CANADIAN JOURNAL OF PHILOSOPHY
Volume 16, Number 1, March 1986, pp. 1-10

Self-Subverting Principles of Choice

MICHAEL PERKINS
4348 Malin Street
Columbus, OH 43224

DONALD C. HUBIN
Department of Philosophy
The Ohio State University
Columbus, OH 43210-1365

The thesis that rationality consists in the straight-forward maximization of utility has not lacked critics. Typically, however, detractors reject the Humean picture of rationality upon which it seems based; they seek to emancipate reason from the tyranny of the passions. It is, then, noteworthy when an attack on this thesis comes from 'within the ranks.'

David Gauthier's paper 'Reason and Maximization' (1975) is just such an attack; and for this reason, among others, it is interesting. It is not successful, though. In defense of this conclusion, we shall begin by relating the essentials of Gauthier's argument. Then we shall examine in some detail Gauthier's claim that the principle of straighforward max-

imization fails to be self-supporting. We shall argue that Gauthier's defense of this claim is at best incomplete. Finally, we shall show that the fact that a normative principle is self-subverting or non-self-supporting does not entail that the principle is defective.

I

Gauthier's argument is based on the claim that the conception of rationality which requires individual utility maximization in all circumstances is not self-supporting. A conception of rationality, R, is self-supporting if, and only if, R entails that one ought (rationally) to choose R as one's conception of rationality.

Gauthier attempts to demonstrate this claim by considering the situation in which one is to choose between the two conceptions of rationality characterized by the following conditions:

> Condition of Straightforward Maximization (CSM):
>
> > A person acts rationally only if the expected outcome of his action affords him a utility at least as great as that of the expected outcome of any action possible for him in the situation (Gauthier, 418).
>
> Condition of Constrained Maximization (CCM):
>
> > A person acting interdependently acts rationally only if the expected outcome of his action affords each person with whom his action is interdependent a utility such that there is no combination of possible actions, one for each person acting interdependently, with an expected outcome which affords each person other than himself at least as great a utility, and himself a greater utility (Gauthier, 427).

Interdependent action is 'action in a manner on which all agree' (Gauthier, 424). Thus, while CSM requires unrestricted utility maximization, CCM prohibits a person from violating an agreement in order to increase his utility if by doing so he diminishes the utility of those with whom he has the agreement.[1] Gauthier argues that in a situation involv-

---

1 Strictly speaking, CSM and CCM offer only necessary conditions for rational action. By themselves, then, they do not rationally require any action. However, on the assumption that, in each of the cases considered, there *is* some rational action, CSM and CCM will present rational requirements. We shall make this assumption throughout; we think that it is evident that Gauthier does so as well.

ing a choice between the two conceptions of rationality characterized by CSM and CCM, CSM will recommend that one choose the conception characterized by CCM. Consequently, CSM is not self-supporting. It should be noted that Gauthier's argument, if successful, would establish not only this conclusion, but also that CSM is self-subverting. A principle R is self-subverting if, and only if, R entails that one ought *not* (rationally) to choose R as one's conception of rationality. Self-subversion seems a more serious charge against a principle of choice than does failure to be self-supporting. A principle may fail to recommend itself because it restricts itself to cases which do not involve choices between principles of choice. But, it seems, if the principle does give an answer to the question of what principle of choice one should adopt, the answer should not be that one should not adopt that principle itself. It is one thing to go to an advisor and be told that she doesn't give advice about whom to choose as an advisor. It is quite another for her to tell you not to choose her as an advisor. Should you follow her advice or not?

Gauthier defends the claim that CSM recommends that one choose the conception of rationality characterized by CCM by arguing that proponents of CCM do better in prisoner's dilemma situations than do proponents of CSM; in all others they do the same.

> We find that in all those situations in which individual utility-maximization leads to an optimal outcome, the expected utility of each is the same, but in those situations in which individual utility-maximization does not lead to an optimal outcome, the expected utility of straightforward maximization is less. In these latter situations, a constrained maximizer, but not a straightforward maximizer, can enter rationally into an agreement to act to bring about an optimal outcome which affords each party to the agreement a utility greater than he would attain acting independently (Gauthier, 429).

Of course, how well the proponent of CCM does in these situations will depend on the actions performed by those individuals with whom he is interacting. Gauthier claims that a constrained maximizer is committed to carrying out an agreement which has been reached only in the context of 'mutual expectations on the part of all parties to the agreement that it will be carried out' (Gauthier, 429).

> Nevertheless, since the constrained maximizer has in some circumstances some probability of being able to enter into and carry out, an agreement, whereas the straightforward maximizer has no such probability, the expected utility of the constrained maximizer is greater (Gauthier, 430).

Thus, according to Gauthier, CSM requires that a rational person adopt CCM.

II

Let us now look closely at those situations in which, according to Gauthier, the constrained maximizer does better than the straightforward maximizer. Consider the following prisoner's dilemma situation:

	b_1	b_2
a_1	$O_{11}(1,1)$	$O_{12}(10, 0)$
a_2	$O_{21}(0,10)$	$O_{22}(9,9)$

Since the constrained maximizer performs the same action as the straightforward maximizer if no agreement is reached,[2] let us suppose that A and B agree to perform a_2 *and* b_2 respectively, and that A has good expectations that B will keep the agreement. Thus, the principle of constrained maximization will require that A perform a_2.

Note that it does not follow that the principle of straightforward maximization requires that A perform a_1. Suppose that A's actions are related to B's actions in a way reflected by the following probability matrix:

	b_1	b_2
a_1	.99	.01
a_2	.01	.99

In such a case, CSM will prescribe that A perform a_2.

$$EU(a_1) = (.99)(1) + (.01)(10) = 1.09$$

$$EU(a_2) = (.01)(0) + (.99)(9) = 8.91$$

Thus,

$$EU(a_2) > EU(a_1).$$

---

2 '[T]o act independently is to act interdependently with oneself alone' (Gauthier, 427). Thus, independent action becomes a special case of interdependent action with the result that when one is acting independently, CSM and CCM are extensionally equivalent.

Let us assume, then, that B's actions are not related to A's in such a way that CSM requires A to perform a_2.

Now we have a case in which individual utility maximization does not lead to an optimal outcome (O_{22}), but there is some chance that constrained maximization will lead to an optimal outcome. Gauthier's claim is that in situations like this, the expected utility of constrained maximization is greater than the expected utility of individual maximization.

It should be clear, however, that the utilities afforded by a choice of each principle are reflected in the following matrix:

	b_1	$b2$
A chooses CSM	$O_{11}(1,1)$	$O_{12}(10,0)$
A chooses CCM	$O_{21}(0,10)$	$O_{22}(9,9)$

Ex hypothesi, if A chooses CSM, he will perform a_1; and if A chooses CCM, he will perform a_2. Assuming that the utility of choosing a conception of rationality is a function solely of the utility produced by the actions required by that conception of rationality,[3] we can take the utilities for such choices from our original specification of the game.

Now, either B's action is causally independent of A's choice of principle or it is not. Suppose that it is. Then a corollary of CSM, the principle of dominance with causal independence, prescribes that A choose CSM. (For an account of the principle of dominance with causal independence and its relation to what we are calling CSM, see Gibbard and Harper [1978].) As the constrained maximizer's chances of securing an optimal outcome (O_{22}) increase, the individual maximizer's chances of securing an even better outcome (O_{12}) increase. Insofar as the constrained maximizer is able to secure O_{22}, the straightforward maximizer is *not* condemned to O_{11}. Thus, it is not the case that the expected utility of choosing CCM is greater than that of choosing CSM.

It follows, then, that the only cases in which the constrained maximizer has a greater expected utility are cases in which the actions of the

---

3 Gauthier must be making this assumption. If this assumption is not made, Gauthier's argument for the superiority of CCM over CSM fails. It is possible that someone attaches very high utility to choosing CSM. For such a person, selecting CSM may well maximize expected utility regardless of the sorts of considerations Gauthier adduces against it.

person with whom he is interacting are causally dependent on his choice of principle. In particular, the constrained maximizer has greater expected utility just in case the probability matrix looks something like the following:

	$b1$	$b2$
A chooses CSM	.99	.01
A chooses CCM	.01	.99

Then:

$$EU(CSM) = (.99)(1) + (.01)(10) = 1.09$$

$$EU(CCM) = (.01)(0) + (.99)(9) = 8.91$$

Thus,

$$EU(CCM) > EU(CSM).$$

Now, what do these result show us? First note that when we speak of self-support or self-subversion, we must qualify our claim. Both notions are decision-theoretic – they are defined in terms of what it would be rational to choose. As such, they must be relativized to a situation. To ask whether a principle of choice is self-supporting or self-subverting without qualification is like asking whether it is better to draw for a straight or for a flush. It all depends on the situation you are in. Our results above, then leave us with the following position: In situations where there exists a causal relationship which yields a probability matrix like that above, CSM is self-subverting; in situations where there is causal independence between A's choice of a rational principle and B's action (or if A's choosing CSM renders it more likely that B will perform b_2), CSM is self-supporting.

Gauthier's claim may be that CSM is self-subverting in some possible choice situation. If so, then his argument is sound. But, as we shall show in Section III, all teleological normative principles *including CCM* are self-subverting in some possible situation. (This claim is too strong. The appropriate qualifications are made in Section III. See, especially, footnote 5. The weakening of the claim does not harm the argument we present in that section.) The more interesting claim is that CSM is self-

subverting in the choice situation in which humans actually find themselves. This appears to be the claim that Gauthier is trying to demonstrate. If it is, however, his argument is incomplete. He needs to give us some reason to believe that there is a causal relationship of the relevant kind between a person's choice of a rational principle and the actions of others.[4] He has not done so.

III

Gauthier has failed to show that in the situation in which we find ourselves CSM is self-subverting (or even non-self-supporting). But let us suppose that it is. What are we to make of this? One might think that a conception of rationality which is not self-supporting is, eo ipso, defective. Judgments of deficiency seem even more plausible when the conception in question is self-subverting. There is an apparent analogy here with the self-subversion of the verification theory of meaning. The verification theory of meaning, if stated as a necessary condition for the meaningfulness of any sentence, seems to entail that the theory itself is meaningless and, hence, not true. Since a necessary condition for such a theory to be true is that it not be true, it cannot be true.

But such an analogy is only apparent. The correctness of CSM does not imply that CSM is incorrect – even in the circumstances in which

---

4 At present we can only speculate about what sort of causal relationship someone might allege obtains here. We have not been able to construct a particularly plausible causal hypothesis of the requisite sort. Perhaps the most plausible hypothesis is that over time the straightforward maximizer's chosen principle of rationality will be revealed in his behavior and that other people will refuse to cooperate with a person they believe to be a straightforward maximizer. Note that the only way in which a straightforward maximizer's behavior will differ from that of a constrained maximizer will be in the keeping or breaking of agreements designed to secure optimal outcomes in prisoner's dilemma situations. The straightforward miximizer, however, may have a number of reasons to keep such agreements. He may attach intrinsic utility to keeping agreements; or he may want to foster a reputation of trustworthiness so that he will be able to engage in cooperative activities in the future. Note, moreover, that insofar as a constrained miximizer has good reasons to believe that the straightforward maximizer with whom he is acting interdependently will keep the agreement, CCM requires that the constrained miximizer make and keep the agreement.

CSM is self-subverting. In order to see this, consider a case which is structurally analogous. Utilitarians have often wondered whether espousing utilitarianism is utilitarian. Given the difficulties people have in applying the theory to specific cases, it may well produce greater utility to advise people to follow some other ethical theory. If their belief that utilitarianism is true hampers their ability to follow this advice, it may be one's utilitarian duty to convince others that utilitarianism is false. Thus, if utilitarianism is a correct conception of morality, it may be our duty not to espouse it. But notice that a parallel argument could be made of the act of *choosing* utilitarianism as one's conception of morality. If, for example, acceptance of the principle of utility leads people to attempt to calculate utilities of actions and act on those calculations, and if our calculations are less reliable (or lead to more disastrous consequences when faulty) than acting upon another conception of morality, then the principle of utility may require that one not accept it. The principle of utility is, then, self-subverting given certain factual assumptions.

This result strikes some as paradoxical and even more as undesirable; it is neither. It is a consistent and desirable feature of any teleological normative principle.[5] Such principles deem actions (broadly construed so as to include mental acts) as correct in virtue of their tendency to promote some determinate end. So long as the relation between the act of accepting the principle in question and the end is contingent, there will be some conceivable situation is which that act does not promote the end.

Consider, for example, the following doxastic principle:

> D. One ought to believe a proposition if, and only if, doing so maximizes one's expected utility.

Now, imagine that some errant epistemologist, incensed by the growing popularity of such an expedient view, proposes to turn the philosophical tide by direct action. He purchases an I.B.M. Brainstate Scanner (the Delta model which reads dispositional beliefs as well as occurrent ones) and a 44 caliber pistol and begins to search for adherents of D. (Perhaps, as Nozick might say [1981] 4), he is simply trying to give a knockdown argument for his position.) So long as such a philosophical fanatic is on

---

5 Actually, this is too broad. The logical possibility of self-subversion by a teleological normative principle requires two things: (1) that the action of accepting or rejecting the principle falls within the scope of the principle; and, (2) that achieving the end does not logically require the action of accepting the principle.

the loose, it may well be that if D is correct, one ought not to believe that it is.

There is no paradox here. The correctness of D is one matter; the correctness of accepting it, quite another. D is essentially a criterion of what beliefs ought to be held; it is a criterion for us *to accept* only contingently, if at all. With the necessary changes, the same point must be made about other teleological normative principles like the principle of utility and the condition of straightforward maximization.

One might deny that there is any sense to be made of a teleological normative principle being correct over and above its being correct to accept it. But, at least with regard to principles of rational choice, such a position would be difficult to sustain. What we mean by calling a principle of rational choice correct is that all and only actions which in fact conform to it are (objectively) rationally correct. Accepting a principle of choice is itself an action. This action is (objectively) rationally correct if and only if it in fact conforms to a correct principle of rational choice. Thus, if it does not make sense to speak of a principle of rational choice being correct in the sense we do, it is impossible to argue that the action of accepting any principle of rational choice is (objectively) rationally correct.[6]

If there is lingering doubt as to the possibility of a self-subverting conception of rationality (or any other teleological normative theory, for that matter) being correct, consider the following unpleasant turn of events. Suppose that our misguided epistemologist takes an interest in the theory of rational behavior. He is convinced of the truth of CSM and becomes outraged that young philosophers, in their formative years, are being seduced into choosing CCM by what he sees as bad arguments. Armed as before (but with the optional 'choice reader' attachment for his brain scanner), he searches out those who choose CCM and employs his pistol to perform a rather crude form of psychosurgery so that they no longer choose CCM (or anything else, as it happens). In such a situation, it seems, CCM would endorse not choosing CCM.

The point of these examples – which some may, perhaps, find fanciful – is simply this: If we limit our attention to teleological conceptions of rationality in which the summon bonum is only contingently tied to

---

6 For a recent and interesting discussion which explores and employs the distinction between a principle being correct and its being correct to accept it in a different context, see Peter Railton's 'Alienation, Consequentialism, and the Demands of Morality'.

the choice of a conception of rationality, all conceptions of rationality are self-subverting in some choice situations – but this is no indictment of such conceptions of rationality.

IV

As we have argued above, Gauthier has not shown that CSM fails to be self-supporting in our real-life situation. Any demonstration that it does requires establishing a specific causal connection between our choice of a conception of rationality and the actions of others. But Gauthier doesn't provide a reason for believing that there is this connection. If it were to turn out that such a causal connection does, in fact, obtain (and, hence, that CSM is self-subverting in our real-life situation), this would not show that CSM is incorrect. Rather, we would say, it would show that one ought not to accept CSM; and it would show this precisely because CSM *is* correct.[7]

Received November, 1983
Revised May, 1984

References

Gauthier, David, 'Reason and Maximization,' *Canadian Journal of Philosophy*, **4** (1975) 411-33

Harper, William L. and Alan Gibbard, 'Counterfactuals and Two Kinds of Expected Utility,' in C.A. Hooker, et. al., eds. *Foundations and Applications of Decision Theory* (Boston: D. Reidel 1978) 125-62

Nozick, Robert, *Philosophical Explanations* (Cambridge, MA: Harvard University Press 1981)

Railton, Peter, 'Alienation, Consequentialism, and the Demands of Morality,' *Philosophy and Public Affairs*, **13** (1984) 134-71

Sobel, J. Howard, 'Interaction Problems for Utility Maximizers,' *Canadian Journal of Philosophy*, **4** (1975) 677-88

---

7 We are indebted to George Schumm and Daniel Farrell for helpful comments on an earlier draft of this paper.