Irrationality in Game Theory

Yamin Htun

Dec 9, 2005

Abstract

The concepts in game theory have been evolving in such a way that existing theories are recasted to apply to problems that previously appeared not to fit in. One of the most scrutinized concepts is backward induction. In extensive form games, given common knowledge of rationality, the outcome is backward induction. However, experiment results are contradictory from what game theory predicts. Several attempts have been made in the literature to solve this problem. The present paper describes and discusses some of those attempts and their limitations.

1 Introduction

The history of game theory has been evolutionary rather than revolutionary - some of the most important developments consisted of novel ways of recasting theories to apply to problems that previously appeared not to fit in [2]. One of the most debatable issues in game theory is rationality. Ironically, rationality is what game theory is all about; almost all of the theories are based on the assumption that agents are rational players who strive to maximize their utilities. However, economists have long expressed dissatisfaction with the strict assumption about rationality with several objections. A common objection is that laboratory experiments indicate that people often fail to conform to some of the basic assumptions of rational decision theory. Furthermore, experiments also indicate that the conclusions of rational analysis sometimes fail to conform to reality. Additionally, the conclusions of rational analysis sometimes seem unreasonable and counter intuitive even on the basis of simple introspection [3].

1.1 Backward Induction

A game theory concept closely related to rationality is backward induction. Backward induction is an iterative process for solving extensive form games. In such games, the player who makes the last move of the game, chooses an action that maximizes his payoff. Taking this as given, the next-to-last moving player makes a choice that maximizes his own payoff in his turn. The process continues in this way backwards in time until the beginning of the game is reached [4]. Effectively, the subgame perfect equilibrium, in which players'

strategies constitute a Nash equilibrium in every subgame of the original game is determined.

Figure 1 shows an extensive form of a well-known finitely repeated prisoner's dilemma game with length two. The end of each round of the game is marked by a dotted line. The payoff to each of the two players is obtained by adding their payoffs for the two rounds and is listed at the end of each terminal node the tree. The payoffs are stated in terms of R, S, T and P where they can have any value with $T > R > P > S$ and $2R > (T + S)$. Therefore, if we apply backward induction concept to this twice played prisoner dilemma, the players should always play D in every round, adopting non-cooperative behaviors.

However, the experiments with human players have shown the contradictory results; players do cooperate at least for some time until near the end of the game. In the process, the outcomes end up with payoffs that are strictly greater than they would obtain under equilibrium play [5]. Therefore, practically and also intuitively, backward induction is implausible or unreasonable, though it is a game-theoretically correct concept.

2 Solutions

Since 1980s, economists have explored solution concepts, which can reflect the reality and explain the observed results of the experiments [5, 2, 3, 4, 7]. In this section, some explanations or solutions to the backward induction paradox are described.

2.1 Failure of Common Knowledge of Rationality

One of the most straightforward solution concepts was in recognizing that neither player is in a position to run the backward induction required [7]. They claim that it is mistaken to assume that a player is in a position to run these arguments before making his first move, or having his opponent's move made. The mutual belief of rationality of the players can exist, but it does not entitle the player to believe that in subsequent rounds his opponent will still believe he is rational, irrespective of how he has acted in the interim. Hence, the premise about players' rationality at later rounds is not available. In other words, it describes the breakdown of the common knowledge of rationality. Common knowledge of rationality means all players know that they are rational, all know that all know it, and so on *ad infinitum* (or at least, for a number of levels no less than the maximum duration of the game).

2.2 Reputation and Imperfect Information

The approach by [5] attempted to solve the paradox by admitting a "small amount" of incomplete information, while maintaining rationality of players. They described equilibrium of the repeated PD in which two rational players

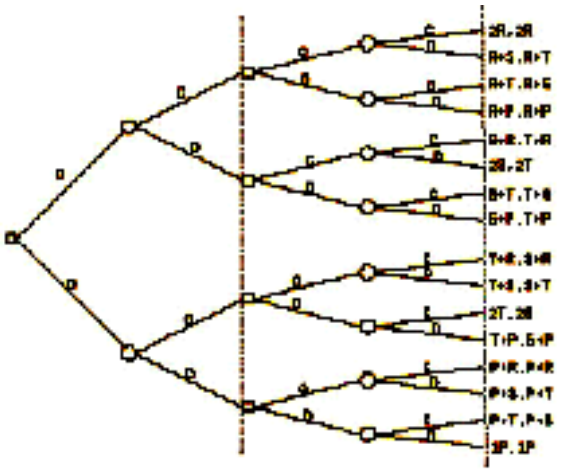

Figure 1: Twice repeated prisoner's dilemma

both believe that there is a small probability δ that the other is "irrational." Two models of irrationality proposed are as follows:

> **Model 1:** The opponent might be playing a tit-for-tat strategy, in which a player starts the game by cooperating, and at subsequent rounds *j+1*, the player chooses the action that the other player chose in round *j*.

> **Model 2:** The opponent may get extra utility from mutual cooperation by being an altruistic type, such as cooperation is the best response to cooperation. Therefore, the utility U_i of a player i with such reciprocal altruistic type can be described as

$$U_i = p_i + \alpha$$

> where $\alpha \geq 0$ whenever player i and his opponent(s) $-i$ cooperate and p_i being the original payoff of the game for the agent i

In either case, a sufficiently high δ can lead the players to adopt a cooperative strategy until round T or until the opponent defects and to defect thereafter. The approach does not require 'irrationality' or 'altruism' to exist, but only the sufficient beliefs about such existence. Hence, one can play "irrational" strategies to entertain some doubt about the irrationality or to build the reputation that he or she is altruistic. By doing so, he or she can motivate the other player to play in some specific way (e.g., a mutually beneficial way). In other words, the rational players disguise themselves as irrational; they make others believe they are altruistic, thus forcing others to play cooperatively. Therefore, this approach is also well known as "crazy perturbations."

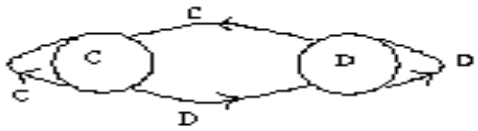

Figure 2: Finite Automata for Tit-for-Tat strategy

2.3 Bounded Complexity

Unlike the above two, the solution concept proposed by Neyman justifies cooperation in the finitely repeated PD without deviating from rationality or complete information [6]. The fundamental assumption is existence of bounds to the complexity of the strategies that the players may use.

Theorem 1 states that if the players are restricted to using some mixtures of pure strategies, which can be represented by finite automata of a fixed size l, for a sufficiently large number of repetitions N, there exists an equilibrium that yields a payoff close to the cooperative outcome.

> **Theorem 1**: *For any integer k, there is an* N_0 *such that if* $N > N_0$ *and* $N^{1/k} \leq min(l_1,l_2) \leq max(l_1,l_2) \leq N^k$*, then there is a (mixed strategy) equilibrium in which the payoffs to each player are at least 3 - 1/k* [6]

Hence, the tit-for-tat strategy can be modeled with the finite automata with size 2 as the figure 2 shows. In this case, l is only 2, but the theorem remains true even when the automaton size l is very large compared to N as long as it satisfies the constraints of the theorem.

3 Discussions

Though the solutions explained the "irrationality" (at least to some extent), they made different assumptions of the game model being analyzed. In the first one, the breakdown of common knowledge of rationality is assumed. In the second, the informational asymmetries among the players are introduced. In the final one, the strategies available to players are restricted. The latter two concepts introduced more formal and specific cases and strategies involving "irrationality", while the former one simply gave the reason for cooperation.

The first concept is logical that players are not necessarily in a position to run the backward induction because the necessary condition or premise of common knowledge of rationality (CKR) is not fulfilled. However, the concept is very broad and it can refer to many cases where CKR breaks down (e.g., the opponent plays a cooperative action, the opponent is irrational or even the opponent is disguising himself or herself as irrational). While it is very easy to point to the failure of CKR and claimed that one has solved the paradox, due to

the breadth covered by the concept, it is hard to discuss about how to develop or justify one's strategies under such circumstances. Moreover, even if the CKR is fulfilled in the model, it is arguable that players will run the backward induction because of limits on human memory or computational ability. In other words, human have the limited foresight - the process of predicting all the possible future states of a game.

The second concept takes on a very different approach. The game being analyzed is changed to a different game model, from the extensive form games to the Bayesian game with uncertainty about the player's type (and hence utility). Taking the alternate utilities and the priors of doubts about the opponent's irrationality or taking the opportunity to build the reputation of "altruism" so that the opponent will cooperate, there is nothing irrational about players anymore; one can easily develop a strategy to maximize his or her own utility within the rational concept of game theory. Therefore, irrationality is used just as one of the strategies. In a way, it goes beyond the basic utility maximization that is inherent in Nash equilibrium. The concept demands rationality off the equilibrium path while the Nash equilibrium demands rationality only on the equilibrium path [3]. The concept has been tested with the experiments and results have been shown to support the concept [1].

The third concept is very close to human cognition and reasoning, especially when the game involves numerous and complex strategies. However, if the game has a simple set of possible strategies and a small number of rounds left, restricting the strategy solely to the automata might fail to reflect human behavior.

4 Conclusions

A backward induction concept has limitations in its applicability. Ironically, the solutions proposed to solve it also have their own limitations. However, it does not mean that we should discard the concept of backward induction or the proposed solutions. Though not common in other areas of studies, substantive conditionals and counterfactuals are necessary in game theory; without those, one really cannot discuss decision making. Making a decision means choosing among alternatives. Thus one must consider hypothetical situations. Though the solution concepts presented do not cover and explain all the possible shortcomings of players' rationality, they made significant contributions of explaining irrationality to the game theory which *used to be* all about rationality.

References

[1] Andreoni, J. & Miller, J. Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence *The Economic Journal* 103 (1993) 570-585.

[2] Aumann, R. Irrationality in Game Theory *Economic Analysis of Markets and Games, Essays in Honor of Frank Hahn.* (1992) 214-227.

[3] Aumann, R. Rationality and Bounded Rationality *Games and Economic Behavior*21 (1997) 2-14.

[4] Aumann, R. Backward Induction and Common Knowledge of Rationality *Games and Economic Behavior*8 (1995) 6-19.

[5] Kreps, D., Milgrom, P., Roberts, J., & Wilson, R. Rational Cooperation in the Finitely Repeated Prisoners' Dilemma *Journal of Economic Theory*27 (1982) 245-252.

[6] Neyman, A. Bounded Complexity Justifies Cooperation in the Finitely Repeated. *Economics Letters* 19 (1985) 227-229.

[7] Pettit, P. & Sugden, R. The Backward Induction Paradox *The Journal of Philosophy*86 (1989) 169-182.