RATIONALITY AND THE PRISONER’S DILEMMA IN DAVID GAUTHIER’S *MORALS BY AGREEMENT*

In *Morals by Agreement*, David Gauthier\(^1\) defends "the traditional conception of morality as a *rational constraint* on the pursuit of individual interest" (MA, 2; my italics). His defense is set against the background of a prominent view of rational conduct, the view that Gauthier calls "the utility maximizing conception of rationality" (MA, 6f). The maximizing conception in its standard interpretation holds that

an act of a person \(p\) is rational if and only if it is expected to maximize \(p\)'s own interest or preferences.\(^2\)

This principle codifies the idea of immediate connection between rationality and the direct pursuit of individual interest—the idea that "there is a connection between [practical] reason and interest—or advantage, benefit, preference, satisfaction, or individual utility" (MA, 6).

The traditional conception of morality Gauthier attempts to defend is clearly at odds with the maximizing view of rationality. The conception of morality does not deny the apparent fact that morality may and often times does conflict with the direct pursuit of individual interest; and it holds that in some cases of such conflicts it is rational to comply with moral commands at the cost of one's own interest or preferences. Thus, its defense requires justifying why it is rational at all to comply with moral commands even when they require sacrificing one's own interest.

Yet Gauthier wants to base his argument on the maximizing conception of rationality itself. He somehow thinks that the maximizing conception, in one version or another, is "the only one capable of critical examination" (MA, vi). Whether or not one agrees with this, his attempt, if successful, would result in a strong and desirable defense of morality. The opponents of (the rationality of) morality typically adhere to the maximizing conception;\(^3\) so the justification

---

\(^*\) I am grateful to David Gauthier, Gregory Kavka, and Philippa Foot for their comments on the earlier versions of this paper.


\(^2\) Cf. MA, 6: "a person acts rationally if and only if she seeks her greatest interest or benefit."

\(^3\) In particular, I take Cauthier’s opponent, Hobbes’s Foole (cf. fn. 8), as an avowed promulgator of what he takes to be the correct implication of rationality as maximizing of individual interest.
of morality in terms of their own conception of rationality, if available, will turn out to be the most conclusive argument against them. The strategy of his argument might be best explained with an analysis of the prisoner’s dilemma. Recall a standard description of the (two-person) prisoner’s dilemma, the possible outcomes of which are summarized in the following table, where numbers indicate the years the two suspects, he and I, will live in jail:

<table>
<thead>
<tr>
<th></th>
<th>Confess</th>
<th>Not</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>(5, 5)</td>
<td>(1, 10)</td>
</tr>
<tr>
<td>Not</td>
<td>(10, 1)</td>
<td>(2, 2)</td>
</tr>
</tbody>
</table>

Table 1

In this situation, adherence to the maximizing rationality by both agents—i.e., their unrestrained pursuits of their respective individual interests—yields them a suboptimal outcome: 5 years in jail for each. If, however, they can effectively cooperate, they can reap a better outcome of two years in jail for each, an outcome which is fair and more advantageous to each than that of their individually “rational” actions.

Gauthier’s approach is to view morality as a code agreed upon to enable a fair and mutually advantageous social cooperation, which, as the prisoner’s dilemma is apt to show, may sometimes be at variance with the direct pursuit of individual interest. To justify the constraint that the contractarian morality imposes on the direct pursuit of individual interest, Gauthier argues that when the agents in a situation characterized by “a conflict as well as identity of interests” (ibid., p. 126) (such as the prisoner’s dilemma) attempt to reach the optimal cooperation outcome on the basis of the maximizing conception of rationality, they find it rational (in the maximizing sense) to agree not to confess. Furthermore, he argues, they find it rational to form a disposition to abide by the agreement so made—given certain provisos. That is, straightforward pursuers of individual interests (“straightforward maximizers”) have the reason (by their own standards) to change into “constrained maximizers.”

---

namely, those who restrain the pursuits of their respective interests in terms of, say, the contractarian morality.

In order to elucidate the nature of his argument, I shall follow Gauthier in concentrating on the analysis of what it is to act rationally in the prisoner’s dilemma. There I shall attempt to show why the rationality of (the act of) forming a disposition is one thing and the rationality of acting in conformity with the disposition quite another. Thus, I argue, in justifying the rationality of complying with morality viewed as a code enabling social cooperation, Gauthier has as yet to vindicate the principle connecting the two:

If it is rational to be disposed to restraint in the direct pursuit of individual interest, then it is rational to act in accordance with the disposition.

Finally, reflecting on the implications of adopting this principle, I shall argue that Gauthier’s justification, properly understood, turns out to presuppose, rather than to derive, that there exists an implicit constraint on the direct pursuit of individual interest in what he should argue is the proper interpretation of “the maximizing conception” of rationality.

I

Stage I. Recall table 1 of payoffs proposed to two prisoners, he and I. Given the maximizing conception (in its standard interpretation), it is rational for me to confess under that circumstance. In this context it is convenient to talk in terms of the expected years during which I shall live in jail depending on whether I choose to confess or not—assuming that the longer the expected years, the less the expected utility (alternatively, the less it is in my interest or the less I prefer it). Now, depending on whether or not I confess, the measure, Y, of expected years in jail may be calculated as follows:

\[
Y \text{(confess)} = 5 \times p + 1 \times (1 - p)
\]

\[
Y \text{(not confess)} = 10 \times p + 2 \times (1 - p)
\]

where \( p \) (0 ≤ p ≤) is the probability that I initially assign to the chance that he will confess. Now, Y (not confess) is greater than Y (confess) whatever the initial probability, \( p \), of his confessing may be;\(^5\) for 10 × \( p \) is greater than 5 × \( p \) and 2 × (1 − \( p \)) is greater than 1 − \( p \). That is, confessing is rational for me.

I also realize that the situation is symmetrical between him and me; given that he is a rational agent with a perfect power of reason-

\(^5\) One may reason that whether or not he is assumed to confess I shall be better off if I confess. This reasoning takes into account the two limiting cases when \( p \) is respectively 0 and 1.
ing, it is also clear to me that he, too, will choose rationally to confess. Consequently, my reasoning continues, I am doomed to live 5 years in jail; that is, in this situation the measure of the expected years in prison is 5 when everything, including his rational action, is taken into account. Thus, I naturally want to change the situation so as to make him not to confess; and it is the same with him. But, alas, that option is not available to us; for we cannot coordinate our actions by way of, say, communication.

Stage II. Suppose now that the situation has slightly changed; that is, we are allowed to communicate once. In the meeting I initially try to dissuade him from confessing, but I shortly find out that I can dissuade him only if I also promise not to confess and vice versa. Consequently, we make an agreement not to confess. Assuming, following Gauthier (MA, 156), that the bargaining is cost-free, it seems rational to make such an agreement; for it renders my expected utility at least as high as otherwise.\(^6\) In fact, the actual discussion might not have been essential. Suppose that, fortunately, we previously read Gauthier’s book together and that, therefore, we were convinced of, say, Gauthier’s principle of “minimax relative concession,” which specifies terms of a fair and mutually advantageous cooperation.\(^7\) Furthermore, each of us knows that the other has the same understanding. Then I can conclude that we would have agreed not to confess were there a chance to discuss. And I, knowing his acumen, can be certain that he should be reasoning in the same way. In sum, it is exactly as if we made an actual agreement; so we might rely on the hypothetical agreement.

Does the agreement, hypothetical or actual, make a substantial difference to the situation? Here Hobbes’s Foole\(^8\) enters into the stage, arguing that it makes no essential difference to the situation. Given the maximizing conception, the Foole argues, the agreement falls short of introducing an intrinsic reason to comply with the

---

\(^6\) I shall argue that in the end the agreement fails to upgrade my expected utility. Thus, as Kavka correctly noted, I as a perfectly rational agent may not find it the only rational action to make the agreement since I can see that it is pointless. (Insofar as maximizing rationality is concerned, on the other hand, it is not irrational, either, to make the agreement given that the bargaining is cost-free.) Yet the major point in this section remains valid: even given the agreement, it is still rational, in the maximizing sense, for me to confess.

\(^7\) For Gauthier’s account of the principle of minimax relative concession, see MA, 141–5. In this context, however, the specific content of the principle is not crucial. What is important is that we both recognize the optimal outcome as the one based on a/the fair and mutually advantageous principle.

\(^8\) See Hobbes, Leviathan, C. B. Macpherson, ed. (New York: Penguin, 1968), ch. 15, p. 203: “The Foole . . . does not therein deny, that there be Covenants; . . . but he questioneth, whether Injustice . . . may not sometimes stand with that Reason, which dictates to every man his own good.”
agreement. Consequently, he continues, the agreement is not sufficient to secure the cooperatoral benefits to us, although the agreement seems rational to make and is necessary for cooperation; for I still have a choice between complying with and deserting the agreement. Here is the table of possible outcomes:

<table>
<thead>
<tr>
<th></th>
<th>He</th>
<th>Not</th>
<th>Comply</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>(5, 5)</td>
<td>(1, 10)</td>
</tr>
<tr>
<td></td>
<td>Comply</td>
<td>(10, 1)</td>
<td>(2, 2)</td>
</tr>
</tbody>
</table>

Table 2

It shows that I am facing the same situation as the original dilemma; so I find it rational not to comply with the agreement, i.e., rational to confess. The agreement turns out to be of no avail, although it initially seems rational to make. The Foole, as Gauthier notes, is challenging “what would seem to be the ordinary view that given neither unforseen circumstances nor misrepresentation of terms, it is rational to comply with an agreement if it is rational to make it” (MA, 165; my emphasis).

Stage III. My dilemma persisting in stage II lies in the fact that I did not succeed in making him not confess even by reaching an agreement with him; consequently, the agreement fails to upgrade my expectation. What is needed is not making an agreement but ensuring his compliance.

Suppose now that we are granted another chance to meet. Before the meeting I am considering how to make it certain that he will comply. To be sure, the best result of the meeting would be to ensure his unilateral compliance regardless of my action, but this is simply impossible. So I look for a way of achieving the outcome that is the second best result for me: making it certain that he will comply by way of ensuring him of my own compliance. Otherwise, I have no way to shorten 5 years in jail. The way to ascertain my compliance is to form and make him check, with absolute certainty, my disposition to constrained maximization, i.e., the disposition to comply with an agreement based on the fair and mutually advantageous principle in case I am certain that he will also comply with it.9 He reaches the

9 This is a simplified version of the disposition to constrained maximization, the version which I think is adequate for the simplified situation at issue here. Cf. MA, 167f.
same conclusion on his own, and both of us succeed in forming the requisite disposition. In the meeting we "see" each other's disposition, and reach again or confirm the agreement not to confess.

Now I can reliably expect to gain the benefit of 2 years in jail. By forming the disposition, I have succeeded in lowering the ultimate measure, $Y$, of expected years in jail to 2.

II

In stage III, the dilemma facing the prisoners is given a practical solution; they succeed in reaping the benefit of cooperation through an agreement backed up by the disposition to constrained maximization. And it is rational in the maximizing sense to be so disposed in such a situation. Yet, I argue, the problem of justifying the rationality of acting as the disposition leads one to is far from being solved.

In my analysis of the solution of the dilemma in stage III, it is not necessary at all to assume that it is rational for him or "me" to act in accordance with the disposition to constrained maximization; there it is sufficient to observe that the prisoners (he and "I") are now reliably expected to act in accordance with the disposition—that is, not to confess—because the disposition is firmly entrenched. In their ("our") own deliberation, too, their expectations about the other's action are entirely based on the available information concerning the other's dispositions. Although it is argued that it is rational to form the disposition to constrained maximization, no appeal needs to be made to the rationality of acting in accordance with the disposition. In sum, the rationality of internalizing a fair and mutually advantageous principle of cooperation is not yet shown to imply that it is rational to act in accordance with it.

Thus, Hobbes's Foole, the staunchest promulgator of the maximizing conception (in its standard interpretation), will come back to the stage and continue to argue against Gauthier's defense of morality as follows. Consider again, the Foole argues, the situation where we have found each other's disposition to constrained maximization with absolute certainty. I, the Foole, grant that it was rational to be so disposed. But the question remains whether it is rational to act in accordance with the disposition—to be sure, granting that the agreement is fair$^{10}$ and mutually advantageous, and that neither unforeseen circumstances nor misrepresentation of terms is found, and that "the rationale for [the disposition] does not appeal to any weakness or imperfection in [our] reasoning" (MA, 186).$^{11}$

---

$^{10}$ I understand this as including fairness of the initial position of the agreement, which Gauthier argues is ensured by satisfaction of what he calls "the Lockean proviso."

$^{11}$ All these assumptions are satisfied by description of the situation itself.
There are indeed two kinds of differences between the present situation (i.e., stage III) and the dilemma in the stage II. On the one hand, there are differences based on the fact that I am myself disposed to constrained maximization; on the other hand, those deriving from the fact that I now have different expectations about the other’s action because I know that he has a firm disposition to constrained maximization. The first kind of differences, I argue, do not count when the rationality of the conforming action is at issue. The rationale for this consists in the distinction between questions concerning rationality and those concerning motivational strength. I, the Foole, concede that by forming the disposition I made myself strongly disposed not to confess in the present situation. Consequently, I know that I will not confess. But this is a far cry from saying it is rational to act so.

Imagine that I am reflecting on the situation several years later after not confessing: I came out of the jail after 2 years with complacence in the successful cooperation. At this point, there is no longer any doubt about which motivation was the strongest when I was deciding whether or not to confess. But the question still remains whether it was rational (by the standard of that time) to act in accordance with the disposition, and the answer is not automatically affirmative. We often meet with another kind of situation in which the distinction is typically made between questions concerning rationality and those concerning motivational strength. The fact that I failed to act in a certain way is not automatically regarded as implying that it was irrational for me to act in that way; the discrepancy is often explained with an appeal to what is called “weakness of the will.” If this is a real phenomenon, why should motivational strength of a certain disposition directly imply the rationality of acting in accordance with it, even granting that the disposition itself was formed on a rational basis?

Then what is rational for me to do in stage III? I, the Foole, again suggest a simple calculation of the expected utilities. Look at table 3 below:

<table>
<thead>
<tr>
<th></th>
<th>He</th>
<th>Not</th>
<th>Comform</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Not</td>
<td>(5, 5)</td>
<td>(1, 10)</td>
</tr>
<tr>
<td></td>
<td>Conform</td>
<td>(10, 1)</td>
<td>(2, 2)</td>
</tr>
</tbody>
</table>

Table 3
Table 3 is exactly like table 2, but in calculating expected utilities I may also take it into account that I now expect him to act in conformity with the disposition. Consequently, only the second column in table 3 really matters. Here are the measures, $Y$, of the expected years in jail, depending on whether I also act in conformity with the disposition or not:

$$Y\ (\text{Not conforming}) = 1$$
$$Y\ (\text{Conforming}) = 2$$

This shows that not conforming, i.e., confessing, is rational in spite of my entrenched disposition not to confess.

To emphasize again, this does not mean that I shall not act in conformity with the disposition. Rather, I take this as a situation in which I absolutely feel like not confessing, although I am aware that it is not rational. That is exactly what he, the other party, wanted to make certain; it did not concern him at all whether I would find it rational to conform to the disposition once it was made absolutely certain that I would conform.

III

I presented the Foole's argument as consisting of two parts. In the first part, he argues that the following principle is yet to be justified:

1. Connecting principle: (given that typical provisos\textsuperscript{12} are met) if it is rational to be disposed to constrained maximization, then it is rational to act in conformity with the disposition.

The latter part consists in persistently applying the maximizing conception of rationality to the action of conforming to the disposition, which leads to the conclusion that confessing is still rational. To be sure, Gauthier agrees with neither of these. I shall begin by discussing his disagreement with the first point. The discussion will eventually lead to his rejection of the second point.

\textsuperscript{12} For some of the provisos, see above.

\textsuperscript{13} I understand Gauthier's perfect actor as the actor who is without "any weakness or imperfection in the reasoning." Note that the Foole's argument for the first point takes no advantage of such imperfections. On the other hand, if he means by a perfect actor a person who chooses to do something if and only if it is rational to do so, then he begs the question by assuming that the perfect actor will choose to act in conformity with the rationally grounded disposition.

Gauthier might respond that in his analysis of the prisoner's dilemma the agents are assumed to be perfect actors in the latter, stronger sense. It is indeed true that the "straightforward maximizers" are perfect actors, given the maximizing conception. Yet the issue is whether they remain to be so when they change themselves into "constrained maximizers." This is not ensured by the rationality of the metamorphosing actions. The Foole's straightforward answer, to be sure, is in the negative.
He affirms that "if her [the perfect actor's] dispositions to choose are rational, then surely her choices are also rational" (MA, 186).\textsuperscript{13} By this he seems to endorse principle (1). But the view itself needs to be justified, and the justification should be based on (some interpretation of) the maximizing conception of rationality; otherwise, he has yet to make his case that he is giving "a utility-maximizing rationale" (MA, 158) for compliance. Principle (1) might be granted to be part of our common conception of rationality, but that is another matter.

Imagine that principle (1) might be justified by the maximizing conception of rationality as it is formulated in the beginning of this paper:\textsuperscript{14}

2. \textit{Action-maximizing principle}: an act is rational if and only if it is expected to maximize the agent's own interest.

Or, suppose that it is justified as a kind of conceptual truth linking the concept of the rationality of a disposition to that of the rationality of conforming action or that it is justified, independently of principle (2), as reflecting a fixed point in our intuitions about the concept of rationality. Then it is rather worse than better for principle (2) because it results in a \textit{reductio} of (2): by utilizing Gauthier's argument for the rationality of being disposed to constrained maximization and the implication of the principle (2) drawn by the Foole, one can show that, given relevant assumptions built into the description of the prisoner's dilemma, principle (2) together with principle (1) leads to a contradiction. Let me explain.

Here is the conclusion of Gauthier's argument, which I grant: in the prisoner's dilemma,

3. It is rational to be disposed to constrained maximization.

(1) and (3) imply, as is intended, that

4. It is rational to act in conformity with the disposition to constrained maximization.

Now, apply principle (2) to the stage III quasi dilemma. Then, as the Foole is apt to show, it follows that:

5. It is rational not to act in conformity with the disposition to constrained maximization.

and that:

6. It is not rational to act in conformity with the disposition.

\textsuperscript{14} It was based on Gauthier's formulation in MA, ch. 1. Cf. fn. 2.
which is due to the "only if" part of principle (2). (6) contradicts (4). The prisoner's dilemma, it turns out, helps to elucidate that principles (1) and (2), each of which taken independently seems adequate, lead to a contradiction when applied to the realistic situation marked by "conflict as well as identity of interests."

Gauthier may be anticipating this problem when he says that he is working with a different interpretation of the maximizing conception, which "identifies practical rationality with utility-maximization at the level of dispositions to choose." And he "carries through the implication of that identification in assessing the rationality of particular choices" (MA, 187). Thus, while rejecting (2) he might be countenancing the following principle:

7. **Principle of disposition maximizing**: a disposition is rational for a person \( p \) to have if and only if it is expected to maximize \( p \)'s own interest.

In addition, he certainly needs principle (1), whatever its justification may be, to show the rationality of acting in conformity with the disposition to constrained maximization.

It is quite obvious that principle (1) should be independent of the disposition-maximizing principle since this has nothing whatsoever to say about the rationality of particular actions. Gauthier might continue that it is so much the better for his disposition-maximizing conception of rationality. Principle (1) may be simply added as an independent principle to the conception. One might motivate the move in the following way: we begin by reflecting on the maximizing conception in its inarticulate form, and arrive at a correct formulation of it; and then we can appropriately extend that conception to reach a coherent and comprehensive concept of rationality. Now think about the present situation: the Foole's persistent argument makes it clear that principle (1) needs to be incorporated into the concept of rationality to launch Gauthier's justification of morality, and it turns out that adding it to the disposition-maximizing principle can have no bad consequences; then, it is rather mandatory to add it (or some other principle that implies it). Or so he might argue. This might exemplify a respectable methodology, but it is still worth noting that he is, in effect, implicitly working with at least two independent principles of practical rationality to show the rationality of morality; he no longer undertakes to derive the one from the other.

Finally, insofar as he wants to endorse the "maximizing conception" in its correct interpretation, he will certainly agree to adding the following principle:
8. An action is rational if it is expected to maximize the agent's interest, on the condition that it does not conflict with an action that is ruled to be rational by virtue of principles (1) and (7).

To prepare for the concluding remarks, let me present the ultimate conception of rationality, which I have just reached, as consisting of the following principles:

I. A disposition is rational for \( p \) to have if and only if it is expected to maximize \( p \)'s own interest.
II. If it is rational to be disposed to constrained maximization, then it is rational to act in conformity with the disposition.
III. An action is rational if it is expected to maximize the agent's interest.
IV. In case of conflict, rulings based on principle (II) together with principle (I) take precedence over those based on principle (III).

I am now in a position to look back to assess how far Gauthier should have come away from the naive understanding of the maximizing conception of rationality in the attempt to justify the rationality of following the commands of morality. Let me quote Gauthier:

But Hobbes . . . must revise his conception of rationality, breaking the direct connection between reason and benefit with which he began to reply. Hobbes needs to say that it is rational to perform one's covenant even when performance is not directly to one's benefit, provided that it is to one's benefit to be disposed to perform. But this he never says. And as long as the Foole is allowed to relate reason directly to benefit in performance, rather than benefit in the dispositions to perform, he can escape the refutation (MA, 162; my emphasis).

This more than adequately makes my point: when one introduces the principle of disposition maximizing while rejecting the principle of action maximizing, one should be breaking "the direct connection between reason and interest." I made a further point that may have also been implicit in Gauthier's discussions: rejection of the principle of action maximizing in its full scope requires the introduction of the principle governing the connection between the rationality of dispositions and the rationality of actions as an independent principle; the latter can not be a consequence of the core of the maximizing conception.

Suppose now that adding principle (1) is acceptable to Gauthier. Then the previous point about the connection between reason and individual interest needs an elucidation; for it is not merely that the direct connection from individual interest to rationality is not available in an adequate conception of rationality. Rather, the correct picture may be that a constraint to "the direct pursuit of individual
"interest" is found to be embedded in the concept of rationality itself, however tenuous its force may appear before serious deliberation. The constraint lies in restricting the application of the (weakened) action-maximizing principle (III) in terms of the connecting principle (II) and, perhaps more explicitly, the precedence principle (IV).

If this is a correct analysis, I may come to another elucidation of the nature of Gauthier's moral contractarianism. Again, let me begin by quoting some passages in MA, chapter 1, where Gauthier presents an overview of his theory:

Morals by agreement begin from an initial presumption against morality, as a constraint on each person's pursuit of her own interest (8f). Let us suppose it agreed that there is a connection between reason and interest (8). Let it be further agreed that in so far as the interests of others are not affected, a person acts rationally if and only if she seeks her greatest interest or benefit (6f).

These statements indicate that Gauthier believes that a constraint on the direct pursuit of individual interest, which is absent in the concept of rationality in itself, can be derived when the direct pursuits need to be compromised in view of the complexity of the situation in which we are put. I grant that I might be wrong in diagnosing Gauthier's own view of his project; for instance, the third passage might have been introduced just to explain the action-maximizing conception of rationality which Gauthier rejects in due course. Still, it may be worth noting that the project as I possibly mistakenly understand it does not necessarily conform to what is actually going on in his contractarian justification of morality. What the contractarian moral theory shows may be that what initially does not look like much of a constraint turns out to be substantial and powerful when applied to so complex a situation as we are put in, a situation in which identity and conflict of interests are hard to sort out. If this understanding is correct, I doubt it can be significantly claimed that "[m]orals by agreement begin from an initial presumption against morality" (MA, 8f).

BYEONG-UK YI

University of California/Los Angeles