Title
Incorporating Behavioral Assumptions Into Game Theory

Permalink
https://escholarship.org/uc/item/54f523db

Author
Rabin, Matthew

Publication Date
1992-07-01

Peer reviewed
Incorporating Behavioral Assumptions Into Game Theory

Matthew Rabin
Economics Department
University of California at Berkeley

July 1992

Key words: behavioral assumptions, coordination, Nash equilibrium, Pareto efficiency, rationalizability

JEL Classification: C72

This paper is based in part on Chapters 1 - 3 of my dissertation (Rabin [1989]) completed at MIT. I thank Eddie Dekel-Tabak, Joe Farrell, Oliver Hart, Joel Sobel, and especially Drew Fudenberg for their contributions, as well as the National Science Foundation and the Alfred P. Sloan Foundation for financial support during that period.
Abstract

By now, most game theorists are familiar with an apparent shortcoming of non-cooperative game theory: Using even the most strained arguments about what rationality implies, the analysis of many games does not yield sharp predictions. I believe that game theorists must make their peace with these limits, and incorporate "behavioral assumptions" — assumptions that restrict players’ beliefs due to a shared perception of how people behave.

In this paper I discuss an approach to formulating solution concepts that combines the assumption of rationality with such behavioral assumptions. To do so, I suppose players share the belief that they will behave consistently with some limited set of strategies. I then analyze their behavior as if these focal strategies constituted the actual game. For such behavioral restrictions to be rational, I check that no player would wish to deviate given the real, unrestricted game. I call any set of predictions that can be constructed in this way a Consistent Behavioral Theory.

I illustrate my approach by constructing equilibrium and non-equilibrium solution concepts that incorporate the behavioral assumption that players tend to focus their behavior and beliefs on Pareto-efficient Nash equilibria.
I. Introduction

The standard approach to making predictions in non-cooperative game theory is to invoke internal consistency—behavior is only ruled out when we can argue that if players came to believe in the behavior, then at least one player would wish to deviate. Such internal-consistency arguments clearly underlie the solution concept rationalizability, and arguably underlie most prevalent game-theoretic solution concepts.\(^1\) By now, most of us perceive an apparent shortcoming of this approach to non-cooperative game theory: Using even the most strained arguments about what rationality implies, analyses of many games do not yield sharp predictions.

I believe that game theorists must make their peace with these limits. We ought incorporate "behavioral assumptions"—assumptions about how play might be limited not solely by the structure of a game, but also by players' shared assumptions about what is standard behavior. Thus, we might rule out some internally consistent behavior because we believe it is implausible that the players will come to believe in such behavior.

Recently, the use of behavioral assumptions has become widely accepted in the literature on "cheap talk."\(^2\) When communication is not costly, arguments

---

\(^1\) By prevalent, I include solution concepts such as Nash equilibrium, Selten (1975), Kreps and Wilson (1982), Cho and Kreps (1987), and Kohlberg and Mertens (1986), and Bernheim (1984) and Pearce (1984). Harsanyi and Selten (1988) argue for a theory of rational play that predicts a unique outcome in every game. In my view, theirs is a combination of a theory of rationality and a behavioral theory. Minimally, a solution concept purporting to rely on rationality alone can never rule out strict Nash equilibria. If players believed with common knowledge that a particular strict Nash equilibrium would be played, no theory of individual rationality could dictate that any player would wish to deviate. Harsanyi and Selten's algorithm for equilibrium selection rules out strict Nash equilibria.

from the point of view of internal consistency can never guarantee communication. This literature studies how we can make stronger predictions about outcomes in many situations by making the behavioral assumption that people expect truthful communication.

Consider Example 1, which represents a simple, simultaneous-move coordination game preceded by an opportunity for player 1 to make a suggestion on the play of the game. Most of us would have little trouble believing that the outcome here will be that player 1 suggests either (U,L) or (D,R), and the players play the equilibrium that player 1 suggests. Moreover, if we realistically assumed a much richer vocabulary, we still would believe that the players would coordinate on one of the two efficient equilibria. Yet no solution concept from rationalizability to Nash equilibrium to strategic stability guarantees any communication here. To predict that communication and coordination would occur in Example 1, we must invoke behavioral assumptions.

Likewise, in applications of game theory more generally, we may be able to invoke simple behavioral regularities to realize that certain internally consistent outcomes in an economic situation are likely to occur. In this paper, I discuss one means of formulating such behavioral assumptions. While my purpose is ultimately to incorporate empirically valid behavioral assumptions, my focus in this paper is on how one adds behavioral assumptions without abandoning standard assumptions about rationality. Just as we have developed a language for what types of structural assumptions are coherent

---

3 Could such a conclusion be squeezed out of some internal-consistency arguments based on the payoffs and the structure of the game? I would argue No, and I believe there is a simple "proof" of this perspective. Importantly, nothing in the structure of the game or in the payoffs says that player 2 speaks English. Suppose he does not. Would our prediction be that the players would coordinate? No—we invoke the behavioral assumption of meaningful communication only if we believe that the players share a common language. But whether or not the players share a common language appears nowhere in the traditional description of the game; we must invoke it as an additional assumption, and its validity cannot be inferred from the payoffs.
(e.g., perfect recall, common priors), so too I believe that we can develop a language for what types of behavioral assumptions are coherent.

My approach is simple. First, I assume that, while players enter into a strategic situation with a set of possible strategies, they contemplate their choices in terms of some set of focal strategies, which they imagine are the only strategies that are likely to be played. Once they so imagine, they play the game as if this focal set of strategies is the real game. In Example 1, for instance, we could postulate that only strategies in which player 2 will behave according to player 1's suggestion will be focal. Then, if player 1 chooses his strategy believing that player 2 is behaving this way, he will surely propose one of the efficient equilibria.

To illustrate this approach in a non-communicational example, consider Example 2. Among the set of pure- and mixed-strategy Nash equilibria in this game, two Nash equilibria—(U,L) and (M,C)—are Pareto-efficient.\(^4\) It is common to suppose that players will focus especially on those equilibria that are Pareto efficient. Suppose we argue that such a focus is likely in this game, but that we do not believe that the players perceive either (U,L) or (M,C) as being uniquely focal. It is therefore natural to believe that player 1 will select a strategy from \(\{U,M\}\), and player 2 will select a strategy from \(\{L,C\}\). Formally, we might hypothesize that the players would play this game "as if" there were no option to play either D or R.

My approach to formalizing this intuition is to formally invoke rationalizability in the restricted game. In Example 2, for instance, we can assume that players will act as if the game is restricted to the strategies \(\{U,M\}\) for player 1 and \(\{L,C\}\) for player 2. Then, they will choose their behavior from the set of rationalizable predictions on this hypothetical

---

\(^4\) Throughout the paper, when I refer to "Pareto-efficient equilibria" I mean Pareto-efficient among the set of Nash equilibria.
Are all such solution concepts that apply rationalizability to a hypothetical game sensible? Example 3 helps illustrate that sometimes such a solution concept is not reasonable. As in Example 2, the two Pareto-efficient Nash equilibria in Example 3 are \((U,L)\) and \((M,C)\). Suppose that, as in Example 2, we think it natural for the players to focus in on the hypothetical game in which they cannot play \(D\) or \(R\).

Applying rationalizability to this hypothetical game, we would predict that any combination of \(U\) and \(M\) might be played by player 1, and any combination of \(L\) and \(C\) might be played by player 2. For instance, player 1 might believe that player 2 is playing \(L\) with probability .5, and \(C\) with probability .5. In the hypothetical game, player 1 would respond to such beliefs by playing \(U\). Yet in the real game—where player 1 is allowed to choose the strategy \(D\)—he would in fact choose \(D\) in response to these beliefs. Using similar arguments for player 2, we realize that even if the players will "tend to focus" their beliefs and behavior on \((U,L)\) and \((M,C)\)—but do not know which of these equilibria is most reasonable—we cannot rule out the outcome \((D,R)\).

Thus, to check that a behavioral theory is consistent with rationality, I impose the conditions 1) that it be equivalent to rationalizability on some hypothetically restricted game, but also 2) that it be consistent with rationality on the actual game. In Sections II and III, I develop the concept of Consistent Behavioral Theories to formalize these notions of consistency.\(^5\) The set of predictions \(\{U,M\} \times \{L,C\} \) constitute a consistent behavioral theory.

---

\(^5\) The consistency criteria I define correspond to those I have defined earlier in Rabin [1989], and are quite similar to those defined in Basu and Weibull [1990] and Gul [1991]. Moreover, they are the criteria used in Rabin [1990, 1991a, 1991b] in defining communicational solution concepts. Papers that incorporate behavioral restrictions along with rationalizability in non-communicational settings include Cho [1992] and Watson [1992a, 1992b].
in Example 2, but not in Example 3.

Consistent behavioral theories are behavioral analogs of rationalizability, as defined by Bernheim [1984] and Pearce [1984]. The question arises: What are the behavioral analogs to Nash equilibrium, perfect equilibrium, etc.? This is harder to say, because many solution concepts such as Nash equilibrium have been motivated and interpreted by different people in different ways. Nonetheless, I propose in Section IV the definition Behavioral Equilibrium Theory as the behavioral analog to Nash equilibrium. I contrast behavioral equilibrium theories to equilibrium concepts that seem to incorporate restrictions beyond both the Nash equilibrium hypothesis and assumption of common-knowledge behavioral restrictions.

In Section V, I discuss a means of developing consistent behavioral theories using an algorithm which—beginning with a particular set of focal strategies—iteratively expands the set of strategies by adding in all strategies that are rational responses to beliefs by the players. In Section VI, I use this method to develop a consistent behavioral theory that incorporates the idea that players tend to focus their beliefs and behavior on Pareto-efficient Nash equilibria. I also discuss more generally the role of behavioral assumptions in helping players coordinate on efficient outcomes.

I conclude in Section VII with a discussion of possible extensions and refinements of the approach, highlighting the fact that the model of this paper focuses solely on behavioral analogs to rationalizability and Nash equilibrium, rather than on the well-known refinements of these concepts.
Consider a two-player, normal-form game \( G \), where players 1 and 2 have the set of pure strategies \( S_1 \) and \( S_2 \) with \( n_1 \) and \( n_2 \) elements, respectively.\(^6\) For each ordered pair of strategies, \((s_1, s_2) \in (S_1, S_2)\), let \((u^1(s_1, s_2), u^2(s_1, s_2))\) be the payoffs for each of the two players. Let \( \Sigma_1 \) and \( \Sigma_2 \) be the sets of mixed strategies for players 1 and 2. The expected payoffs over mixed strategies \((\sigma_1, \sigma_2) \in (\Sigma_1, \Sigma_2)\) can be represented by the functions \((v^1(\sigma_1, \sigma_2), v^2(\sigma_1, \sigma_2))\), derived from the utility functions \( u^1 \) and \( u^2 \).

The approach I develop will incorporate beliefs more explicitly than is conventional; because of this, I now develop some non-standard notation and jargon. Define a plan \( \Theta_i \) for each player \( i \) as a pair \((\sigma, \mu)\), where \( \sigma \in \Sigma_i \) is a \( 1 \times n_i \) vector representing a mixed strategy for player \( i \), and \( \mu \in \Sigma_j \) is a \( 1 \times n_1 \) vector representing player \( i \)'s beliefs about the strategy employed by player \( j \). A player's plan includes both the strategy he chooses, and his beliefs about the other player's strategy. Let \( \Theta_i \) be the set of all possible plans for player \( i \); \( \Theta_i \) can be represented as a set of points in \( \mathbb{R}^{n_1 + n_2} \), representing both strategies and beliefs as probability distributions over \( S_1 \) and \( S_2 \). A prediction for player \( i \), \( p_i \), is a probability distribution over the set of all possible elements in \( \Theta_i \). Let \( P_1 \) be the set of all possible predictions over \( \Theta_i \). For every \( p_i \in P_1 \), and for every \( \theta \in \Theta_i \), define \( p_i(\theta) \) as the probability that \( p_i \) places on \( \theta \). Whereas \((P_1, P_2)\) are sets of probability distributions over plans, it will frequently be useful to separately examine the implied distribution over strategies. For a given \( p_i \in P_1 \), define \( p_i^S \) as the  

---

\(^6\) The formal approach I outline in this paper is for two-person, finite-action games. It could readily be extended to a broader setting, with some complications. Also, while I shall not concentrate on such an interpretation, the approach outlined here can readily be used for incomplete-information games, interpreting strategies as as type-contingent strategies.
probability distribution over player i's strategies \( s_i \in S_i \) derived from \( p_i \):
\[
p_i^S = \sum_{(\sigma, \mu) \in \Theta_i} p(\sigma, \mu) \cdot \sigma.
\]

I say that a pair of sets \((P_1, P_2)\) of predictions is a theory. Because predictions over the strategies of the players will often be of interest, I define the sets \((P_1^S, P_2^S)\), where \( P_i^S = \{p_i^S| p_i \in P_i\} \). These are the more traditional data of analysis.

A central tenet of game theory is that players behave rationally: For all plans \( \theta_i = (\sigma, \mu) \) given positive weight by some prediction, the strategy \( \sigma \) ought to be optimal given the beliefs \( \mu \). Formally, \( \sigma^* \) is a best response for player 1 to the beliefs \( \mu \) if \( \sigma^* \in \arg\max_{\sigma \in \Sigma_1} v_1(\sigma, \mu) \). Define the sets of predictions \((P_1^*, P_2^*) \subseteq (P_1, P_2)\) as the set of predictions that put positive weight only on plans for which the strategies are best responses to beliefs. \((P_1^*, P_2^*)\) are the sets of predictions consistent with players being rational.

To define those predictions that are consistent with common knowledge of rationality, I shall construct the concept that is the analog within this framework of the solution concept rationalizability developed by Bernheim [1984] and Pearce [1984]. The construction here parallels one of the two equivalent constructions of rationalizability in Pearce [1984]. I begin with the definition of sets that have the best-response property:

Definition 1:

The sets of predictions \((A_1, A_2) \subseteq (P_1^*, P_2^*)\) have the best-response property iff for \( i=1,2 \), for \( j \neq i \), for all \( \theta_i = (\sigma, \mu) \in \Theta_i \) with \( p(\theta_i) > 0 \) for some \( p \in A_i \), there exists a \( p_j \in A_j \) such that \( \mu = p_j^S \).

Sets of predictions for players have the best-response property if all beliefs the theory allows for one player correspond to some behavior that the theory allows for the other player. If a strategy belongs to any sets with the
best-response property, then it is consistent with common knowledge of rationality. This is because any strategy in $A_1$ is a utility-maximizing response for player 1 to some beliefs over the plans in $A_2$ employed by player 2, each of which are optimal responses to some beliefs over the plans in $A_1$ employed by player 1, each of which are optimal responses to some beliefs over the strategies in $A_2$ employed by player 2, etc. So long as two sets meet the best response property, each prediction for each player corresponds to some such infinite sequence of rational strategy choices.

From Definition 1, it is straightforward to characterize the set of all predictions consistent with common knowledge of rationality.

**Definition 2:**

The set of Rationalizable Predictions is $(R_1,R_2)$, where $R_1 = \{ p_1 | p_1 \in A_1$ for some $(A_1,A_2)$ with the best-response property $\}$.

Pearce illustrates that this definition of rationalizability is equivalent to the process of iterated deletion of dominated strategies.

Clearly, the set of rationalizable strategies has the best-response property. It can also be shown that the set of rationalizability has several other features. Because these other features are useful for some of the definitions below, I now present them:

**Definition 3:**

The sets of predictions $(A_1,A_2) \subseteq (\mathcal{P}_1,\mathcal{P}_2)$ have the covering property if for $i=1,2$, for $j \neq i$, for all $p_j \in A_j$, there exists some $p_i \in \Theta_i$ such that for all $(\sigma,\mu)$ for which $p_i(\sigma,\mu) > 0$, $\mu = p_j$.

Sets of predictions have the covering property if, for any behavior that
the theory allows for one player, the theory allows that the other player believes with certainty in that behavior. This is sort of the converse of the best-response property.

**Definition 4:**

\((A_1, A_2)\) is said to be convex if for all \(p^*_i \in P_i\), and for all \(\lambda \in [0,1]\), then \(p^{**}_i \in A_i\), where for all \(\theta_i \in \Theta_i\), \(p^{**}_i(\theta_i) = \lambda \cdot p_i(\theta_i) + (1-\lambda) \cdot p_i(\theta_i)\).

By imposing convexity of predictions for a player, we essentially impose convexity on the sets of permissible beliefs for both the game theorist and the other players.

**Definition 5:**

For the sets \((A_1, A_2)\), say that \(\sigma_i\) is a component strategy for player \(i\) if there exists a prediction \(p^*_i \in A_i\) and \(\mu \in P^S_j\) such that \(p^*_i(\sigma_i, \mu) > 0\).

\((A_1, A_2)\) is said to be regular if, for \(i = 1, 2\), then \(p_i \in A_i\) if and only if it puts positive weight on plans \((\sigma, \mu)\) such that

1) \(\sigma\) is an optimal response to beliefs \(\mu\).

2) \(\mu \in P^S_j\), and

3) \(\sigma\) is a component strategy for player \(i\).

This essentially says that a strategy is employed against beliefs for which it is optimal if and only if it is employed against all beliefs in the theory for which it is optimal. As the name suggests, this will used as a convenient regularity condition.  

---

7 In the next section, I will present theories in which—unlike with rationalizability—we may exclude some optimal responses by players. In such
It relatively straightforward—and useful for later purposes—to show that the set of rationalizable strategies have all of these properties:

Lemma 1:

The set of rationalizable predictions are convex and regular, and have the best-response and covering properties.\(^8\)

III. Consistent Behavioral Theories

In this section, I present the notion of Consistent Behavioral Theories (CBTs), which formalizes the criteria for behavioral theories that I outlined in Section I.

First, consider some closed, convex subsets of mixed strategies \(Q = (Q_1, Q_2)\), \(Q_k \subseteq \Sigma_k\) for \(k = 1, 2\). Let us consider the implications if the players restrict their beliefs and behavior to only strategies in \(Q\). I restrict attention to convex sets because that is the natural implication of people's belief-formation—as in considering real games, we allow players to have any conceivable mixes of beliefs.

Let \(\Gamma(G, Q)\) be the hypothetical game with strategy sets \(Q\) and payoffs \(U^Q(s)\) such that, for all \(s \in Q\), \(U^Q_k(s) = U_k(s)\). This game is well-defined, with strategy spaces that are subsets of the strategy spaces in \(G\), and payoffs that are identical to \(G\). We can therefore define any solution concept in this hypothetical game. A natural starting place is the solution concept with the

---

cases, the force of this condition is that we only exclude certain strategies for some beliefs only if we exclude them for all beliefs.

\(^8\) The proofs of most of the propositions in this paper follow very closely from the definitions, and from known features of rationalizability and Nash equilibrium. When this is the case, formal proofs are omitted.
most intimate connection to the rationality assumption—rationalizability. For a game G and chosen convex subsets of strategies, Q, let \((R_1(G,Q), R_2(G,Q))\) be the set of rationalizable strategies on the game \(\Gamma(G,Q)\). If we let \(P^*_1(G,Q)\) be the set of rational predictions in the game \(\Gamma(G,Q)\), then \(R_1(G,Q)\) is by construction a subset of \(P^*_1(G,Q)\). Importantly, however, \(R_1(G,Q)\) is not necessarily a subset of \(P^*_1\); this is because we do not know that best responses for player 1 when restricted to strategies in \(Q_1\) are also optimal when he can employ any strategy in \(\Sigma_1\).

I say that any sets of strategies that correspond to rationalizability in some such hypothetical game is a behavioral theory:

**Definition 6:**

The sets of predictions \((P_1, P_2)\) constitute a Behavioral Theory if there exists \((Q_1, Q_2) \subseteq (\Sigma_1, \Sigma_2)\) such that \((P_1, P_2) = (R_1(G,Q), R_2(G,Q))\).

A behavioral theory is a set of predictions that would be consistent with common knowledge of rationality if the players were restricted to employ strategies only in \((Q_1, Q_2)\). Are all behavioral theories reasonable? It is quite possible they are not, if we consider that the real game does not restrict the players to choose from sets of strategies \(Q_k\), but rather from sets \(\Sigma_k\).

Example 3 illustrates the problem: the set of rationalizable predictions on the hypothetical game consisting of strategies \((U,M) \times (L,C)\) are not rational in the actual game. Thus, in such a behavioral theory, not all predictions involve a player responding rationally given his beliefs. The next definition rules out irrational predictions:
Definition 7:

A behavioral theory \((P_1, P_2)\) is a Consistent Behavioral Theory (CBT) if 
\((P_1, P_2) \preceq (P_1^*, P_2^*)\).

This definition simply says that the theory thus derived must have each player responding rationally given his beliefs, given his unrestricted choice of strategies. The idea of a CBT essentially characterizes behavioral theories which can be common knowledge to the players, in the sense that if the players hold any beliefs consistent with the theory, they do not (strictly) prefer to deviate from the theory. In this way, they are similar to rationalizability—if players believe with common knowledge in the set of rationalizable strategies, then no rational player would want to deviate to play a non-rationalizable strategy.

From this relatively simple construction, much can be implied about these sets. Indeed, it is straightforward to show that every CBT has the same properties as the set of rationalizable strategies, as outlined in Section II:

Theorem 1:

If \((P_1, P_2)\) is a CBT, then it is convex and regular, and has the best-response and covering properties.

It turns out that the converse is true: if a pair of sets \((P_1, P_2) \preceq (P_1^*, P_2^*)\) and have these four properties, then they constitute a CBT.
Theorem 2:

If \((P_1, P_2)\) is convex and regular, and has the best-response and covering properties, then \((P_1, P_2)\) is a CBT.⁹

Proof:

Let \((Q_1, Q_2) = (P^S_1, P^S_2)\). Then \((P_1, P_2) = (R_1(G, Q), R_2(G, Q))\). Because \((P_1, P_2) \subseteq (P^*, P^*)\), these sets constitute a CBT.

The proof simply observes that, if we create a hypothetical game from the set of strategies implied by a set of predictions with the four properties, then applying rationalizability on this game will yield the set of predictions we started out with. Then, clearly, this set will be a CBT, because the sets of predictions are rational given the real game.

Thus, these four conditions are necessary and sufficient for a set of predictions to be a CBT. In constructing CBTs, we can therefore look for sets of predictions that meet these criteria, rather than applying the two-step, hypothetical-game formulation.

IV. Equilibrium Theories

The previous section outlined a general approach to formulating solution concepts that are behavioral analogs of rationalizability. In this section, I propose a test for whether an equilibrium theory is a behavioral analog to Nash equilibrium.

Suppose that the players believe with common knowledge that they will

---

⁹ Together, these properties constitute the definition of a regular public theory, as defined in Rabin [1989].
play strategies from some hypothetical game \((Q_1, Q_2)\) meeting the consistency criterion from the previous section. Then we can predict that they will play strategies from the corresponding CBT. Now suppose that the only further hypothesis we have about behavior is that only Nash equilibria will be played. Then an outcome is plausible if and only if it is a Nash equilibrium contained in the CBT. I call such an equilibrium concept that combines Nash equilibrium with some common-knowledge behavioral restrictions a Behavioral Equilibrium Theory (BET).

If an equilibrium solution concept is not a BET, then it must involve a motivation beyond the common-knowledge behavioral restrictions, or beyond the basic equilibrium hypothesis. Suppose—as I will shortly demonstrate can be the case—that some subset of Nash equilibria in a game does not correspond to the set of Nash equilibria for any CBT. Then if we invoke such a subset of Nash equilibria as a solution concept, then we must be invoking some restriction on behavior beyond a common-knowledge behavioral restrictions and the basic equilibrium hypothesis.

Of course, such restrictions might make sense. There exist internal-consistency arguments for eliminating Nash equilibrium, as well as many dynamic stories that attempt to model in strategic situations the implications of evolution or learning over time.\(^\text{10}\) I discuss in the concluding section how one might combine such arguments with behavioral restrictions.

The concept of a BET essentially constitutes a specific method of refining Nash equilibrium based solely on common-knowledge behavioral assumptions. Namely, we should use the behavioral assumptions to first refine

\(^{10}\) Selten [1975], Kreps and Wilson [1982], and Kohlberg and Mertens [1986] are examples of internal-consistency refinements. Fudenberg and Kreps [1988] is an example of a learning-based refinement of Nash equilibrium. My view is that signaling refinements such as Cho and Kreps [1987] and Banks and Sobel [1987] are implicitly dynamically-motivated refinements.
rationalizability, and then use as our equilibrium concept all the Nash equilibria contained in the corresponding CBT.

To illustrate the idea of a BET, consider Example 2 again, and consider also the behavioral assumption that players have a tendency to focus on strategies consistent with Pareto-efficient Nash equilibria. From such a theory, we can consider the CBT that includes all rationalizable predictions in the game excluding beliefs and behavior focused on the strategies D and R.

The set of Nash equilibria consistent with this CBT are the two Pareto-efficient Nash equilibria, and the inefficient mixed-strategy Nash equilibrium \((2/3U,1/3M;1/3L,2/3C)\). That is to say, if we incorporate the idea that players tend to focus on Pareto-efficient Nash equilibria into a common-knowledge behavioral assumption, we discover that the players may still play an inefficient Nash equilibrium. Interestingly, it can be shown that any CBT that contains both the equilibria \((U,L)\) and \((M,C)\) in Example 2 also contains the mixed-strategy Nash equilibrium over these strategies.\(^{11}\) This means that, in this game, the solution concept "Pareto-efficient Nash equilibrium" is not a BET based on any CBT.

Example 3 illustrates the limited powers of behavioral assumptions even more strikingly than Example 2. It can be shown that any CBT containing the two efficient Nash equilibria also contain all Nash equilibria in this game. The assumption that people tend to focus on Pareto-efficient Nash equilibria has no behavioral implications in this game.

I now present a definition and a theorem allowing us to define BETs formally:

---

\(^{11}\) This follows immediately from the convexity condition incorporated into CBTs, because the mixed strategy involves each player believing that the other is playing a convex combination of the other player's pure-strategy Nash-equilibrium strategies.
Definition 8:

For a CBT \((P_1, P_2)\), define its set of equilibrium predictions as those strategy pair \(P^e \subseteq \Sigma_1 \times \Sigma_2\) such that:

\((\sigma_1, \sigma_2) \in P^e\) iff there exists a \((p_1, p_2) \in (P_1, P_2)\) such that

1) for \(i = 1, 2\), \(\sigma_i = p^s_i\); and

2) for \(i = 1, 2\), \(j \neq i\), for all \((\sigma, \mu)\) such that \(p_1(\sigma, \mu) > 0\), \(\mu = \sigma_j\).

The following theorem is an immediate corollary to the fact that the set of rationalizable predictions always includes a Nash equilibrium:

Theorem 3:

If \(P = (P_1, P_2)\) is a CBT, then \(P^e\) is non-empty.

I can formally define a a Behavioral Equilibrium Theory:

Definition 9:

An equilibrium concept \(E \subseteq \Sigma_1 \times \Sigma_2\) is a Behavioral Equilibrium Theory (BET) if there exists a CBT \((P_1, P_2)\) such that \(E = P^e\).

The above argument shows that the theory "Pareto-efficient Nash equilibrium" is not a BET—it must involve, in addition to a behavioral assumption, some theory of coordination beyond the Nash-equilibrium hypothesis.\(^{12}\) I discuss in Section VI a CBT and BET we can formulate that assumes that Pareto-efficient Nash equilibria are focal in a game.

In all of these examples, we could of course argue that players are likely

---

\(^{12}\) Rabin [1989] provides an algorithm for checking as to whether a set of equilibria corresponds to all the Nash equilibria from an CBT. The method involves a generalization of the algorithm developed in the next section.
to focus on a particular Nash equilibrium. It would then be a consistent behavioral theory to simply predict this equilibrium as the outcome. Indeed, any time we propose as our theory in a game that a particular Nash equilibrium will occur, then that pair of predictions meets the criteria of a CBT:

Theorem 4:

For any Nash equilibrium \((\sigma_1, \sigma_2)\), the theory \((\{p_1\}, \{p_2\})\), where \(p_1(\sigma_1, \sigma_2) = 1\) and \(p_2(\sigma_2, \sigma_1) = 1\), is a CBT and a BET.

If the players know which equilibrium will obtain, then predicting this equilibrium can be a sound behavioral theory—it is both a CBT and a BET. Thus, every Nash equilibrium in every game is contained in at least two BETs—the prediction by itself, and the entire set of Nash equilibria. The "restrictiveness" of the BET approach therefore is that it says which combinations of equilibria can together exclusively and exhaustively constitute a solution concept.

V. Constructing CBTs from Focal Sets

While the formal argument for CBTs essentially involves beginning with a set of strategies, and then "reducing" this set, much of the intuition in the examples have worked in the opposite direction—they began with a particular focal set of strategies, and expanded the set to include more predictions.

This procedure can be formalized:
Definition 10:

Choose any subsets of predictions \((A_1, A_2) \subseteq (\mathcal{P}_1^*, \mathcal{P}_2^*)\). Then let the Maximal Expansion of \((A_1, A_2)\) be the set \(Z(A_1, A_2) \subseteq (\mathcal{P}_1^*, \mathcal{P}_2^*)\) constructed as follows:

Let \((A_1(0), A_2(0)) = (A_1, A_2)\).

Then, for integers \(k > 0, \ i = 1, 2, \ j \neq i,\)

Let \(\mathcal{P}_j(k-1)\) be the convex hull of \(A_j^S(k-1)\).

Let \(A_i(k) = \{p_1 \in \mathcal{P}_1^* \mid \forall (\sigma, \mu) \text{ such that } p_1(\sigma, \mu) > 0, \mu \in A_j^S(k-1)\}\).

Then \(Z(A_1, A_2) = \lim_{k \to \infty} (A_1(k), A_2(k))\).

This procedure involves beginning with a set of rational predictions for each player. We then iteratively add all predictions for each player that involve best responses to beliefs that are consistent with the behavior of the other player. We do so until we add in no more predictions for either player. The use of this procedure is indicated by the following theorem:

Theorem 5:

If a set of predictions \((A_1, A_2) \subseteq (\mathcal{P}_1^*, \mathcal{P}_2^*)\) has the best-response property, then its Maximal Expansion \(Z(A_1, A_2)\) is a CBT.

Proof:

The maximal expansion will clearly itself have the best-response property; moreover, by the fact the maximal expansion is constructed as a limit, it will have the other three properties.

Note that the set of predictions for players corresponding to any subset of Nash equilibria constitute a set with the best-response property. This

---

13 It is fairly straightforward to see that this process will end in a finite number of iterations in any finite-action game.
immediately implies that if we apply a maximal expansion to any subset of Nash equilibria, we will end up with a CBT. I apply this method to the set of Pareto-efficient Nash equilibria in the next section.

VI. Coordinating on Efficient Outcomes

The examples in this paper have emphasized the role of behavioral assumptions in helping players coordinate on efficient outcomes. In Example 1, I argued that the existence of a meaningful, common language can guarantee that players will coordinate on an efficient equilibrium. In Examples 2 and 3, I discussed the possibility for increased coordination in games without communication, by assuming that Pareto-efficient equilibria are natural focal points. I showed that--unless there is a single Pareto-efficient equilibrium that players find focal--this focalness does not itself guarantee that the players will in fact play one of the Pareto-efficient Nash equilibria. To consider the implications of this behavioral assumption generally, I now define a solution concept that assumes that players focus in on Pareto-efficient equilibria:

Definition 11:
For \( i = 1,2 \), let \( A_i \) be the set of predictions for player \( i \) consisting of the strategy-belief pairs consistent with the set of Pareto-efficient Nash equilibria in the game. Then Pareto-Focal Rationalizability is the set of predictions corresponding to the maximal expansion of \((A_1, A_2)\).

14 This theme underlies most attempts to incorporate behavioral assumptions. My papers on communication (Rabin [1990, 1991a, 1991b]), and papers such as Cho [1992] and Watson [1991] also seem to relate to the theme of efficiency.
We know that Pareto-focal rationalizability is a CBT because of the theorem in the previous section. In Example 2, Pareto-focal rationalizability refines rationalizability. In Example 3, Pareto-focal rationalizability is equivalent to rationalizability.

We can obviously define the BET based on this CBT:

**Definition 12:**

A Nash equilibrium is a Pareto-Focal Nash Equilibrium if it is an equilibrium prediction in the set of Pareto-focal rationalizable predictions.

Because Pareto-focal rationalizability does not refine rationalizability in Example 3, Pareto-focal Nash equilibrium is obviously the same as Nash equilibrium in this game. Essentially, we cannot with a behavioral assumption rule out any inefficient Nash equilibria in this game unless we also rule out one of the efficient Nash equilibria. By contrast, while Pareto-focal Nash equilibrium does not guarantee us full efficiency in Example 2, we do rule out the most inefficient Nash equilibria. In fact, Example 2 well illustrates the fact that the consistency criteria incorporated into BETs in general allow us to refine our predictions in many games, but also tend to restrict us from choosing arbitrary sets of Nash equilibria as theories. The games for which Pareto-focal Nash equilibrium clearly has the most power are those in which there are multiple Nash equilibria, but in which there is a unique Pareto-efficient Nash equilibrium.

Applications of CBTs to communicational issues help us see how assuming that players share a common language can lead us to refine Nash equilibrium in such a way that we can predict greater efficiency. Consider my model in Rabin [1991a] (which generalizes an example presented in Farrell [1987]). This paper
posits that two players communicate extensively before they play any given complete-information game. It then posits an assumption about how players use language to focus their behavior. I show that any Nash equilibrium in which players use language in this way will be inconsistent with certain inefficient outcomes. But I also show we cannot guarantee that only fully efficient equilibria are played—players might agree to, and play, a Pareto-inefficient Nash equilibrium.

Indeed, from my papers on pre-game communication, and from examining the CBT Pareto-focal rationalizability, a theme emerges. Basically, the general disposition of players to play efficient equilibria does not translate into the ability to play those efficient equilibria, even when they can communicate extensively. Rather, unless there is a uniquely focal efficient equilibrium, we can only guarantee that some of the more inefficient equilibria can be ruled out.

VII. Conclusion

This paper has attempted to outline a set of consistency criteria for incorporating behavioral assumptions into formal game-theoretic analysis. Implicit throughout has been the idea that rationalizability and Nash equilibrium are the appropriate non-equilibrium and equilibrium solution concepts that incorporate basic internal-consistency arguments. Yet such a view ignores the many compelling arguments for stronger internal-consistency criteria. These arguments have yielded solution concepts such as iterated weak dominance, and equilibrium concepts such as trembling-hand perfection, sequential equilibrium, and strategic stability.

My approach allows us to incorporate iterated weak dominance into a
solution concept; we can begin by eliminating from a game all strategies that do not survive the iterated deletion of weakly dominated strategies. We can then form a hypothetical game by further eliminating strategies based on behavioral assumptions, and apply rationalizability to create a behavioral theory. Then this behavioral theory would pass the test for consistency proposed earlier—that all proposed plans be rational given the overall game—if and only if it would pass the same test where we ignored those strategies eliminated in process of iterated deletion of weakly dominated strategies.

How might we combine equilibrium refinements with behavioral assumptions? One possibility might be to simply combine such refinements in the same way as we combined Nash equilibrium with behavioral assumptions—first we can refine rationalizability using our behavioral assumptions, and then we can select those equilibria in the theory that meet the criteria of the proposed refinement. We could, for instance, simply select the set of perfect equilibria among Pareto-focal rationalizable predictions, rather than the set of Nash equilibria.

Though this approach may be promising, it can be problematic. Consider Example 4, and suppose that we construct a CBT by applying rationalizability to the hypothetical game \((U,M) \times (L,C)\). In this game, rationalizability would eliminate nothing. Thus, the corresponding BET would include two Nash equilibria—\((U,L)\) and \((M,C)\).

What would we get if we applied trembling-hand perfection here? We must be careful to specify whether we would allow "trembles" to be over all strategies in the game, or simply over the strategies in the hypothetical game. If we allowed trembles over all strategies in the game, neither of these Nash equilibria are trembling-hand perfect. In general, because our initial internal-consistency test for constructing the CBT did not incorporate the
same test we are applying to eliminate equilibria, there is no reason that existence should be guaranteed.

Of course, we could apply the trembles only to those strategies making up the hypothetical game. In Example 4, this would mean that we would select only \((U,L)\). Yet this is awkward—seemingly we have made a refinement among the behaviorally plausible strategies based on an internal-consistency argument that we have ignored when testing the overall consistency of the behavioral assumption itself.

The proper approach to combining internal-consistency based refinements with behavioral assumptions would be to somehow incorporate the same criteria motivating the equilibrium refinement into the test for the consistency of CBTs. Essentially, we can define sets \((P_{1}^{**},P_{2}^{**})\) that—instead of having the simple best-response property incorporated into the set of rational predictions \((P_{1}^{*},P_{2}^{*})\)—incorporate a non-equilibrium version of the consistency criteria incorporated into the proposed equilibrium refinement. Our consistency criterion for a proposed behavioral theory could then be simply whether it is a subset of \((P_{1}^{**},P_{2}^{**})\).
Examples

1

"Let's play (U,L)"

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>2,2</td>
<td>0,0</td>
<td>-5,-5</td>
</tr>
<tr>
<td>M</td>
<td>0,0</td>
<td>1,1</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>-5,-5</td>
<td>0,0</td>
<td>2,2</td>
</tr>
</tbody>
</table>

"Let's play (D,R)"

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>2,2</td>
<td>0,0</td>
<td>-5,-5</td>
</tr>
<tr>
<td>M</td>
<td>0,0</td>
<td>1,1</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>-5,-5</td>
<td>0,0</td>
<td>2,2</td>
</tr>
</tbody>
</table>

"Let's play (M,C)"

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>2,2</td>
<td>0,0</td>
<td>-5,-5</td>
</tr>
<tr>
<td>M</td>
<td>0,0</td>
<td>1,1</td>
<td>0,0</td>
</tr>
<tr>
<td>D</td>
<td>-5,-5</td>
<td>0,0</td>
<td>2,2</td>
</tr>
</tbody>
</table>

Example 1
### Example 2

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td>2,1</td>
<td>0,0</td>
<td>-5,-5</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>0,0</td>
<td>1,2</td>
<td>-5,-5</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>-5,-5</td>
<td>-5,-5</td>
<td>-2,-2</td>
</tr>
</tbody>
</table>

### Example 3

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td>30,10</td>
<td>0,0</td>
<td>-5,9</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>0,0</td>
<td>10,30</td>
<td>-5,9</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>9,-5</td>
<td>9,-5</td>
<td>-2,-2</td>
</tr>
</tbody>
</table>

### Example 4

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>C</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong></td>
<td>3,1</td>
<td>1,0</td>
<td>0,1</td>
</tr>
<tr>
<td><strong>M</strong></td>
<td>0,0</td>
<td>1,3</td>
<td>1,0</td>
</tr>
<tr>
<td><strong>D</strong></td>
<td>3,0</td>
<td>1,0</td>
<td>1,2</td>
</tr>
</tbody>
</table>
References


Banks, Jeffrey S. and Randall L. Calvert, "Communication and Efficiency in Coordination Games", Department of Political Science, University of Rochester, Working Paper No. 156, August 1989


Milgrom, Paul and John Roberts, "Rationalizability, Learning and Equilibrium in Games with Strategic Complementarities," mimeo, Stanford University, 1989.


July 7, 1992

Working Paper Series
Department of Economics
University of California, Berkeley

Individual copies are available for $3.50 within the USA and Canada; $6.00 for Europe and South America; and $7.00 for all other areas. Papers may be obtained from the Institute of Business and Economic Research: send requests to IBER, 156 Barrows Hall, University of California, Berkeley CA 94720. Prepayment is required. Make checks or money orders payable to "The Regents of the University of California."


