Assuring, Threatening, a Fully Maximizing Theory of Practical Rationality, and the Practical Duties of Agents

Author(s): Duncan MacIntosh

Source: *Ethics*, Vol. 123, No. 4, Symposium: David Gauthier's Morals by Agreement (July 2013), pp. 625-656
Published by: The University of Chicago Press
Stable URL: http://www.jstor.org/stable/10.1086/670248
Accessed: 11/07/2013 13:35

Assuring, Threatening, a Fully Maximizing Theory of Practical Rationality, and the Practical Duties of Agents*

Duncan MacIntosh

Theories of practical rationality say when it is rational to form and fulfill intentions to do actions. David Gauthier says the correct theory would be the one our obeying would best advance the aim of rationality, something Humeans take to be the satisfaction of one's desires. I use this test to evaluate the received theory and Gauthier's 1984 and 1994 theories. I find problems with the theories and then offer a theory superior by Gauthier's test and immune to the problems. On this theory, it is rational to treat something different as the aim when doing so would advance the original aim. I argue that the idea that this would be irrational bad faith entails contradictions and so is false, as must be theories saying that rationally we must always treat as the aim the bringing about of objectively good states of affairs or obeying a universalizable moral code.

I. INTRODUCTION

Theories of practical rationality say when it is rational to form and fulfill intentions to do actions. David Gauthier says the correct theory would be the one our obeying which would best advance the aim of rationality, something Humeans take to be the satisfaction of one's desires. I shall use this test to evaluate the received theory and Gauthier's 1984 and 1994

* For general discussion, thanks to my students in my classes on the theory of rational decision, and to Richmond Campbell, David Copp, Louise Daoust, Claire Finkelstein, Alice MacLachlan, Greg Scherkoske, Heidi Tiedke, Terry Tomkow, and Lisa White. I'm especially grateful for meticulous and probing comments from referees of previous versions, especially two referees and the Editor and two associate editors for *Ethics*. Thanks also to the audience at the conference, "Contractarian Moral Theory: The 25th Anniversary of *Morals by Agreement*," in particular, Chrisoula Andreou, Michael Bratman, Stephen Kuhn, Malcolm Murray, Robert Sugden, and Bruno Verbeek. Finally, my thanks to Susan Dimock for her organizational and editorial work on the conference and on this symposium. Early work on this article was supported by the Social Sciences and Humanities Research Council of Canada.

Ethics 123 (July 2013): 625–656

theories. I find problems with the theories and then offer a theory superior by Gauthier's test and immune to the problems.

The received theory counts an agent's action as rational only if she believes that doing it will advance the aim. Gauthier's theories claim that actions she believes will not advance the aim can so count provided she believes that forming intentions to do them will advance the aim by making other agents advance it for her. Gauthier's 1984 theory supposedly counts as rational, forming and fulfilling assurance intentions to cooperate in Prisoner's Dilemmas, and conditional threat intentions to retaliate in Deterrence Paradoxes; his 1994 theory, only assurance intentions. I argue that his theories will count the same actions rational as the received theory, making neither assurances nor threats rational. But all theories pronouncing such intentions irrational entail contradictions and are unfollowable in the foregoing situations. And the only possible noncontradictory, always-followable theory that would pass Gauthier's test would count both as rational. A theory can do this only if it has it that, when it would evidently be aim-advancing to form such an intention, it would be rational first to revise one's aim into one that would be advanced by fulfilling it. (For example, in Prisoner's Dilemmas, one must come to care more about keeping promises than about one's original goals.) Many philosophers will think revising the aim on this pretext to be bad faith and so irrational. But this view too entails contradictions and so is false, as are theories claiming that we rationally must always treat as the aim such things as bringing about objectively good states of affairs, or obeying a universalizable moral code.

II. GAUTHIER'S ADEQUACY CRITERIA FOR A THEORY OF PRACTICAL RATIONALITY

Gauthier thinks the correct theory must fit the following principles.[1]

(i) The rationality of intentions and actions is dictated by the correct style of rational deliberation.

(ii) This style is dictated by the aim of rational choice, which is that one's life go best (perhaps as measured by whether one's desires get satisfied);[2] so the correct style is the one the following of which makes one's life go best.[3]

1. David Gauthier, "Assure and Threaten," *Ethics* 104 (1994): 690–721.

2. The content of the desires is left open (ibid., 690–91): they may be desires for (or even against) one's own welfare or that of others; or for supposedly objectively good states of affairs, or to obey universalizable moral laws; or just for this and that. So for one's life to go best is not necessarily for one to have welfare; it is just for one to attain the ends one desires. Thus whenever, following Gauthier, I speak of one's life going better, or of one being better off, I mean only that things are more as one desires them to be; or that the world is going better relative to one's desires; or that one is better attaining one's aims.

3. Ibid., 719.

(iii) It is rational to intend to do an action only if one can expect it will be rational to do it.
(iv) It is never rational to act on an intention if one expects this to leave one's life worse than had one never so intended.
(v) So it is never rational to form an intention to do such an action.[4]
(vi) But the way a rational person deliberates about what to intend and how to act need not be taken directly from the aim of rational choice in this way: that she does only actions that make her life go best and intends to do only such actions.
(vii) She rationally should use a more indirect style if this will make her life go better.[5]

How then do the above theories fare by these criteria? I begin with the received theory.

III. THE RECEIVED THEORY AND GAUTHIER'S FIRST ALTERNATIVE

Before Gauthier it was thought that one's actions are rational (or at least instrumentally rational) just if one thinks they are the means to one's ends; and one's intentions are rational just if they are intentions to do such actions. More precisely, and couched in terms of decision theory, one's actions are rational if one thinks they maximize one's expected utility ("maximize," for short), one's intentions, just if one takes them to be intentions to do maximizing actions.[6] (An agent who exclusively so reasons is a "straightforward maximizer.") This is plausible for parametric choice situations where only one's actions influence which outcomes one will experience. But a theory must also be plausible for strategic situations where the outcomes partly depend on choices of other agents, agents influenceable by the predictions they make of one's choices with their knowledge of what factors one will find decisive in choosing. Examples are the one-shot, multi-stage Prisoner's Dilemma (PD) and the Deterrence Paradox (DP). In these situations, the possibility of one's gaining depends on one's being able to commit to doing an action it would disadvantage one's self to do, the commitment giving others evidence that one will so act. Consider the one-shot, multi-stage PD: as in the one-shot, single-stage PD, the agents prefer the outcome of their defecting (from any agreement

4. Ibid., 717.
5. Ibid., 692.
6. Assuming one's utility is higher as states of affairs obtain that one more highly prefers, an action maximizes one's expected utility if doing it compared to doing any other makes as high as possible the sum of the products of the utility of each outcome which doing the action might cause, and the odds of that outcome's obtaining.

to cooperate) and the other cooperating; second, both cooperating; third, both defecting; and least, their cooperating and the other defecting. But unlike in the single-stage PD where the agents must choose knowing only each others' preferences, beliefs, and the fact that each is rational, in the multi-stage variant agents first choose strategies for how to choose actions; then they internalize dispositions to follow the strategies; then they meet to try to read each others' dispositions; they then have the option of promising cooperation (or of forming a commitment, intention, plan, or resolution to cooperate); then they are separated and given the choice of cooperating or defecting. Prior to Gauthier it was thought rationally obligatory to defect; for no matter what the other agent does one does better by defecting—it secures a chance at one's best outcome and saves one from one's worst, while if one cooperates one has at best a chance only at one's second-best outcome and risks one's worst. Thus it seems pointless to promise, intend, or give any other form of assurance that one will cooperate. For it is known that one will find it rationally necessary to defect. This also seems to make assuring irrational, at least if assuring is like intending, so that it is rational sincerely to issue such an assurance only if it is rational to fulfill it; for fulfilling it is never rational.

Now the Deterrence Paradox (DP): you head a superpower facing an enemy who threatens nuclear attack. Your only hope is to try to reduce the odds of him attacking by sincerely intending to retaliate if attacked. (He can see your character so you cannot bluff.) You rank outcomes by the number of harms to all parties, most preferring the outcome yielded by him not attacking (zero harms), second most, him attacking and you not retaliating (for then you have not added pointless retaliation harms), third, him attacking and you retaliating (everyone is harmed). This is a one-time only interaction; so there is no deterrent benefit in retaliating, only in intending to retaliate. Prior to Gauthier it was thought irrational to retaliate, for it yields one's worst result, while not retaliating preserves one's second-best result; and so it was thought irrational to intend to retaliate. Thus one rationally cannot avail one's self of the advantages of threats in DPs.

Some philosophers thought these results falsified the received theory. For if only agents in the above PD could find it rational to make and keep resolutions to cooperate with those likewise inclined they could have the advantages of cooperation, securing their second-best outcomes instead of being doomed to their third. Further, that PD models interactions in which we have moral duties not to exploit other agents by defecting from agreements; and if we wish to prove the rational obligatoriness of moral restraint we must find a way to see cooperation as rational. As for the DP, it can seem bizarre that an agent who most wants to minimize harms and who knows her best chance of this is to form a sincere intention to retaliate, rationally cannot so intend because the action condi-

tionally intended is nonmaximizing, however unlikely it is that she will ever have to do it. Evidently the received theory fails to be such that agents deliberating by it will probably have their lives go best because it requires agents to intend and do only maximizing actions, even as their lives would likely go better if they deliberated in a more indirect fashion.

So Gauthier wrote a theory which exploited the fact that forming an intention (or disposition, etc., or making a commitment) to do an action can itself be maximizing because it may make other agents act to one's advantage. On Gauthier's theory an intention is rational just if adopting it maximizes, as where this would make the other agent cooperate or lower the odds of him attacking. Meanwhile an action is rational just if it maximized to intend to do it.[7] And this yields the desired results: it is rational (since maximizing) to intend to cooperate (with just those who, upon seeing one's intention—and only then—would probably cooperate in turn).[8] And it is rational to intend to retaliate (against those who, upon seeing one's intention, would be less likely to attack). So it is rational (even if nonmaximizing) to cooperate and to retaliate (in the event—made unlikely by one's so intending—that one is attacked anyway). This theory seems to beat the received one; for agents who deliberate by it will probably do better, precisely because they are not bound to form and fulfill only intentions to do maximizing actions, instead deliberating more indirectly in light of the aim of rational action.

We may find it implausible to say that cooperating rationally expresses or is justified by a PD agent's values given that defecting would yield her greater advantage; and that retaliating expresses a DP agent's values given that it adds to the harms she hates. In both cases the theory calls it rational to do actions directly frustrating one's values—counterpreferential actions. And we might well ask why agents would so act. How is that consistent with their values? At best their actions express their values only indirectly—by expressing intentions whose adoption directly expressed their values. But Gauthier thought that a theory must see such actions as rational in order to have the advantages he sought over the received theory, namely, being able to represent intending to cooperate and retaliate as rational, permitting rational agents to form and so reap the benefits of such intentions. And if we accept that it is a desideratum that a theory have these advantages, then we could object to his theory for its having this feature only if we could find a theory with all the advantages of his theory but lacking the feature.

On to Gauthier's second revisionist theory.

7. See David Gauthier, "Deterrence, Maximization and Rationality," *Ethics* 94 (1984): 474–95, and *Morals by Agreement* (Oxford: Oxford University Press, 1986).

8. This is not an argument for cooperating in single-stage, one-shot PDs, where one lacks the option of influencing other agents by means of adopting dispositions or intentions to cooperate.

IV. GAUTHIER'S RECANTATION AND SECOND THEORY

Gauthier's old theory says that it is rational to form an intention if this maximizes. But then contra criterion (v), above, it can be rational to conditionally intend to act in a way expected to leave one worse off than had one not so intended provided adopting the intention maximized. (Gauthier calls this a "threat" intention.) And contra (iv), it can be rational to act on that intention. In the DP (Deterrence Paradox), if one's retaliatory intention does not prevent an attack, one is committed to retaliating; and retaliating will cause the outcome one most hates where everyone is harmed. Gauthier came to find this intention irrational (for violating [iv] and [v]). So he proposed a new theory:[9] forming an intention is rational just if, in the circumstance one expects forming it to create, one expects that acting on the intention would leave one better off than had one not formed it and had instead done the most advantaging action that would have been available in what would then have been the circumstance. (These are "assurance" intentions.) And the action of fulfilling the intention is rational just if one still expects doing it to leave one better off than one would be in doing any other action one could have done had one not so intended.[10]

Note the contrast with Gauthier's old theory on which one forms an intention just if it maximizes to form it and fulfills it just if it maximized to form it. In the new theory one forms an intention just if, compared to never having formed it, one expects to gain from having formed it even if one fulfills it. And one fulfills it just if one expects that, even fulfilling it, compared to having never formed it, one will still gain by having formed it. If the gains were to result from one's intention or commitment inducing another agent to give one a gain, then the old theory says it is rational to act on one's commitment if by committing one bettered the odds of him giving the gain; the new theory, only if one still expects him to give it.

Both theories can justify intending and doing a nonmaximizing (i.e., counterpreferential) action—the old, if the intention to do the action was maximizing to form; the new, if refraining from maximizing still has one do better than had one never intended to refrain. For example, in a version of Gauthier's multi-stage PD, although defecting maximizes, it is rational on the new theory to intend to cooperate second with those who will cooperate first just if they expect one to reciprocate. For if things go as expected one will be cooperating with someone who, due to her having seen one's conditional intention to cooperate, will have just cooperated; and this yields one's second-best result, while had one not formed the intention one would be defecting against a defector for only one's third best. But on the new theory intending to retaliate is irrational; for fulfill-

9. See Gauthier, "Assure and Threaten."
10. Ibid., 705.

ing the intention would have one do worse than had one not formed the intention—retaliating only adds to harms.[11]

This seems the best possible result; for while Gauthier's first theory is attractive for seeing moral restraint as rational in the PD, some thought Gauthier wrong about its being rational to intend to retaliate and to retaliate: it is rational to give and fulfill assurances of cooperation in the PD but irrational to make and fulfill threats in the DP. Indeed, some saw his DP result as a reductio against his PD result (because of the moral monstrousness of retaliation, and because of retaliation's huge utility cost to the retaliator given her desires, something the witting incurring of which seems blatantly irrational). His new theory happily separates the scenarios, denying the rationality of retaliating but affirming that of cooperating.

There are, however, forms of the PD with the same structure as the DP: imagine another agent will make it 80 percent likely he will cooperate with you first provided you make it 100 percent likely you will cooperate with him second. You may get unlucky and have to cooperate knowing he defected, ending worse off than had you not formed the intention to cooperate. This situation superficially has you offering an assurance of cooperation; but structurally this is really a threat in Gauthier's sense. And Gauthier admits his theory forbids intentions of this sort.[12]

So while Gauthier's new theory avoids the odd result of his old that it can be rational to fulfill and so to make commitments one knows fulfilling which would leave one worse off than had one never committed, it has its own odd result: in DPs, those who want the chance of massive harm reduced cannot rationally deter harm, precisely and strangely because of their aversion to harm. For intending to retaliate is irrational. So deterring, if it can be done only by sincerely intending to retaliate, is rationally impossible. Similarly, in the above DP-like PD, those wanting to reduce the chance that they will be defected upon cannot rationally do what is needed to reduce it, again, because of their very aversion to being defected upon; for intending to cooperate here is irrational; so reducing the chance of being defected upon, if it can be done only by forming a sincere intention to cooperate, is impossible. So there are many fewer situations in which one can advantage one's self with commitments than on Gauthier's old theory.

But Gauthier stands by this. For since acting on threats would certainly be ultimately self-damaging and so, he thinks, irrational, and since one cannot rationally intend to do an irrational action, it cannot be rational to form threat intentions.

However frustrating this may be for those wishing to deter in DPs or to induce cooperation in DP-like PDs, the constraints on rationally possi-

11. Ibid., 707, 709–13.
12. Ibid., 713–17.

ble commitments seem to make such things impossible. So for all we have seen this is not the fault of Gauthier's new theory, but a fact about rationality which his theory has the virtue of recognizing and explaining. Nevertheless, there are, I think, fatal problems with Gauthier's theory, the first of which is the reversion problem.

V. THE REVERSION PROBLEM

This problem also afflicts Gauthier's old theory. On that theory a disposition to follow a certain style of deliberation is rational just if it maximizes to adopt the disposition, and it supposedly maximizes to adopt dispositions to refrain from maximizing in PDs and DPs because, supposedly, when certain other agents see that one has these dispositions, the agents will probably be induced to cooperate and not attack, respectively. But then in situations where you are to choose your actions after other agents have chosen theirs on the basis of the disposition the agents saw in you, and so after your disposition has had whatever effect it can have on others, it will be rational for you to revise your style back to that of someone who maximizes directly in choosing actions (and so who would defect in PDs and not retaliate in DPs); for that will now be the style it maximizes for you to adopt. So the theory would not really rationalize compliance with assurances or threats. For if it is rational to act only on rational dispositions, then since the dispositions most recently rational to adopt prior to choosing actions will be dispositions to fail to comply, complying is irrational.[13]

The same for Gauthier's new theory: it asks us to form and fulfill just those intentions we expect the fulfilling of which will correlate with us doing better than had we never formed them, and this supposedly justifies our forming assurance intentions to cooperate with agents who we expect will first be induced to cooperate with us by our having such intentions. But after one's intention has induced cooperation from other agents (if it can), the theory will justify one in forming an intention to defect, for that will then be the intention forming and fulfilling which one can expect will correlate with one's doing better. Once again, Gauthier's apparatus divides through and recommends defection. Since other agents can foresee this they will not cooperate.

13. I moot this criticism in my "Preference's Progress: Rational Self-Alteration and the Rationality of Morality," *Dialogue: Canadian Philosophical Review* 30 (1991): 3–32, and "Retaliation Rationalized: Gauthier's Solution to the Deterrence Dilemma," *Pacific Philosophical Quarterly* 72 (1991): 9–32. Gauthier once suggested in correspondence that an agent whose commitments would revert would not in fact have the style of deliberation it would most advantage her to have, since her style would make her commitments ones known to be empty. I agree. My worry is that a style described the way Gauthier describes it—as deliberating to a different way of choosing given one's aims—will not be one most to her advantage, because it will face the reversion problem. I try below to specify a style truly immune to that problem.

Gauthier saw that if the basis upon which actions are to be evaluated is whether they advance one's desires, the basis will require one not to fulfill assurances and threats. So he sought to have it that the basis changes upon adopting dispositions or intentions whose adoption would advance one's desires. One evaluates prospective actions differently afterward, finding an action rational if dictated by an intention it advanced one's desires to adopt rather than by whether the action is directly desire-advancing. The problem is that one's desires persist, continuing to be the basis for the evaluation of intentions as the circumstances in which one holds them change. And this means the intentions rational to form and fulfill will vary as one moves through the stages of assurance and threat scenarios—the intentions advantageous to have before other agents choose actions are different from the ones advantageous to have after. This undermines the rationality and so the efficacy of the intentions one originally adopted, with the net result that the same intentions and actions are recommended by Gauthier's new theories as by straightforward maximization on one's desires. For there has not been a sufficiently fundamental change in the evaluative basis. It is still one's life going well by the measure of one's initial desires. I will return to this point. But now I consider problems with Gauthier's views on the rationality of threats.

VI. THE RATIONALITY OF THREATS

Neither of Gauthier's theories succeeds in representing forming and fulfilling assurance or threat intentions as rational. But ought the correct theory to so represent them? Gauthier's first theory purported to have both as rational, his second, only assurances. And in offering each theory he offered rationales for these verdicts. I now assess the rationales.

Gauthier says a style of deliberation is rational just if using it makes one's life go best. His new theory's style saves an agent from knowingly doing something certain to be ultimately self-damaging in fulfilling a threat. But it also asks her to do things certainly courting self-damage: she must refrain from making threat commitments to the possible doing of self-damaging actions the making of which would almost certainly massively advantage her or save her from massive disadvantage. For she must refrain from commitments that have any chance of resulting in her having to do things she would know to be ultimately self-damaging. (Recall that a threat is a conditional commitment to do something one would know at the time of action will have one do worse than had one never committed. True, the odds are that one will never have to act on the threat. But if one ever does have to act on it one will know this is making one's life worse.) So Gauthier's new theory saves the agent from very unlikely possible disasters, but only by exposing her to guaranteed disasters: if she does not threaten she will be attacked—a disastrous result. And surely her life can be expected to go

better (it is a better bet) if she can deliberate in a way that allows her to be very likely to fend off such otherwise certain catastrophes—in this case by forming a threat intention to retaliate. But this recommends Gauthier's old theory over his new.

Gauthier's reply is that, while a person's being rational should not result in her knowingly doing actions that would defeat the rational aim (that her life go best), she is not necessarily able to adopt an intention it would advantage her to have; for rationally she cannot adopt intentions to do irrational actions.[14] An agent may not rationally do an action she knows to be ultimately self-damaging, but she may rationally fail to adopt a self-advantaging intention to do, under certain conditions, an ultimately self-damaging action. For an intention to do an irrational action is not one that rationally can befall a rational agent.

But all of this depends on what counts as a rational action. Those who deliberate by Gauthier's old theory think an action rational if it maximized to intend to do it; those who deliberate by his new theory think an action rational only if they expect doing it still to correlate with a gain compared to never having intended to do it. So absent an independent test of the correctness of a conception of rational action, the idea that one cannot rationally intend to do an irrational action is not decisive on whether it is rational to form a threat intention.

Is there reason to favor Gauthier's new conception? Threat-fulfilling actions are not irrational on his old theory; for it says actions are rational if it maximized to intend to do them. So what was wrong with that style of deliberation? Well, one cannot justify fulfilling the commitments by saying one expects to do better even fulfilling them than had one never made them. Still, on his old theory there is a justification one can give: one is doing a self-damaging action because one had to commit to doing it in order to reduce the odds of catastrophe, and one had to become the kind of person for whom a commitment is binding else one could not have advantaged one's self with commitments.

Yet that rationale for doing something ultimately self-damaging now sounds lame: for one now knows that being a person who keeps any commitment it maximized to make will prove catastrophic because keeping this one will be catastrophic; so maybe it is time to become a different kind of person. When one made the commitment one thought it unlikely one would ever have to act on it and so hoped to benefit from being the kind of person who would keep such commitments (i.e., from having that disposition), a benefit one attains only if one does not have to act on the commitment (i.e., never has to express the disposition). That it was unlikely one would have to act on it was part of what made forming it maximizing. But one has since learned one was never going to gain with

14. Gauthier, "Assure and Threaten," 707.

this kind of commitment, nor by being this kind of person. And since one can foresee that the only condition in which one would have to act on a threat would be a condition under which it might be thought rational to cease being the kind of person who would act on it (and so who would not find acting on it rational), arguably one cannot rationally make threat commitments; for one cannot rationally commit to irrational actions.

But none of this matters to someone deliberating by the old theory. She sees her intention as one that she should fulfill just if she still thinks she was right about the odds of its preventing attack. (She refrains only if she learns that her enemy's likelihood of attacking could never have been reduced by her threatening.) Gauthier's new deliberator, by contrast, thinks an intention should be fulfilled just if she expects this to leave her better off than had she never formed it, and since this is false of fulfilling a threat, forming a threat intention is irrational.

Is the rationality test of Gauthier's new deliberator more intuitive? Asking this is in effect asking whether Gauthier's main test for the correctness of a theory—that following it would make one's life go best—is properly implemented by or most consistent with his claims that it is never rational to act on an intention if one expects this to leave one's life worse than had one never so intended, and so never rational to form an intention to so act. The latter two claims are the core of his second theory. But each of his theories can be defended by appeal to his main test; for each describes a style of deliberation that can make one's life go better, his first theory, by letting one form threat commitments likely to make one's life go better; his second theory, by saving one from the unlikely but disastrous possibility that one's threat fails and then requires one to do an action making one's life much worse.

So, is the rationality test of Gauthier's new deliberator more intuitive? This is a tough call. Threatening is very attractive because it has a high chance of causing one's best result, just as assuring is attractive because it is certain to cause one's second-best result over one's third. But if one's threat fails one must cause one's worst outcome, and it may seem rationally unimaginable to do this and so rationally impossible to intend to do it. And yet Gauthier's new deliberator can assure even though she knows that when she must comply there will be an action open to her that would have her do even better. How can she imagine doing an action which will have her do worse than another action she could then do? While if she can imagine this, why not fulfilling a threat? True, in fulfilling an assurance she will have the consolation that at least she is doing better than she would be had she never committed, while in fulfilling a threat she would know she was doing worse. But the advantage forgone in not defecting differs only in degree from that forgone in retaliating. The mystery is how a rational agent can do an action she knows involves forgoing an advantage, and so how she can commit to it either. In both

cases she must imagine putting herself in a situation where it would advantage her to violate her commitments.

So suppose she does not threaten: what consolation does she have? She is forgoing a superb chance at her best outcome, and she knows she is doing that right now. In this respect her situation would be like that of someone who did not issue an assurance. Maybe her consolation is that at least she is ensuring her second-best outcome, avoiding her worst. Looked at one way, then, the positions of assurers and nonthreateners are the same: both guarantee their second-best result over their third. And both must forgo an advantage: assurers, later, when complying; nonthreateners, up front, when refraining from threatening. Finally, both have a consolation: assurers, that in complying they will be doing better than had they never committed; nonthreateners, that they will never have to cause their worst outcomes.

Perhaps the foregoing reveals Gauthier's new rationality test to be this: an intention is irrational if it commits one to unconsoled compliance. So assuring is rational because complying is consoled; threatening is irrational because only refraining from threatening is consoled.

And yet how to measure and balance these regrets and consolations? In deciding whether to threaten, how is one to balance the certitude that one will not have one's worst outcome if one does not threaten—the consolation for forgoing the advantages of threatening—against the near certitude of getting one's best outcome if one does threaten—the possible advantage forgone in not threatening? The standard way is by expected utility: the expected utility of threatening is very high, that of not threatening, very low. Why is that not decisive so that one should threaten? True, the expected utility of fulfilling the threat is exceedingly low. But the expected utility of fulfilling an assurance is also low compared with that of defecting from it. If the latter is no objection to assuring, why is the former an objection to threatening?

One difference is this: assuring guarantees one's second-best outcome, and in an assurance situation there is no way to get one's best. Meanwhile, threatening does not guarantee one's best outcome, but not threatening guarantees one's second-best. But if this is what is in play, then in selecting intentions (although not actions) it seems Gauthier's new deliberators are to use the principle that one should minimize the maximum possible disaster; his old, that one should maximize expected utility. And in deciding whether to fulfill intentions both ask whether the circumstance they now face is the one they expected when forming their commitments: his new deliberator asks, did the other agent cooperate as expected? His old, did my commitment increase the odds of the other agent doing what I wanted him to do (cooperating, not attacking)? Both deliberators comply just if the answer to these questions is yes.

One can see the attractions in both views—Gauthier's old view gives one extra opportunities; his new view saves one from possible catastrophes. But it is not obvious that one view is superior. We have conflicting intuitions of rationality (or at least conflicting desiderata). I now propose a test decisive as between the two conceptions of rationality in actions, namely, the test of whether the conceptions are coherent. Gauthier's new theory is not.

VII. CONFLICTS IN GAUTHIER'S SECOND THEORY

The problem begins like this: even if one rationally cannot directly form an intention to do an irrational action, there are actions one can do to get one's self to form such intentions, even intentions possibly to do ultimately self-damaging actions: for example, one could act to arrange to be brainwashed into having desires that would make forming the intentions rational even by Gauthier's standard (the standard that one can rationally intend to do only actions that will leave one better off by the measure of whatever desires one has than had one never formed the intentions). In the case of threat intentions in the DP, these might be desires to retaliate against those whom one had to threaten in order to reduce the odds of them attacking. And the actions of arranging to form threat intentions would be advantageous, actions to avoid arranging this, disadvantageous. And if it is irrational to do a self-damaging action, or to refrain from one the refraining from which would be self-damaging, then it would be irrational by Gauthier's new theory for one not to act to arrange to have threat intentions. So rational agents should form and fulfill intentions to do actions to arrange to have threat intentions.

But this leaves Gauthier's view conflicted: it recommends the actions because not doing them would make one's life worse, but it also forbids them because they would induce intentions that the theory finds irrational, intentions to do ultimately self-damaging actions. Thus there being different rationality rules for intentions and actions makes Gauthier's theory incoherent. For his explicit rules for action rationality then conflict with implicit rules for action rationality deriving from his rules for intention rationality. So the theory is inconsistent in its advice: it says to do and not to do the actions that would give one threat intentions. Indeed, the theory is unfollowable in DPs because it is impossible jointly to fulfill both parts of conflicting advice.[15]

15. For developments of this sort of criticism of theories with structures like Gauthier's, see my "Prudence and the Reasons of Rational Persons," *Australasian Journal of Philosophy* 79 (2001): 346–65, and "Prudence and the Temporal Structure of Practical Reasons," in *Weakness of Will and Practical Irrationality*, ed. Christine Tappolet and Sarah Stroud (Oxford: Clarendon, 2003), 230–50.

This is a problem for any theory rationally requiring one to do actions that stand in a certain relation to one's desires—for example, that of advancing them—but forbidding intending to do an action that would not stand in that relation. For PDs and DPs are so structured that one's having a forbidden intention would make other agents certain or likely to better advance one's desires than one could do on one's own. So one's doing an action to arrange to have the intention would be required—for standing in the right relation to one's desires—but also forbidden—for resulting in one's having an intention standing in the wrong relation. This is obvious of the received theory, which requires one to do desire-advancing actions, while forbidding intentions to do non-desire-advancing actions. It is even a problem for Gauthier's first theory. For the reversion problem means the theory really finds non-desire-advancing actions irrational, and so, since it is irrational to intend to do irrational actions, it really finds intentions to do such actions irrational too.

Any theory that rationally requires agents to do actions advancing their desires, but rationally forbids forming desire-advancing intentions to do non-desire-advancing actions, will give conflicting advice about whether to do actions that would result in such intentions. No fully followable theory can combine these two edicts. Which edict to drop? Our test for the correctness of a theory is that following it should make one's life go better than following any other. One's life would likely go better if one assures, and one's assuring cannot result in one knowingly acting to make one's life worse. Meanwhile one's life would certainly go worse if one does not threaten and will probably go better if one does threaten. True, there is a chance that it will go much worse. But it is a better bet that it will go well if one threatens. Thus a theory that recommends assuring and threatening beats a theory that does not—beats it because adopting the theory's style of deliberation has a higher expected utility than adopting one that forbids such intentions. So we must drop the edict against having desire-advancing intentions to do non-desire-advancing actions. Still, one cannot rationally intend to do an irrational action. Thus a theory can have it that it is rational to form assurance and threat intentions only if the theory will also see it as rational to do the actions that would fulfill them. But how can this count as rational given that it means doing actions against one's desires? In the next section I argue that this is possible, but only if, when it would advance one's desires to form such intentions, it is first rational to revise the desires into ones that would be advanced by fulfilling the intentions. The resulting theory will be coherent and so always followable. And it will beat the earlier theories in giving desire-advancing advice in paradoxical situations where the other theories are conflicted and so silent, while giving the same advice as the other theories to form and fulfill intentions to do maximizing actions in normal situations, advice that cannot be bettered.

VIII. A RADICAL PROPOSAL

Gauthier thinks it a goal of rational choice theory to have an account on which it is irrational for one ever to knowingly ultimately disadvantage one's self with one's intentions or actions. But he sees no way for these always to be advantaging looked at from all points in one's life.[16] Faced with a choice between a theory that asks an agent to damage herself by being unable to make threats, and one which lets her make them, but at the expense of her possibly having to act on them, whence she would ultimately damage herself in so acting, he chose the former. For it is irrational to intend to do an irrational action, and one's action is irrational if one knows it to be ultimately self-damaging.

I claimed that his theory fails by its own measure; for in asking us not to form threats, it asks us to refrain from the actions needed to arrange forming them, and so asks us to do actions (refrainings) which we know are self-damaging; or—just as bad—the theory is conflicted: it asks us both not to do and to do these actions. But I now offer a theory which never asks an agent to be knowingly ultimately self-damaging in choosing whether to form or fulfill intentions, and so which lets her intentions and actions always be advantaging seen from all points in her life; a theory also meeting Gauthier's desideratum that rationality should never ask us to commit to doing actions that would leave us worse off than had we never committed, and avoiding the problems we found with other theories.

I say it is rational to maximize on the desires held when acting and to intend to do only maximizing actions. But in multi-stage PDs (both assurance and threat versions) and the DP (which calls for a threat), it maximizes on one's desires to replace them with ones on which it would maximize to cooperate and to retaliate, respectively. For then one would find it rational and so possible to fulfill and so to form the assurance and threat intentions the forming of which would be the best bet of making one's life go best. One would find these things possible because one would then experience the intentions as intentions to do maximizing actions, actions maximizing on the new desires. Accordingly, replacing one's desires is rationally required. So instead of most desiring the outcome one originally most desired, *O*, one would immediately come to be such that one most desires *O* except where one had to commit to acting to prevent *O*, should certain unlikely conditions obtain, in order to increase the odds of someone else helping to bring about *O*, for which unlikely conditions one prefers more to do the actions to which one has committed than one prefers *O*.[17] This means that we need not be self-damaging (in the

16. Gauthier, "Assure and Threaten," 694, 720–21.

17. For example, in the DP, I should go from desiring the minimization of all harms, to desiring the minimization of all harms except those that would have to be inflicted in

sense of going against the desires we have) in choosing intentions; for we can find it rational to form the intentions there is advantage (given our desires) to forming. And then when it is time to fulfill the intentions it will be rational to do so since this would maximize on one's new desires, the ones adopted just before forming the intentions.[18] Thus we need not be self-damaging in acting on intentions (again, in the sense of going against the desires we will have at the time of acting), this solving the difficulty Gauthier sought to avoid in his latest theory. And we need never call rational, actions retarding desires one has when acting; for by the time one is to cooperate or retaliate, one's desires will have rationally changed to favor those actions—one's desires changed just before committing to do the actions. And by the measure of the new desires, neither action is self damaging. Nor is there any sense in which one's actions do not express the desires one has at the time of acting (an unintuitive feature of Gauthier's two theories). For fulfilling a commitment is now directly desire-expressing—it directly expresses one's revised desires.

Thus there is no sense in which, at any given time in one's life, one rationally must make a choice, whether of intention or of action, that cannot be seen as self-advantaging by the measure of the desires one has at the time of that choice. One rationally gets to form the intentions it advances the desires one has at the time of intention-formation to form, and one rationally gets to do the actions it advances the desires one has at the time of acting to do. Of course, looking down the road, in forming, say, a retaliatory intention, one realizes that one may wind up acting on the intention, going against the desires with which one began. But that is a risk one is willing to take in changing one's desires to make such intentions possible, and so taking the risk advances the original desires. And after the change, one does not care that the action the new desires justify would go against the ends of the original desires, for one now has new desires, and they are advanced by the action.[19]

---

fulfilling a presumptively deterrent threat. But how can I both favor harm-minimizing generally and favor harm-infliction in retaliation? The answer is that I no longer favor harm-minimizing generally. Instead I favor only harm-minimizing when the harms are not in fulfillment of a threat—the latter harms have the distinguishing property that I had to threaten to inflict them, and so I can take a different attitude to them than to other harms. Thus I need not have become a total monster. For more on this, and for a discussion of the psychological plausibility of the proposal, see my "Preference-Revision and the Paradoxes of Instrumental Rationality," *Canadian Journal of Philosophy* 22 (1992): 501–27, and "Persons and the Satisfaction of Preferences: Problems in the Rational Kinematics of Values," *Journal of Philosophy* 90 (1993): 163–80.

18. I criticize Gauthier and argue that intending to retaliate and retaliating are rational only after value revision in my "Retaliation Rationalized," and "Preference's Progress."

19. Matters might be different if one would look forward and rightly think one would later be having the wrong sorts of reasons for action, or would look back and rightly think one shouldn't have come to one's current views about what count as good reasons. But as I

Further, there is no good reason to revert to one's precommitment stance after one's commitment has had whatever effect it can have on one's situation. For given one's new values, values which prioritize keeping certain kinds of commitments, it maximizes to cooperate and to retaliate, while reverting would prevent one from doing these actions, and so reverting would be nonmaximizing, irrational. One would go back on one's commitment only if one discovered that assumptions made in forming it were false (e.g., the assumption that making the commitment would improve the odds of other agents acting to one's advantage),[20] or if one in the meanwhile faced a situation in which the new values would be best advanced by one's undergoing yet another change in values. In the former case, one goes back because one's new values only demand that one fulfill commitments with agents of the right sort (namely, agents in some degree susceptible to being influenced to one's advantage by one's intentions), and so fulfilling the commitments would not be maximizing; in the second case, because it would have since become maximizing on one's new values to adopt values that demand different actions. Thus while one's values, intentions, and commitments are always up for reconsideration, the mere fact that these things have already influenced other agents will not be a reason to reconsider, and so the values, and so forth will be stable into the endgames of PDs and DPs. The reversion problem is solved.

Moreover, the theory never asks us to do actions (e.g., arranging to form an intention to retaliate) that will result in our having intentions (e.g., the intention to retaliate) to do other actions (e.g., retaliate) that the

---

in effect argue in Sec. X, below, the conceptions of good reasons necessary to sustain these worries are contradictory and so false, e.g., conceptions according to which one's initial desires, or desires thought to be objectively correct, should be used throughout one's life to evaluate one's choices. Relatedly, one might think—as did an editor for this journal—that all of this is a refutation of the desire-fulfillment conception of a life's going well, and that there is a proper conception by whose measure the life of someone who must fulfill a retaliatory threat is going horribly, a conception that should guide the choices of intentions and actions of a rational agent. But this implies that there could be a standard of a life's going well that could coherently serve always as an agent's rational aim even in situations like the Deterrence Paradox; and this too is something in effect argued against in Sec. X, below.

20. If I made the enormous change in my desires necessary to make this commitment rationally possible for me, how can I go back on the commitment (and, presumably, renounce the acquired desires) merely upon the apprehension of this fact? Hume is instructive here: when we form a desire on the assumption that a certain fact holds, the desire disappears (and rationally should disappear) upon the discovery that the fact does not hold. Alternatively, we might say that I don't renounce my new desire—I still desire that, if I've been attacked by someone whose odds of attacking were reduced by me sincerely threatening to retaliate, I retaliate. It's just that the person I'm actually interacting with doesn't have the property that my desire targets, and so my desire gives me no reason to retaliate against him—I have learned that he could never have had the odds of his attacking reduced by a threat. For more on this, see my "Preference-Revision."

theory asks us not to do. Recall that Gauthier's second theory does ask this, making it incoherent; for it demands that we arrange for the problematic intention on pain of our life going worse for allowing a nuclear attack, but also that we not arrange for this intention, since the theory finds the intention irrational for being an intention to do an action it would be irrational to do, irrational because it would worsen our lives by causing retaliation harms. Thus the theory must call the intention-arranging action both rational and irrational (a problem had by all the foregoing theories). My theory avoids this because it sees that, whenever we would find it advantageous, by the measure of our current desires, to form a commitment to do an action against them, we would also find it advantageous and so rational to change the desires into ones by whose measure the action to which we commit is not one against our desires. So it cannot happen that a commitment is advised by our desires but the action to which we commit is not. Thus the theory can never conflict with itself by advising doing an action because doing it will yield a commitment advantageous to form, while also advising against doing the commitment-yielding action because it advises against the action to which one is to commit. If the theory advises one to do a commitment-forming action it will also advise that one do the action to which one is to commit.

Finally, the theory is consistent with its being irrational to form intentions to do ultimately self-disadvantaging actions. To see why, imagine a situation where you are not choosing future actions by choosing intentions but instead are choosing the locking in of future behaviors. (You could arm a doomsday device in the DP, a device certain to retaliate on your behalf if you are attacked; perhaps the device is an irremovable exoskeleton that will force your finger onto the retaliate button.) Is it rational to lock in any of the behaviors that Gauthier's new theory (and every other theory we considered) finds it irrational to intend to do? Surely it is provided not locking in would be self-damaging and provided the device will not force your cooperating/retaliating upon the other agent defecting/attacking if it is discovered that locking-in could never have increased the odds of his cooperating, or lowered the odds of his attacking, in the degrees needed for locking in to have been maximizing.

Gauthier's criteria for the adequacy of a theory accept this. For by the time one is to behave, say, as a retaliator, one's behavior will not be an action, because it will not be voluntary—it will be compelled by the locking-in mechanism. So any scruples against the rationality of fulfilling threats will not apply to the rationality of locking in, even though such scruples would object to intentions to fulfill them since, qua action, fulfilling them is not rational and so cannot be rationally intended. But it is not a condition on rationally locking in a behavior that, were the behavior to be done as an action rather than as a mere behavior, it would be a rational action; only the action of locking in need be rational.

But now suppose you are offered preference-revision as the only available means of locking in (you are offered a hypnotist, a brain-washer, a pill, whatever, whose effect would be to change your desires). Now, the behaviors the new desires generate will be actions because they will be desire-governed, governed by your new desires; but the actions will be rational because they will be self-advantaging given the new desires and the measure of advantage is always the desires one holds at the time of acting. So there can be no objection to doing this on any of the purely instrumental theories of rationality we considered. And surely one will rationally take this option even where one cannot deter using Gauthier rationality and no preference-revision. For this allows one to deter, and to bypass Gauthier rationality, and still be all-in rational. Adopting the new desires is advantaging by the measure of the old; acting on the new desires is advantaging by the measure of the new. So one is always doing only advantaging actions, a desideratum in Gauthier's conception of rational deliberation.

But desire-change technology may not be available; so for agents lacking it, will we not have back the problems that plagued Gauthier? That depends on whether a rational agent would need technology in order to undergo the change. Suppose there are people who would not. Their desires automatically change whenever what they are desires for would become more likely to obtain if the desires were replaced with certain new desires. Are these people in any way irrational? They always do whatever they know will advance the desires they have; they never intend to do actions it will not advantage them to do, never have self-damaging intentions, never do self-damaging actions. Arguably there is nothing irrational about them. Indeed, they may look with pity upon those of us whose desires do not respond to pragmatic reasoning in this way, for we cannot advantage ourselves with threats and assurances—our lives are doomed to go worse than their lives will probably go; our choices (certainly of intentions, and probably of actions) will have lower expected utilities. And if such agents took steps to become agents whose desires would not change like this, arguably they would be doing something irrational (since this would not be maximizing), and they would be making themselves irrational (since they would be making themselves incapable of undergoing maximizing alterations to their desires).

This suggests that a truly, fully rational agent's desires would change automatically. Consider two analogous forms of rational change of attitude: first, one automatically comes to desire to do an action upon seeing that doing it would cause an end one desires; second, one's beliefs undergo revision upon new evidence. If one is rational, no technology is needed to make one desire to take means to one's ends, nor to make one's beliefs respond to evidence, and nor, I claim, to make one's old desires be supplanted with ones advantageous to have as a means of advancing the

ends of the old. For if it is demonstrated to be rationally obligatory to undergo a change of attitude, whether conative or cognitive, one simply will undergo it as a consequence of being rational: no further explanation is needed of how this is done. To argue that an attitude change is rationally required is to argue that nothing external to rationality is needed to effect the change. So if it really is rationally obligatory to undergo the change, then desire-change is, as Edward McClennen would call it, an endogenously available solution to the problem (i.e., nothing outside to the situation, like technology, is needed); it is the rationally required response. People think technology is needed only because they doubt that the usefulness of forming new desires as a means of advancing the old proves the rationality of having new desires; at best it makes retaining the old unfortunate, not irrational; so we need something outside of rationality to make ourselves change. But if I am right that the change is rationally required,[21] then it should occur as a nonactional response to the facts, one nonetheless under the control of rationality, and so one whose occurrence should be ensured by the mere fact of one's rationality.[22]

IX. THE ORDER OF RATIONAL DELIBERATION

The style of deliberation in the received theory evaluates actions in the first instance—actions are rational if they advance the aim of rationality (e.g., if the aim is to satisfy one's desires, then actions are rational if they are expected to cause satisfaction of one's desires). Meanwhile, the style evaluates intentions derivatively—intentions are rational if they are intentions to do rational actions, ones that advance the aim. Since fulfilling assurances and threats never advances the aim, intending to fulfill them can never be rational, so agents cannot exploit the effects on others of their intentions. Gauthier's first theory evaluates intentions first—intentions are rational if having them advances the aim—and actions derivatively—actions are rational if they are rationally intended. Gauthier's second theory evaluates intention-action pairs first—intentions are rational

21. Are desires really under the control of reason? Is desire-change psychologically possible? Is there any point to bringing something about by desire-revision if one won't desire it by then? I've tried to deal with these and other issues surrounding the idea elsewhere. See my "Preference-Revision," "Persons," "Retaliation Rationalized," "Preference's Progress," "Prudence and the Reasons," "Prudence and the Temporal Structure," and "Moral Paradox and the Mutability of the Good" (unpublished manuscript, Dalhousie University, 2012). In the main text, below, I begin mooting a major objection to the idea—that it would be bad faith. And I hope one day to treat of the arguments against the idea in Elijah Millgram, *Practical Induction* (Cambridge, MA: Harvard University Press, 1997).

22. A given agent might be irrational in this respect, so that her desires would not automatically respond to this kind of reasoning. And unless she has technology to help her, she will be unable to issue effective assurances and threats, and probably her life will go less well. But this will be due to her irrationality, not to a conflict in the theory of rationality.

just if, even if one acts on them, one expects the aim to be better advanced than had one never formed them—and actions derivatively by whether the intentions to do them are still such that the aim is better advanced by one's having formed and acted on them than had one never formed them. These theories were supposed to give agents the advantages of assurances, and, on his first theory, threats, since in evaluating intentions first agents were supposed to be able to exploit the effects on other agents of holding intentions. But the theories fail in this because they have it that the aim evaluates intentions by whether it advances the aim to hold them, or to hold and act on them, respectively; and the advantageousness of intentions by this measure changes over the course of the relevant scenarios, this inducing the reversion problem and netting out to advising the same actions as the received theory.

On my theory deliberation evaluates one's very aim of rationality first—the aim one has is rational just if nothing else is such that taking *it* as the aim would better advance the current aim—and actions and intentions derivatively—actions are rational just if they advance a rationally currently held aim, intentions, just if they are intentions to do actions advancing that aim. Here one gets the advantages of both assurances and threats because one is enjoined to exploit the effects on others of one's having aims. When it would be to one's advantage as measured by one's current aim to form assurance and threat intentions, one is enjoined to take as one's aim whatever would be advanced by one's doing the actions involved in fulfilling the intentions (while otherwise demanding the same actions as the previously most recently and rationally held aim). This in turn makes it rationally possible to form intentions to do these actions as actions it would advance the new aim to do. And this allows one to form the assuring and threatening intentions whose effect on other agents is to make them more likely to advance one's original aim. Because one is always to deliberate using whatever aim one has at the time of deliberating, there is no reversion problem; for reverting would prevent one from advancing one's new aim and so would be irrational. And one is allowed the benefits of both assurances and threats because the rationality of one's forming their constituent intentions depends only upon the advantages of one's forming them as measured by one's previously most recently rationally held aim, not on their fulfillment leaving one better off by the measure of that aim than had one never formed them; thus the difference Gauthier noticed between what gains correlate with fulfilling assurances but not threats is irrelevant.

No style of deliberation could be expected to make one's life go better than this one; for this is the only style letting one maximize in every situation, that is, both *ex ante* when choosing commitments and *ex post* when choosing whether to act on them; and letting one fully exploit the effects of one's aims, intentions, and actions. There is nothing else to ex-

ploit. And, by definition, no strategy can be expected to beat always maximizing expected utility.

X. DIFFERENT THEORIES OF THE AIM OF RATIONALITY

If the aim of rational choice is to satisfy one's desires, then my proposal says rational deliberation requires one to undergo their revision in assurance and threat scenarios. And for a Humean about desires such as Gauthier—someone who thinks there is nothing rational or irrational about holding any given desires, or at least nothing based in the properties of the things desired, and who thinks one's only rational duty to one's desires is to take means to their satisfaction—there should be no objection to undergoing a change in desires when this would be the best means. Instrumental rationality—the only kind of practical rationality there is on the Humean view—dictates that one change one's desires; one is only obliged to retain them when this would best advance them (as in strategic contexts where retaining them justifies and motivates one in doing actions making their satisfaction more likely). In this theory the aim is identical to the ends of one's desires; the aim is to satisfy one's desires, whatever they are. And nothing makes a given end the aim except someone's desiring it, and that, only for the agent doing the desiring. When such an agent changes her desires for good reason—for example, on the pragmatic pretext I've offered—then the aim changes, at least for her.

But there are other theories of practical rationality, theories distinguished by their positing different kinds of thing as the rational aim, for example, that the rational aim is to bring about states of affairs ranked as highly as possible by their inherent, objective, immutable goodness; or to be in maximal compliance with universalizable moral principles, those one could will without contradiction that all agents always follow; or to advance only those of one's desires rationally validated by being desires for good states of affairs, or by being desires action on which would fit with a universalized moral code. Call these "objectivist" theories, because they have it that it is an objective and permanent fact whether something is the rational aim.

These theories, like the received theory and Gauthier's second theory, make the following claims. First, something, *X*, is such that the rational aim is always attaining *X*—this is always the measure of whether one's life goes well. Second, one must always act to attain *X*. Third, one may not rationally form even conditional intentions to do certain kinds of actions against *X*—for the received theory, actions not maximizing on *X*; for Gauthier's second theory, actions which would leave one worse off than had one never intended to do them. For all of these theories Gauthier's general issue can arise: what is it to be instrumentally rational in action and intention given the aim? And the puzzles that concerned Gauthier

arise no matter what the aim. This is because, for any such theory, one can write a situation where the only way to make attaining *X* more likely is to form a conditional intention to act against *X*. Since the theory rationally forbids such intentions, agents deliberating by it will typically do less well at attaining their *X* than agents deliberating by a theory positing a different *X*, *X*1, that lets them form the intentions (e.g., the Humean theory, above), because the former agents cannot avail themselves of threats, or assurances, or both.

This will not worry objectivists. For even if there is a theory positing another aim, and even if agents who deliberated by it would likely do better in the terms of their aim than objectivists would by deliberating in terms of the objectivist-approved aim, since objectivists would think that the competitor aim is not the true aim of rationality, who cares if someone aiming at things irrational to aim at could better attain their aims?

It is generally thought that these theories are consistent. And while agents deliberating by them, and so bound by their constraints on rationally possible intentions, cannot have the benefits of assurances and threats, this fact, however consternating, is not thought fatal to the theories.

But these theories are not consistent. For while on their account a rational agent cannot rationally directly form assurance and threat intentions, we can always imagine an action, *A*, whose effect would be to induce the (supposedly) non- or ir-rational forming of an advantageous conditional intention to do an action against *X*; and the theories give conflicting and so unfollowable advice about whether to *A*: the duty to act so as to attain *X* says to *A*, since *A*-ing makes attaining *X* more likely; but the duty never to form an intention to act against *X* says not to *A*, because *A*-ing would result in a conditional intention to act against *X*. The theories were thought consistent because evaluating intentions seemed to be one thing, actions, another, so that there could be no conflict between the respective evaluative principles. But in fact the theories' rules about permitted intentions imply restrictions on permitted actions, restrictions that conflict with enjoinments concerning actions issued by the theories' rules about actions.

Might we then conclude that any such theory is necessarily false on the grounds that it would ascribe to *A* the properties that *A* is both to be done and not to be done (rationally recommended and not, rationally correct and not), which is a contradiction, something no true theory can entail? Perhaps not. For if we read the theories as enjoinments (to do certain actions and to refrain from forming certain intentions) rather than descriptions (of certain actions and intentions as rational or irrational), then since it is only descriptions that can be contradictory, the theories do not entail contradictions. They do engender dilemmas—practical dilemmas that are sometimes moral dilemmas (as when they

are about whether to form and fulfill intentions forming which would have morally good effects, fulfilling which, bad). But many philosophers tolerate irresolvable moral dilemmas. Perhaps these philosophers would also countenance nonmoral practical dilemmas. I am less sanguine of the possible correctness of theories that give dilemmatic advice. For one thing, saying the theories aren't contradictory is no help; for they still don't unambiguously tell us what to do, and so they are still unfollowable. Besides, the objectivist theories are usually offered as truths, not mere enjoinments. So they do entail contradictory descriptions of certain actions and therefore simply cannot be the truth about the duties of practical rationality.[23] But it is one thing to reduce a theory to absurdity in this way, another to figure out what has gone wrong with the theory.

These theories take it that to be practically rational is to advance whatever is the aim of rationality, they posit such an aim, they conjecture that it cannot change, they enjoin acting to advance it, and they forbid intending to do actions against it, this implicitly forbidding actions that would advance it if they would result in intentions to act against it, this resulting in the contradiction. Now if it is irrational to intend an irrational action, and if an action is made rational by advancing the aim, then the only way to make a coherent theory is to have it that the aim changes when advancing it requires forming an intention to act against it.

But on these theories of the aim it is difficult to see how it could change; for the theories expressly or implicitly conceive it as unchangeable. If the aim is to bring about the objectively best states of affairs one can, it is presumed that the same things are always good or bad for everyone; while if the aim is to obey whatever laws of conduct universalize, then if a law universalizes, by hypothesis it applies to all people always.

True, one can always write scenarios in which the theories' aims would be better met if one came to treat something different as the aim—as where someone else will likely bring about a morally good state of affairs, one that one could not bring about on one's own, only if one forms an intention to bring about a bad one under some unlikely conditions; or where one will likely be permitted to obey the moral laws only if one conditionally commits to breaking them should certain unlikely conditions ob-

23. What if in a given case there happened to be available to the agent no action of the *A* sort, i.e., one able to induce in her non- or ir-rationally the intention to go against X? Then she could follow these theories in this case, for here no action is one they would leave her conflicted about whether to do—she is just to watch helplessly as someone else, e.g., an attacker in a DP, acts against *X*. But while this saves her from conflict in the case, it doesn't save the theories. They are sunk if there is even a logically possible such action; for then they will contradictorily describe it and so be false. It would be heroic—even if salvational for these theories—to argue that it is logically impossible for there to be such an action, although one can read Millgram as coming close to trying in *Practical Induction*.

tain. But how can the pragmatic value of aiming at something different as a means of advancing the original aim effect a change in whether a state of affairs is good, or a proposed action, concordant with a universalizable law? And if it cannot, then how can it mandate a change in the rational aim?

Well, maybe these scenarios do not prove that the timeless aim must change, only that we were wrong about what it was, something even objectivists must admit is a possibility. But this will not work because for any posited aim one can write an assurance or threat scenario; so there can be no aim such that no circumstance could demand its revision. Even if we thought of the aim as given in a list of timeless conditionals to the effect that, if the situation is *s*1, the aim is *a*1, if *s*2, *a*2, and so on, we could write a scenario where anyone disposed to follow the list would do a better job of it by conditionally intending not to follow it under certain unlikely conditions.

Perhaps we could have it that the aim stays fixed but the rational way to deliberate in light of it changes—Gauthier's proposal. But since the only argument for forming assurance and threat intentions, and so the only argument for deliberation being such that it should recommend them, is the advantage of having them due to their aim-advancing effects on other agents, such proposals will always have a reversion problem; for after the intentions have had whatever effect they could have on others it will be advantageous, because aim-advancing, to have intentions not to fulfill the former intentions, canceling the benefit of the whole exercise.

This leaves two options. One is to hold that, while objectivist theories might be right about what kind of thing the aim is—causing good states of affairs, obeying moral laws—they were wrong about their immutability. Instead, the character of the thing posited as the aim really does change in assurance and threat scenarios. For example, if one was supposed to aim at bringing about good states of affairs, then what counts as a good state changes. Whether a state is a good one is relative to the situation of agents seeking to bring it about—nothing is necessarily good for all possible people in all possible circumstances. This is not entirely implausible. It is widely granted that, for at least some sorts of states of affairs, whether they are good depends solely on the attitudes one takes to them, that is, on whether one would find them welcome, something that could vary as an agent's aim varies. These attitudes are surprisingly flexible; so what counts as a good of this sort might be contextual and relative to the person by whose attitudes they are being evaluated. And perhaps other sorts of good will prove more like this than we thought. So if an agent in a DP had to undergo an alteration in her attitudes to make a threat intention possible, for example, perhaps this brings about a change in what states of affairs are good relative to her—maybe for her, retaliation harms become good. Meanwhile, if one was supposed to aim at complying with correct

moral laws, assurance and threat scenarios would show that no law is one we could will without contradiction be followed by everyone always—for any law there can be a situation in which, if behavior in it is properly law-governed, it is properly governed by a different law.

The other option is to say that while what counts as a good state or a correct law can't change, that doesn't entail that a rational person would always aim at them.

Neither option is initially palatable. For even if there is strategic advantage in adopting something different as the aim, the whole reason the original things were thought the proper aim of rational choice is that it was thought there were things to be said in favor of bringing them about—if they were supposedly good states of affairs—or for obeying them—if they were moral laws. Thus we have good reason always to treat them as the aim, good reason of a different sort sometimes to cease so treating them, and now we have an antinomy of practical reason—apparently decisive and incommensurable arguments for and against the same conclusion.

Our conception of practical rationality seems conflicted. On the one hand, contra J. L. Mackie, we think there are facts whose recognition should and would move agents to act. Perhaps these are facts about which states of affairs would be good, for example, states in which everyone's needs are met. We imagine a given rational agent to have recognized these facts and to have become moved to bring about these states. She has rationally acquired the correct aim of rationality. Then she encounters a paradoxical situation, one calling for a threat. Here, in order to make it likely that a good state will come about, she will have to form a conditional intention to bring about a bad one should some unlikely event occur. She seems now to have reason to do whatever actions would be needed for her to form this intention. But since it is an intention to bring about something bad under certain conditions and since acting on it would consist in knowingly bringing about something bad, this would mean her ceasing to aim at the good. And now we see her as having conflicting reasons: she should do what would form the intention for its good consequences, but should not do it because it would result in her having an intention to do bad things, and possibly in her having to fulfill the intention. We imagine that if she manages to form the intention she will be guilty of bad faith; likewise were she to act on it. But now we find ourselves with conflicting standards of rationality: an agent should always act to advance the good and should never even conditionally intend to bring about the bad, and yet she cannot fulfill both duties where advancing the good requires conditionally intending the bad.

Well, why can't she just do what the aim requires, namely, act to advance the good, that is, do the action that would result in the intention to go against the aim, even if this would put her in bad faith? Because this

theory of reasons also tells her never to intend to do an action against the aim and so implies that she should not do the action that would result in her having such an intention. That is, the theory both demands and forbids bad faith. The same problem arises if we go the other way: suppose we say that facts about what would be a good state of affairs would make it irrational for her to follow through on any conditional intention she might form to bring about a bad state; so such intentions are irrational and she should not do the actions necessary to form them, instead restricting her intentions to ones that facts about the good would give her reason to fulfill. The problem is that these same reasons would justify her in doing what is necessary to form one of these supposedly irrational intentions, since this would likely cause a good state. Again, the theory both forbids and demands bad faith. This tells us that our theory of practical rationality—of the practical duties of rational agents—is contradictory and so false.

Is there a way to resolve the conflict? Yes. The conflict derives from the assumption that there are contra-Mackie facts dictating the aim of rationality, directly regulating the rationality of intentions and actions, and with the power to move agents to intend and act. But I shall argue that the very idea of such facts is contradictory, so there can be no contra-Mackie facts.

Note first that, for some fact to have the power to move a person, she must be susceptible to being moved by it. There is no such thing as a fact having the dispositional power to move a person unless she has the dispositional property of tending to be moved upon seeing that fact.

Now a story about how things would work if there were no contra-Mackie facts permanently regulating what counts as a rational aim and regulating the rationality of intentions and actions. To have something as an aim is just to regard evidence that some action will bring that thing about as a reason to do the action, and to be moved by this to do it. In childhood we have the aims of tiny, selfish gorillas—we desire food, air, water, love, approval. But then our parents and teachers begin our moral education. As children we have little power to dissemble—mothers see everything, including our desires, character traits, aims in short—and little power to advance our aims—we are dependent upon adults for everything. They of course want us to have different aims—to internalize a moral code, to become considerate, altruistic. Our weakness and transparency as children in effect put us in one-shot, multi-stage PDs with our betters: they won't reward us—with their approval and trust to have control over the things for which as children we have appetites, like treats, which should be shared with others—until they see that we have come to want for their own sakes things like fairness to others and have become disposed to take the means to them. Our transparency means pressure can be applied not

just to our actions, but to the basis upon which we do them, so that our original aims will be best advanced only if we come to take different things as aims—we are given pragmatic reasons to adopt different aims. And so we do. We come to be disposed to take it as a reason to do something that it would help others or would conform to a moral code.

But now suppose that as adults we find ourselves in a Deterrence Paradox (DP). To advance our altruistic aims we must form threat intentions, something we cannot do unless we revise our aims. Since to have an aim, *X*, is just to see as a reason for doing something the fact that doing it would advance *X*, in changing our aims we come to see different facts as reasons to act, or old facts as reasons to behave differently than before. And we will persist in these new aims unless and until the new aims would be best advanced by our adopting yet different aims.

This parable suggests a reconciliation of our conflicting conceptions of reasons and of practical rationality. Things seemed intractable because it seemed that there is good reason always to aim at certain things, yet also good reason not to aim at them in paradoxical situations where having different aims would advance the originals. The things imagined in the former description were thought to be inherently the aim of rationality. But perhaps the truth is that nothing has that status inherently. Something is an aim only *for* someone, and only if she sees facts about means to its end as motivating reason to act. Perhaps, then, we can have what counts as a morally good state of affairs be one thing, what counts as a reason to bring it about, another. As children we had no reason to bring these things about; it had not yet become rational for us to be responsive to the fact that certain actions would yield morally good outcomes. As adults we have become so responsive. But DPs put us in situations where it advances the aim of causing morally good outcomes to cease to be responsive to facts about which actions would cause them. No fact is inherently such that we rationally ought to be responsive to it. The reasons why something is good are not necessarily reasons why a rational person ought to bring it about; and the reasons a rational person can have to bring something about are not necessarily reasons that make it good. And just as we can rationally move from being immoral children to being moral adults, so we can rationally move to being immoral in certain desperate circumstances.[24] And then we are no longer responsive to arguments that

24. Actually, when morality and practical rationality require someone with morally approved aims to adopt different aims—perhaps ones prima facie morally monstrous—in order to advance the former aims, it seems a mistake simply to call her, her aims, or the actions they might make her do immoral, monstrous. This seems unfair and even contradictory. For then the agent could have escaped condemnation only by both advancing and retaining the original aims. In the Deterrence Paradox, the agent would have counted as monstrous had she done nothing to try to prevent the destruction of half the planet, so she had to change her aims to make this possible, but then she would also count as monstrous

doing a certain action would be immoral. That is now, for us, no more a reason than it is for the world's atrocity-committers who as children were never put in a situation where it would have been pragmatically rational, given their prior aims, to become responsive to moral considerations. True, the facts which formerly made it irrational to, say, fulfill a threat are still in place—retaliating will still cause massive gratuitous harms. But what has changed is whether those facts are reasons not to retaliate. For an agent in a paradoxical situation like a Deterrence Paradox, those facts are no longer reasons, no longer relevant; for she has justifiedly ceased to have the minimization of harms as an aim.

If the above picture of the nature of reasons were correct, our problem would be solved. For we would then not be violating a duty of practical rationality if we came to aim at something bad in order to advance the good, nor if we formed and fulfilled the intentions the new aims would rationalize; for no standard properly regulating the rationality of these things both persists and condemns them. But is the picture correct? Yes. Reasons why something is good and reasons to be responsive to its being good must be separate matters. Otherwise we arrive at contradiction. For suppose the factor that made something, *S*, good also made *S* something we rationally had to be responsive to, that is, gave us conclusive reason to be disposed to advance *S*, and so conclusive reason to advance *S*. And suppose the only way to advance *S* was to do an action, like *A*, above, which would have the effect of making us no longer responsive to *S*, that is, no longer disposed to advance *S*. Then *A* would be such that it rationally should be done—in order to satisfy the duty to advance *S*—and not done—for doing it would result in our failing our rational duty to have the disposition to be responsive to *S*. Contradiction. Assuming *S* always remains good, the only way out is if we are rationally required in paradoxical situations to cease to be responsive to *S*'s being good. But then *S*'s being good isn't necessarily conclusive reason to be responsive to *S*.

The problem we've been treating derived from the posit that certain considerations make it that it is always rational to have a certain thing as the aim of rational choice, and that this means one must always intend and act to advance that aim. We saw that in paradoxical situations, the

---

for fulfilling any threat she had to make to avoid the former charge of monstrousness, meaning that, to avoid being called a monster, she should have retained the original aim. So she would somehow have been obliged to both revise and retain the original aim. This is logically impossible and yields the contradictory description of changing aims here as both morally good and not, meaning that any theory of morality that entails it is necessarily false. To avoid this we must see moral obligation (and moral monstrousness) as relativized, like we are seeing for the rational aim. See my "Moral Paradox" for more. And for the inception of this issue, see Gregory Kavka, "Some Paradoxes of Deterrence," *Journal of Philosophy* 75 (1978): 285–302.

duty to act to advance the aim would require one to do something that would result in one's having an intention not to advance the aim, and so in one ceasing to have the aim; and here, these posited duties yield a contradiction, namely, that doing the intention-forming action both is and is not rational. It follows then that the posit about our duties is false. And this entails the falsity of any thesis that could be true only if the posit was true. For example, it entails that no state of affairs can be one we are always rationally obliged to act to bring about and to intend to bring about; no law is one we can be rationally obliged always to intend to follow and to follow; no reason can be such that it makes it conclusively rational always to intend to do and do any of the foregoing things; no disposition can be coherently described as a disposition always to intend to do and to do the foregoing things; and no fact can have the power to make it that we always do or should intend to do and do any of those things. Further, no coherent analysis of any duty of a person—for example, the duty to be rational, to be morally well disposed, to have correct desires or aims, to be sane, to be nonmonstrous, to be a person of good faith—can entail that the duty's expression would require us always to intend to do and do those things. And since a contra-Mackie fact that does and should always motivate us to do and intend to do certain things can exist only if there can exist a disposition to conform to such a norm, and only if there can be such a norm, and since on pain of contradiction neither can exist, no such fact can exist.[25]

Paradoxical situations are therefore proved not to be occasions on which it is logically inevitable that we either fail to do morally and rationally required actions or fail to have morally and rationally required intentions and aims. Instead, they are occasions to deliberate about which

25. Couldn't there be an aim—and so a corresponding anti-Mackie fact—immune to these arguments provided it embeds side constraints? Consider the aim of always minimizing harms except never using an intention to cause harms as the means to the minimizing. Surely there will not be the contradiction generated by an aim being such that the duty to advance it and the duty to intend to advance it collide, since for this aim, the former duty is qualified to respect the latter duty. But I don't think this will work. For suppose you have the aim, but an Evil Demon will make you immediately violate it unless you do an action, *A*, that would result in you having a conditional intention to violate the aim should some unlikely event occur. That is, if you don't do *A*, the demon will certainly and immediately induce you to fail to do an action needed to minimize harms even though it would not result in an intention the aim would find problematic; or he will induce in you a problematic intention. Surely in rightly having the aim, you also have a duty to act to make it that you comply with the aim as often as possible. So you have a duty to do the action, *A*, that will result in the conditional intention. But now you have conflicting duties: you must do *A* on pain of failing the duty to make it that you comply with the aim as often as possible, but you must not do *A* on pain of failing always to have only intentions to comply with the aim. For *A* will result in you forming a conditional intention to form a problematic unconditional intention forbidden by the aim, or a conditional intention to fail to do an action that the aim requires and permits. All our problems resurface.

facts to be responsive to, that is, which aims to have. Deliberation is sometimes rationally obliged to be about aims; and properly understood, it is a consequence of all the theories at hand that such deliberation is, where rational, pragmatic. This is because on these theories there is no vantage from which to deliberate about which aims to have except the aims one already has, as on the Humean theory, or the aims one is initially posited to be rationally obliged to have, as on the objectivist theories. These determine what one has reason to be responsive to. For every theory sees practical rationality as advancing some aim, and so sees what there is reason to intend to do and do as dictated by it. And since whatever the initial aim, ironically, each aim, on pain of contradiction, dictates its own revision in paradoxical situations, these aims also dictate what facts one should come to be responsive to going forward, that is, what aims one should come to have.[26] Because the aims rational to have dictate the correctness of intentions, actions, and even new aims, when we undergo a change in our aims on the pragmatic pretext that this would serve our current aim, we are doing what good faith respecting this aim requires, and when we then intend to do and do the actions required by the new aims, we are doing what good faith respecting these new aims requires. In all cases we are doing exactly what practical reason demands. And as we've seen, there can be no permanent standard respecting which any of this constitutes bad faith or irrationality—the very idea that there could be such a thing proved to be contradictory and so false.

Now, for any aims there will be all manner of objections to their being revisable—that changing them would be in bad faith (something addressed above), or would be rationally motivated irrationality, or would amount to insanity (assuming being sane entails having certain values or aims), or would exemplify lack of integrity, or moral monstrousness; or that one may rationally aim only at what is on a list of supposedly permanently objectively good things and that only evidence of states of affairs having different nonnormative properties than those formerly thought can be a good reason to change in whether one aims at them; or that changing would go against intuitions of what is worthy of having as an aim; or that experience would continually teach us the inappropriateness of aiming at the new things; or that, even if a true theory of the aims of rationality dictated that one proceed as if the aims were different, that wouldn't make the new aims the correct aims; or that if changing were appropriate, this would, absurdly, confound universalization and deny categorical imperatives.

26. Must all theories of practical rationality have this structure and so these issues? J. David Velleman might think his theory an exception, although I suspect paradoxical situations pose a reversion problem for it, one whose solution will require a change in what his agents would count as practical reasons. See his "Deciding How to Decide," in *The Possibility of Practical Reason* (Oxford: Clarendon, 2000), 221–43.

Obviously I cannot here individually address all of these claims. But note that some are just different ways of saying that rational aims can't be changed—for example, that changing them would be in bad faith or would be rationally motivated irrationality—and so aren't really arguments against changing them. And all are such that, whatever purported justification there may be for them, they supposedly entail that rational agents should always have certain aims, always advance them, and always intend only what would advance them. But we have seen that this view entails contradictions and so must be false; so the foregoing arguments must all contain mistakes. This is the master argument against the rational unchangeability of aims on pragmatic pretexts. It is a project for another time to investigate where exactly the arguments go wrong.[27]

27. I have elsewhere tried to deal with the arguments of Thomas Nagel in his *The Possibility of Altruism* (Oxford: Clarendon, 1970), for example. See my "Prudence and the Reasons," and "Prudence and the Temporal." Millgram (*Practical Induction*) is also relevant. Meanwhile, the charge of insanity is of special interest. My view entails, on pain of the foregoing contradiction from the "master argument," that being sane does not necessarily involve having certain approved values or aims, or that the ones required are relative to one's situation—e.g., in a DP, a sane person would acquire values that would be advanced by retaliating—or that a rational person as such would not always be sane. Given the adjacency of the concepts of sanity and rationality, the last option may be the least plausible. But the matter needs more discussion.