A Note on the Consistency of Game Theory[*]

by

Itzhak Gilboa[**]

Abstract

It has been claimed in the literature that classical game theory is inconsistent, since it (implicitly) assumes that all players are rational and that this is common knowledge among them, while these two assumptions seem to be contradictory. The purpose of this note is to suggest a framework which allows the formalization of these implicit axioms in a consistent way.

The main idea is to distinguish between conceivable and possible states of the world, while both exist as formal objects in the theory. Thus we may require that the players would make rational choices only at possible states of the world, and that this fact be common knowledge at all (conceivable) states, where the impossible ones are present in the model for the sole purpose of formally presenting the players' reasoning.

It seems that the new concept of possible states of the world is an analytical tool which may have further (theoretical) applications.

[*] I wish to thank Philip Reny and Cristina Bicchieri for the discussions which motivated this note, and the encouragement to write it down. I am also grateful to Robert Aumann, Ehud Kalai, Dov Samet, Moshe Vardi and Lenore Zuck for helpful discussions.

[**] KGSM/MEDS, Northwestern University, 2001 Sheridan Road, Leverone Hall, Evanston, Illinois 60208.

1. Motivation

Let us consider the two-person game given in figure 1. Reny (1988), following Kreps, Milgrom, Roberts and Wilson (1982) claimed that one cannot assume that rationality is common belief, let alone common knowledge, at every node of the game, since should player II arrive at node 2 his/her belief that player I is rational would be inconsistent with player I's actual behavior at node I, if rationality of both were indeed common belief at this node.

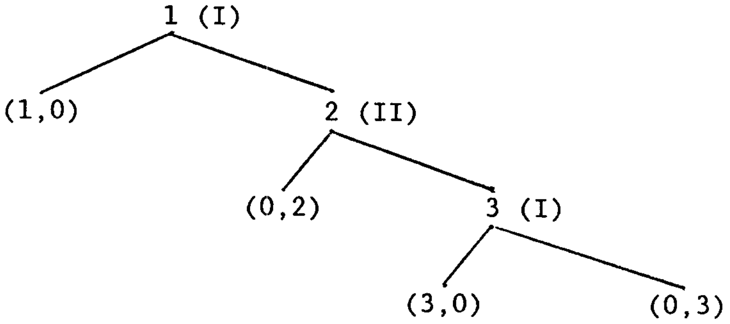

Figure 1
(Arabic numbers denote decision nodes, Latin ones - players)

Bicchieri (1988a,b,c and 1989) has carried this argument one step further to claim that traditional game theory, implicitly assuming common knowledge of rationality, is inconsistent. (See also Bonanno (1988).)

While game theorists seem to be willing to admit that common knowledge of rationality is not a very likely assumption in certain contexts (see Selten (1978), Rosenthal (1981), Kreps et al. (1982), Aumann (1988), Aumann-Sorin (1989) and the literature on bounded rationality), most of them would probably find the claim of inconsistency rather disturbing, partly because there are situations in which these assumptions are quite reasonable, and partly because being unrealistic is hardly as strong a flaw of a mathematical model as being inconsistent (in fact, the former sometimes seems to be a merit.) We are therefore challenged to come up with a viable mathematical model that would solve the problem explained above.

We introduce here two possible answers; the first, which we find somewhat unsatisfactory, is along the lines of traditional game theory, while the second uses the new concept of "possible" states of the world.

2. The Traditional Reply

We can simply write down the axioms of rationality being common knowledge at any state of the world--following Aumann (1974,1976,1987) in some formalized form as suggested, say, in Gilboa (1988) or Kaneko (1987). The meaning of "rationality" here may be somewhat unclear, but let us assume, for simplicity, that it consists of a set of axioms specifying what would each player do given any possible knowledge, and that this specification is, indeed, what we expect it to be (Later on we will present a slightly more sophisticated version of this axiom.) Then we can translate the well-known backward-induction argument to a theorem stating that in the game above there is a single state of the world at which player I plays left at node 1, thus terminating not only the game but also the discussion. The questions of why did player I play that way, what he/she thought player II would think should I play right and so forth are meaningless in this model; they may be very interesting as questions **about** the model, suggesting the unreasonability of its assumptions, but they cannot imply that the model is inconsistent.

The difficulty with this answer is that the model one ends up with does not seem to capture the players' reasoning, even if the scenario dictated by it is actually followed by them. It is somewhat disturbing that there are no states of the world corresponding to the other nodes of the game tree; after all, it seems **conceivable** that they would materialize, even if we ended up convincing ourselves that this cannot be the case. Indeed, one has to think about those possibilities in order to exclude them. Without having a state of the world as a "name" for each such possibility we cannot claim to have formalized the implicit assumption of game theory; in the model presented above it was only the outside observer who reasoned and understood the backward induction; the players themselves did not.

3. A Digression: Counterfactuals and Proofs by Negation

Noting that the (standard) model described above, flawed as it is, was rich enough to allow for a proof of the "rational" outcome, we may ask what distinguishes this perfectly valid proof-by-negation from the counterfactual argument we would like our players to conduct. The answer suggested by the discussion above is quite simple: both a counterfactual argument and a proof by negation may be described as considering a statement (not p) in the presence of the statements (not p implies q) and (not q), from which p is deduced. However, we may distinguish between them as follows: if the latter two statements are tautologies, that is, if they are true at each and every conceivable state of the world--then this is a mathematical proof by negation; if, on the other hand, there are conceivable states of the world at which this argument does not hold--this is a counterfactual reasoning.

4. An Alternative Reply

The obvious reply which the above discussion seems to suggest is the following: we have to begin with a set of conceivable states of the world S, which is large enough to describe any outcome of the game. Thus, if we analyze a game in an extensive form, there will be at least as many states of the world as there are terminal nodes ("leaves"), and in a normal-form analysis--as many as entries in the (super-)matrix of the game. (In fact, one would have to have more states of the world, since for each player a state of the world has to specify what would occur as a result of every possible action of this player, and not only of the one actually chosen at it. However, for simplicity we may ignore the counterfactual elements in the states of the world.) Events, surprisingly enough, will be subsets of states of the world, and in particular, every node in the game tree will correspond to an event containing all the leaves that may be reached from it.

The main novelty is the following: we will use a non-empty subset P of S, interpreted as the set of **possible** states, as a formal object in our model, which may be an object of players' knowledge. Thus, in a model such as Gilboa (1988), it will be meaningful to ask, say, whether at a certain state of the world a certain player considers another state (or the same one) to be possible or not, and so forth. Adhering to Savage's (1954) principle of a state-of-the-world "resolving all uncertainty," the answers to all such questions will be a part of the description of a state, including states which are conceivable but not possible according to some states (possibly - not even in the possible set according to themselves.)

We will also find it useful to extend the definition of possibility to events. Again, in a very general model each state may define each event to be possible or not, without any relationship to the possible states of the world. However, we will extend the consistency assumptions in Gilboa (1988), which say, for instance, that every pair of states of the world define any third one in the same way, to include the following:

1. According to every state of the world, an event is possible if and only if it has a non-empty intersection with the set of possible states.

2. The definition of the set P is identical across states. (Note that this assumption also implies that the definition of P is common knowledge.)

3. The event P is common knowledge at every state in it. (One may require that P be common knowledge over all S, thereby making the states of the world outside it inconsistent since what players know at those states need no longer be true at them; this inconsistency should not pose a problem: one simply has to modify the axiom saying that what a player knows is true to apply only to possible states of the world.)

A set P with the above properties still does not enjoy the special status we would like it to have: so far, "possible" is just a word that we --the game theorists--and the players can use in the same way. The meaning of this word will be given to it by the actual behavior of both the players and ourselves. (Although, to a certain extent, assumption (3) already says something about the set P beyond its mere definition.) As we will shortly see, the set P will be incorporated into game-theoretic axioms to make it meaningful to the players; however, we, as outside observers, will have to be interested in this set P as well. Namely, a fact that is proven to hold at each state of the world in P would have to be construed as "true," i.e., as an implication of the theory.

Assuming that the players are endowed with a reasonable reasoning ability, which is also common knowledge, we deduce that whatever we can prove to be impossible (namely, whatever state of the world we can prove is not in P) can also be proven by the players, and has to coincide with their knowledge of the definition of P. However, the definition of P--as known to the players--may be more restrictive. There may be sets satisfying our axioms, the occurrence of which cannot be proved from game-theoretic axioms. To avoid unwarranted exclusion of states we have to explicitly assume that all players consider as possible every state that cannot be proved impossible by game-theoretic assumptions. (Which are also assumed to be common knowledge, as in Gilboa-Schmeidler (1988).) We will assume that these assumptions are parameterized by an event A: for every A (to be thought of as a candidate for the set of possible states P) each axiom may have a different meaning. (This will hopefully become clearer in the example of the "common sense" axiom below.)

We are therefore led to the following axiom on the set P: there is a finite sequence $S = A_0 \supset A_1 \supset \ldots \supset A_k = P$ such that A_i can be proven from A_{i-1} and the game-theoretic assumptions for A_{i-1} $(1 \leq i \leqq k)$, and this chain is maximal with respect to this property. (I.e., no proper subset of P can be proved from P and the game-theoretic assumptions corresponding to P.)

We can now describe a notion of rationality which may be common knowledge without leading to a contradiction: this assumption is close to the "common sense" assumption in Gilboa-Schmeidler (1988), and it basically says that players do not choose dominated strategies. We have to be more specific here and understand domination in the following way: an action x (of player i) is dominated by an action y (of the same player) at a decision node n with respect to a set A if at n player i knows that at every state of the world compatible with n (i.e., contained in the event associated with n), which is also in A, the action y guarantees him/her a strictly higher payoff than x. With this definition in mind, let us now define the axiom of **common sense with respect to A** to be the following: for every player i, and every node n, **if** (the event associated with) **n is A-possible** (namely, intersects A) and if an action x is strictly dominated by an action y at n with respect to A, then if i reaches n, i will not choose x.

Let us assume that the only game-theoretic assumptions are common sense and common knowledge (of all the model's assumptions, hence also of itself.) Then it is quite straightforward to translate the well-known backward induction arguments to show that the undominated solutions dictated by them constitute a valid P. Moreover, in the game given above, as in every finite game, P is unique. (In perfect-information extensive-form games without "ties," the unique P is a singleton.)

Thus, we are able to reconstruct the backward induction argument by proving (outside the model, but if you will--also inside it) that the only consistent set P is the singleton at which player I plays left at node 1. The main point is that the contradiction disappears since nodes 2 and 3 are impossible, (and this is common knowledge,) so that the rationality assumption is vacuously satisfied at these nodes. Finally, the advantage this admittedly more cumbersome model has over the previous one is that it is rich enough to describe every relevant and conceivable aspect of each player's decision problem, and thus to formally describe the players' reasoning.

5. Possible Applications of the Possible Set

5.1 Weak Domination

Let us consider weak domination, instead of strict one, in the definition of "common sense." This will serve as another example of the possibility set, showing that it may not be unique, and suggesting a certain refinement of it. The following discussion can also be viewed as another theoretical application of the notion of the possibility set, in which it provides some further insight on the elimination of weakly dominated strategies.

Consider the following normal-form game:

		player II	
		L	R
player I	T	(1,1)	(1,1)
	B	(1,1)	(0,0)

Figure 2

Obviously, player I's B is weakly dominated by T, and similarly player II's R is weakly dominated by L. There are three valid possibility sets: $P_1 = \{(T,L),(T,R)\}$, $P_2 = \{(T,L),(B,L)\}$ and $P_3 = \{(T,L)\}$. (In all three cases the chain of reasoning is of length 1. However, one may decide at the first stage whether to use the common sense assumption to eliminate both B and R, or only one of them. At any rate, after the first elimination the resulting strategies define a possibility set.)

At first glance, P_3 seems to be the most appealing one, as it is the only symmetric set (which is a natural choice for a symmetric game,) and the outcome it predicts seems plausible. However, upon a more careful scrutiny its validity as exhausting all that may possibly occur appears somewhat dubious: precisely if the players know (as opposed to "believe with a high probability") that (T,L) should be the outcome of the game, this knowledge cannot be justified: if player I, for instance, actually knows that player II is about to play L, there is no reason for him not to play B. In a way, the very knowledge of the "theory" represented by P_3 casts a shadow of doubt on this theory. (Note that this is a considerably less fundamental flaw than being a "self-refuting theory" in the sense of Bicchieri (1989): here the theory is logically consistent; it only seems somewhat arbitrary.)

If we wish to avoid this type of problem we can, in general, propose the following definition: a possibility set is **maximal** if it is maximal with respect to set inclusion. Obviously, P_3 is not maximal, while P_1 and P_2 are.

The fact that common knowledge of common sense--with weak domination rather than strict one--does not yield a unique maximal possibility set, nor a symmetric one for symmetric games, may be considered a theoretical disadvantage of weak domination. Naturally, this is closely related to the fact that weak domination is not inherited by subgames, which also implies that reduced games (with respect to it) are not unique. See Gilboa-Kalai-Zemel (1989) for further discussion.

5.2 Imperfectly Rational Players

The framework described above may be used to describe other models, and, in particular, ones which make less stringent rationality assumptions. For instance, by relaxing the assumption that the possibility set is common knowledge, one may assume that all players are actually rational, and even that this fact is known by all to a certain degree, but that it fails to be common knowledge. Introducing probability into the model may also allow us to quantify the extent of irrationality by the subjective probabilities of those states of the world (which are, in fact, impossible) and the level of knowledge at which they occur. (For instance, one may argue that a system of beliefs in which every player is actually rational, but suspects the others of being irrational, is, on the whole, "more rational" than one in which the players actually are irrational.)

These ideas have, of course, been suggested before (see, for instance, Aumann (1988)), but it seems it would be difficult to formally capture the distinction between violations of rationality occurring at different levels of knowledge without having the distinction between "possible" and "conceivable" states of the world as a basis.

References

Aumann, R. J. (1974), "Subjectivity and Correlation in Randomized Strategies," **Journal of Mathematical Economics 1**, 67-95.

Aumann, R. J. (1976), "Agreeing to Disagree," **Annals of Statistics 4**, 1236-1239.

Aumann, R. J. (1987), "Correlated Equilibrium as an Expression of Bayesian Rationality," **Econometrica 55**, 1-18.

Aumann, R. J. (1988), "Irrationality in Game Theory," a paper presented at the OSU conference on game theory, July 1988.

Aumann, R. J., and S. Sorin (1989), "Cooperation and Bounded Recall," **Games and Economic Behavior 1**, 5-39.

Bicchieri, C. (1988a), "Strategic Behavior and Counterfactuals," **Synthese 76**, 135-169.

Bicchieri, C. (1988b), "Common Knowledge and Backward Induction: A Solution to the Paradox," in the **Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge**, edited by M. Vardi, Morgan-Kaufmann, 381-393.

Bicchieri, C. (1988c), "Backward Induction without Common Knowledge", in the **Proceedings of the Philosophy of Science Association Meeting**, Oct. 1988.

Bicchieri, C. (1989), "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," **Erkenntnis 30**, 69-85.

Bonanno, G. (1988), "The Logic of Rational Play in Games of Perfect Information," mimeo.

Gilboa, I. (1988), "Information and Meta-Information," in the **Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge**, edited by M. Vardi, Morgan-Kaufmann, 227-243.

Gilboa, I., E. Kalai and E. Zemel (1989), "On the Order of Eliminating Dominated Strategies," Northwestern University discussion paper.

Gilboa, I. and D. Schmeidler (1988), "Information-Dependent Games: Can Common Sense be Common Knowledge?" **Economics Letters 27**, 215-221.

Kaneko, M. (1987), "Structural Common Knowledge and Factual Common Knowledge," RUEE Working Paper No. 87-27, Hitotsubashi University.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson (1982), "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," **Journal of Economic Theory 27**, 245-252.

Reny, P. (1988), "Rationality, Common Knowledge and the Theory of Games," mimeo.

Rosenthal, R. W. (1981), "Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox," **Journal of Economic Theory 25**, 92-100.

Savage, L. J. (1954), **The Foundations of Statistics**, Wiley, NY.

Selten, R. (1978), "The Chain-Store Paradox," **Theory and Decision 9**, 127-1589.