THE RATIONALITY OF CONDITIONAL COOPERATION

ABSTRACT. In *Morals by Agreement*, David Gauthier (1986) argues that it is rational to intend to cooperate, even in single-play Prisoner's Dilemma games, provided (1) your co-player has a similar intention; (2) both intentions can be revealed to the other player. To this thesis four objections are made. (a) In a strategic decision the parameters on which the argument relies cannot be supposed to be given. (b) Of each pair of a-symmetric intentions at least one is not rational. But it is impossible to form symmetric intentions to cooperate conditionally. For the condition on which the decision depends cannot be fulfilled without deciding. (c) If one's intention has to be ascertained on the basis of information about one's past performance, it is straightforwardly rational to intend to cooperate, but there is no reason to do so in a single-play PD. (d) The argument cannot be extended to n-person games which are Gauthier's principal concern.

1. GAUTHIER'S ARGUMENT

"The jury is . . . still out on the question of whether constrained maximization is rational for individuals to adopt".¹ In this paper I will attempt to reach a comprehensive verdict on the plausibility of this most fundamental thesis of *Morals by Agreement*.

Let me start by briefly summarizing Gauthier's argument as I understand it. Contexts of social interaction can be divided into two types. In the first type of context people aiming at the maximization of their personal utility arrive at outcomes which satisfy the following description: no person can improve his own position by unilateral action, and all persons together cannot improve the positions of each at the same time by adjusting their choices to each other. In other words: all equilibrium outcomes are Pareto-optimal ones as well. This type of context can be identified as the perfectly competitive market. Here rational agents cannot be expected to subject their utility maximizing decisions to any "moral" constraints:² the market is a "morally free zone". But the conditions of perfect competition are not always satisfied; externalities may be present leading to suboptimal equilibrium outcomes. The Prisoner's Dilemma game (PD) is the best-known, but certainly not the only example of this second type of context. Gauthier
wants to argue that in this type of context rational agents trying to maximize their personal utility will have reason to subject their decisions to some constraints resembling moral requirements, provided some conditions are met. Such constrained maximizers will be able to arrive at Pareto-optimal outcomes, even if these are no equilibria. By applying the constraint to her decision each of them will adjust her choice to the choices of the others in such a way that the optimal outcome results. (This criterion determines the content of the constraint.) Such a choice is to be called a cooperative one.

The conditions of constraint are the following:

(a) The cooperative surplus (the collective gains made in exchanging the equilibrium for the optimum outcome) are distributed according to one supremely rational bargaining principle.

(b) The distribution of initial endowments (for market and non-market interaction) also satisfies a unique criterion of rationality.²

The constraint to be accepted is the principle of Conditional Cooperation (CC): be prepared to make the cooperative choice, iff you have reason to expect your co-player(s) to do so as well. What I want to discuss is whether Gauthier succeeds in showing that it is rational to accept this constraint.

His argument can be most easily presented in terms of the one-shot 2-person PD game (or the play of the iterated 2-person game which is mutually known to be the last one), but Gauthier means to extend it to all n-person interaction patterns in which the equilibrium outcomes are sub-optimal. Suppose people are transparent to each other, every one being able to perceive, failinglessly, any other person's intentions. If I intend to play D, you will predict that I will do so, and so, rationally, will choose D yourself. If I intend to play C unconditionally, you will equally find that you have no reason to refrain from choosing D. But if we both intend to play C, provided we know the other to do so as well, you will perceive my intention, and this perception will cause you to choose C. And so we will both improve our position.

Your decision will determine whether I end up with one of my two most or with one of my two least preferred outcomes. It is true that as soon as you have made your decision irrevocably, I will always get the best outcome of the pair you have given me, by defecting. But which