The Journal of Socio-Economics 38 (2009) 549–556

Contents lists available at ScienceDirect

The Journal of Socio-Economics

journal homepage: www.elsevier.com/locate/soceco

Commitment and weakness of will in game theory and neoclassical economics

Roger A. McCain*

Drexel University, Le Bow College of Business, Room 503 "O" Matheson Hall, Philadelphia, PA 19104, United States

ARTICLE INFO

Article history:
Received 21 August 2007
Received in revised form 25 August 2008
Accepted 30 August 2008

JEL classification:
C7
D03

Keywords:
Intertemporal inconsistency
Weakness of will
Cooperative game theory
Noncooperative game theory

ABSTRACT

This paper draws on theoretical and experimental studies in game theory and on the neoclassical concept of intertemporal inconsistency in choice to argue that the motivational theory shared by neoclassical economics and noncooperative game theory is mistaken in assuming that commitment never takes place in human decisions. The paper first gives two parallel examples, one of intertemporal inconsistency in a financial decision, and the other of noncooperative (subgame perfect) equilibrium in a game in extensive form. So far as one decision-maker is concerned, the two decisions are isomorphic, and both can be associated with weakness of will. By contrast, the cooperative analysis of the game (along the lines originally suggested by von Neumann and Morgenstern) predicts a different decision associated with commitment to a particular conditional sequence of "behavior strategies", i.e. a particular pure strategy. In effect the cooperative analysis assumes perfect strength of will. The paper then argues that strength of will and rationality of decisions are independent dimensions of a decision process and reviews some experimental evidence that suggests that both traditions are mistaken in their extreme assumptions about commitment or strength of will: neoclassical economics and noncooperative game theory in assuming that commitment never takes place, and cooperative game theory in assuming that it always does. The evidence indicates moreover that commitment is more likely in at least one context of value judgments than in its absence. The value context is reciprocity. It is suggested that a focus of research on the circumstances that favor commitment – rather than on modifications of an assumed utility function to accommodate non-self-regarding motivations – might lead to a more fruitful behavioral economics and behavioral game theory.

A decision-maker may be said to be rational if his decisions fulfill two conditions: first, they are consistent with his preferences in the light of his beliefs, and second, his beliefs are consistent with the available evidence. Either of these conditions may be fulfilled exactly or only approximately. Selten writes (1975 p. 320) "... game theory is concerned with the behavior of absolutely rational decision-makers whose capabilities of reasoning and memorizing are unlimited ...." The same is true of neoclassical economics. But it is not an open question whether this "concern" is descriptive of actual human decision-makers. There is ample evidence to reject it. No rational decision-maker, choosing among theories, can choose one that is based on absolute rationality. Real human rationality is bounded[1], and although boundedly rational decisions may approximate those that would be made by one of Selten's decision-makers, theories based on the postulate of absolute rationality can be quite misleading. (Akerlof and Yellen, 1986; Akerlof, 2001).

Nevertheless, absolute rationality retains some intellectual interest. On the one hand, the "possible world" inhabited by absolutely rational beings is a possible world that derives its interest from our interest in our own (bounded) rationality and the implications of our attempts to improve it. If we adopt a model of boundedly rational learning, absolute rationality can define attractors and stable points that may (or may not) be observed in the learning process. In a related way, noncooperative game theory (at least) can be a useful problem-finding tool. (McCain, forthcoming) Finally, to the extent that individual boundedly rational decisions do approximate absolutely rational ones, difficulties or ambiguities in the meaning of absolute rationality carry over to real rational behavior. This paper is concerned with one such ambiguity.

In "The Intimate Contest for Self-Command" (1980) and other publications, Thomas Schelling explored the implications of weakness of will for rational choice and behavior. Noting that weakness of will creates within an individual a conflict not altogether unlike the conflict of objectives assumed in noncooperative game theory, he drew on game theory to point out strategies by which an individual might overcome weakness of will and carry out commitments

* Tel.: +1 215 895 2176.
E-mail address: mccainra@drexel.edu.
URL: http://faculty.lebow.drexel.edu/mccainr/.

[1] Simon (1955, 1995). It should be stressed that bounded rationality (artificial intelligence) is a theory of rational, not irrational behavior.

1053-5357/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.socec.2008.08.010

that could create such a conflict. Schelling subsumed weakness of will to *bounded* rationality, but regarded it as very widespread (if not universal) in imperfect human behavior. His examples make it clear that weakness of will is an obstacle to commitment. But what is the relation of absolute rationality to weakness of will? This is the ambiguity in the concept of absolute rationality with which we will be concerned.

1. Weakness of will

The issue does not arise in most of neoclassical economics because neoclassical economics excludes, by assumption, many of the circumstances in which it might arise. Consider the case of intertemporal inconsistency in choice. We adopt the neoclassical convention of expressing time preference by a discount rate. Most economic literature assumes that this discount rate per unit time is the same regardless of the delay before the payment is made. This assumption of a uniform rate of time preference has no basis in empirical observation, but is made in order to reconcile the theory of rational choice, as it is understood in modern economics, with the assumption of time preference. The difficulty is that a non-constant rate of discount can result in what are called intertemporal inconsistencies in decision-making. What this means is that a rational, maximizing decision-maker would make one decision at one point of time, but at a later point of time would rationally prefer the alternative he has initially, rationally rejected. (There has been some recent research on alternatives to constant rates of time preference, such as hyperbolic discounting, but it has been directed to a different issue.)

This can be illustrated by an example. Suppose that the decision-maker discounts any prospect delayed by more than six months at 18%, but that his rate of discount for prospects delayed six months or less is zero. Now the decision-maker must choose at t_0 between two alternatives. Alternative Alt 1 is a payment of \$5000 at $t_0 + 1$ year. Alternative Alt 2 is a payment of \$10000 at $t_0 + 5$ years, but Alt 2 has a cancellation clause: at any time during the first year, for a cancellation fee of \$100, the decision-maker can cancel his decision for Alt 2 and receive the payment of \$5000 at $t_0 + 1$ year.

At t_0, the discounted present values are

Alternative Alt 1 \$4, 237

Alternative Alt 2 \$4, 371

Accordingly, the decision-maker chooses alternative Alt 2. However, at $t_1 = t_0 + 6$ months and one day, the payoff for alternative Alt 1 is less than six months away, and so is not discounted, and is valued at \$5000. To obtain this payment, however, the decision-maker must pay the cancellation fee of \$100. The net values discounted to t_1 are

Alternative Alt 1 \$4, 900

Alternative Alt 2 \$4, 748

Therefore, the rational decision-maker reverses his decision.

This is a one-person game. Suppose we express these decisions as plans of action for the successive stages like the pure strategies as understood by von Neumann and Morgenstern[2] (2004 originally published 1944). The decision-maker has three pure strategies:

(1) Choose Alt 1.
(2) Choose Alt 2, then do not cancel.
(3) Choose Alt 2, then cancel.

The payoffs of these strategies, discounted to t_0, are

(1) \$4,237.
(2) \$4,371.
(3) \$4,127.

Why, then, does our rational decision-maker not simply choose strategy 2 and stick with it? Suppose that the decision-maker has a weak will, in Schelling's sense, and knows that he does. Then he can anticipate that if he chooses Alt 2, he will indeed cancel it after six months and in fact carry out strategy 3. Because of his weakness of will, strategy 2 simply is not available to him. That being so, in the spirit of Ulysses and the Sirens (note Elster, 1977) the rational but weak-willed decision-maker will choose strategy 1 and alternative Alt 1.

Strength of will may be very rare, but I can say from my own experience that it does exist. I recall that in 1960 my late father gave up smoking, after having smoked about 40 cigarettes a day for two decades. Many people have quit smoking, and I would say that all who succeed possess strong wills, even if they have taken steps (like avoiding places where they were in the habit of smoking) to reduce temptation. My father, however, carried an unopened pack of cigarettes in his pocket every day for a year after he quit. He explained to me that he wanted to prove to himself that he was not a slave to the habit—and he never started again.

This is not to say that intertemporal inconsistency does not exist. No doubt a strong-willed decision-maker, having chosen strategy 2, will feel some subjective tension in the nature of regret or temptation during the time interval t_1 to $t_2 = t_0 +$ one year. Does rationality require him to act on the temptation? Well—perhaps it does. Selten writes (1975 p. 328) that the decision-maker "... should not be guided by his payoff expectations in the whole game but by his conditional payoff expectations," at the moment the decision is made.

Weakness of will may also be a factor in interactive decisions. Consider the following two-person game in extensive form, shown as Fig. 1. All decisions are close enough together in time that there is no need to discount payments to present value.

First we note that the subgame perfect equilibrium for this game is for decision-maker A to choose alternative 1 for a payoff of 4,237. However, when we express this game in terms of von Neumann–Morgenstern contingent strategies, we have, for decision-maker A,

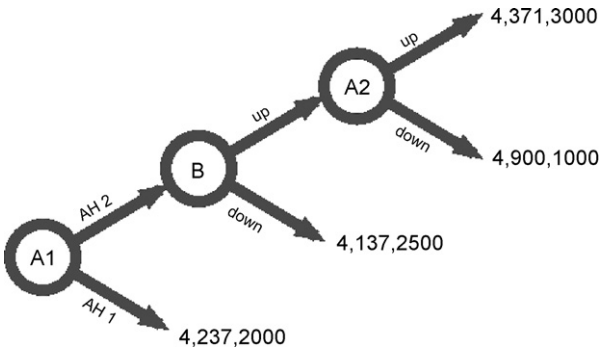

Fig. 1. Two-person game in extensive form.

[2] "Imagine now that each player . . ., instead of making each decision as the necessity for it arises, makes up his mind in advance for all possible contingencies . . . We call such a plan a *strategy*." (von Neumann and Morgenstern, 2004, p. 79).