TOWARDS A THEORY OF EQUILIBRIUM BEHAVIOUR IN PUBLIC GOODS GAME

ALEXIS BELIANIN

This paper aims at offering a completed theoretical explanation to the widely documented cooperation in experimental public goods (PG) games (Ledyard, 1995), together with some implications for the general theory of collective interaction, cooperation and trust. The paper is complementary to an ongoing empirical research joint with Marco Novarese (Univ. Piedmont, Italy). A typically observed pattern of strategies in the PG game consists in rather high (50% and more) contribution to public account at the initial stages of the game, which only gradually declines. This observation contradicts the equilibrium prediction of zero contribution, which is the only the Nash equilibrium strategy.

Several theoretical explanations to this deviation have been proposed in the literature, including random errors of the players (Anderson e.a., 1998), altruism or warm-glow (Palfrey and Prisbrey, 1997; Goeree e.a., 1999; Carter e.a., 1992), conditional cooperation (Keser and van Winden, 2000; Levati, 2002), utility of reciprocity or fairness (Rabin, 1993), inequality aversion (Fehr and Schmidt, 1999) or justice considerations (Fehr and Gachter, 2000). None of these, however, views cooperative behaviour in the game as rational in the basic, non-augmented interpretation of the game. This, however, can easily be done subject to a very natural interpretation of what is "rational" behaviour in this strategic context. If, in the spirit of Herbert Simon, we understood as *rational* the behaviour which results in reaching optimal outcome, without imposing any prior restrictions on the choice procedures used, then rational might be the behaviour which achieves good outcome given the behaviour of the others[1]. To get the intuition, assume the game of 10 participants lasts for 10 periods and the payoff function is $10 - c_i + \frac{1}{3}\bar{c}$. Suppose in the first period the nine players choose $c_i = 10$ (contribute everything to public account), and the last player assumes they would do it now and will do the same thing forever provided everybody sticks to that strategy, while if someone defects by a single unit, they would immediately switch to defection forever. If this last player is convinced the others use such trigger strategy, it is optimal to do the same because $33.33 \cdot 10 = 333.3 > 40 + 10 \cdot 9 = 130$, the defection payoff. By contrast, if the beliefs of the last player are such that in every period except the first one, at most half of the total capital of the other players will be invested, it is better to defect in full rather than contribute anything to public account, because $40 + (\frac{1}{2}\frac{90}{3} + 10) \cdot 9 = 265 > 33.33 + \left(\frac{1}{2}\frac{90}{3} + \frac{10}{3}\right) \cdot 9 = 198.33$.

Rationality of this strategy when others' behaviour is unobservable has, in fact, been recognized by game theorists a long ago[2]. The peculiar feature of this optimization is, however, that the public goods game is a dynamic game of incomplete information, which fact has been stressed in Kreps et.al. (1982) paper. The solution suggested in this paper, however, treats individual types as exogenously given, which limits the scope of substantial analysis of specific strategies of empirical relevance. In particular, boundedly rational individual may fail to figure out the exact conditions under which cooperative strategies will be substantially rational, as in Kreps et.al. A deeper look at these strategies is possible within the framework of 'epistemic games', as pioneered by Aumann and Brandenburger (1994). In particular, this framework allows for an explicit account for each player's beliefs about beliefs of the opponents on the grounds of summary statistics of the opponents' past strategies (which are reported to the players in the course of the game). This view allows to figure out the following core observation:

Observation 1. *The inference of player i about future actions of the opponents is based on i's* perception *of the rationality of the other players, which need not the be the same as the* actual *rationality of these other players.*

In other words, when making their inference, player i would naturally think that the other players' motives and intentions are those which she (player i) thinks these are — colloquially speaking, she puts herself in place of the opponents.

We use these considerations to build the model of optimal strategic behaviour in the public goods game as a standard extensive form game Γ with incomplete information (e.g., Kreps and Wilson, 1982). History of play h_t up to period t in this game is observed only imperfectly, yet the set of all possible past histories H^t

[1] The example that follows is clearly akin to the familiar paradoxes of rationality, such as Newcomb's paradox (Nozick, 1974) or the story of rational Rachel and irrational Irene presented by Martin Hollis (1979).

[2] "[If] the state of the world, as perceived by A, is uncertain, he must construct some assessment of B's action and optimize accordingly." (Bernheim, 1984, p.1011)

is always well-defined, common and finite. In each time period this set plays the role of underlying (or zero level) uncertainty space X_0^t. Using the spaces of beliefs of each player about this space, beliefs about each others' beliefs etc., construct the standard types space Ω^t as the set of all infinite hierarchies of beliefs of all levels (Brandenburger and Dekel, 1993), endowed with the usual canonical properties. In particular, the space is Polish (complete separable metric) and is endowed with the Borel σ-algebra.

Rational individual strategy on this space consists is an expected utility maximizing mapping $s_i^{t*} \in S_i^t$ from each state of Ω^t to the set of available actions $S_i^t \equiv \{s_i^t : \Omega^t \to A_i^t\}$. Our observation 1, however, precludes usage of the standard mental model (Dekel and Gul, 1996): the beliefs held by player i in state ω^t about other players' types (and hence their actions) need not be the same as the actual beliefs which guide the opponents' actions. We use an alternative route to define this model: for each player i define, alongside with the space $(\Omega^t, \mathscr{F}_t)$ of the *factual*, or *true* uncertainty, we take another copy $(\mathcal{E}^t, \mathscr{E}^t)$ of that space, interpreted as *believed* uncertainty. Continuous images $\tau_i^t : (\Omega^t, \mathscr{F}^t) \to (\mathcal{E}^t, \mathscr{E}^t)$ establish correspondence between the possible actual states and the possible perceptions of these states by individual i, hence they will be called *theories* of player i at t, and denoted τ_{ik}^t. The set $E^t \subset \mathcal{E}^t$ of continuous images of all Borel sets in Ω^t, or the set generated by the projections τ_i^t of all subsets B of a Polish space $\Omega^t \times \mathcal{E}^t$ onto $\mathcal{E}^t$, is called *Souslin*, or *analytic*.

Lemma 1. *Any theory τ_{ik}^t of player i defines an analytic set on the space of individual beliefs $(\mathcal{E}^t, \mathscr{E}^t)$.*

Definition 1. *A player is called* rational learner *if she rejects as impossible any observed set of histories H_t which are inconsistent with her prior beliefs about the states of the world.*

Lemma 2. *Inverse mapping from the individual belief space $(\mathcal{E}^t, \mathscr{E}^t)$ to the state space $(\Omega^{t+1}, \mathscr{F}^{t+1})$ constitute an isomorphism.*

These properties are used to prove several theoretical results, in particular:

Proposition 1. *For any kernel of any system $\mathcal{A}^t$ of A-sets, there exists a sequence of regular hierarchical sets $B_{k_1\ldots k_n k_{n+1}} \subset B_{k_1\ldots k_n}$ which determines the set of strategies consistent with these strategy-type pairs.*

Proposition 2. *There exists a collection of individual beliefs compatible with rational individual learning, but not necessarily compatible with common knowledge.*

Proposition 3. *The Souslin types space with rational individual learning in a general game of incomplete information is compatible with a form of generalized rationalizability (Pearce, 1984) for incomplete information games (Battigali-Siniscalchi, 1997).*

One of the conjunctions of these results is that coordination of actions (on cooperative outcome) is the consequence of this miscoordination of rationalities in epistemic game of incomplete information. Further development of these arguments, including dynamic model of learning in the public goods game, as well as its implications for general theory of cooperation, ethics and political science, remain to be investigated.

International College of Economics and Finance (ICEF) of Higher School of Economics, and IMEMO RAS, Moscow, Russia, ph. 623 5055, 8916 159 1329, email: icef-research@hse.ru.