Twentieth Century Values, ed. by Kenneth F.T. Cust

The Toxin and the Tyrant: Two Tests for Gauthier's Theory of Rationality

BEN EGGLESTON[1]
DEPARTMENT OF PHILOSOPHY
UNIVERSITY OF KANSAS

1. Introduction

Hume famously said that "Reason is, and ought only to be the slave of the passions."[2] Let us assume, with Hume, that reason does not, because it cannot, tell a person which ends to pursue. In other words, let us assume that although reason can apprise a person of the availability of various ends and of the costs and benefits likely to attend the pursuit of those ends,[3] it cannot judge the desirability of those ends themselves. Assuming all this—assuming, in short, a purely *instrumental* view of rationality—it is natural to think that at least the following (if only this) can be said on reason's behalf: the more rational a person's choice of conduct is, the more will it further her ends, whatever they may be. And from this it is natural to infer that what is fully rational is for a person to choose whatever conduct will further her ends the most.

This conception of rationality—the idea that it's rational for a person to choose whatever conduct will further her ends the most—is as simple as it sounds, and I think it's no exaggeration to say that it enjoys the status of orthodoxy among rational-choice theorists, game theorists, and other people who traffic in such things. But like any orthodoxy, this one has its heretics, and one of these is David Gauthier. As an alternative to

[1] My sincere thanks go to David Gauthier, for encouraging my work on this paper and for commenting extensively, and with unflagging patience and energy, on successive versions of it. I also want to thank Donald Bruckner, for providing comments that strengthened this paper in many places, but especially in the crucial sections 6.2 and 6.3; Adrian Staub, for pointing out several errors in an early draft of this paper and for helpful discussion on its topic; and the members of the audiences to whom I read this paper, at the University of Pittsburgh philosophy graduate-student colloquium, on April 2, 1999, and at the 27th Conference on Value Inquiry, at Central Missouri State University, on April 23, 1999.

[2] David Hume, *A Treatise of Human Nature*, second ed., rev. by P. H. Nidditch (Oxford: Clarendon Press, 1978), p. 415.

[3] *Ibid.*, p. 416.

this conception of rationality, which he calls straightforward maximization, Gauthier proposes a conception of rationality that he calls constrained maximization.[4] In this paper, after reviewing Gauthier's reasons for rejecting straightforward maximization in favor of constrained maximization (in parts 2, 3, and 4), I'll discuss (in part 5) Gauthier's recent attempt to refine the theory underlying constrained maximization so that it ceases to have a certain implication that he regards as objectionable. In particular, I'll argue (in part 6) that the refinement Gauthier introduces may be initially appealing, but actually does his theory more harm than good.

2. How Straightforward Maximization Is End-Frustrating and, Thus, Self-Defeating

2.1 Gauthier rejects straightforward maximization because it's self-defeating in the following sense: a person who deliberates in the way that straightforward maximization prescribes (choosing in every case whatever conduct will further his ends the most) ultimately does not further his ends as much as he would if he were to deliberate in some other way. We can see Gauthier's point by considering three hypothetical situations of choice.

2.2 First, consider a case that Gauthier[5] adapts from Hume, in which we are to suppose that you and I are farmers. I want you to help me with my harvest this week, and you want me to return the favor next week. Each of us would most like to get help without giving it, but each of us still prefers cooperative harvesting to independent harvesting. And because you in particular prefer cooperative harvesting to independent harvesting, you'll be willing to help me this week if, but only if, I sincerely assure you that I'll return the favor next week. (I'm a terrible bluffer, so only a *sincere* assurance will work.) But let us suppose, in addition, that I'll have no incentive to help you next week: by then I'll have gotten the help from you that I wanted (or won't have gotten it, and won't then be able to get it), I am not much moved (and won't be much moved next week, either) by moral considerations favoring reciprocation, and I will soon be retiring to Florida, never again to cross paths with you or the neighbors to whom you'll surely report my rude conduct if I fail to reciprocate. The key point is that what will further my ends the most next

[4] David Gauthier, "Reason and Maximization", in David Gauthier, *Moral Dealing: Contract, Ethics, and Reason* (Ithaca: Cornell University Press, 1990), pp. 209–233 (originally published in *Canadian Journal of Philosophy* vol. 4, no. 3 [March 1975], pp. 411–433), p. 227–228.

[5] David Gauthier, "Assure and Threaten", *Ethics* vol. 104, no. 4 (July 1994), pp. 690–721, pp. 692–693.

week—what will be *optimal* for me—is to *refuse* to help you, whether you have helped me this week or not.

Now, if I'm a straightforward maximizer, then I have a problem. For I can already see, this week, that I'll definitely refuse to help you next week. And this foreknowledge that I have, of my own straightforwardly maximizing behavior, prevents me from sincerely assuring you that I'll help you. As result, you will (quite reasonably) refuse to help me this week, and our cooperative venture won't even get off the ground. More generally, I can never be admitted to *any* cooperative venture in which I would be called on to contribute after benefiting, because I cannot give the sincere *assurance* of contributing that is needed for people to admit me to such ventures in the first place. (I can be admitted to ventures in which I benefit *after* contributing, and my contribution will be verified, since then my contributing is optimal: it's a means to my benefiting.) Of course, I would *like* to be able to give such assurances, even at the cost of following through (remember, I prefer cooperative harvesting to independent harvesting); indeed what's especially frustrating about such cases is that I'm not even trying to *exploit* other participants in the venture—I just want to get in and get my fair share of the fruits of cooperation, along with everyone else. But as a straightforward maximizer, I'm excluded from such ventures. Gauthier sums up my trouble succinctly: "his very way of choosing affects the situations in which he may expect to find himself. And the effects are to his disadvantage."[6]

2.3 Giving a sincere assurance to perform non-optimal conduct, even when giving such an assurance would itself be optimal, is not the only potentially optimal conduct that the straightforward maximizer is unable to perform. A central problem in recent work on deterrence is that straightforward maximizers cannot make sincere *threats* to perform non-optimal conduct.[7] Borrowing another example from Gauthier,[8] let us suppose that I buy widgets from you in large quantities. I know that what I pay you far exceeds what you need to be paid in order for our transactions to be worth your while, and I know that you would give in to my demand for a discount if you feared that I would take my business

[6] David Gauthier, "The Incompleat Egoist", in Gauthier, *Moral Dealing*, *op. cit.*, pp. 234–273 (originally published in Sterling M. McMurrin [ed.], *The Tanner Lectures on Human Values*, vol. 5 [Salt Lake City: University of Utah Press, 1984]), pp. 67–119, p. 263.

[7] See David Gauthier, "Deterrence, Maximization, and Rationality", in Gauthier, *Moral Dealing*, *op. cit.*, pp. 298–321 (originally published in *Ethics* vol. 94, no. 3 [April 1984], pp. 474–95), pp. 298–302. Also see Anthony Kenny, *The Logic of Deterrence* (London: Firethorn Press, 1983), pp. 21–22 and pp. 37–38.

[8] David Gauthier, "Commitment and Choice: An Essay on the Rationality of Plans", in Francesco Farina, Frank Hahn, and Stephano Vannucci (eds.), *Ethics, Rationality, and Economic Behaviour* (Oxford: Clarendon Press, 1996), pp. 217–243, pp. 224–225.

elsewhere. So you'll give me the discount if, but only if, I sincerely threaten to take my business elsewhere. (Again, I'm a terrible bluffer.) But, in addition, taking my business elsewhere will be non-optimal for me: even if you do not give me the discount, the price you charge for widgets will remain (as it is now) the lowest price I can find, and (as above) there will be no other consequences to worry about. In short, what will be optimal for me is to continue to buy widgets from you, whether you have given me the discount or not.

If I am a straightforward maximizer, then I have the same problem as before: I can't sincerely tell you now that what I do second will depend on what you do first, since I know that's not true. And so you won't give me the discount that I want. Admittedly, in this case my motives are harder to sympathize with, because instead of trying to participate in a venture that helps us both, I'm trying to exploit you, and we might be glad that I can't. But however socially desirable this outcome may be, it's still a failure of instrumental rationality: it's a failure to use the means at my disposal—in this case my strong bargaining position—for the maximal furtherance of my ends. In a sense, only my *thoughts* (my thoughts about what it will be rational for me to do, if you ignore my threat) stand in my way. It is as if I have the opportunity to pick your pocket, but am somehow handcuffed from the inside.[9]

2.4 In both the assurance case and the threat case, what being a straightforward maximizer prevents me from doing is *forming certain intentions*: forming intentions to perform acts that I know will be non-optimal, such as contributing to a cooperative enterprise when I've already benefited, or following through on a failed threat. To focus more directly on the straightforward maximizer's inability to intend to perform non-optimal acts, consider Gregory Kavka's much-discussed toxin puzzle. An eccentric billionaire—without whom many a philosophical thought experiment would be utterly insolvent—offers me the following deal:

> He places before me a vial of toxin that, if I drink it, will make me painfully ill for a day, but will not threaten my life or have any lasting effects. . . . The billionaire will pay me one million dollars tomorrow morning if, at midnight tonight, I intend to drink the toxin tomorrow afternoon. He emphasizes that I need not drink the toxin to receive the money; in fact, the money will already be in my bank account hours before the time for drinking it arrives, if I succeed. . . . All I have to do is . . . intend at

[9] Some months after writing this sentence, I discovered that Gregory Kavka had already used similar language, or at least had evoked a similar image, in order to describe an agent similarly situated. Kavka writes of that agent, "He is a captive in the prison of his own virtue." See Kavka's "Some Paradoxes of Deterrence", *The Journal of Philosophy* vol. 76, no. 6 (June 1978), pp. 285–302, p. 291.

> midnight tonight to drink the stuff tomorrow afternoon. I am perfectly free to change my mind after receiving the money and not drink the toxin.[10]

Now of course I am a lot more interested in getting that million dollars than I am in avoiding a day's illness. So although I would love to get the million dollars *without* drinking the toxin, I would still much prefer getting the million dollars—even *along with* drinking the toxin—to any outcome in which I do not get the money. The money matters most. And all I have to do to get the money is form the intention to drink the toxin. But actually drinking the toxin will be non-optimal: when it's time to drink, I'll either have the money or not; and the only thing left for me to decide will be whether to drink and thus make myself painfully ill. Naturally, I would prefer not to.

You know what is coming next. If I'm a straightforward maximizer, then I'm out of luck: I can see now, and will continue to see at midnight tonight, that I will decline tomorrow afternoon to drink the toxin. This keeps me from forming the intention to drink, and so it keeps me from getting the million dollars. Obviously, if the order of events were reversed, so that my getting the money were contingent on my actually drinking, then I'd be all set: for then drinking would be optimal, and I'd gladly do it. But the situation is not like this. It calls for me to benefit first and pay a cost second. Paying the cost will be non-optimal, and because I know I do not knowingly perform non-optimal acts, I can't form the million-dollar intention. As before, what stops me from taking full advantage of the resources and opportunities at my disposal, or doing what will further my ends the most, is just my *knowledge* that this is the kind of agent I am: an agent who always chooses the act that furthers his ends the most. When the eccentric billionaire makes his offer, I can't plan to take advantage of it; the best plan I can make is to try to just forget about the money. "At least I have my health," I'll console myself.

2.5 Hollow assurances, idle threats, and second-rate plans: these seem to be the hallmarks of the straightforward maximizer. Note that the straightforward maximizer's problem is not that he fails to take account of all of his options, or fails to foresee some of the consequences of some of those options, or fails to see how some of those consequences will affect the furtherance of his ends; we are considering an idealized straightforward maximizer, one who acts with perfect information about the world around him. His problem is deeper. His problem is that his very way of choosing—the mere fact that he's a straightforward maximizer, along with his awareness of this fact—prevents him from

[10] Gregory Kavka, "The Toxin Puzzle", *Analysis* vol. 43, no. 1 (January 1983), pp. 33–36, pp. 33–34. For continuity with sections 2.2 and 2.3, I have replaced Kavka's second-person pronouns with first-person pronouns.

intending to perform non-optimal acts. And this prevents him from securing the benefits that certain situations offer.

But are these difficulties—which seem so alien to the very idea of rational choice—genuinely indigenous to straightforward maximization, or have they been planted there by its opponents? In defense of straightforward maximization, it is tempting to point out that the sort of straightforward maximizer we have been considering—who handles the situations under discussion so ineptly—does not represent straightforward maximization in its best light; and that a sophisticated straightforward maximizer, instead of forgoing the benefits to be gained from forming plans that require him to perform non-optimal acts, will secure those same benefits by establishing precommitment devices such as side bets, the existence of which may make the required acts optimal, and external structures in which we "deposit our will"[11] so that the desired consequences of the required acts can be effected without subsequent action by the agent. But this reply does not vindicate straightforward maximization. For it is not always possible in the cases under discussion to establish precommitment devices, and even when it is, doing so has costs: tying oneself to the mast (either literally, like Ulysses, or figuratively, in any of a number of ways) consumes resources, as does contriving a penalty for one's future self if one deviates from the plan one makes today. And side bets, if they do not consume resources, at least make them temporarily unavailable, insofar as one's share must be placed in escrow. Moreover, as Gauthier points out, the strategy of precommitment "fails to face the real issue—that taking my reasons for acting directly from my aim [which is how the straightforward maximizer deliberates] is in certain situations counter-productive and, indeed, self-defeating in relation to that aim."[12] If only the agent could intend to perform non-optimal acts, he'd be a lot better off.

Admittedly, this is a bit counterintuitive—this idea that a person would be better off if he could intend to perform non-optimal acts. For if a person could *intend* to perform non-optimal acts, then he might occasionally *perform* non-optimal acts, thereby frustrating his ends rather than furthering them. And it might seem unlikely that a person can end up better off, if he has this non-optimal conduct flaring up from time to time. The answer is that what makes him better off is not the performing of any non-optimal act, but the intending to perform it. Intentions can

[11] Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality* (Cambridge: Cambridge University Press, 1979), p. 43.

[12] Gauthier, "Assure and Threaten", *op. cit.*, p. 696. For more on the inherent costliness of establishing precommitment devices, see Edward F. McClennen, *Rationality and Dynamic Choice: Foundational Explorations* (Cambridge: Cambridge University Press, 1990), pp. 196–98; and Edward F. McClennen, "Pragmatic Rationality and Rules", *Philosophy and Public Affairs* vol. 26, no. 3 (Summer 1997), pp. 210–258, pp. 233–234.

have what Kavka called "*autonomous effects*"—effects that are "independent of the intended act's actually being performed."[13] In the cases I discussed earlier, the benefits I had to forgo—your cooperation in the assurance case, your capitulation in the threat case, the million dollars in the toxin case—were all effects of this kind. They were contingent on my intentions, not on my acts. Thus, merely by limiting the range of things than an agent can intend to do, straightforward maximization can keep that agent from maximally furthering his ends. And since maximally furthering one's ends is precisely what straightforward maximization says it's rational for an agent to do, straightforward maximization ends up being self-defeating.[14]

3. How *Constrained* Maximization Is Less End-Frustrating and, Thus, More Effective

3.1 As an alternative to straightforward maximization, Gauthier proposes *constrained* maximization. Like the straightforward maximizer, the constrained maximizer is concerned to further her ends, but her deliberative procedures differ from those of the straightforward maximizer in a crucial way. Recall that for the straightforward maximizer, optimality of acts is everything: acts deemed non-optimal are *ipso facto* ruled out. But for the constrained maximizer, matters aren't so simple. The constrained maximizer chooses an act not solely on the basis of its optimality, but also on the basis of such considerations as its compatibility with the plan (or plans) the forming and execution of which, considered as a one continuous course of action, can be expected to further her ends the most. Now this portrayal of constrained maximization omits much of the detail of Gauthier's conception of it,[15] but it is sufficient for our purposes because our concern is simply with how the constrained maximizer fares (compared to the straightforward maximizer) in the three situations discussed earlier.

3.2 When confronted with the farming situation, a constrained maximizer would realize that the best plan she can make is one calling

[13] Kavka, "Some Paradoxes of Deterrence", *op. cit.*, p. 201.

[14] Some defenders of the orthodox conception of rationality dispute this conclusion. (See, for example, R. Eric Barnes, "Constraint Games and the Orthodox Theory of Rationality", *Utilitas* vol. 9, no. 3 [November 1997], pp. 329–49.) But I shall not address this question here, as my aim in this paper is to identify a problem that arises for Gauthier regardless of how this question gets settled.

[15] For some of Gauthier's specifications and modifications of constrained maximization, see his "Reason and Maximization", *op. cit.*, pp. 227–30; "The Incompleat Egoist", *op. cit.*, pp. 264–67; "Introduction", in his *Moral Dealing*, *op. cit.*, pp. 1–8, pp. 4–5; "Assure and Threaten", *op. cit.*, pp. 702–707; and "Rethinking the Toxin Puzzle", in Jules L. Coleman and Christopher W. Morris (eds.), *Rational Commitment and Social Justice: Essays for Gregory Kavka* (Cambridge: Cambridge University Press, 1998), pp. 47–58, pp. 48–53.

for reciprocation (i.e., helping you if you help her, and not helping you if you don't help her) because that enables her to give you the sincere assurance that results in her getting your help; and when she proceeds to reciprocate, the result is cooperative harvesting, which she prefers to independent harvesting. Admittedly, her best *outcome* would be to get your help but then to refuse to reciprocate; but this is not an outcome that she can *plan on*: her getting your help is contingent on her giving you a sincere assurance of reciprocation, which is contingent on her genuinely planning to help you if you help her; and she cannot both plan to reciprocate (as needed for the sincere assurance) and plan not to reciprocate (as needed for her to plan on her best outcome). Seeing that reciprocation is part of the best *plan* she can make, she judges that reciprocation will be rational—and, more to the point, she expects that she'll reach this same judgment next week, too. As a result, she can now give you the sincere assurance on which your helping her this week depends, and the mutual-benefit game works out perfectly.

Now it may be objected that the constrained maximizer doesn't take *full advantage* of the situation in which she finds herself, because she voluntarily forgoes an opportunity to exploit you (by getting your help but not reciprocating); but that is beside the point. For the point is that the constrained maximizer's outcome (cooperative harvesting) is preferable to the straightforward maximizer's outcome (independent harvesting). As Gauthier writes, "Although the [constrained maximizer] refrains from making the most of her opportunities, yet she finds herself with opportunities that the [straightforward maximizer] lacks and so may expect payoffs superior to those that he can attain."[16] And so in this case, at least, constrained maximization proves to be more effective than straightforward maximization, in the furtherance of the agent's ends.

3.3 Now consider the threat case. Our constrained maximizer would see that the best plan she can make is one calling on her to take her business elsewhere if you don't capitulate, because if she makes this plan, then she can issue the sincere threat that, by hypothesis, results in her best outcome. Of course, this would be the best plan for a *straightforward* maximizer to make, too, if he were capable of making it; but because it would call on him to perform a non-optimal act (or, at least, to be prepared to perform a non-optimal act, depending on whether you capitulate), he cannot adopt it. But because a constrained maximizer can plan to perform acts that are not optimal (if they are called for by plans that are optimal, as this one is), she can issue the sincere threat that

[16] Gauthier, "The Incompleat Egoist", *op. cit.*, p. 265. In the passage from which this quotation comes, Gauthier refers to the straightforward maximizer as an "egoist" and to the constrained maximizer as a "conditional cooperator." For terminological consistency with other parts of my discussion, I have replaced Gauthier's names as indicated.

results in your capitulation, outdoing the straightforward maximizer yet again.

3.4 Finally, recall the case of the toxin. The constrained maximizer's approach should be clear: she sees that the best plan she can make is to drink the toxin tomorrow afternoon, because that will result in her intending, at midnight tonight, to drink the toxin; and even when she drinks the toxin, at least she has secured the money. (Remember, the money matters most.) As in the farming case, she does not attain her *best* outcome (getting the money and then not drinking), because this is not something she can *plan* to attain (for if she plans on not drinking, then she won't intend to drink, and she won't get the money). Seeing, then, that drinking *is* called for by the best plan she *can* make, she judges that drinking will be rational, and she expects that she'll reach this same judgment when it's time to drink. As a result, she intends to drink, earning the money that the straightforward maximizer is denied.

3.5 In each of the three situations, then, a constrained maximizer can adopt plans and intentions that a straightforward maximizer cannot, enabling her to avoid the inconveniences of straightforward maximization and to secure the benefits that we would expect from truly rational choice. As a result, constrained maximization is more effective than straightforward maximization, in the furtherance of the agent's ends.

4. Why, According to Gauthier, the Results of Parts 2 and 3 Matter

4.1 It might be objected that the results of parts 2 and 3 are irrelevant to the question of whether constrained maximization is a better conception of rationality than straightforward maximization. For it might be thought that situations sometimes arise in which irrational agents happen to fare better than rational ones, and that the situations discussed above are of this kind—meaning that the correctness of a conception of rationality is not a function of how effective it is. But according to Gauthier, it matters very much that constrained maximizers fare better than straightforward maximizers, because the furtherance of the agent's ends is what instrumental rationality is all about:

> If a person's reasons take their character from her aim, then it is surprising and troubling if acting successfully in accordance with her reasons, she must sometimes expect to do less well in relation to her aim than she might. . . . If the orthodox account of the connection between aim and reasons were correct [in other words, if straightforward maximization were the best conception of rationality], then sometimes I should not expect success in

> acting on my reasons to lead to my life going as well as possible. And so I propose to rethink the connection.[17]

The connection that Gauthier proposes is what I call the *maximal-effectiveness* theory of rationality:

> I conclude that deliberative procedures are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible.[18]

To compare two rival sets of deliberative procedures, he explains, we are to consider the situations in which they yield different choices, and if one set "would sometimes access inferior prospects [relative to the other], and never access superior prospects, then reject it as less than fully rational."[19] It follows that we ought to reject the deliberative procedures prescribed by straightforward maximization as less than fully rational, since we saw in parts 2 and 3 that the effects of employing them are, for the agent, sometimes worse than, and apparently never better than, the effects of employing the deliberative procedures prescribed by constrained maximization. In short, the deliberative procedures prescribed by constrained maximization are more conducive to the furtherance of the agent's ends, or to her life's going well, than are the deliberative procedures prescribed by straightforward maximization. And the maximal-effectiveness theory of rationality says that this matters: it says that it's this feature of constrained maximization—its pragmatic effectiveness—that makes it a better conception of rationality than straightforward maximization.

4.2 Before proceeding, a clarificatory note: The fact that straightforward maximization is self-defeating is actually something of a red herring. What matters is that the deliberative procedures prescribed by straightforward maximization are not *pragmatically effective*, in that straightforward maximizers fare worse than constrained maximizers. The fact that straightforward maximization sets pragmatic effectiveness as the *aim* of conduct is, at the end of the day, a coincidence. After all, a conception of rationality can easily be pragmatically ineffective (and, thus, rejected by the maximal-effectiveness theory of rationality) without being self-defeating. For example, an agent who always chooses whatever conduct will minimize her exposure to the sun may actually succeed in minimizing her exposure to the sun (meaning that her conception of rationality is not self-defeating), but unless she has a very unusual set of ends, with avoiding exposure to the sun being pre-eminent

[17] Gauthier, "Assure and Threaten", *op. cit.*, p. 694.

[18] *Ibid.*, p. 701.

[19] David Gauthier, "Resolute Choice and Rational Deliberation: A Critique and a Defense", *Noûs* vol. 31, no. 1 (March 1997), pp. 1–25, p. 22.

among them, then on the whole the effect of employing such deliberative procedures is likely to be worse than the effect of employing deliberative procedures that are somewhat more responsive to the ends that she actually has. Such a "heliophobic" conception of rationality would not be self-defeating, but it would clearly be pragmatically ineffective, and it is on this basis that the maximal-effectiveness theory would reject it. Similarly, the problem with straightforward maximization is not that it's self-defeating (though it is), but that it's pragmatically ineffective. For our purposes, it's just a coincidence that the pragmatic ineffectiveness of straightforward maximization also makes it self-defeating.

5. How the Maximal-Effectiveness Theory Goes Too Far (the Case of the Tyrant)

5.1 So the maximal-effectiveness theory of rationality says that deliberative procedures are rational if and only if they are the deliberative procedures the having of which is optimal, or most pragmatically effective. This comes from papers by Gauthier published in 1994 and 1997. But more recently, he has qualified this claim. Consider the following passage, from a paper of his published in 1998:

> To guard against misunderstanding my account of deliberation, it is essential to emphasize that deliberative reasons relate to effective direction. They are not simply whatever considerations would need to weigh with someone if he is to realize his concerns.[20]

And to this Gauthier adds the following footnote:

> Thus, what I said in another essay – 'deliberative procedures are rational if and only if the effect of employing them is maximally conducive to one's life going as well as possible' – needs emendation. As a first approximation, we might say that deliberative procedures are rational if and only if they are effectively directed to making one's life go as well as possible.[21]

And in the text Gauthier goes on to say that in order for my deliberation to be rational, it must be "directed effectively at the realization of my concerns."[22] This is what I call the *directed-effectiveness* theory of rationality.

5.2 To motivate this refinement, Gauthier asks us to consider a case in which deliberation that's effective for the realization of my concerns is

[20] Gauthier, "Rethinking the Toxin Puzzle", *op. cit.*, p. 49.

[21] *Ibid.*, p. 58, n. 5.

[22] *Ibid.*, p. 49.

not *directed at* the realization of my concerns. Specifically, suppose that I am under the control of a being (whom I call the tyrant) who will punish me if I deliberate in the usual way (i.e., with a view to furthering my ends) but will reward me if I "take her directives as reasons for acting in themselves, independently of how they relate to my concerns."[23] In such a situation, it would clearly be maximally effective—maximally conducive to the furtherance of an agent's ends—for the agent to just forget about her ends and simply take the tyrant's commands *in themselves* as reasons for acting.

Reflecting on this case leads Gauthier to repudiate the claim that the rational agent is simply the agent who does what it takes to succeed (or, more precisely, who thinks in the way in which an agent has to think in order to succeed). For the realm of the tyrant is such a weird place that successful deliberative procedures will be so bizarre—taking the tyrant's commands *in themselves* as reasons for acting, independently of how they relate to independently held ends—that Gauthier recoils from endorsing them as *rational*. Instead, claiming that mere effectiveness is not enough, Gauthier says that in order to be rational, an agent's deliberation must also be *directed at* the furtherance of her ends. It is unfortunate that in a case such as this one, deliberation that is directed in this way would not be very effective, but there are bizarre cases in which the link between effectiveness and rationality cannot be sustained. In such cases rational deliberation is ineffective, while effective deliberation is irrational. Gauthier concludes,

> The pragmatic standard . . . that I embrace does not lead to the absurd view that rationality is simply a matter of what pays.[24]

In this way, Gauthier settles on the directed-effectiveness theory of rationality.

6. What the Directed-Effectiveness Theory Amounts To (and How It Really Supports Straightforward Maximization)

6.1 Gauthier is not very explicit about what it takes for an agent's deliberation to be *directed at* the furtherance of her ends, as opposed to being merely *effective for* the furtherance of her ends. Still, as a first approximation, it seems reasonable to infer (from Gauthier's discussion of the tyrant case) that deliberation directed at the furtherance of an agent's ends is deliberation in which the agent is consciously concerned to advance her aims: she deliberates *about* the furtherance of her ends, if not necessarily in a way that *results* in the furtherance of her ends.

[23] *Ibid.*, p. 49.

[24] *Ibid.*, p. 50.

Though not single-mindedly committed to the furtherance of her ends in the way that a straightforward maximizer is, the rational agent eschews blind obedience—e.g., of the sort that is effective in the tyrant case—and keeps her ends in view.

6.2 But when we consider the constrained maximizer, it becomes clear that she succeeds in cases of assurances, threats, and toxins only to the extent that her deliberation is *not* directed at the furtherance of her ends. She succeeds only because, instead of keeping her ends in view, she gives absolute priority to making sure that she follows through on her assurance, or her threat, or her plan to drink the toxin. We can count on her to make good on her word in these cases only because we can assume that when it is time for her to act, she will be decisively moved to make good on her word, even though doing so will require her to *neglect* the pursuit of her ends. For at the time of acting, her ends oppose doing what she chooses to do; she has no end-based reason, no outcome-oriented reason, to do any of them. If these things strike her as worth doing, it's because she's lost sight, however temporarily, of her ends.

We can appreciate the depth of the constrained maximizer's neglect of her ends by imagining how she might reply if we were to ask her, at the time of decision, why she is choosing to follow through. She might appeal to the future, past, or current benefits of following through. First, she might say something about the future benefits of following through, such as the pangs of conscience that would attend defaulting and the reputation effects of defaulting, but any reply along these lines would be a non-starter, since in each case it is stipulated that the benefits of following through are outweighed by the costs: in each case, when all the costs and benefits are counted, following through is unambiguously non-optimal. Second, she might mention the benefits that her having been a constrained maximizer has enabled her to secure in the past: your cooperation in the assurance case, your capitulation in the threat case, a million dollars in the toxin case. But at the time of decision, she has already secured these benefits, and no decision then available to her can put them at risk. Third, she might dwell on the benefits that her having being a constrained maximizer enables her to enjoy in the ongoing present: she might fully admit that following through only frustrates her ends, but she might add that even after she frustrates her ends in this way, she will be doing better than if she had never intended to act in this way in the first place. But this reply misses the point. She has already secured the benefits of so intending; they are not at risk. So when she takes these benefits to be reasons for following through, then her deliberation is essentially backward-looking, and thus not truly directed at the furtherance of her ends. If her deliberation were truly so directed, then what would matter to her would be the fact that following through is now just a deadweight loss, and she would see no reason to follow through.

6.3 An objection that naturally arises here is that we must have just stumbled into a misunderstanding of what it means for an agent's deliberation to be directed at the furtherance of her ends. For if we think that deliberation so directed cannot be backward-looking, then we are led to think that deliberation so directed is simply that of the straightforward maximizer; and we acknowledged, at the end of section 6.1, that this is not what Gauthier means for his notion of directedness to entail. After all, Gauthier proposes the directed-effectiveness theory precisely so that, in giving up the maximal-effectiveness theory, he is not thereby driven to retreat to the orthodox theory. To see this clearly, it may be helpful to imagine the theories arrayed along a spectrum. At the poles would be the orthodox and maximal-effectiveness theories, representing exclusive concern with directedness and with effectiveness, respectively. The directed-effectiveness theory would lie somewhere near the middle, since it originates in a concern with effectiveness but is tempered by some concessions to directedness. With this picture in mind, it is clear that in retreating from the maximal-effectiveness theory, Gauthier does not mean to retreat all the way back to the orthodox theory, but to stop well short of that. And so it may appear obviously erroneous to claim, as I did in section 6.2, that deliberation directed at the furtherance of an agent's ends is necessarily that of the straightforward maximizer.

The objection, then, is that the discussion of section 6.2 goes off track somewhere, because its understanding of Gauthier's notion of directedness ascribes to that notion more than Gauthier has in mind when he proposes it. But this objection presupposes that there is some coherent notion of directedness that entails what Gauthier wants it to entail (the rejection of tyrant-placating deliberation) while not entailing what he does not want it to entail (a complete retreat to straightforward maximization). And the discussion of section 6.2 can be read as questioning this presupposition: it can be read as claiming that there is no coherent notion of directedness that entails giving up the maximal-effectiveness theory of rationality without entailing a commitment to the orthodox theory. I claim, then, that if Gauthier wants to retreat from the maximal-effectiveness position, then there is no principled stopping point short of the orthodox position. The compromise position that Gauthier seeks to occupy is just conceptually unavailable.

What follows from all this is that Gauthier's claim that rational deliberation must be directed at the furtherance of the agent's ends inevitably pushes him all the way back to the orthodox theory and, thereby, to straightforward maximization. Of course, it is open to Gauthier to recant this claim, so that he can continue to reject straightforward maximization in favor of constrained maximization. But then he must allow that rationality is, in fact, simply a matter of what pays—not only in the case of the toxin, but also in the case of the tyrant.