과학철학 17-3(2014) pp. 23-51

Game Theory, Rational Choice Theory, and the Prisoner’s Dilemma: Some Clarifications

Hun Chung*

There seems to be critics who think that game theory can provide very little insights in doing empirical social scientific research or normative political theory/political philosophy. This is because these people tend to think that game theory is committed to some highly contestable theory of human psychology; namely, that human beings either are or should be primarily motivated by their own exclusive self-interest. From this, critics tend to think that game theory is defective both as a normative theory of action as well as a descriptive theory of action. After explaining the basics of game theory, I will try to show that such criticisms are mostly based on a general misunderstanding of game theory. In the end, I will argue that game theory is simply a mathematical tool that could be used to model any strategic interaction for many different purposes, and is not committed to any substantial theories of human nature.

【Key Words】Game theory, Rational choice theory, Prisoner’s dilemma, PD game, Philosophy of the social sciences

* Department of Philosophy, Rochester Institute of Technology, hunchung1980@gmail.com.

1. Two Common Objections against Game Theory

In this paper, I would like to clarify some common misunderstandings of game theory (or, more generally, “rational choice theory” of which game theory is conceived to be a part) which I find quite a few political philosophers, political theorists, and political sociologists, who are mostly non-specialists in game theory, share. The misunderstandings generally stem from (mistakenly) thinking that game theory is committed to some highly contestable theory of human nature or human motivation; that human beings either *are* or *should be strictly egoistic and self-interest-maximizing beings*.

Based on such general assumption about game theory, critics have gone far as to claiming that game theory, and, more generally, rational choice theory have been used as a major vehicle to serve a particular political aim; namely, the expansion of American neoliberal capitalism.[1)]

I will not specifically comment on these grand sociological criticisms – namely, that there is some kind of political agenda behind the expansion of game theory and rational choice theory in many empirical and theoretical disciplines. I will simply say that I think these sociological criticisms are exaggerated and are not based on a correct understanding of game theory.

What I instead will do is to comment on two of the most basic forms of objections on which most of these grand sociological criticisms of game theory and rational choice theory seem to be based. The two basic objections can be summarized as follows:

- **Objection 1.** Game theory is defective as a *normative theory of action*; it urges one to care only about one's own self-interest

[1)] See Amadae (2003), Archer and Tritter (2000).

when one ought to care about other things - such as morality, good citizenship, the common public good - as well.

- **Objection 2.** Game theory is defective as a *descriptive theory of action*; it assumes that people, as a matter of fact, care only about their own self-interests even when they apparently do not.

From this, critics tend to think that game theory can provide very little insights in doing empirical social scientific research or normative political theory/political philosophy. The main purpose of this paper is to show that such criticisms are based on general misunderstandings of game theory. To do so, I will first explain the basics of modern game theory, and introduce the famous *Prisoner's Dilemma* game. I will, then, comment on each of these two criticisms and try to explain precisely in what way they are misplaced.

2. What is Game Theory? − A Short Introduction

So, what is game theory? We can say that game theory is a set of mathematical tools designed to model or represent a *strategic* or *interactive* interaction among two or more individuals. An interaction is *strategic* when the outcome that results from the interaction depends, not merely on one's own actions, but also on the actions performed by others as well. For instance, in a penalty shoot-out in a soccer game, whether or not the kicker scores depends, not merely on the direction towards which the kicker shoots, but also on the direction towards which the defending goalkeeper throws himself to block. Since the outcome depends on both the kicker as well as the goalkeepers actions, we can say that a penalty shoot-out in a soccer game is a *strategic* situation.

By contrast, consider a situation in which one is trying to decide whether to drink orange juice or milk. Here, the outcome does not depend on any other person's action besides the action performed by the person who is trying to decide which beverage he/she should drink. Therefore, a person deciding whether to drink orange juice or milk is not a situation in which strategic interaction occurs.

Game theory aims to model, not just any situation, but a situation in which strategic interaction occurs among two or more individuals. Then, what does it mean to *model* a strategic interaction? A representative case of a model is a map. A map is a model of a given geographical territory. As a model, a map *represents* a given geographical territory by focusing on what the mapmaker deems to be the essential features of the geographical territory while simplifying and ignoring other information that is considered nonessential. For instance, a typical map will include the names of the roads and show how different roads are interconnected in various junctions and intersections. It will also include the names of various buildings and important landmarks. However, a typical map will not include the colors or the materials from which the various buildings are made. Analogously, as a model, a game represents a given strategic interaction by focusing on what the game theorist deems to be the essential feature of such strategic interaction while simplifying and ignoring other information that is considered nonessential. What features of a given strategic interaction that a game theorist regards essential will become more apparent soon.

One major assumption that game theorists make is that people act according to their *preferences* and that these preferences conform to a minimum set of consistency requirements that render each individual's preference relation an *ordering*. When we are taking people's *weak preferences* as our theoretical starting point, the two consistency requirements that are usually invoked are *completeness* and *transitivity*.

We say that a person's weak preferences are *complete* if, for any two outcomes *x* and *y*, the person either weakly prefers *x* to *y* or weakly prefers *y* to *x*; if the person happens to weakly prefer both *x* to *y* and *y* to *x,* then we say that the person is *indifferent* between the two outcomes *x* and *y*. Intuitively, a person's weak preferences are complete if the person is able to find no two outcomes that are incomparable.

We say that a person's weak preferences are *transitive* if, for any three outcomes *x*, *y*, and *z,* the fact that the person weakly prefers *x* to *y* and weakly prefers *y* to *z* implies that the person weakly prefers *x* to *z*. This is quite plausible from a practical standpoint as it would be strange if a person weakly preferred beef to pork and weakly preferred pork to chicken, but strictly preferred chicken to beef.

When a person's preferences are both complete and transitive, then it is possible for the person to *order* any set of outcomes from best to worst. What this disallows is for a person to have what are known as "preference cycles" in which there exists some natural number *n* such that a person strictly prefers x_1 to x_2, x_2 to x_3, $\cdots$, $x_{(n-1)}$ to x_n, and, yet, strictly prefers x_n to x_1. An example of such preference cycle would be a person who likes beef more than pork, pork more than chicken, and chicken more than beef. One might think that there is something *odd* about such person's preferences. And, we might characterize such oddness as exemplifying some sort of irrationality of the person's preferences.

So, the basic assumption of game theory as well as rational choice theory is that people act according to their preferences and that these preferences are *rational*. Here, saying that a person's preferences are rational simply means that his/her preferences satisfy the two properties of completeness and transitivity, which allows him/her to order any set of objects from best to worst according to his/her preferences. It is very important to understand that, in rational choice and game theory, the requirement that one's preferences be rational does not concern nor restrict *the specific contents* of the person's preferences. One person may prefer beef to chicken, while another person may prefer chicken to beef; from the perspective of rational choice or game theory, neither person's preferences are irrational as long as they are both complete and transitive.

One operational convenience that results from a person's preferences forming an order is that, now, such preferences can be represented by what is known as an (ordinal) utility (or a payoff) function. A utility (or payoff) function represents an individual's preferences by assigning greater numbers to outcomes that are higher ranked in the individual's preference-ordering. For example, if individual *i* happens to strictly prefer beef to pork and pork to chicken, then we might represent *i*'s preferences over the three types of meat by a utility function u_i such that $u_i(beef)=3$, $u_i(pork)=2$ and $u_i(chicken)=1$.

Here, the numbers represent only the *order* of individual *i*'s preferences, and conveys no information about the *intensity* with which *i* prefers each type of meat. So, statements like, "individual *i* likes pork two times more than he/she likes chicken" or "individual *i* likes beef three times more than he/she likes chicken" are *meaningless*, as assigning $u_i(beef)=13$, $u_i(pork)=8$ and $u_i(chicken)=3$, or assigning $u_i(beef)=100$, $u_i(pork)=50$ and $u_i(chicken)=0$ would be an equally valid way to represent *i*'s preferences. Note that the two statements that concerned the intensity of the individual's preferences will not remain true in these other assignments of payoffs. In the language of game theory, we say that individual *i's* utility (or payoff) function is *unique up to strictly increasing (monotonic) transformation.*[2)]

I have explained that game theory assumes that people act according to their preferences; that is, whenever people act, game theory assumes that they will choose an action that generates an outcome that is located highest in their preference-ordering. We have just seen that, in game theory, people’s preferences are represented by their ordinal utility (or payoff) functions which assign higher numbers, i.e. payoffs, to those items that are higher ranked in one’s preference-ordering. This means that, in game theory, the assumption that people act according to their preferences is translated into the assumption that people maximize their payoffs whenever they act.

I believe that this is the part that many non-specialists seem to misunderstand. It is true that game theory assumes that people maximize their payoffs whenever they act. However, when a non-specialist hears this, it is very easy for him/her to interpret an individual’s “payoff” as denoting some measure of the individual’s personal benefit, and, thereby, interpret the assumption that people maximize their payoffs as saying that game theory assumes that people always tries to maximize their own self-interest. This is untrue. As explained, a “payoff” simply represents a person’s well-ordered preferences. And, a person’s preferences can be both altruistic and self-sacrificing; game theory does not deny nor affirm the existence of such preferences. As long as a person’s preferences are well-ordered, even a perfectly altruistic saint, who is always acting in self-sacrificing ways, is maximizing his/her payoffs, according to game theory. The fact that people maximize their payoffs does not imply that these people are selfish or egoistic.

[2] A transformation F is strictly increasing (or monotonic) if $u(x) > u(y)$ implies $F(u(x)) > F(u(y))$. Saying that a utility function is *unique up to strictly increasing (or monotonic) transformation* means that if the utility function u represents an agent’s preferences, then performing any strictly increasing (or monotonic) transformation to u (say, $F \circ u$) will also be an equally good way to represent the agent’s preferences.

We now have the basic tools to construct a finite game with simultaneous moves. A finite game with simultaneous moves consists of:

(1) A set of players: $N=\{1,2,...,n\}$

(2) A set of actions for each player *i*: $A_i=\{a_1, a_{2,...,}a_k\}$

(3) Each player *i*'s preferences over the set of action profiles represented by a payoff function u_i: $A_1 \times A_2 \times ... \times A_n \rightarrow \mathbb{R}$.[3)]

Now, suppose that one hears the following story:

The Story of Two Suspects: Two suspects are arrested by the police. The police believe that the two suspects have jointly committed an egregious crime. However, the police lack sufficient evidence to charge the two suspects with the egregious crime that they quite confidently believe that the two suspect have committed; the police only have enough evidence to charge the two suspects with a minor offense. So, in order to charge the two suspects with the egregious crime, the police would need to receive confession from the two suspects. In order to induce confession, the police put the two suspects into two separate interrogation rooms rendering communication between the two suspects impossible. The police propose the following deal to each of the suspects: "If both of you remain silent, then both of you are going to each serve 1 year in prison for the minor offense that we are able to charge with our available evidence. However, if one of you confesses while the other remains silent, the one who confesses will get parole and will

[3)] The product set $A_1 \times A_2 \times ... \times A_n$ is defined as: $\{(x_1, x_2, ..., x_n) \mid x_1 \in A_1, x_2 \in A_2, ..., x_n \in A_n\}$. The symbol $\mathbb{R}$ denotes the set of real numbers. So, u_i is a function from the product set $A_1 \times A_2 \times ... \times A_n$ to the set of real numbers.

> be immediately released for cooperating with the investigation, while the other who remained silent will be fully charged with the egregious crime and serve 10 years in prison. If both of you confess, then both of you will each serve 5 years in prison, which is a slightly reduced sentence for the egregious crime that you both have jointly committed. So, what are you going to do; confess? Or remain silent?" Here, the two suspects care only about the number of years that each serves in prison.

Suppose that one wishes to represent the situation depicted in the story of two suspects via a game-theoretic model. How should one proceed? We might model the situation by the following game:

(1) A set of players: $N=\{1,2\}$

(2) A set of actions:

$$A_i=\{Cooperate(RemainSilent), Defect(Confess)\}$$

(*for* $i=1,2$)

(3) Preferences:

$$u_1(Defect, Cooperate)=0$$
$$u_1(Cooperate, Cooperate)=-1$$
$$u_1(Defect, Defect)=-5$$
$$u_1(Cooperate, Defect)=-10$$

$$u_2(Cooperate, Defect)=0$$
$$u_2(Cooperate, Cooperate)=-1$$
$$u_2(Defect, Defect)=-5$$
$$u_2(Defect, Cooperate)=-10$$

For a more intuitive understanding of the structure of the game, we can represent the game into a matrix form, which is known as the normal (or strategic) form of the game:

<Table 1: Prisoner's Dilemma>

Player 1 \ Player 2	Cooperate (Remain Silent)	Defect (Confess)
Cooperate (Remain Silent)	-1, -1	-10, -0
Defect (Confess)	0, 10	-5, -5

Many people will realize that this game is the well-known game of *Prisoner's Dilemma* (PD Game).

3. The PD Game and Nash Equilibrium

Let's look at Table 1 which represents the PD game in matrix form. How should we read the game matrix? Here, each row represents an action that player 1 can perform. Each column represents an action that player 2 can perform. We can see that each player has two actions; cooperate and defect. This results in a total of four action combinations (or action profiles) which is represented by the four cells in which the payoffs are written. Each player receives a payoff for each action profile. The payoff written on the left-side of the comma represents player 1's payoff, while the payoff written on the right-side of the comma represents player 2's payoff. So, now, we know how to read the PD game in Table 1.

Then, what conclusions can we draw from the PD game? Or how would the strategic interactions between the two players in the PD game eventually unfold? In order to know this, we would have to *solve* the game by using a particular *solution concept*. There are many solution concepts that apply to finite simultaneous move games.[4] However, the solution concept that is most widely used in solving such class of games is the solution concept called, *Nash equilibrium* (NE for short).

Then, what is a Nash equilibrium? A Nash equilibrium is an action profile in which no individual has an incentive to unitarily deviate to another action given that the actions of the other individuals remain the same; this is so whenever such unitary deviation will give a no greater payoff than what one is receiving in the current action profile. Since nobody has any private incentive to deviate, we can say that a Nash equilibrium, once reached, is a very *stable state*.

To formally define a Nash equilibrium, consider any arbitrary finite game with simultaneous moves with *n* players. Let $a = (a_{1,}a_{2,} \ldots, a_n)$ denote an action profile in which player *i* plays action $a_{i.}$. Let $(a'_{i,}a_{-i})$ denote an action profile in which all players other than player *i* play action a_j (for $j = 1, 2, \ldots, i-1, i+1, \ldots, n)$), while player *i* plays action a_i'. Then, we may formally define a Nash equilibrium as follows:

- **Definition (*Nash Equilibrium*)** *An action profile* $a^* = (a_1^*, a_2^*, \ldots, a_n^*)$ *is a Nash equilibrium if and only if for every* $i \in N$ *and every* $a_i \in A_i$, $u_i(a_i^*, a_{-i}^*) \geq u_i(a_{i,}a_{-i}^*)$.

To explain the definition in words, an action profile $a^* = (a_1^*, a_2^*, \ldots, a_n^*)$ is a Nash equilibrium if and only if, for each player *i*, playing his/her strategy in the action profile $a^* = (a_1^*, a_2^*, \ldots, a_n^*)$ (i.e. playing a_i^*) will give him/her a payoff of as least as great as playing any other strategy

[4] Such as: nobody uses strictly dominated strategies, iterative elimination of strictly dominated strategies, nobody uses weakly dominated strategies, iterative elimination of weakly dominated strategies, and so on.

available to him/her, given that every other player also plays the strategies in $a^{*}=(a_1^{*},a_2^{*},\ldots,a_n^{*})$. In the PD game in Table 1, it is easy to verify that (Defect, Defect) is the *unique* Nash equilibrium of the game.

Proposition *In the PD game in Table 1, (Defect, Defect) is the unique Nash equilibrium.*

Proof. Consider the action profile (Defect, Defect). Here, both players each receive a payoff of -5. If any player unitarily deviated to Cooperate, then his/her payoff becomes -10. So, neither player has an incentive to unitarily deviate and play a different action. This proves that the action profile (Defect, Defect) is a Nash equilibrium. Now, to show that (Defect, Defect) is the *unique* Nash equilibrium, consider any other action profile, say, (Cooperate, Cooperate). Here, both players each receive a payoff of -1.Given that the other player Cooperates, any player can receive a payoff of 0 by Defecting, which is higher than what he/she would receive by Cooperating. Therefore, both have incentives to unitarily deviate from (Cooperate, Cooperate) and play Defect. Hence, (Cooperate, Cooperate) is not a Nash equilibrium. Now, consider the action profile (Cooperate, Defect). Here, player 1 receives a payoff of -10, while player 2 receives a payoff of 0. Given that player 2 Defects, player 1 can increase his/her payoff to -5 by Defecting. Therefore, player 1 has an incentive to unitarily deviate to play Defect, and, hence, (Cooperate, Defect) cannot be a Nash equilibrium. An analogous reasoning shows that (Defect, Cooperate) cannot be a Nash equilibrium as well (as, now, player 2 will have an incentive to unitarily deviate and play Defect.) Therefore, (Defect, Defect) is the unique Nash equilibrium of the game.

We may signify that (Defect, Defect) is the unique Nash equilibrium of the PD game by putting an asterisk on top of the payoffs in Table 1.

<Table 2: Nash equilibrium of PD game>

Player 1 \ Player 2	Cooperate (Remain Silent)	Defect (Confess)
Cooperate (Remain Silent)	-1, -1	-10, -0
Defect (Confess)	0, 10	**-5^{*}, -5^{*}**

The fact that (Defect, Defect) is the unique Nash equilibrium of the PD game suggests that when a real life strategic interaction has the structure of the PD game, it is very likely that all players will defect and fail to cooperate. The interesting thing about (Defect, Defect) being the unique Nash equilibrium of the PD game is that it *is sub-optimal* (i.e. *Pareto inferior*.) That is, if both players Cooperated, both of them would have achieved an outcome (viz. each receiving a payoff of -1) which is strictly better than what they get in (Defect, Defect) (i.e. each receiving a payoff of -5.) And, this is precisely why the Prisoner's Dilemma is called a *dilemma*; it is a dilemma in the sense that optimal behavior from each individual's own perspective can fail to achieve a collectively efficient and optimal outcome.

It is an irony that two individuals cannot achieve a collectively efficient outcome even when both individuals simultatneously strictly prefer such outcome to the sub-optimal outcome in which they eventually end up discovering themselves. However, before one actually writes down the model and solves it, it is far from intuitively evident that (Defect, Defect) will be the unique Nash equilibrium of the PD game. And, this is precisely where I believe one of the major theoretical values of game theory lies. Game theory can, in many cases, show how the strategic interaction of different individuals can eventually lead to a rather unexpected collective outcome which may have been unpredictable by a mere reflection in one’s armchair.

4. Two Basic Objections of Game Theory

Now, let us return back to the two objections that were introduced in section1. To remind ourselves, the two objections were:

- **Objection 1**. Game theory is defective as a *normative theory of action*; it urges one to care only about one's own self-interest when one ought to care about other things - such as morality, good citizenship, the common public good - as well.

- **Objection 2**. Game theory is defective as a *descriptive theory of action*; it assumes that people, as a matter of fact, care only about their own self-interests even when they apparently do not.

Again, the two basic objections stem from the thought that game theory is committed to the view that human beings either *are* or *should be strictly egoistic and self-interest maximizing beings*. Let me comment on each of these objections in turn.

4.1. Game Theory is Defective as a Normative Theory of Action

When one is first introduced to game theory and the PD game, it is very easy for one to understand game theory as recommending a certain prescription in the PD situation; that is, game theory might seem to be saying that, in a PD game, defection is rational, and this might seem to imply that game theory *recommends* defection in the PD game.

Understood in this way, it seems that game theory is recommending people to be *selfish*; that is, it seems that it is urging people only to care about their *narrow self-interest* rather than to cooperate with other people even when such cooperation is possible. For instance, in his article, "The Rational Choice Approach to Politics: A Challenge to Democratic Theory", Mark Petracca claims,

> In the main, proponents of rational choice theory "assume that it is egoistically, individualistically, irrational not to maximize one's satisfactions and seek one's own greatest good."(Petracca 1991, p. 296)

Many political theorists find such conclusion rather distasteful. To them, even if it is true that defecting in the PD game would maximize one's self-interest, there could be other considerations, such as a *moral reason*, that dictates one to cooperate rather than to defect in the PD game. (For instance, maybe, the two players in original the prisoner's dilemma story made a *promise* not to confess if they happen to get interrogated by the police beforehand.) Some people might think that such *moral reason* should override any reason that stems from purely egoistic considerations. To them, game theory ignores such moral reasons or any other considerations that are not directly relevant to maximizing one's own self-interest by relying on a very narrow conception of rationality. According to Petracca,

> The influence of values, ethics, and ideas on individual motivation are alien to rational choice theories of human nature. By this account, public spirited behavior or behavior motivated by other regarding motives is not only irrational, but highly unlikely.(Petracca 1991, p. 297)

We can clearly see here that Petracca understands game theory as saying that there is something *intrinsically irrational* about behaviors that stem from public-spirit or other-regarding motives. Therefore, according to Petracca's understanding of game theory, if one is truly rational, one should ignore these other conflicting moral motives and should always try to maximize one's exclusive self-interest even at the expense of others'.

However, for people like Petracca, urging people to become somebody who only cares about his/her own narrow self-interest is not a proper way to cultivate democratic citizenship to people who should rightfully care about things such as democratic deliberation and the common public good. And, hence, they think game theory is defective as a *normative theory of action* on which any normative political theory or philosophy should be based.

Such objection against game theory and the PD game is misplaced.

First of all, such objection misconstrues what type of behavior game theory deems to be rational or irrational. As we have seen, game theory assumes that people act according to their preferences, and that these preferences are rational whenever they conform to the two properties of completeness and transitivity, which renders the individual's preference relation a preference-ordering, which, in turn, makes it possible for the individual's preferences to be represented by a real-valued (ordinal) utility (or payoff) function. So, rational behavior, according to game theory, is simply behavior that stems from rational preferences. But, whether or not one's preferences are rational have nothing to do with whether they stem from narrowly self-interested selfish motivations.

It is important to understand that game theory is agnostic about where people's preferences come from. Some people might be motivated by ethical considerations, moral values, a sense of public good, or even altruism; others might be motivated exclusively by their own self-interest. However, as long as the final preferences of these people satisfy the formal requirements of completeness and transitivity that render their preferences an ordering, none of these preferences are, from a game-theoretic point of view, intrinsically irrational. This is because, according to game theory, the rationality of preferences is related to the formal/logical properties of the preference relation *itself*, rather than the *specific contents* of those preferences.

Second, it is true that game theory considers it rational for one to defect in the PD game. However, we should be very careful not to interpret this as claiming that one should defect whenever one encounters a PD-game-like real-life situation or that there is something intrinsically irrational about cooperating in PD-game-like real-life situations. When game theory deems it rational for each player to defect in the PD game, this is so given that the two players *already have the preferences that they are assumed to have in the PD game*. In the PD game, each player can achieve an outcome that he/she strictly prefers by unitarily deviating from cooperation to defection regardless of what the other player does. In game-theoretic language, in the PD game, defection *strictly dominates* cooperation, and, hence, it would be rational for one to defect regardless of the other player's action. However, here, the rationality of defection hinges on the two players *already* having the type of preferences that they are already assumed to have in the PD game. And, game theory does not claim that, outside the PD game, people should have PD-game-like preferences.

This is similar to saying that it would be rational to choose an action that gives one an apple instead of an action that gives one an orange *given that one prefers having an apple to having an orange*. However, one should be clear that this is not to say that one should prefer an apple over an orange in the first place; or that there is something intrinsically irrational about preferring an orange over an apple.

Similarly, game theory does *not* claim that any two individuals *should* *have* or that it would be *irrational* for any two individuals *not to have* the type of preferences that would render their interaction a PD game in the first place. Concerning the question of what sort of *substantive preferences people should have*, game theory does not take any stance.

So, game theory is *not*, as many people mistakenly believe, a normative theory of action that claim that people *should be selfish* or that *it is rational to care only about one's exclusive self-interest*. Other than requiring one's preferences to be consistent enough to form an ordering, game theory does *not* suggest what type of preference people *should* have in the first place; it only tells us what would happen if people do have the type of preferences that they are already assumed to have in a given model. And, as we have seen, it is up to the game theorist's own discretion to specify the type of preferences each player has when modeling a given strategic interaction.

If there are any substantive normative conclusions that we might be able to draw from the PD game, it would be that there can be certain social situations where the social structure itself could generate a sub-optimal social equilibrium, and that whenever we confront a social situation that resembles the structure of the PD game it might be recommendable to alter the incentive structure of the individuals in order to restore the social optimum and achieve a Pareto-improvement on some predefined welfare criteria.[5] The possibility of suggesting such normative conclusions is one way, I believe, that game theory could contribute to normative political philosophy/theory.

4.2. Game Theory is Defective as a Descriptive Theory of Action

To this, the objector might raise another objection of the following

[5] Such things are usually done in the field that is now known as "mechanism design."

line: regardless of whether or not game theory is intended to be a normative theory of action, it is even defective as a *descriptive theory of action*. This is because, according to these critics, empirical evidence has shown that considerations of self-interests play a very marginal role in actual human beings' real-life actions and motivations.

> … a growing body of empirical research in a variety of social science disciplines shows the explanatory limits of the rational choice approach to human nature. … Tom Tyler's recently published study of why people obey the law shows that normative values about distributive and procedural justice matter in the motivation of individual behavior. In a study of randomly selected citizens in Chicago, Tyler made this important discovery:
> "People obey the law because they believe that it is proper to do so, they react to their experiences by evaluating their justice or injustice, and in evaluating the justice of their experiences they consider factors unrelated to outcome, such as whether they have had a chance to state their case and been treated with dignity and respect. On all these levels people's normative attitudes matter, influencing what they think and do.(Tyler 2006, *Why People Obey the Law*, p. 178)" (Petracca 1991, pp. 300-1)

Similar empirical findings have been found in the study of PD games in real-life situations: that is, it has been confirmed by many experiments that people participating in a game that mimics the structure of the PD game tend to cooperate far more often than what game theory predicts.[6]

However, what the results of these empirical experiments really show is *not* that there is any fault in game theory's predictive or descriptive power, but merely that many people do not have the preferences that would make their interaction in the experiments instances of a PD game.

[6] See Dawes and Thaler (1988), Cooper et al. (1996).

For example, suppose that an experimenter randomly picks two people from a group and makes them play the following game:

> **The Money Game**: Each player can choose either to "cooperate" or "defect." When one player cooperates while the other defects, the person who cooperated pays $1 while the person who defected receives $2. If both players cooperate, then both players receive $1. If both players defect, then both players receive nothing. Moves are made simultaneously. The situation can be summarized by the following payoff matrix.

<Figure 3: The Money Game>

Player 1 \ Player 2	Cooperate	Defect
Cooperate	$1, $1	$1, $2
Defect	$2, $1	**$0*, $0***

Given that the two players care only about the amount of money they receive, we can see that the experiment has exactly the same structure as the PD game; that is, defection strictly dominates cooperation for both players and, thereby, the unique Nash equilibrium is mutual defection. However, suppose that, after many trials of the experiment, it turned out there were very many cases where the two players chose to cooperate rather than to defect.

It is very easy to think that such experiment falsifies a major assumption as well as a general prediction of game theory; that people care only about promoting their own self-interest which would render the unique Nash equilibrium of the situation to be universal defection. On the contrary, what the experiment really shows is merely that money is not the only thing that people in general care about. And, the claim that people *do care* or *should care* only about money is *not* a part of game theory.

People participating in the experiments might care about their reputation, etiquette towards strangers, public humiliation etc., and they might have thought that winning an extra dollar is not worth compromising any of these things. If this explanation is correct, then this means that the preference-orderings of the people who were participating in the experiments might very well be the following:

1. Display good manners and win $1 (*Mutual Cooperation*)
2. Display good manners and lose $1(*Unitary Cooperation*)
3. Display bad manners and win $2 (*Unitary Defection*)
4. Display bad manners and win nothing (*Mutual Defection*)

Again, if we assigned (ordinal) utilities to the outcomes associated with each action-pair of the two participants representing each player's preference-orderings (in the sense that option *a* is strictly preferred to option *b* if and only if the utility assigned to option *a* is greater than the utility assigned to option *b*), their mutual interaction could be represented by the following payoff matrix:

<Figure 4: The Game that the Experimental Subjects may be playing>

Player 1 \ Player 2	Cooperate	Defect
Cooperate	$\mathbf{4^*}$, $\mathbf{4^*}$	3, 2
Defect	2, 3	1, 1

Here, cooperating strictly dominates defection for both players, and, thereby, the unique Nash equilibrium is mutual cooperation. Clearly, this is not a PD game. In other words, the participants might not actually be playing a PD game even when the experimenter deliberately tries to mimic the structure of the PD game in designing the experiments. This shows that there could always be a gap between the experimenter's intention behind experimental design and how the experiment is actually perceived by the subjects.

As I have explained, game theory is a set of mathematical tools that could be used to model a strategic interaction among two or more individuals. What game theory assumes is that people generally choose according to their preferences and that these preferences conform to a minimum set of consistency requirements that render them an ordering. However, as already explained, game theory, by itself, is silent on the issue of what specific preferences people do have or should have.

What types of preferences are assigned to each player in a given game-theoretic model is up to the modeler's own discretion. Different set of preferences (among the players) results in a different game. If people's real-life preferences happen to roughly conform to the preferences of the players in a specific game-theoretic model, then the equilibrium of that specific game is a good predictor of what type of social situation will eventually emerge as a result of those people's interactions. However, (unsurprisingly) if people's preferences are misrepresented, then the resulting game-theoretic model will very likely give false predictions. This does not show that there is any intrinsic fault in game theory; it merely shows that we have chosen the wrong game to represent the situation.

5. Concluding Remarks: Seeing Game Theory as a Mathematical Tool

In short, game theory, by itself, is not committed to any substantial theory of human psychology; specifically, game theory does not claim that people *are* or that they *should be selfish* (e.g. that they (should) care only about money, reducing their years in prison, and so on.) Game theory need not deny that people's preferences can be based on other things – such as their moral or religious convictions, their sense of right and wrong, and certain types of other-regarding desires – as long as their preferences form an order.

People might have different views on how much value the application of game theory (or rational choice theory) in empirical social scientific research or normative political theory or political philosophy has. However, one should at least not try to throw away game theory on the grounds that it is defective either as a normative or a descriptive theory of human action for the reasons explained in the previous sections.

Before I conclude this short essay, I want to say a little bit more about the usefulness of game theory as a methodological tool. Note that a tool can be used for many purposes. This means that the usefulness of a tool can only be evaluated in light of the specific purpose such tool was intended to serve. Also, for any given purpose, a tool can either be used proficiently or poorly, depending of the proficiency level of the wielder. I believe that exactly the same thing applies to game theory.

Not all game-theoretic models are constructed to serve a single unitary purpose. Some people might want to use the tools of game theory to *predict* some social-political-economical phenomenon. Other people might want to use game theory to *explain* a given social-political-economical phenomenon by providing a possible causal mechanism that underlies and could have possibly generated such pattern.

Clarke and Primo (2007) have explained that any given (game-theoretic) model could serve at least five distinct purposes:

- A *foundational model* provides theoretical insights into a general class of problems,

- A *structural model* provides empirical generalizations of known facts,
- A *generative model* produces non-obvious and rather counter-intuitive results from a set of assumptions that are widely believed to be well-known,
- An *explicative model* explores causal mechanisms that could explain a given social pattern, and
- A *predictive model* forecasts events or outcomes.[7]

This is why the well-known criticisms against rational choice theory presented by Green and Shapiro (1994) are not entirely fair. Consider Green and Shapiro's criticisms against rational choice theory of which game theory is a major part:

> Our focus here is on *the empirical power* of rational choice theory. [⋯] in our view the case has yet to be made that these models have advanced our understanding of how politics works in the real world. To date, a large proportion of the theoretical conjectures of rational choice theorists have *not been tested empirically*. Those tests that have been undertaken have either failed on their own terms or garnered theoretical support for propositions that, on reflection, can only be characterized as banal⋯(Green and Shapiro 1994, p. 6)

The reason why Green and Shapiro's criticisms are unpersuasive is because not all game-theoretic models are designed to generate empirically testable observable predictions. In fact, a vast number of game-theoretic models hardly produce any predictions at all.

Among the five different purposes Clarke and Primo claim that a given game-theoretic model can serve, only a predictive model aims at generating empirically testable predictions. The value of the other four

[7] Clarke and Primo (2007), pp. 743-4.

types of game-theoretic models lies in how well these models serve the specific purpose that *they* aim to accomplish. And, one cannot criticize game theory on the ground that it did poorly on some given aspect for which the specific game-theoretic models did not even try to do well.

Even if a game-theoretic model that purports to provide an empirically testable prediction did poorly precisely on this respect, this is not a sufficient reason to conclude that game theory is a poor tool to conduct empirical research. This is because the reason why such game-theoretic model produced false predictions might have nothing to do with the tools of game theory themselves, but rather with how the specific researcher used the tools to model the specific strategic interaction in question (e.g. the researcher might have falsely attributed PD-game-like preferences to the players when cooperation was actually the dominant strategy for these players as was the case in the example introduced in the previous section.)

As I have explained, just like a map, a game is a model of reality. Just like there are good maps and bad maps, there are good game-theoretic models and bad game-theoretic models. And, just like it would be nonsensical to completely deny the usefulness of maps simply because one had encountered a number of bad maps in the past, it would be nonsensical to completely deny the usefulness of game theory simply because there have been a number of bad game-theoretic models used in theoretical/empirical research.

We should remember that any tool is good for certain types of tasks while being not so good for other types of tasks. Game theory, as a mathematical tool, can be good for certain types of academic inquiry while being not so good for other types of academic inquiry. It is particularly useful when, for whatever purpose, one wants to rigorously analyze strategic interactions. However, even when game theory is bad for a specific academic purpose, we should remember that it is not because it is committed to some substantial theory of human psychology that claims that human beings *are* or *should be* narrowly self-interested.

References

Amadae, S. M. (2003), *Rationalizing Capitalist Democracy: The Cold War Origins of Rational Choice Liberalism*, University of Chicago Press.

Archer, M. S. and Tritter, J. Q. (2000), *Rational Choice Theory: Resisting Colonization*, Routledge.

Binmore, K. (2009), *Rational Decisions*, Princeton University Press.

Campbell, R. and Sowden, L. (eds.) (1985), *Paradoxes of Rationality and Cooperation: Prisoner’s Dilemma and Newcomb’s Problem*, UBC Press.

Chwe, M. S. (2013), *Jane Austen, Game Theorist*, Princeton University Press

Clarke, K. and Primo, D. (2007), “Modernizing Political Science: A Model-Based Approach”, *Perspectives on Politics* 5: pp. 741-53.

Cooper et al. (1996), “Cooperation without Reputation: Experimental Evidence from Prisoner’s Dilemma Games”, *Games and Economic Behavior* 12(13).

Dawes, R. and Thaler, R. (1988), “Anomalies: Cooperation”, *The Journal of Economic Perspectives* 2(3).

Dixit, A. and Skeath, S. (2004), *Games of Strategy (second edition)*, W.W. Norton & Company.

Dutta, P. (1999), *Strategies and Games: Theory and Practice*, The MIT Press.

Fishburn, P. (1968), “Utility Theory”, *Management Science* 14(5).

Green, D. and Shapiro, I. (1994), *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*, Yale University Press.

Osborne, M. (2003), *An Introduction to Game Theory*, Oxford University Press.

Petracca, M. (1991), “The Rational Choice Approach to Politics: A Challenge to Democratic Theory”, *The Review of Politics* 53(2).

Resnik, M. (1987), *Choices – An Introduction to Decision Theory*, University of Minnesota Press.

Date of the first draft received	2014. 10. 05
Date of review completed	2014. 11. 02
Date of approval decided	2014. 11. 06

게임이론, 합리적 선택이론, 그리고 죄수의 딜레마에 대한 해명

정 훈

게임이론에 대한 비판자들은 게임이론이 실증적인 사회과학 연구나 규범적인 정치철학/정치사상 연구에 거의 도움을 주지 못한다고 주장한다. 이들이 이렇게 생각하는 가장 큰 이유들 중 하나는 이들이 게임이론이 인간의 본성과 심리에 관한 매우 잘못된 가정 위에 세워진 이론이라고 생각하기 때문이다. 그 잘못된 가정이란 바로 인간들이 본성적으로 이기적이거나 이기적이어야만 한다는 생각이다. 이러한 인식으로부터 게임이론에 대한 비판자들은 게임이론이 인간행위에 대한 기술적인 이론으로서나 규범적인 이론으로서 커다란 결함을 가지고 있다고 생각한다. 본 논문에서 필자는 이러한 비판들이 대개의 경우 게임이론에 대한 일반적인 오해에 바탕하고 있다는 것을 보일 것이다. 그러기 위해서 필자는 논문의 전반부에서 게임이론의 기초에 대해서 설명을 한 후에, 논문의 중-후반부에서 게임이론이 인간본성에 관한 어떤 특정한 전제를 가정하고 있는 것이 아니며, 그것은 단순히 다양한 전략적인 상황들을 다양한 목적을 위해 모형화하고 분석할 수 있는 것을 가능하게 해주는 매우 유용한 수학적인 도구라고 주장할 것이다.

주요어: 게임이론, 합리적 선택이론, 죄수의 딜레마, PD 게임,
사회과학의 철학