative action. Nevertheless, it is rational to make such a commitment, and to restrict my subsequent deliberation to actions intentionally compatible with it, provided that in so doing I act in a way that I expect will lead to my life going better than I reasonably believe that it would have gone had I not made any commitment. As a rational agent I shall not be able to commit myself to actions if I believe at the time of commitment that performing them would leave me worse off than had I not committed myself, but I shall be able to offer and honor assurances when it is advantageous for me to do so.

VI

Consider Gregory Kavka's well-known "toxin puzzle" in the light of this account of rational deliberation.²⁰ Suppose that a million dollars will be deposited in your bank account at midnight tonight by a wealthy scholar studying choice behavior, if at that time he believes that you

19. For example, suppose an unexpected side effect of an intentionally restrictive act leads to my life going less well as a consequence of having performed this act, than had I not performed it. But I gain the expected benefit from my act and this benefit exceeds the cost of conforming to the intentional restriction. Does the additional cost of the unexpected side effect give me reason not to conform? I am grateful to an anonymous referee who raised this problem. I think that the more complex intentional structures (policies) that I introduce in Sec. IX below give me the resources necessary to resolve it, but I shall leave discussion to another occasion.

20. Gregory S. Kavka, "The Toxin Puzzle," *Analysis* 43 (1983): 33–36. I note that Michael Bratman and I disagree about the resolution of the puzzle; see his *Intention, Plans, and Practical Reason*, pp. 101–6.

sincerely intend to drink a glass of toxin tomorrow morning at eight. You know that the toxin will make you miserably ill for a day, but will have no lasting effects. Suppose also that the prospective donor is notorious for his shrewdness in discerning his fellows' intentions. You have little hope of outsmarting him; you expect that if at midnight tonight you sincerely intend to drink the toxin, he will believe that you do and will deposit the million dollars, whereas if at midnight tonight you do not sincerely intend to drink the toxin, he will believe that you do not and will not deposit the money. But what you actually do tomorrow morning will have no effect; what matters is only the donor's belief tonight about what you intend to do tomorrow.

If you consider that your life will go better with an extra million dollars, even though you must also be sick for a day, you have good reason to form the intention to drink the toxin. You will be better off if you gain the million and drink the toxin, than if you remain as you are now. But you certainly do not want to drink the toxin. You would prefer to form the convincing but insincere simulation of an intention to drink the toxin if you could, but you realize that, alas, only a sincere intention is likely to convince the prospective donor. However, if you take yourself to be rational, then you can form such an intention only if you suppose that tomorrow morning it will not be irrational for you to drink the toxin. But tomorrow morning you will gain only a day's misery by drinking the toxin, and so to many persons, including Kavka, it seems that it could not be rational for you to drink it. On my view they are quite wrong. The rational outcome of your deliberation tomorrow morning is the action that will be part of your life going as well as possible, subject to the constraint that it be compatible with your commitment—in this case, compatible with the sincere intention that you form today to drink the toxin. And so the rational action is to drink the toxin.

To some this seems utterly mad. How, an objector will ask, can it be rational to drink the toxin, when you will only make yourself sick by drinking it? But someone who asks this will surely agree that it would be rational, even at some cost, to ensure before midnight that tomorrow morning at eight, like it or not, you will drink the toxin, if such an arrangement would convince the donor of your intention. Suppose for example that just before midnight you could be laid on a couch, your hands and feet securely bound to prevent your moving or freeing yourself, a tube inserted into your mouth and connected with a vial containing the toxin. A time release lock controls a stopper on the vial; it is set to release the toxin into your mouth at eight in the morning. Another time release lock controls your bonds; it will release them at a suitable interval thereafter. Now if the donor were to accept your so binding yourself as evidence of your intention to drink the toxin (and not as a dodge to avoid having any intention in

the matter), and if no other less unattractive device were at hand, then surely it would be rational for you to employ it. Unless you happen to be a bondage freak, you would no doubt spend a very unpleasant night in addition to suffering a subsequent day's illness, but wouldn't you consider it worth one million dollars? And if it would be rational to inflict a night's misery on yourself in addition to a day's illness, were that the best means to gain one million dollars, then surely it would be rational simply to make up your mind to drink the toxin and avoid the night's misery, since this is a better means to the million.

The objector will agree but insist that this misses the point. Making up your mind today to drink the toxin tomorrow is perfectly rational—the best thing for you to do, if you can do it. But drinking the toxin is still irrational. This of course complicates the task of making up your mind to drink the toxin. Maybe it will prove excessively costly or even impossible. And in admitting this, the objector reveals the weakness that I have already noted in the account of rationality that he is presupposing. Treating the rationality of each action discretely, and as determined strictly by its consequences, the objector finds that in certain situations rationality is a hindrance rather than a help to one's aim that one's life go as well as possible. On his account, the toxin puzzle faces the rational agent with the task of outwitting her own rationality. Rationality must undermine itself.

The objector thinks it mad to drink the toxin. I, on the other hand, think it mad not to be the sort of person who would drink the toxin. And I see no ground for treating sanity, manifested here in the willingness to drink the toxin, as anything other than rational. An agent who grasps the relation between intending and acting, and who reflects on situations in which an intention affects one's situation in ways independent of the intended action, will understand how her reasons for performing an action can derive from her reasons for forming the preceding intention, rather than the other way round. She will drink the toxin. And for her, though not for Kavka and those who refuse to drink, rationality will give coherent expression in her actions to her aim.

VII

Or so one might wish. But I must turn now to situations of a type that may not seem to be well accommodated by my proposed characterization of rational deliberation—situations involving threats. A threat expresses a conditional intention: if you act in a certain way, then I shall respond with a retaliatory act (or omission). But of course not all expressions of conditional intention constitute threats. For me to threaten you, I must first of all suppose that the costs you would expect from my retaliatory act outweigh the benefits you would expect from ignoring my threat, so that on balance you prefer giving in to the

threat over ignoring it and facing retaliation. But this is not sufficient; pointing out to you that, should you not help me harvest my crops, I shall in turn not help you harvest yours, is not threatening you, even though I suppose that you prefer mutual assistance to harvesting alone. The act that I threaten to perform must be one that is in some way inappropriate on prudential or moral or legal grounds for me as a response to your behavior. Here my concern is only with the first of these, morality and law falling outside the scope of the discussion, and so with threats in which the retaliatory act is one that it would be costly for me to perform, were you to ignore my threat. If I were to retaliate, then I should not expect my life to go as well for me thereafter as it might. Thus for present purposes, a threat (if sincere) must be supposed by its issuer to commit her conditionally to a retaliatory act that would make the threatened party's life go less well than if he were not to incur it, and her own life to go less well than if she were not to execute it.

To illustrate a threat situation I shall relate another farming story. Our farms are connected to the public road by a longish lane, yours lying a short way beyond mine. In the past you have maintained the portion of the lane between our farms which only you use, and we have cooperated in maintaining the remainder. But now you propose to cease cooperating, reasoning that I shall need to maintain the lane from my farm to the public road whether or not I have your assistance, so that you can be a free rider. I respond by threatening to let the lane go unmaintained if you cease to cooperate. If, as we may reasonably suppose, the costs to you of not having the lane maintained outweigh the benefits you would expect from not participating in its maintenance, and the benefits to me of maintaining the lane exceed the costs of being the sole maintainer, then the conditions that characterize a threat are satisfied. A diagrammatic comparison of the structure of this situation with the structure of the harvesting situation reveals perspicuously the deep differences between threats and assurances (fig. 1).²¹ One threatens to avoid being put in an undesirable position; one assures to gain access to a desirable position. The threat succeeds if one does not find oneself faced with and committed to the choice of an action that does not lead to one's life thenceforth going as well as possible; the assurance succeeds if one does find oneself faced with and committed to such a choice.

21. The two correspond to the "pure threat" and "pure promise" distinguished by Daniel Klein in formalizing the pioneering work on commitment by Thomas Schelling. See Daniel B. Klein, "A Game-Theoretic Rendering of Promises and Threats," Irvine Economics Paper no. 90-91-21 (University of California, Irvine, 1991); and Thomas C. Schelling, *A Strategy of Conflict* (Cambridge, Mass.: Harvard University Press, 1960), chaps. 2, 5, 7.

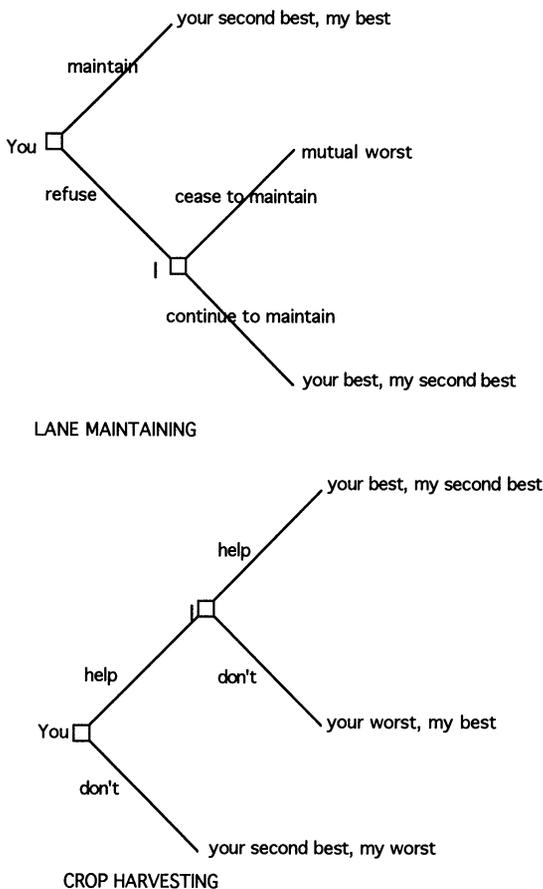


FIG. 1

In threatening not to maintain our lane without your cooperation, I seek to affect your expectation about your benefits and costs should you cease to share in its maintenance, in such a way that you will choose not to cease. A successful threat will be most advantageous for me, and so if I may reasonably expect a threat to be successful, issuing it would seem to be my best option. And we may suppose that, just as in harvesting I could not reasonably expect to gain your help with an insincere assurance of reciprocal help, so here I cannot reasonably expect to deter your ceasing to share in the maintenance of our lane with an insincere threat to cease myself. I must be prepared to carry out my threat should you cease, and so I must consider whether ceasing to maintain the lane would be rational for me.

Let us then consider whether deliberating rationally about whether to carry out a failed threat is parallel to deliberating about

carrying out a successful assurance. Having made a sincere threat which has failed to deter you, I recognize that it has become an actually intentionally restrictive act, and so I consider those actions now open to me that are intentionally compatible with it. If my choices are to continue maintaining the lane or to cease, clearly only the latter is intentionally compatible with my threat, and so it is by default the best. I now compare how I expect my life to go were I to cease maintaining the lane, with how I should have expected my life to have gone had I not issued a threat—or issued one not as a commitment, but only as a bluff. Since I should have expected you to withdraw your assistance in response to no threat or to a bluff, the comparison is between leaving the lane unmaintained and maintaining it without your assistance. I prefer the latter, and indeed, must prefer it if what I issued was a genuine threat. Thus I believe that acting in a manner intentionally compatible with my failed threat would lead to my life going less well than I reasonably believe that it would have gone had I not performed any potentially intentionally restrictive act. I should now be better off had I not issued a threat, and I shall be better off if I do not carry it out than if I do. But if this is so, then can it be rational for me to carry it out—to make my life go, not only less well than it could given the alternatives now open to me, but less well than it would have gone had I not committed myself in the first place? Not according to the account of rational deliberation that I proposed in Section V above. I cannot rationally carry out my threat, and once I realize this, I am unable to seek to deter you from withdrawing your assistance in maintaining our lane by threatening to retaliate by not maintaining it myself.

We may generalize from this particular example. In discussing assurances, I distinguished those that are rationally honored from those that are foolish. I took an assurance to be foolish if honoring it would be expected by the agent to be more costly than not offering it. And I proposed an account of deliberation that would make it rational to honor assurances if one thereby stood to benefit in comparison with not having made an assurance, but not rational to honor ones that prove foolish. Suppose now that we seek similarly to distinguish threats that are or would be rationally executed from those that are foolish, following the characterization of rational deliberation that I generalized from the case of assurances. One threatens to do something disadvantageous—something that one believes would leave one worse off than were one instead to acquiesce in what one seeks to deter the other from doing. Thus an agent must expect that executing a threat, should it fail to deter, will impose a cost on her greater than she would have expected to bear had she not made it. (Indeed, executing it will only prove less costly if the agent finds herself to have underestimated the cost of acquiescence or overestimated the cost of

executing her threat. And this she cannot have expected.) A threat is a potentially intentionally restrictive act. If it is ignored, then it becomes actually intentionally restrictive, and in such a way that maintaining intentional compatibility with it—carrying it out—leaves the agent worse off than she would expect to have been had she not issued it, and so had she performed no potentially intentionally restrictive act. And since carrying out her threat also leaves the agent worse off than not doing so, then according to my account of rational deliberation it is not rational for her to carry it out. If we extend my account of foolish assurances to threats, we find that threats are normally foolish. The irrationality of honoring foolish assurances seems then to extend to the irrationality of carrying out normal threats.

If one's aim is that one's life go as well as possible, then one will want to offer an assurance or issue a threat if by so doing one expects one's life to go best. But although carrying out an assurance may be part of a course of action that leads to one's life going better than any other one could have undertaken, carrying out a threat cannot be expected to be part of a best course of action. One may offer as one's reason for carrying out an assurance, that one's life will go better than if one had not made the assurance, but one cannot offer a parallel reason for carrying out a threat. Without such a reason, one would act irrationally in doing what one did not expect would thenceforth make one's life go best, and so one would act irrationally in carrying out the threat. And if one did not expect to have such a reason, one could not rationally do what one realized might intentionally restrict one to acts that would be irrational without it. As a rational agent I am able to offer sincere assurances, but it seems that I am unable to issue sincere threats.

VIII

One might be happy with this result. Assurance behavior is frequently rational; threat behavior is never rational. We may think that it would be nice if this were so. I began my discussion of threats by suggesting that they seem not to be well accommodated by my account of rational deliberation, but one might see that as a virtue of the account. It would be personally somewhat embarrassing, since I have on more than one occasion defended the rationality of deterrence, which is a form of threat behavior.²² If the account of rational deliberation that I have sketched here is correct, then my defense of deterrence is mistaken. Even if embarrassed, I should not be altogether unhappy to discover this. But if my proposed account of rational deliberation is correct,

22. See "Deterrence, Maximization, and Rationality," and "War and Nuclear Deterrence," both cited in n. 4 above.

there seem to be other less pleasing consequences. Here is an example that reveals them (see fig. 2).

In this situation chance determines whether mutual help is or is not more beneficial to me than no help. And chance determines this, I am supposing, in such a way that the estimated probability that mutual help is a benefit is .8; the probabilities of the alternative chance results are shown on the diagram. My initial expectation is that mutual help will benefit me; I determine this by calculating the probability-weighted measure of how well my life may expected to go, which is $(.8 \times 8) + (.2 \times 0)$, or 6.4, and comparing it with the measure if you do not help me, 5. So assuring you that I shall return your help seems, on the face of it, rational; I expect to do better making and honoring a commitment than making no commitment.²³ But suppose that I give you the assurance and find myself unlucky; if I help you, I shall end up with 0. Then I should clearly have been better off not to have committed myself to helping you, and if it is not rational for me to carry out a commitment that, as it turns out, leaves me worse off than had I not made it, I shall conclude that I should not honor my assurance. It will prove rational for me to honor my assurance if and only if helping you yields me a payoff of 8. Thus the only sincere assurance I could give you would be conditional; I shall return your help if but only if chance is favorable, so that mutual help will leave me better

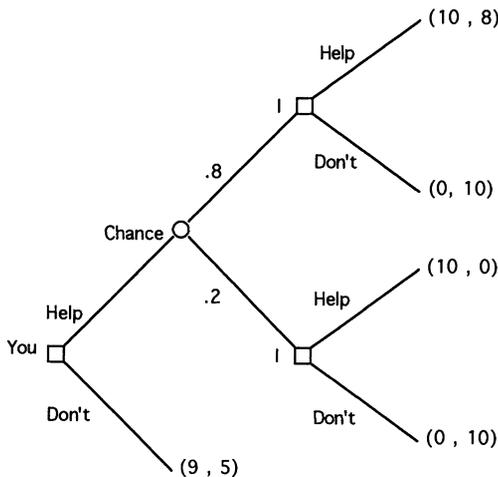


FIG. 2.—The parenthesized numbers are measures of how well one's life goes; yours appear first.

23. I am here assuming the orthodox account of rational decision making under risk as involving the maximization of expected utility. I am no longer as convinced as I once was that this is the uniquely rational procedure, but the problem that I am examining arises for other plausible procedures for risky choices.

off than had you not helped me. But such an assurance will not suffice to elicit your help. For if you calculate the probability-weighted measure of how well your life will go should you help me, given such a conditional assurance, you find it to be $(.8 \times 10) + (.2 \times 0)$, or 8, which is less than the measure if you do not help me, 9. Only an unconditional assurance would elicit your help, by giving you an expected measure of 10 should you help me.

Yet surely it is to my advantage to be able to give you an effective assurance. It may seem evident that I should revise my account of rational deliberation to accommodate this. And this revision may seem easily done. What is needed is to relate posterior deliberation, about whether to honor an assurance or to execute a threat, to prior deliberation, about whether to offer an assurance or to issue a threat. In my example, calculating on the basis of estimated probabilities I determine that the expected net benefit of gaining your assistance by assuring you of mine in return is positive, in relation to the outcome if neither of us assists the other. And so I consider it rational to offer a sincere assurance of reciprocal help. When I come to consider whether or not to honor my assistance, I simply refer back to my previous deliberation; if I reaffirm my judgment that in the circumstances as I then knew them, it was rational for me to issue an assurance, then I conclude that I should honor it, even if circumstances as I now know them are such that I should in fact have done better never to have given the assurance. Similarly, in considering whether to threaten you with ceasing to maintain our lane should you terminate your participation, I estimate the probability of my threat being effective, and calculate whether, summing the probability-weighted benefit of securing your continued participation in lane maintenance should my threat succeed with the probability-weighted cost of having the lane unmaintained should my threat fail, I may on balance expect my life to go better than it would go were I to maintain the lane alone. And if I expect my life to go better, then it is rational for me to issue a threat and I do so. Should it fail, I refer back to my previous deliberation; if I reaffirm my judgment that, in the circumstances as I then knew them, it was rational for me to issue a threat, then I decide to execute it, even though circumstances as I now know them are such that I should in fact have done better never to have made it.

I have in the past defended this account of deliberation. I have claimed that it allows an agent to establish a coherence in her actions and her life that is precluded by the orthodox account, which requires her on each occasion to choose the action which at that time she expects to make her life go as well as possible. The agent who accepts orthodoxy's canon of reason cannot, except by indirection, interrupt the relentlessly forward-looking character of her deliberations to commit herself in ways that would plainly be advantageous to her. The

alternative that I have just sketched allows such commitment, by licensing deliberation that is backward-looking in its appeal to previously adopted intentions or plans, to assurances given and threats issued. And it is not mindlessly backward-looking; the agent need not respect her past intentions if she finds herself in circumstances quite different from those she envisaged in forming them, or if she comes to believe that she did not have good reason *at the time* to form them. But the phrase that I have italicized should give us pause.

I issue a threat; you are undeterred by it. I am now, we may suppose, in exactly the situation I envisaged should my threat fail. Furthermore I may even have thought its failure likely. For it is not simply the likelihood of success, but that likelihood weighted by the benefits of success and the costs of failure, that determines whether making a threat has greater expected value than any alternative. If a successful threat would afford me a great benefit, whereas a failed threat would subject me only to a small loss, then I may have judged it worthwhile to threaten you even though I anticipated failure. So let us suppose that my threat having failed, I am in the situation I expected to be in on balance, and it is as I expected it to be. Nevertheless, my beliefs must be significantly altered by the failure of my threat.²⁴ For in making it I believed threatening to be the best course of action open to me, and I now know that belief to have been mistaken. I issued the threat on the basis of my prior estimate of its likely success, and however reasonable that estimate may have been, nevertheless issuing the threat was in fact not to my advantage. And this I did not know, given that I acted rationally in issuing the threat. I did not know that my life would in fact go worse as a result of making the threat, than had I not made it. If nevertheless I carry out my threat, then it would seem that I am acting in direct disregard of this new knowledge, which establishes the failure of my course of action to result in my life going as well as possible. How then can carrying out the threat be rationally related to my aim?

If we suppose that a rational agent will update her assessment of the rationality of threat behavior in terms of her present knowledge of the effect of her threat on how her life goes, then we shall conclude that it is normally not rational to carry out a threat. Equally, we shall conclude that it is not rational to honor an assurance, if one finds that doing so would leave one worse off than had one given no assurance. But we shall not conclude that it is not rational to honor an assurance in those more typical cases in which doing so leaves one better off

24. I begin here the argument that, it now seems to me, undercuts the defense of the rationality of deterrence, and more generally of threat behavior, that I offered in my earlier papers about deterrence, referred to in n. 4 above.

than one would expect to be, had one given no assurance. For finding oneself in the situation in which one is called upon to honor one's assurance gives one no new knowledge showing that contrary to expectation one's course of action has not led to one's life going as well as possible. If I sincerely assure you that I shall return your help, and you then help me, I surely have every reason to think that my assurance has yielded exactly the result I desired—that you have helped me because I offered it, and would not have helped me otherwise.

Thus the argument that I have sketched against the simple subordination of posterior deliberation to prior deliberation, does not return us to the strictly forward-looking view of deliberation which I have been seeking to undermine. Rather, it returns us to the idea that, in deliberating rationally, one considers whether one's course of action is best conducive to one's life going as well as possible, where a course of action is distinguished and demarcated by its intentional structure. One acts rationally in doing what, among those possible actions intentionally compatible with one's previous behavior, will lead to one's life going best, provided one expects to do better than one would have done had one not performed any potentially intentionally restrictive acts that have proved relevant to one's choice. And so it seems that one would not act rationally in carrying out threats and assurances that, however great the benefits one would anticipate in issuing them, prove to require making oneself worse off than had one not issued them. And if this would indeed be irrational, then a rational agent could not knowingly and sincerely issue such threats and assurances; she would be unable to perform the necessary potentially intentionally restrictive acts.

IX

In performing a potentially intentionally restrictive act, an agent creates an intentional structure for his future conduct. To this point I have considered two very simple forms of intentional structure, assurances and threats. I have shown that intentional structures create problems for the orthodox account of deliberation, which insists that rational actions are those that directly promote the agent's aim, taking as illustrative the aim that one's life go as well as possible. I have proposed an alternative account of deliberation that avoids some of these problems but, as I have just argued, seems to leave the rational agent unable to threaten sincerely and able to assure sincerely only when his assurance doesn't require him to risk having to choose to act in a way that would lead to his life going less well than he believes that it would had he not offered it. But this apparently severe limitation on my account of what is rationally possible arises only because I have examined a very restrictive range of intentional structures.

In addition to particular assurances and threats, an agent may have policies that require him to offer assurances or to issue threats

should he find himself in various specified kinds of circumstances. These policies are themselves potentially intentionally restrictive. A firm can sincerely adopt a policy of retaliating against price-cutting by its competitors only if it intends to retaliate should its competitors cut prices. Failure to retaliate against price-cutting would be intentionally incompatible with its policy, just as failure to reciprocate your assistance in harvesting would be intentionally incompatible with my assurance. Now it may be that in a particular situation retaliation is disadvantageous to the firm, even taking such long-term considerations as reputation effects into account. Nevertheless, the firm, or its officers, may reasonably believe that the policy of retaliation is beneficial overall, so that retaliating in the present situation leaves the firm better off than it would have expected to be, had it not adopted a policy requiring it to perform potentially intentionally restrictive acts. And so, given the account of deliberation that I have offered, it is rational for the firm to retaliate.

Consider an agent who adopts a policy of issuing and enforcing threats. She plans to issue threats that maximize her expectation of benefit, taking the probability-weighted deterrent effect and enforcement cost together. In this way she provides an intentional structure for future threat-issuing behavior. And she forms the general intention of executing these threats, thus providing a further intentional context for each particular threat that she issues, over and above the intention involved in its issuance. Faced with the failure of one of her threats to deter, she considers whether she is better off executing it than if she had never created the intentional structure of which it forms part. Now she must consider not only what she would have expected had she not issued the threat in the first place, or perhaps issued it but only as a bluff. For her policy requires not just the enforcement but also the issuance of threats. If she issued the threat as part of a general policy, then not fulfilling it is intentionally incompatible with that policy. She must compare the outcome of fulfilling it with what she would have expected had she not adopted a threat policy. And even though executing her failed threat does make her life go worse in comparison with not having issued it, it need not make her life go less well than had she not adopted that general policy. The benefits that she has gained, or may expect to gain, from her threat policy, may well result in her life going better than had she not adopted it, even though she must face the costs of executing her failed threats.

Embedding particular threats, or of course particular assurances, in a policy of threatening or assuring, makes it possible for an agent rationally to expose herself to greater risk than if she formed intentions about her future behavior merely on a case-by-case basis. She can sincerely threaten, or assure despite risk, knowing that it will be rational for her to carry out her threat or assurance as long as she expects,

at the time of performance, that she is better off so doing than had she formed no potentially intentionally restrictive policy. But for her policy to pass rational scrutiny, it must incorporate a significant limitation. Suppose she considers it on balance advantageous to issue what I shall call an apocalyptic threat—one that, should it fail, would require her to bring utter disaster on her head. She reasons that the probability of her threat failing is sufficiently small that, despite the enormous cost of failure, she would maximize her expectation of how well her life would go by issuing it. So she might take issuing an apocalyptic threat to be part of her threat policy. But should it fail, she would find herself faced with a cost that would outweigh all of the benefits she had gained, or might expect to gain, from her overall policy. At that point she would expect her life to go less well, were she to enforce her threat, than it would have gone had she not embarked on any policy of issuing and enforcing threats. And so she would not consider it rational to fulfill her apocalyptic threat. But then the making of such a threat could not be required by any policy that she could rationally and sincerely adopt. However advantageous in prospect a threat might be, the possibility of sincerely issuing it, and so of adopting a policy requiring or permitting its issuance, must depend on its expected costs, should one be called upon to execute it, being offset by the overall expected benefits of the policy, so that on balance adopting the policy is more advantageous than adopting no policy. A rational agent cannot sincerely and wittingly issue an apocalyptic threat. Rational deterrence is limited in ways that I have previously failed to recognize.

Enlarging the account of deliberation that I have offered in this paper to embrace policies, I have been able to show that it does not distinguish sharply between assurances and threats in a way that licenses the former but forbids the latter. There are contexts in which an agent would find a threat policy rational. Such contexts may involve the irrationality of other persons, since the rational response to the would-be threatener may be a policy of threat resistance. In a world of fully rational persons threats might prove altogether irrational; perhaps this distinguishes them from assurances. But I cannot enquire into these matters further here. My concern is only to illustrate the role of an expanded intentional structure in broadening the range of actions that, even though they may fail to make the agent's life go as well as possible, nevertheless may prove to satisfy the standards of deliberative rationality.

X

Have I succeeded in offering an account of deliberation in which my reasons for acting are taken from my aim, and, if I act successfully in accordance with my reasons, I do as well as possible in relation to my

aim? Have I succeeded in overcoming the problem that I found in Kavka's treatment of the toxin puzzle—that on his view the rational agent faces the task of outwitting her own rationality? It may seem not. For I have claimed that it would not be rational for a person to execute an apocalyptic threat, even if she had reasonably expected that her life would go best were she to issue such a threat. Since a rational agent cannot intend what she believes she will not have reason to do, there are intentional structures that she is unable to erect, even though she would expect to benefit from erecting them. Limits on the rationality of carrying out intentions entail limits on the rational capacity to form intentions. I may then act successfully in accordance with my reasons, but the limits that rationality sets on my capacity to intend may prevent me from doing as well as possible in relation to my aim. I may be in a situation in which a sincere apocalyptic threat would in fact save me from serious misfortune, yet be unable to issue the threat. Or less dramatically, I may be in a situation in which an unconditional assurance would in fact benefit me greatly, yet be unable to offer it, because I recognize that at the time of performance I might find conditions such that honoring it would make my life go less well than had I adopted no intentional structure of which it was part.

One might then suppose that I should return to my earlier revisionist account, in which posterior deliberation is simply subordinated to prior deliberation. Despite the argument I sketched against this view in Section VIII, it might seem to offer an account of reasons such that an agent who acted successfully in accordance with these reasons would do as well as possible with respect to his aim. But this is not so. Suppose that posterior deliberation is subordinated to prior deliberation, so that an agent may rationally issue an apocalyptic threat in circumstances in which it offers her better expected prospects than any alternative. An agent who supposes that she has reason, taken from her aim, to create an intentional structure because she expects such a structure to lead to her life going better, must also suppose that she has reason to act compatibly with that structure even when she recognizes her expectation to be mistaken. But an agent who takes herself to have reason to execute an apocalyptic threat fails, in acting on this reason, to do as well as possible in relation to her aim. In subordinating her posterior deliberation entirely to her prior deliberation, she persists in a course of action that she recognizes is leading her to disaster and so is directly contrary to her aim. In order that she may act so that her life would go best overall as judged from her prior expectation, she must be willing to act in ways incompatible with her life going best overall as judged from her posterior realization.

Reasonable expectations about the benefits and costs of different actions or courses of action may vary with time. Any account of rational deliberation must accommodate this variation. There is no way to

characterize reasons for acting so that an agent may be certain that, acting successfully in accordance with her reasons, her life may be expected to go best as judged from every temporal perspective. I have supposed that an agent's reasons must relate her actions to her aim as judged at the time of performance. But I have also supposed that an agent's reasons must relate her actions to the intentional structures that enable her better to fulfill her aim. Trying to accommodate both of these suppositions has led me to offer the present account as a beginning toward an adequate theory of rational deliberation.

I act to determine how my life will go. In deciding how to act I look forward, to the effects different possible actions may be expected to have. I take reasons for acting from those expected effects. But when I look forward, I see more than the discrete effects of particular actions. How my life will go will depend in part on the intentional structures I create. For my life to go as well as possible, I require a mastery over my future choices that enables me to give assurances and, at least in dealing with those less than fully rational, to make threats. I must be able to take reasons for acting from the intentions that these commitments embody, over and above those reasons that arise directly from expected effects of my actions. And so I have been concerned here with the role of intentional structures in enabling me to "stand as my own guarantor"—the phrase is adapted from Nietzsche.²⁵ An adequate theory of rational deliberation, did we have one, would articulate fully what I have only begun to sketch—the interplay of intention and reason that makes possible the realization of this ideal.

25. See Friedrich Nietzsche, *On the Genealogy of Morals* (1887), 2d essay, sec. 2, trans. W. Kaufmann and R. J. Hollingdale, ed. W. Kaufmann (New York: Vintage Books, 1967), p. 59.