Rationality and Coherent Theories of Strategic Behavior[†]

Faruk Gul
Northwestern University

November 1999

[†] This paper relies heavily on the work of Douglas Bernheim, David Pearce and Phil Reny. I have also benefited from discussions with Dilip Abreu, Douglas Bernheim, Ken Binmore, Eddie Dekel-Tabak, David Kreps, David Pearce, Andrew Postlewaite, Phil Reny, Hugo Sonnenschein and Robert Wilson on Nash equilibrium and subgame perfection. Geir Asheim, Chris Avery, Outi Lantto and Sonia Weyers have also provided many valuable comments and criticisms. Financial support from the Alfred P. Sloan Foundation and the National Science Foundation is gratefully acknowledged.

Running head: Rationality and Coherent Theories

Faruk Gul
Department of Economics
Northwestern University
Evanston, IL 60208-2600

Abstract

A non-equilibrium model of rational strategic behavior that can be viewed as a refinement of (normal form) rationalizability is developed for both normal form and extensive form games. This solution concept is called a τ-theory and is used to analyze the main concerns of the Nash equilibrium refinements literature such as dominance, iterative dominance, extensive form rationality, invariance, and backward induction. The relationship between τ-theories and dynamic learning is investigated.

JEL classification number C72

1. Introduction

In their work on rationalizability, Bernheim [5] and Pearce [19] have shown that Nash equilibrium behavior can not be deduced solely from assumptions regarding the rationality of players and their knowledge of the rationality of their opponents. In particular, they have shown that all rationalizable strategies, and only rationalizable strategies, are consistent with the assumption that rationality is common knowledge.

Identification of the implications of the common knowledge of rationality is undoubtedly a most significant contribution to the theory of strategic behavior. Nevertheless, both Bernheim and Pearce have noted that game theory need not restrict itself to this task and that other factors may well be incorporated into the analysis. Specifically, Bernheim has analyzed how learning and dynamics could impose restrictions on the beliefs of rational players about the behavior of other rational players, while Pearce has considered the implications of the extensive form. Both have dealt with the possible impact of rational players' concerns regarding "error" or irrationality on the part of their opponents. Similar ideas have been advanced within the context of Nash equilibrium refinements as criteria for ruling out certain Nash equilibria.[1]

The purpose of this paper is to develop a solution concept or a class of solution concepts that describes how factors that can not be deduced from rationality assumptions might interact with the Rationality Hypothesis to yield predictions about behavior which are more restrictive than rationalizability.[2] That is, I wish to present a general framework for studying and/or developing non-equilibrium refinements of rationalizability.

All of the subsequent analysis will be guided by the following principles:

(1) I wish to distinguish between what is being *assumed* (i.e., exogenous) and what is being *deduced* (i.e., endogenous). In particular, I will take the beliefs of rational players regarding the behavior of their opponents as exogenous and the predictions regarding the behavior of rational agents to be endogenous. I will insist that all conclusions regarding the endogenous variable are implied by (rather than being merely consistent with) the exogenous variables and the Rationality Hypothesis.

[1] Selten [28] makes explicit reference to mistakes and irrationality.

[2] The Rationality Hypothesis is the assertion that rationality (i.e., expected utility maximization given subjective probability assesment about opponents' behavior) is (almost) common knowledge.

The difficulty in maintaining that a particular strategy must be played, even though a continuum of other strategies would yield the same payoff given the conjecture held by the player, is acknowledged within the Nash equilibrium framework as well (see Harsanyi [12]). The first principle above reflects the same concern.

(2) I will insist that the assumptions regarding beliefs should be justifiable by the resulting model. That is, the set of allowable beliefs of player i should include the convex hull of the set of allowable profiles of strategies of player i's rational opponents.

G_1

	x	y
a	4, 1	0, 0
b	0, 0	1, 4
c	5, 0	–5,1

Figure 1.

Both of these principles can easily be illustrated with the aid of the game G_1 in figure 1. Let C_i for $i = 1, 2$ be the set of conjectures and R_i be the set of predictions for player i. Consider the case in which $C_1 = \left\{\frac{1}{5}x + \frac{4}{5}y\right\}$ and $C_2 = \left\{\frac{4}{5}a + \frac{1}{5}b\right\}$. First let $R_1 = C_2$ and $R_2 = C_1$. Note that a model with $(C_i)_{i=1}^n$ as the (exogenous) beliefs and $(R_i)_{i=1}^n$ as the predicted behaviors is ruled out by principle (1) above since with these restrictions on beliefs the Rationality Hypothesis does not enable us to deduce that any other strategy with support $\{a, b\}$ will not be played. Hence, principle (1) rules out the possibility of interpreting any non-degenerate mixed strategy as a singleton prediction of behavior.

Next, let $\hat{R}_1 = \left\{\alpha a + (1-\alpha)b \mid \alpha \in [0,1]\right\}$ and $\hat{R}_2 = \left\{\beta x + (1-\beta)y \mid \beta \in [0,1]\right\}$. Note that the model $(\hat{R}_1, \hat{R}_2, C_1, C_2)$ satisfies the requirements of principle (1) but fails to satisfy principle (2). Thus, principle (2) rules out the possibility of interpreting any mixed strategy equilibrium as an equilibrium in beliefs. Principle (2) reflects the view that it is unreasonable to insist that player 1 must believe $\frac{1}{5}x + \frac{4}{5}y$ given the inability

of the theory to exclude any strategy in $\hat{R}_2$ as a possible rational choice for player 2. The need to relate the predicted behavior to the initial restrictions on beliefs is shared by the current and nearly all (Nash equilibrium and non-equilibrium) approaches to rational strategic behavior. The novelty here is in the nature of this relation. The requirement in principle (1), that conclusions regarding behavior should be *implied* by the exogenous restrictions and the Rationality Hypothesis, together with the requirement in principle (2), that the set of allowable beliefs about rational opponents should include the convex hull of the allowable action profiles, will be called *coherence*.

(3) I will distinguish between rational and irrational players. It is not asserted that all players are rational. However, the conclusions of the theory are only about the behavior of rational players and the coherence principle is imposed only on rational players' beliefs about rational opponents. Irrationality plays a role only because rational players assign some probability to the irrationality of their opponents. Hence, the only beliefs that are considered are the beliefs of the rational players. I will focus on the case in which it is common knowledge that players assign high probability to the rationality of their opponents.

(4) In extensive form games, I will take the position that the Rationality Hypothesis offers no guidance to a player who is in a position to choose an action after his rationally held conjecture is violated.

This final principle is motivated by the work of Basu [2], Reny [23] and others and by what was previously known as the paradox of backward induction.

In sections 2 and 3, I will formally define the notion of a τ-theory (for normal and extensive form games) which results from the four principles above. Every τ-theory is a refinement of rationalizability and shares many of the properties of the collection of rationalizable strategies. The notion of a τ-theory enables a classification of the kind of restrictions that have been employed in the refinements literature. Specifically, I will argue that iterated dominance (Proposition 5) and backward induction (Proposition 8) can be viewed, for two-person games, as restrictions on the nature of irrational behavior. I will show in Proposition 7 that if (trembling-hand) perfection is imposed, then rationality in

the extensive form is equivalent to rationality in the equivalent normal form (invariance); otherwise it is not. I will conclude that many of the apparent paradoxes in game theory arise from game theorists' insistence on interpreting possibly plausible restrictions on the nature of irrational behavior (e.g., assessments of the relative likelihoods of various kinds of errors) as implications of rationality. In section 4, I will analyze the possibility that naive learning might substitute for the Rationality Hypothesis.

The works of Bernheim [5], Pearce [19], and Reny [23] play a central role in the analysis below. Other related work on axiomatic foundations for perfection by Börgers [8] and Dekel and Fudenberg [10], or iterated dominance by Börgers and Samuelson [9] and Samuelson [26], will also be discussed. The section on learning relates to the work of Milgrom and Roberts [17] and Sanchirico [27]. Rabin [20] also explores the possibility of incorporating exogenous restrictions into the analysis of rational strategic behavior. His consistent behavioral theories (CBT's) have features in common with τ-theories for two-person games. However, CBT's may fail the coherence criterion by violating principle (1) above. The section on normal form games relates to Rabin's work. Within the refinements literature, comments similar in spirit to my analysis of normal form games appear in Kalai and Samet [13] and ideas related to my view of extensive form games can be found in Reny [22]. However, the fact that the last two papers have taken Nash equilibrium as their starting point makes any detailed comparison impossible.

2. Normal Form Games

In this section, I will utilize the principles outlined in the introduction to motivate the definition of a normal form τ-theory. I will argue that the first three principles of the introduction lead to the notion of a τ-theory. I will then discuss the relationship between τ-theories and rationalizability, perfection, and iterative (weak) dominance. Before undertaking the formal analysis, some basic definitions and a brief review of rationalizability are in order.

Let $G = (A_i, u_i)_{i=1}^n$ denote a finite n-person game. Hence, for $i = 1, 2, \ldots, n$, A_i is a finite set and $u_i : A \to \mathbb{R}$ is a (von Neumann-Morgenstern) utility function, where $A = \prod_{i=1}^n A_i$. I assume that players have preferences over $S_i \times S_{-i}$ where $A_{-i} = \prod_{j \neq i} A_j$,

S_i and S_{-i} denote the set of all probability distributions on A_i and A_{-i}, respectively, and the actions $a_i \in A_i$ and $a_{-i} \in A_{-i}$ are identified with the appropriate degenerate distributions. For $s_{-i} \in S_{-i}$, let $\pi_j(s_{-i}) \in S_j$ for $j \neq i$ denote the marginal distribution of s_{-i} on A_j. Finally let $U_i : S_i \times S_{-i} \to I\!R$ be defined by $U_i(s_i, s_{-i}) = \sum_{a_i} \sum_{a_{-i}} u_i(a_i, a_{-i}) s_i(a_i) s_{-i}(a_{-i})$. I will refer to S_i as the set of all mixed strategies and S_{-i} as the set of all conjectures of player i. For any set $X \subset S_i$, $\overline{X}_i$ denotes the convex hull of X_i and $\operatorname{Int} X_i = \{s_i \in X_i \mid s_i' \in X,\ s_i'(a_i) > 0 \text{ implies } s_i(a_i) > 0\}$.

The mapping $B_i : S_{-i} \to S_i$ denotes the best response correspondence of i. Hence, $s_i \in B_i(s_{-i})$ iff $U_i(s_i, s_{-i}) \geq U_i(s_i', s_{-i})$ for all $s_i' \in S_i$. (Correlated) Rationalizability will play an important role throughout this paper. A formal definition is presented below.

Definition 1: *For all $i = 1, 2, \ldots, n$, let $R_i(0) = S_i$ and $R_i(t+1) = \{s_i \in S_i \mid s_i \in B_i(s_{-i})$ for some $s_{-i} \in S_{-i}$ such that $\pi_j(s_{-i}) \in \overline{R}_j(t)$ for all $j \neq i\}$. The set $R_i^* = \bigcap_{t=0}^{\infty} R_i(t)$ is called the set of rationalizable strategies of player i. Let $\rho^* := (R_i^*)_{i=1}^n$.*

It is easy to verify that $R_i(t+1) \subset R_i(t)$ for all i and there exists some $\overline{t}$ such that for all $t \geq \overline{t}$ and $i = 1, 2, \ldots, n$, $R_i(t) = R_i^*$ (see Pearce [19]).

The iterative procedure used above to define rationalizability can be interpreted as follows:[3]

> Suppose every player choosing a strategy behaves according to the following axioms of rationality which I will call the "Rationality Hypothesis."
>
> (R1): Every player i has some conjecture s_{-i} regarding the behavior of his opponents. Player i chooses some strategy s_i which maximizes his payoff given his conjecture s_{-i}.
>
> (R2): Every player i knows (R1) above and knows that every player $j \neq i$ knows (R1) above and knows that every player $j \neq i$ knows that every player $k \neq j$ knows (R1) above, etc.; that is, (R1) is common knowledge.[4]

[3] Note that the definition above, and all of the subsequent analysis, allows for correlated conjectures, while Bernheim's and Pearce's original formulation does not. For an argument as to why correlated conjectures may be appropriate, see Aumann [1]. The current formulation not only allows correlation, but makes it impossible to restrict the extent of correlation. Some implications of this are discussed below.

[4] It is possible, and in fact appropriate, to replace the phrase "common knowledge" with "common belief" throughout this paper. However, I will for the sake of simplicity reserve the word "belief" for conjectures about behavior, i.e., the elements of the set S_{-i}.

It is easy to see that (R1) implies that every player i will choose some strategy $s_i \in R_i(1)$, since $R_i(1)$ is the set of all strategies which best respond to some conjecture s_{-i}. But by (R2), every player knows this. Hence by (R1), every player i will choose a best response to some conjecture s_{-i} such that s_{-i} assigns zero probability to any strategy not in $R_j(1)$ for all $j \neq i$. But this is equivalent to saying that every player i will choose a strategy in $R_i(2)$. Repeating the above argument yields that every player i will choose a strategy such that $s_i \in R_i^* = \bigcap_{t \geq 1} R_i(t)$. Thus, (R1) and (R2) imply that every player will choose some rationalizable strategy. A similar argument establishes that, in fact, every $s_i \in R_i^*$ is a choice consistent with (R1) and (R2). Thus, the conclusion that every player will choose a rationalizable strategy is equivalent to the assertion that every player will choose a strategy as if (R1) and (R2) are satisfied.

As I have stated in the introduction, much of the criticism of rationalizability centers on the fact that it rules out only those strategies that are inconsistent with (R1) and (R2). Consider again the game G_1 from figure 1 in section 1 above. Suppose in some context, in addition to (R1) and (R2), it became common knowledge that player 2 believes that player 1 will not play a. Then by (R1), player 2 will play y. But then (R2) implies that player 1 knows that player 2 will play y. Hence (R1) implies that player 1 will play b.

G_2

	x	y
a	1, 0	1, 0
b	0, 2	3, 0
c	0, 2	0, 4

Figure 2.

A second type of restriction which is not captured by rationalizability can be illustrated with the aid of game G_2 in figure 2. It is easy to verify that in G_2, R_1^* consists of all strategies s_i such that $s_i(c) = 0$ and $R_2^* = S_2$. Thus, the only action ruled out by rationalizability is c. Yet many researchers have argued that the only reasonable outcome

of this game is $(1,0)$. Indeed $(1,0)$ is the only payoff pair which is consistent with Nash equilibrium or Pearce's [19] cautious rationalizability. A possible argument for insisting on $(1,0)$ as the only reasonable outcome of this game is the following:

> Suppose we require that both players assign some small probability to the possibility that their opponents might make an error—that is, they might be irrational. Suppose we also assert that irrational players are capable of choosing any strategy. Finally, we assume that even if player 1 is irrational, he is much less likely to play c than a or b (after all, c is strictly dominated). This would imply player 2, if rational, should not play y (since c is much less likely than b). But knowing this and also assigning a high probability to the rationality of player 2, player 1 should play a. Thus we are left with the unique strategy pair (a,x).

The solution concept to be defined in this section will allow for both types of restrictions described with the aid of G_1 and G_2 above. The basic idea is to modify the two axioms (R1) and (R2) so as to incorporate exogenous restrictions on beliefs. Thus, consider the following modified Rationality Hypothesis:

(T1): Every player i, if rational, has some conjecture s_{-i} regarding the behavior of his opponents. According to this conjecture, player $j \neq i$, if rational, will choose some strategy in a set R_j^0; if irrational, in a set Σ_j. Moreover, there is probability at least $1-\epsilon$ $\big($where $\epsilon \in (0,1)\big)$ that each opponent is rational and some positive probability each opponent is irrational. Finally, if rational, player i chooses a strategy s_i which maximizes his payoff given his conjecture s_{-i}.

(T2): Every player i, if rational, knows (T1), knows that every player $j \neq i$, if rational, knows (T1), knows that every rational player $j \neq i$ knows that every rational player $k \neq j$ knows (T1), etc. That is, (T1) is common knowledge among rational players.

The real numbers ϵ and $\tau^0 \equiv (R_i^0, \Sigma_i)_{i=1}^n$ for all i are to be viewed as parameters of the given strategic situation. Given these initial parameters, (T1) and (T2) will enable players to make further deductions regarding the behavior of rational players. The analysis

of this process will yield an iterative procedure similar to the one implied by (R1) and (R2) above. Specifically, (T1) states that every rational player will best respond to some allowable conjecture, where an allowable conjecture places at least a $1-\epsilon$ probability to the rationality of his opponents and some probability to the irrationality of his opponents. Moreover, each player i knows that a rational opponent j chooses a strategy in R_j^0 and an irrational opponent j chooses a strategy in Σ_j. Let R_i^1 denote the set of all best responses to such conjectures. But now, by (T2), each player i can refine his understanding of what a rational player will do and conclude that opponent j, if rational, will choose a strategy in $R_i^1 \cap R_i^0$. But by (T1), this reduces the set of conjectures that a rational player may entertain, and hence further reduces the set of possible strategies that rational players can choose and so on. Definitions 2 and 3 below provide some notation regarding this iterative process.

Definition 2:

(a) Let $\mathcal{P}_i = 2_i^S$, $\Upsilon_i = \mathcal{P}_i \times \mathcal{P}_i$, $\mathcal{P} = \prod_{i=1}^n \mathcal{P}_i$, *and* $\Upsilon = \prod_{i=1}^n \Upsilon_i$.[5]

(b) Let $\succeq$ *be the following binary relation on* $\mathcal{P}$: $(R_i)_{i=1}^n \succeq (R_i')_{i=1}^n$ *iff* $R_i \supseteq R_i'$ *for all* i.

Definition 3:

(a) For any $\tau = (\rho, \rho') = \left((R_i)_{i=1}^n, (\Sigma_i)_{i=1}^n\right)$ *and* $\epsilon > 0$, *define* $C_{-i}^{\epsilon}(\tau) := \left\{s_{-i} \in S_{-i} \mid \text{for all } j \neq i,\ \pi_j(s_{-i}) = \alpha_j s_j^1 + (1-\alpha_j)s_j^2 \text{ for some } \alpha_j \in [1-\epsilon, 1),\ s_j^1 \in \overline{R}_j \text{ and } s_j^2 \in \overline{\Sigma}_j\right\}$. *Let* $C_{-i}(\tau) \equiv C_{-i}(\rho) = \left\{s_{-i} \in S_{-i} \mid \text{for all } j \neq i,\ \pi_j(s_{-i}) \in \overline{R}_j\right\}$.

(b) For any $\tau = (R_i, \Sigma_i)_{i=1}^n$ *and* $\epsilon > 0$, *define* $\mathbf{B}_i^{\epsilon}(\tau) = \left\{s_i \in B_i(s_{-i}) \mid s_{-i} \in C_{-i}^{\epsilon}(\tau)\right\}$. *By convention,* $\mathbf{B}_i^{\epsilon}(\tau) = \emptyset$ *if* $R_j = \emptyset$ *for some* $j \neq i$ *and* $\mathbf{B}^{\epsilon}(\tau) = (\mathbf{B}_i^{\epsilon}(\tau))_{i=1}^n$.

For $\tau = (\rho, \rho')$ we will call $C_{-i}(\tau)$ the set of all τ-allowable, or equivalently ρ-allowable, conjectures and $C_{-i}^{\epsilon}(\tau)$ the set of all $(\epsilon - \tau)$-allowable conjectures. The mapping $\mathbf{B}^{\epsilon}$ describes the rational response to a given set of parameters.

Letting $\epsilon > 0$ and $\tau^0 = (\rho^0, \rho')$, define $\tau^{k+1} = \left(\left(\mathbf{B}_i(\tau^k) \cap R_i^k\right)_{i=1}^n, \rho\right)$ where $\rho^k = (R_i^k)_{i=1}^n$ and $\tau^k = (\rho^k, \rho')$ for all k. Let $R_i = \bigcap_{k \geq 1} R_i^k$ and $\rho = (R_i)_{i=1}^n$. As I have argued above, the implication of (T1), (T2), and the initial parameters ϵ and τ^0 is that every rational agent i must choose a strategy in R_i. Given ϵ and τ^0, the following problems

[5] I will write both $\tau = (R_i, \Sigma_i)_{i=1}^n$ and $\tau = ((R_i)_{i=1}^n, (\Sigma_i)_{i=1}^n)$ to denote a generic element τ of Υ.

could arise. It could be that some $R_i^k = 0$ for some k (and hence, by our convention in defining $\mathbf{B}_i^\epsilon$, $R_j = 0$ for all $j = 1, \ldots, n$). In this case, we conclude that the exogenous restrictions (i.e., parameters) (T1) and (T2) are logically inconsistent. Or it could be that $\rho \neq \mathbf{B}^\epsilon(\rho, \rho')$. Specifically, it could be that $\rho \neq \mathbf{B}^\epsilon(\rho, \rho')$ *and* $\mathbf{B}^\epsilon(\rho, \rho') \succeq \rho$. In this case, we conclude that ϵ and τ are not coherent parameter values as discussed in the introduction: ρ entails restrictions on behavior that can not be justified by (T1) and (T2) *and* the initial restrictions on beliefs. To see an example of this, reconsider the example of incoherence discussed in the introduction: Let $\rho^0 = \rho'$ denote the behavior associated with the mixed strategy equilibrium of G_1 in figure 1. Then $\rho = \rho^k = \rho^0$ for all k. But $\mathbf{B}_1^\epsilon(\rho^0, \rho^0) = \Big(\big\{\alpha a + (1-\alpha)b \mid \alpha \in [0,1]\big\}, S_2\Big) \neq \rho^0$ and hence $\epsilon > 0$ and $\tau^0 = (\rho^0, \rho^0)$ are not coherent values of these parameters.

Proposition 0 below establishes certain basic properties of the map $\mathbf{B}^\epsilon$ and the ρ^k's defined above. All proofs are in the appendix.

Proposition 0:

(i) $\epsilon \geq \epsilon'$, $\rho \succeq \rho'$, *and* $\tilde{\rho} \succeq \tilde{\rho}'$ *implies* $\mathbf{B}^\epsilon(\tilde{\rho}, \rho) \succeq \mathbf{B}^{\epsilon'}(\tilde{\rho}', \rho')$.

(ii) Let $\epsilon \in (0,1)$ *and* $\rho' \in \mathcal{P}$. *Fix* $\rho^0 = (R_i^0)_{i=1}^n \in \mathcal{P}$. *Define* $\rho^k = (R_i^k)_{i=1}^n$ *for* $k = 1, 2, \ldots$ *as follows:* $R_i^{k+1} = \mathbf{B}_i(\rho^k, \rho') \cap R_i^k$. *Then there exists* k^* *such that* $\rho^k = \rho^{k^*}$ *for all* $k \geq k^*$. *Moreover,* $\rho^0 \succeq \hat{\rho}$ *and* $\hat{\rho} = \mathbf{B}^\epsilon(\hat{\rho}, \rho')$ *implies* $\rho^{k^*} \succeq \hat{\rho}$.

(iii) For all $\tau = (\rho, \rho') \in \Upsilon$, *there exists* $\overline{\epsilon} \in (0,1)$ *such that for all* $\epsilon \in (0, \overline{\epsilon})$, $\mathbf{B}^\epsilon(\tau) = \mathbf{B}^{\overline{\epsilon}}(\tau)$. *Moreover, for* $\rho = (R_i)_{i=1}^n$, R_i *is closed for every* i *implies* $\overline{\epsilon}$ *can be chosen so that* $\mathbf{B}^{\overline{\epsilon}}(\tau) \subset \mathbf{B}^{\overline{\epsilon}}(\rho, \rho) = \mathbf{B}^\epsilon(\rho, \rho)$ *for all* $\epsilon \in (0, \overline{\epsilon})$.

Part (iii) of Proposition 0 states that the algorithm implied by T1 and T2 ends in a finite number of steps. The resulting prediction of behavior is the unique maximal (for the binary relation $\succeq$) fixed point of the mapping $\mathbf{B}^\epsilon : \Upsilon_{\tau^0} \to \Upsilon_{\tau^0}$, where $\Upsilon_{\tau^0} = \big\{(\hat{\rho}, \rho') \mid \hat{\rho} \succeq \rho^0\big\}$.

The focus of this paper is on the case where ϵ is arbitrarily small. However, we wish to be somewhat literal about the existence of possible irrationality. Part (iii) of Proposition 0 shows that these two desires are not inconsistent: all $\mathbf{B}^\epsilon$'s are identical for ϵ sufficiently small.

Definition 4: *Let* $\mathbf{B}_i(\tau) = \bigcap_{\epsilon>0} \mathbf{B}_i^{\epsilon}(\tau)$ *and* $\mathbf{B}(\tau) = (\mathbf{B}_i(\tau))_{i=1}^n$. *Then* $\tau = (\rho, \rho') \in \Upsilon$ *is a* τ*-theory iff* $\mathbf{B}(\tau) = \rho$.

Observe that parts (i) and (iii) of Proposition 0 establish that, for any $\tau = (R_i, \Sigma_i)_{i=1}^n$ such that $R_i \neq \emptyset \neq \Sigma_i$ for all i, $\mathbf{B}_i(\tau)$ is non-empty. Also note that in Definition 4, restrictions on the beliefs of rational players regarding the behavior of irrational players are made explicit, but restrictions on the beliefs of rational players regarding the behavior of other rational players are suppressed. This creates no problem. The algorithm described in analyzing (T1) and (T2) (i.e., ρ^k for $k = 1, 2, \ldots, k^*$ as defined in part (ii) of Proposition 0) suggests the following alternative definition: ρ is τ-rational behavior iff there exists $\tau^0 = (\rho^0, \rho') \in \Upsilon$ such that $\rho^k = (R_i^k)_{i=1}^n$, $R_i^{k+1} = \mathbf{B}_i(\rho^k, \rho') \cap R_i^k$ for $k = 1, 2, \ldots, k^*$, and $\rho = \rho^{k^*} = \left(\bigcap_{k\geq 1} R_i^k\right)_{i=1}^n$. Thus, any behavior ρ is τ-rational iff there exist some parameters ϵ and τ^0 such that (T1) and (T2) enable us to *conclude* that all rational players will behave according to ρ. But if such a τ^0 exists, and coherence is satisfied, we have $R_i = B_i(\tau^{k^*}) \cap R_i^{k^*} = B_i(\tau^{k^*})$ so that the same ρ could be reached if we started from (ρ, ρ') rather than $\tau^0 = (\rho^0, \rho')$, which is the motivation behind Definition 4.

The following classification of exogenous restrictions will be useful in understanding many of the ideas of the refinement literature.

Definition 5: *A* τ*-theory* $\tau = (\rho, \rho')$ *imposes no type 1 restrictions (i.e., exogenous restrictions on the beliefs of rational players about the behavior of other rational opponents) iff* $\hat{\tau} = (\hat{\rho}, \rho')$ *is a* τ*-theory implies* $\rho \succeq \hat{\rho}$.

Definition 5 states that, given $(\Sigma_i)_{i=1}^n$, if imposing no exogenous restrictions on beliefs about rational players' behavior does not lead to a τ-theory with ρ as the predicted rational behavior, then the τ-theory is said to impose type 1 restrictions.

Definition 6: *A* τ*-theory* $\tau = (\rho, \rho')$ *imposes no type 2 restrictions (i.e., exogenous restrictions on the behavior of irrational players) iff* $\left(\rho, (S_i)_{i=1}^n\right)$ *is a* τ*-theory.*[6]

The work of Rabin [20] introduces a concept similar to the notion of a τ-theory. Rabin's consistent behavioral theories allow for (exogenous) restrictions on predicted behavior, but rule out type 2 restrictions. One of his motivations for allowing restrictions on

[6] It follows from Lemma 0 of the Appendix and Proposition 0 above that τ is a τ-theory with no type 2 restrictions iff $\mathbf{B}(\tau)$ is an exact set in the sense of Basu and Weibull [3].

predictions is to identify candidates for what may be the best (subjective) assessment of an outside observer. Hence it is not required that all conclusions regarding behavior are deduced from restrictions on beliefs in Rabin's notion of a consistent behavioral theory. In the current framework, predictions consist of the (common knowledge) implications of the assumed restrictions on beliefs and the Rationality Hypothesis. They do not incorporate subjective assessments of any outside observer. Rabin's [20] and [21 work and the related work of Farrell [11] on cheap talk share with this paper the objective of identifying exogenous restrictions on rational behavior/beliefs. Rabin [20] focuses on psychological/cultural factors as encapsulated by the idea of a focal point, while Rabin [21] and the work of Farrell [11] deal mostly with communication as the source of these restrictions.

The remainder of this section will be concerned with establishing the relationship between τ-theories (and their exogenous restrictions) and various basic game theoretic ideas such as rationalizability, (trembling-hand) perfection, and iterative (weak) dominance.

Definition 7: *A τ-theory $\tau = \left(\rho, (\Sigma_i)_{i=1}^n\right)$ is a perfect τ-theory iff $\Sigma_i \subset \operatorname{Int} S_i$ for all i. That is, in a perfect τ-theory, rational players are required to assign some positive probability to every action.*

Note that if a τ-theory is of the form $\tau = (\rho, \rho)$, then irrational players are expected to behave just like the rational players. Thus, rationality is common knowledge in such a τ-theory. Proposition 1 below establishes that assuming the chance of irrationality is sufficiently small (which is implicit in the notion of a τ-theory) and imposing no type 2 restrictions is equivalent to assuming that rationality is common knowledge. That is, if rational players know nothing (or agree on nothing) regarding the nature of irrational behavior other than that it is unlikely, then the resulting behavior is as if rationality is common knowledge.

Proposition 1: *For any game G, (ρ, ρ) is a τ-theory iff $\left(\rho, (S_i)_{i=1}^n\right)$ is a τ-theory.*

Proposition 2 below establishes the strong connection between the notion of a τ-theory and rationalizability. It shows that rationalizability is a (common knowledge) τ-theory and that every τ-theory is a refinement of rationalizability.

Proposition 2: *For any game G, $\tau^* = (\rho^*, \rho^*)$ is a τ-theory. Moreover, $\tau = (\rho, \rho')$ is a τ-theory implies $\rho^* \succeq \rho$.*

One of the more puzzling problems of strategic analysis is the relationship between rationality and (weak) dominance. As noted by Pearce [18] and Samuelson [26], if the sole reason for strategy a's dominance over b is that a does better against some irrational strategy of the opponent, then insisting that (weakly) dominated strategies are never played conflicts with the hypothesis that rationality is common knowledge.

Dekel and Fudenberg [10] have explored the possibility that dominance might be explained by (a small amount of) uncertainty about the payoff of the opponent. They show that this leads to what I will call perfect τ-rationalizability. Proposition 3 below establishes that perfect τ-rationalizability is a perfect τ-theory. Recently, Börgers [8] has independently attempted to provide decision theoretic foundations for perfection (or cautiousness). He shows that the assumption of approximate common knowledge of rationality leads also to perfect τ-rationalizability. While there are some differences in the approaches of Dekel and Fudenberg [10], Börgers [8], and the current paper, it is noteworthy that each ultimately identifies perfect τ-rationalizability.[7] Apparently, being explicit about the source of cautiousness either as uncertainty about an opponent's payoffs or his rationality, in the absence of other restrictions, leads to perfect τ-rationalizability. Proposition 4 below establishes that perfect τ-rationalizability is the weakest perfect τ-theory.

Definition 8: *For any game G, let R_i^u denote the set of undominated strategies for player i. That is, $s_i \in R_i^u$ if and only if, for all $\hat{s}_i \in S_i$, either $U_i(s_i, s_{-i}) = U_i(\hat{s}_i, s_{-i})$ for all $s_{-i} \in S_{-i}$ or there exists $\hat{s}_{-i}$ such that $U_i(s_i, \hat{s}_{-i}) > U_i(\hat{s}_i, \hat{s}_{-i})$. Let A_i^u denote the set of pure undominated strategies, i.e., $A_i^u = A_i \cap R_i^u$. Let G^u denote the game obtained from G by removing all dominated pure strategies, i.e., $G^u = \left\{(A_i^u, u_i)_{i=1}^n\right\}$. The set of perfectly τ-rationalizable strategies R_i^p is defined as $R_i^p = R_i^*(G^u) \cap R_i^u$; that is, R_i^p is the intersection of the rationalizable strategies of G^u with the undominated strategies of G. Let $\rho^p := (R_i^p)_{i=1}^n$.*

[7] Perfect τ-rationalizability entails removing all weakly dominated strategies in the first round and removing only strictly dominated strategies in the subsequent rounds; while iterative dominance entails removing all weakly dominated strategies in every round. Obviously, the former is a more stringent requirement than admissibility (i.e., weak dominance), but less stringent than iterative dominance.

Note that the reason for defining R_i^p as the intersection of $R_i^*(G^u)$ and R_i^u instead of just taking $R_i^*(G^u)$ is that $R_i^*(G^u)$ may contain *mixed* strategies that are dominated in the game G.

It follows from elementary arguments that perfect τ-rationalizable strategies exist for every game G. Börgers [8] has shown that what I have called perfect τ-rationalizability is different from perfect rationalizability in the sense of Bernheim [5].[8]

Proposition 3: *For any game G, $\tau^p = \left(\rho^p, (\operatorname{Int} S_i)_{i=1}^n\right)$ is a perfect τ-theory.*

Proposition 4: *For any game G, $\tau = (\rho, \rho')$ is a perfect τ-theory implies $\rho^p \succeq \rho$.*

As I have stated in the introduction, the position I take in this paper is that while it may be useful to explicitly state the relationship between the nature of the exogenous restrictions and the implied behavior, deciding which kind of restrictions are appropriate in any given context is often not a matter of *a priori* analysis. This is particularly true of type 2 restrictions since these involve the behavior of irrational players. Moreover, both for type 1 and type 2 restrictions, it is very difficult to argue that the normal (or even extensive) form contains adequate or even particularly useful information about the relative merits of various restrictions. Presumably, one of the main motivations of studying a variety of strategic problems within the sparse formalism of normal and extensive games is the desire to concentrate entirely on the strategic aspects and to ignore the institutional complexity, the details of the presentation, the underlying social norms, etc. In most applications, the minor effects that can be attributed to factors such as symmetry of payoffs and the labeling of strategies are sure to be overwhelmed by the kind of factors and information that was suppressed in obtaining an abstract normal form for representation. For a social psychologist or sociologist, normal and extensive form games should constitute very barren territory.

The final task of this section is to identify the relationship between iterative dominance and τ-theories. Since the removal of dominated strategies is taken to be a basic postulate of rationality by some researchers, it has been argued that the same principle should be applied to the game obtained after the first round of removal.[9] Hence one claim is that

[8] I am grateful to Pierpaolo Battigalli and an anonymous referee for pointing this out to me.
[9] Kohlberg and Mertens [14] and Samuelson [26] contain such arguments.

accepting dominance as a common knowledge axiom of rationality inevitably leads to iterative dominance. The work of Dekel and Fudenberg [10] and Börgers [8] cited above should be considered a convincing counter argument against this position. Moreover, it is well-known [see, for example, Kohlberg and Mertens [14]] that iterative dominance is sensitive to the order in which strategies are removed, and for certain games every strategy of a given player can be removed by choosing the order appropriately. The problematic nature of iterative dominance is highlighted in recent papers by Börgers and Samuelson [9] and Samuelson [26].[10] In these papers, the concept of "common knowledge of admissibility" for two-person normal form games (i.e., that players do not choose dominated strategies) is defined, and it is shown that this does not lead to iterative dominance. This result is in agreement with the position that I have taken in this paper that iterative dominance does not follow from the analysis of rationality or common knowledge of rationality, but may follow from very specific restrictions on the behavior of irrational players. For two-person games, Proposition 5 shows that suitable type 2 restrictions that guarantee iterative dominance outcomes (conditional on the rationality of all agents) can always be found.

Definition 9: $(A_i^d)_{i=1}^n$ *is the iterative dominance solution to the game* G *iff* $A_i(0) = A_i$ *and* $A_i(t+1) = \{a_i \in A_i(t) \mid U_i(s_i', a_{-i}) \geq U_i(a_i, a_{-i})$ *for some* $s_i' \in \overline{A_i(t)}$ *and all* $a_{-i} \in \prod_{j \neq i} A_j(t)$ *implies* $U_i(s_i', a_{-i}) = U_i(a_i, a_{-i})$ *for all* $a_{-i} \in \prod_{j \neq i} A_j(t)\}$. *Let* $A_i^d = \bigcap_{t=1}^{\infty} A_i(t)$.

Proposition 5: *For any two-person game* G, *there exists a* τ*-theory* $\tau^d = (\rho^d, \rho) = \big((R_1^d, R_2^d), \rho\big)$ *such that* τ^d *has no type 1 restrictions and* $s_i \in R_i^d$ *and* $s_i(a_i) > 0$ *implies* $a_i \in A_i^d$.

There are three-person games for which no τ-theory guarantees iterative dominance outcomes. This is due to the fact that for any τ, $C_{-i}(\tau)$ does not restrict the extent of correlation in the conjectures. Thus, even though we can find type 2 restrictions that guarantee that every conjecture assigns a higher probability to actions in $A_i(t+1)$ than $A_i(t)$, we can not guarantee that a conjecture assigns a higher probability to *each profile*

[10] In Börgers and Samuelson [9], the notion is called common knowledge of rationality, but the same admissibility requirement is built into the Rationality Hypothesis. I will refer to both this paper and Samuelson [26] again in the next section.

$a_{-i} \in \prod_{j \neq i} A_j(t+1)$ than to any profile $a_{-i} \in \prod_{j \neq i} A_j(t)$, which is needed for generalizing Proposition 5.

Imposing restrictions on beliefs by restricting only the marginals of each s_{-i} enables the relatively simple description of a τ-theory adopted in this paper. However, I am not sure that the inability to impose restrictions on s_{-i} directly (that is, to restrict the extent of correlation permitted) is essential to the approach I have outlined in this paper. Nevertheless, it is noteworthy that such restrictions on the extent of correlation are needed to derive iterative dominance and, as I will discuss in section 3, backward induction whenever $n \geq 3$.

Even for two-person games, it does not follow that the type 2 restrictions needed to guarantee backward induction are always compelling. However, in certain simple games (such as G_2 in figure 2), they may be.

3. Extensive Form Games

The construction of the notion of a τ-theory for extensive form games will proceed in a manner analogous to the construction of τ-theories for normal form games. Some basic notation and definitions involving extensive form games will be needed for the subsequent analysis. A more formal and detailed presentation of finite extensive form games can be found in Kreps and Wilson [16] and Selten [28].

Γ:	finite n-person extensive form game with perfect recall;
A_i:	the set of all pure strategies of player i;
S_i:	the set of all (mixed) strategies of player i;
S_{-i}:	the set of all (correlated) conjectures of player i regarding the strategies of all other players;
S:	the set of all (correlated) strategy profiles;
$I_{i\ell}$:	the ℓ^{th} information set of player i;
Z:	the set of terminal nodes;
$u_i : Z \to \mathbb{R}$:	player i's utility function;
$U_i(s_i, s_{-i})$:	the expected utility associated with the probability distribution on Z induced by the product distribution, (s_i, s_{-i}); hence, $U_i : S_i \times S_{-i} \to \mathbb{R}$.

Each $a_i \in A_i$ specifies an action at every information set $I_{i\ell}$, provided that $I_{i\ell}$ is not precluded by player i's action at some preceding information set. As before, I use, for $j \neq i$, $\pi_j(s_{-i}) \in S_j$ to denote the marginal distribution of $s_{-i} \in S_{-i}$ on S_j. A strategy profile implies a probability distribution on terminal nodes. Associated with each terminal node there is an outcome path. I say that (s_i, s_{-i}) reaches $I_{i\ell}$ if, given (s_i, s_{-i}), there is a non–zero probability that the outcome path will run through $I_{i\ell}$. Similarly, I say that $s_i \in S_i$ reaches $I_{i\ell}$ if there is some $s_{-i} \in S_{-i}$ such that (s_i, s_{-i}) reaches $I_{i\ell}$, and I say that $s_{-i} \in S_{-i}$ reaches $I_{i\ell}$ if there is some $s_i \in S_i$ such that (s_i, s_{-i}) reaches $I_{i\ell}$. It is easy to verify (using perfect recall) that s_i reaches $I_{i\ell}$ and s_{-i} reaches $I_{i\ell}$ imply (s_i, s_{-i}) reaches $I_{i\ell}$. Define $I(i) = \{\ell \mid I_{i\ell} \text{ is an information set}\}$.

The axioms of rationality for extensive form games will be similar to (T1) and (T2). The only novelty is that the following sentence needs to be added to the end of (T1):

> At any information set $I_{i\ell}$ such that s_i (the strategy that i chooses) reaches $I_{i\ell}$, s_i must be a best response at $I_{i\ell}$ (i.e., conditional on $I_{i\ell}$ being reached given $(s_i, \hat{s}_{-i})$) to some conjecture $\hat{s}_{-i}$ that reaches $I_{i\ell}$.

Hence the version of (T1) for extensive form games also requires optimality of information sets $I_{i\ell}$ that can not be reached by the initial conjecture s_{-i} provided $I_{i\ell}$ is not ruled out by s_i. Thus optimality at every reachable information set, given s_i, is being incorporated into the extensive form Rationality Hypothesis. Furthermore, no restriction on $\hat{s}_{-i}$ is being imposed. The idea is that once an initial conjecture that fulfills every requirement of the theory is adopted by i and overturned, the theory offers no further guidance regarding what i should believe. This is the last principle described in the introduction. However, the inability of τ-theories to impose additional (but not all) restrictions at this stage is not as significant as one might think. As I will show below, many additional restrictions at information sets unreached by conjectures on the behavior of rational players can be built into the Σ_i's. The crucial point is that such restrictions are also to be viewed as exogenous and not an implication of rationality. Repeating the analysis of (R1), (R2), (T1), and (T2), it can be seen that the extensive form versions of the axioms (T1) and (T2) also lead to an iterative algorithm. The only distinction is that in extensive form games there is the

additional restriction that rational players choose strategies s_i such that s_i is optimal at $I_{i\ell}$ against some conjecture that reaches $I_{i\ell}$ whenever s_i reaches $I_{i\ell}$.

Definition 10:

(i) *For every s_i and s_{-i} that both reach $I_{i\ell}$, let $U_i(s_i, s_{-i} \mid I_{i\ell})$ denote the expected utility of (s_i, s_{-i}) conditional on $I_{i\ell}$. Since (s_i, s_{-i}) reaches $I_{i\ell}$, the meaning of this conditional expected utility is unambiguous.*

(ii) $R_i^x = \{s_i \in B_i(s_{-i}) \mid$ *s_i reaches $I_{i\ell}$ implies* $\exists\ \hat{s}_{-i} \in S_{-i}$ *such that $\hat{s}_{-i}$ reaches $I_{i\ell}$ and* $U_i(s_i, \hat{s}_{-i} \mid I_{i\ell}) \geq U_i(\hat{s}_i, \hat{s}_{-i} \mid I_{i\ell})$ *for all $\hat{s}_i$ that reach $I_{i\ell}$.*

(iii) *For all* $\tau \in \Upsilon$,

$$\mathbf{B}_i^x(\tau) = \mathbf{B}_i(\tau) \cap R_i^x$$

$$\mathbf{B}^x = (\mathbf{B}_i)_{i=1}^n.$$

(iv) $\tau = (\rho, \rho')$ *is an extensive form τ-theory if and only if* $\rho = \mathbf{B}^x(\tau)$.

Note that in part (ii) above, no restriction on the conjecture $\hat{s}_{-i}$ is imposed. If s_i and the original conjecture s_{-i} reach $I_{i\ell}$, then the fact that $s_i \in B_i(s_{-i})$ will imply that s_i maximizes $U_i(s_{-i} \mid I_{i\ell})$ among all strategies that reach $I_{i\ell}$. If s_{-i} does not reach $I_{i\ell}$, then, as stated earlier, the notion of an extensive form τ-rationality imposes no restriction on what conjectures are allowed at $I_{i\ell}$.

The only new element in the definition of an extensive form τ-theory is the collection of sets $(R_i^x)_{i=1}^n$. These are precisely the strategies that are optimal given any information set they reach against some conjecture which reaches that information set. Also note that every conclusion of Proposition 0 holds if we replace $\mathbf{B}$ with $\mathbf{B}^x$ and $\mathbf{B}_i$ with $\mathbf{B}_i^x$ for all i. As in the case of normal form games, an extensive form τ-theory $\tau = (R_i, \Sigma_i)_{i=1}^n$ will be called a perfect τ-theory iff $\Sigma_i \subset \operatorname{Int} S_i$ for all i. In the remainder of this section, I will explore the relationship between τ-theories and refinement ideas such as invariance and backward induction.

Invariance

Let A_i denote the set of all pure strategies available to player i, in some extensive form game Γ. Let $G(\Gamma) = (A_i, \overline{u}_i)_{i=1}^n$ denote the normal form game where $\overline{u}_i(a)$ is the utility for player i (according to the utility function u_i in Γ) associated with a in the extensive

form game Γ. It is easy to verify that, even if Γ and Γ' are different, it may still be the case that $G(\Gamma) = G(\Gamma')$. Loosely speaking, an extensive form "solution concept" is said to be invariant if it prescribes the same behavior in Γ and Γ' whenever $G(\Gamma) = G(\Gamma')$. Given that two different notions of τ-theory (one for normal and one for extensive form games) have been defined, the question of invariance for τ-theories can be stated as follows: is it the case that, given any extensive form game Γ, $\tau = (R_i, \Sigma_i)_{i=1}^n$ is a τ-theory for Γ iff it is a τ-theory for $G(\Gamma)$? The example in figure 3 (due to Pearce [18]) establishes that the answer to this question is "no."

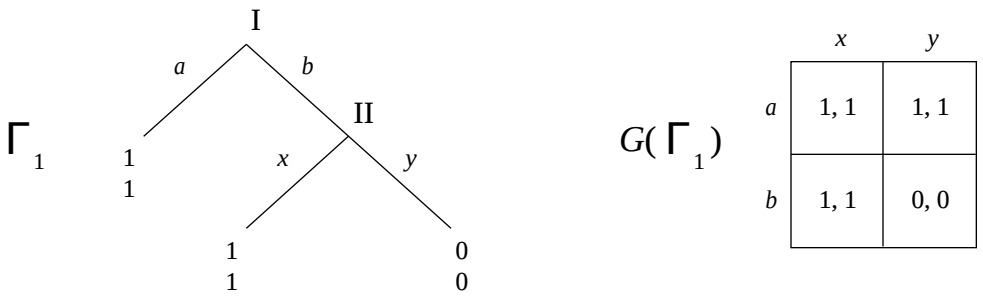

Figure 3.

Observe that $\tau = (S_i, S_i)_{i=1}^n$ is a τ-theory for $G(\Gamma_1)$ but not for Γ_1. The interpretation is the following. In $G(\Gamma_1)$, player 2 can ignore the possibility that player 1 may play b if he is sure that player 1 will not play it; in Γ_1, he cannot. This is due to the fact that, if player 2 is called upon to move in Γ_1, he *knows* that player 1 has played b and, thus, his initial certainty (about player 1 playing a) becomes irrelevant. Two objections can be made to this line of argument.

(1) In the game Γ_1, if player 2 were indeed sure that player 1 would play a and if he is called upon to move, then he can conclude that the joint hypothesis, "Player 1 is rational and player 2 knows the payoff structure in the game Γ_1," has been falsified. But player 1 does not know which part of this hypothesis ought to be abandoned. Hence, we are no longer justified in drawing any conclusions regarding player 2's behavior at his information set.[11]

[11] This seems to be the line of argument in Bonanno [7].

While this argument is logically correct, it does not seem unreasonable to assume that when a conjecture is falsified player 1 merely "forms" a new conjecture consistent with his observation, without questioning his own understanding of the game.

(2) It could be said that y is not a reasonable strategy for player 2, even in the game $G(\Gamma_1)$.[12] After all, he has nothing to lose by playing x and could gain by doing so.

While such (admissibility) requirements do not appear to be unreasonable, as argued in the previous section, they are not consequences of rationality but rather restrictions on the behavior of irrational players. Even if we are ultimately willing to impose admissibility, it is important to understand whether imposing admissibility is sufficient to bridge the gap between normal and extensive form games.

The following propositions attempt to provide an answer to this question and to clarify the necessary restrictions imposed by extensive form rationality. Proposition 6 identifies the maximal extensive form τ-theory which I will call extensive form τ-rationalizability. It is shown that, in general, this theory is a (possibly strict) subset of normal form rationalizability (hence, the extensive form does involve certain restrictions) and a (possibly strict) superset of (normal form) perfect rationalizability. Proposition 7 establishes that, indeed, perfection leads to invariance.

Definition 11: *The set of extensive form τ-rationalizable strategies $\left(R_i^e\right)_{i=1}^n$ is defined as follows: $R_i^e = R^*(G^x) \bigcap R_i^x$ for all i, where $G^x = (A_i^x, u_i)_{i=1}^n$, $A_i^x = R_i^x \bigcap A_i$, and $R_i^*(G^x)$ is the set of rationalizable strategies for player i in the game G^x. Let $\rho^e = (R_i^e)_{i=1}^n$ and $\tau^e = (\rho^e, \rho^e)$.*

Thus extensive form τ-rationalizable strategies are defined by removing all actions not in R_i^x, then computing the rationalizable strategies of the resulting game and removing all mixed strategies not in R_i^x.

Proposition 6: *τ^e is an extensive form τ-theory. Furthermore, if $\tau = (\rho, \rho')$ is an extensive form τ-theory, then $\rho^e \succeq \rho$.*

[12] This argument appears to be at the center of most of the work arguing for invariance (see Kohlberg and Mertens [14]).

It follows from (4) in the proof of Proposition 4 that extensive form τ-rationalizability is a subset of the maximal normal form τ-theory (i.e., correlated rationalizability). Examples such as the game Γ_1 illustrate that indeed the inclusion can be strict. The game Γ_1 also illustrates that extensive form τ-rationalizability may be a strict superset of normal form perfect τ-rationalizability.[13] Proposition 7 below shows that perfect τ-theories satisfy invariance in a very simple and strong sense and enable us to identify precisely the distinction between a normal form and extensive form rationality: the structure of the extensive form game often implicitly imposes some amount of admissibility (or perfection) by providing the opportunity to have players see their conjecture falsified.

Proposition 7: *τ is an extensive form perfect τ-theory for the game Γ iff it is a perfect τ-theory for the game $G(\Gamma)$.*

Backward Induction

In the refinements literature, the basic motivation of backward induction seems to be Selten's [28] insistence that observed deviations be viewed as one-time mistakes that are unlikely to be repeated in the future. It is unlikely that this requirement is compelling as an implication of rationality. Why should a player assume that a particular deviating player will follow the prescriptions of some criterion of rationality in the face of extensive evidence that the same opponent has failed the very same criterion of rationality in the past?

The purpose of this section is to provide support for the following three arguments:

(1) Backward induction is not a necessary consequence of rationality in the extensive form.

(2) Alternative notions of rationality that imply backward induction are likely to encounter problems of existence.

(3) Backward induction may follow from suitable type 2 restrictions in two-person games. In simple games, required restrictions will be intuitively attractive.

[13] Note that b is an allowed strategy for extensive form τ-rationalizability $(R_i^e)_{i=1}^n$, but not for normal form perfect τ-rationalizability $(R_i^p)_{i=1}^n$. Again, a similar argument is made by Pearce [18].

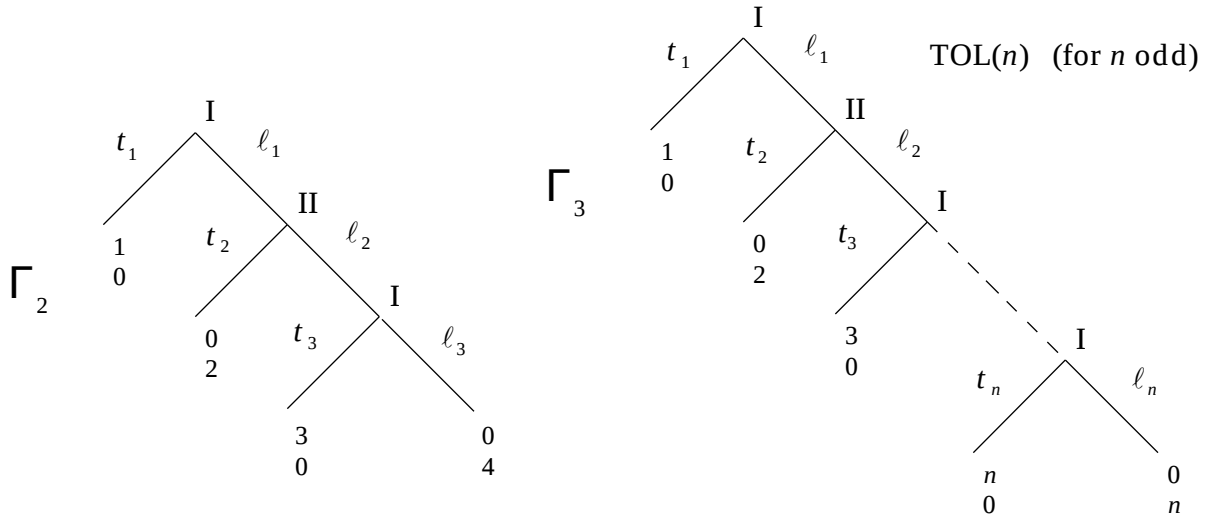

Figure 4.

None of these arguments is entirely new. The purpose here is to see the extent to which the notion of a τ-theory is able to shed light on and provide support for these statements. Thus, I will conclude that many of the paradoxes of rationality stem from game theorists' insistence on viewing backward induction as a consequence of rationality when it is best viewed as a consequence of exogenous restrictions on the behavior of irrational players.

Consider the simple extensive form game Γ_2 in figure 4. Note that this is a version of Rosenthal's [24] "centipede game" and what Reny calls the "take-it-or-leave-it game." The familiar logic of backward induction requires that in this game player 1 take the \$1 immediately (i.e., play at his first information set).

It is well-known that the problem with backward induction is the following: by starting from the end and working backwards, backward induction treats each subgame as if *it* were the game being played. Thus, at player 2's information set, backward induction fails to take into account that it was player 1's failure to take the \$1 which enabled the play to reach this information set. This point is made most forcefully by Reny [23] who formalizes what we might mean by rationality being common knowledge (belief) at a given information set and shows that for Γ_2, rationality cannot be common knowledge (belief) at player 2's information set.

It is easy to verify that every τ-theory $\tau = (R_i, \Sigma_i)_{i=1}^n$ for the game Γ_2 falls into one of the following two categories:

(I) $R_1 = \{s_1 \in S_1 \mid s_1(\ell_1\ell_3) = 0\}$, $R_2 = S_2$, there exists $s_1 \in \Sigma_1$ such that $s_1(\ell_1\ell_3) \geq \frac{1}{2}s_1(\ell_1 t_3)$, $\Sigma_2 \subset S_2$;

(II) $R_1 = \{t_1\}$, $R_2 = \{t_2\}$, there exists no $s_1 \in \Sigma_1$ such that $s_1(\ell_1\ell_3) \geq \frac{1}{2}s_1(\ell_1 t_3)$ and there exists $s_1 \in \Sigma_1$ such that $s_1(t_1) < 1$ and $\Sigma_2 \subset S_2$.

Observe that the τ-theories in category (I) allow non-backward induction strategies for rational players (both player 1 and player 2). For this to be consistent, it must be possible for player 2 to believe, upon being reached, that player 1 is at least as likely to play ℓ_3 as t_3 at his final information set. Furthermore, the theories in this category allow a rational player 1 to play $\ell_1 t_3$ but not $\ell_1\ell_3$.

But doesn't this yield a contradiction? If a rational player 1 is allowed to play ℓ_1, should not player 2, *upon being reached, realize that he is still dealing with a rational opponent* who will choose $(3, 0)$ over $(0, 4)$? The error in this argument is in the phrase in italics. It is only required that player 2 assign a high *initial* probability to player 1's rationality. Thus, if player 2 assigns a high probability to the rationality of player 1, and further, if he assigns a high probability to the event that a rational player 1 is likely to play ℓ_1, then upon being reached he may believe that he is likely to be dealing with an irrational opponent. Furthermore, he may also believe that an irrational opponent is likely to play $\ell_1\ell_3$. Thus upon being reached, player 2 might rationally play ℓ_2. This also explains why a rational player 1 might play ℓ_1 at his initial information set—to lure the rational player 2 into thinking as above—which in turn explains why player 2 may play t_2, because he suspects that a rational player 1 will try to lure him into playing ℓ_2. This in turn explains why player 1, if he is rational, might play t_1, which further supports the belief assigned to player 2 at the beginning of this paragraph and completes the cycle.

Basu [2] also observes the impossibility of maintaining the Rationality Hypothesis at every information set. He considers two different possible assumptions after an observed deviation from rationality. The first is the standard backward induction hypothesis that this is a one-shot deviation from rationality, which has no implication for the future. The second is that nothing can be assumed in the future about the behavior of a person who has taken an irrational action in the past. He argues that the second may in certain cases be more compelling. The approach of this paper is to take the second hypothesis as the

extensive form Rationality Hypothesis, but to allow for additional restrictions as exogenous parameter values and then to impose consistency and coherence.

The τ-theories in category (I) simply state that, if players are rational, any outcome of this game other than $(0, 4)$ may come about. However they do not specify a probability distribution on the remaining endpoints. This is the key difference between the notion of a τ-theory and that of a Nash equilibrium (or subgame perfect Nash equilibrium), and it is the reason why the consistency of the above analysis can be maintained. Also worth noting is the fact that any extensive form τ-theory for Γ_2 with no type 2 restrictions (i.e., rationality is common knowledge) will fail to imply backward induction (i.e., will be in category I). This is consistent with Reny's [23] and Ben-Porath's [4] analysis that rationality cannot be common knowledge at every information set in Γ_2. By contrast, note that the work of Kreps, Milgrom, Roberts and Wilson [15] required *postulating specific restrictions* on the behavior of irrational players to allow for non-backward induction equilibria in extensive form games; whereas, in the current framework, the *absence of (type 2) restrictions* leads to non-backward induction theories, and backward induction can only be justified by a specific type of irrational behavior.

A final important point is that for $G(\Gamma_2)$, the normal form representation of Γ_2, there is a common knowledge normal form τ-theory, namely rationalizability, which predicts the same behavior as theories in category (I). The distinction between normal form rationalizability and theories in category (I) is, however, significant. Rationalizability allows player 2 to play strategy ℓ_2 only because she believes with certainty that player 1 will play t_1. Hence, player 2 is indifferent between t_2 and ℓ_2 and therefore may play ℓ_2. In the extensive form, however, by the time player 2 has to play, the belief that player 1 is rational and, if rational, that player 1 will surely play t_1 is no longer permissible. Thus, player 2 may play ℓ_2, not because of indifference, but due to the fact that she no longer believes player 1 is rational and believes that an irrational player 1 is sufficiently likely to choose ℓ_3.

As I have noted in the discussion of iterative dominance in section 2 above, the definition of a τ-theory, both for normal and extensive form games, does not preclude the possibility of correlated conjectures. More importantly, the notion of coherence does not

permit the possibility of restricting the extent of correlation in rational players' conjectures. In the absence of restrictions on the extent of correlation, for certain extensive form games, one can not find type 1 or type 2 restrictions that imply that every rational action profile must lead to a backward induction outcome. For two-person games, the inability of τ-theories to restrict the extent of correlation in conjectures is costless and hence Proposition 8 below can be proved.

Proposition 8: *For any two-person game of perfect information* Γ *such that distinct terminal nodes yield distinct payoffs for both players, there exists* $\tau = (\rho, \rho')$ *with no type 1 restrictions such that* $\rho = \big(\{a_1\}, \{a_2\}\big)$ *and* (a_1, a_2) *yields the unique backward induction outcome.*

4. Learning

In this section I will address the following question: "How can rationality become common knowledge?" An answer to this question requires a model of rational players who at the outset do not know that their opponents are rational. I will present a naive learning model which is characterized by the modest requirement that all players choose best responses to their conjectures and that their conjectures assign low probability to actions which have been observed infrequently.[14]

Throughout the discussion, I will not be too specific about the actual dynamics of the system. Thus the analysis will be conducted as if a fixed set of players are repeatedly playing the same game. However, the conclusions of this section would hold for random matching models as well. The key implicit assumptions are the following: players ignore the effects of their actions on the future behavior of their opponents; and all players know the entire history of play.[15] The first of these assumptions is appropriate if the discount factors are low or there is random matching from large pools of potential players.

For the remainder of this section, I will discuss an arbitrary but fixed normal form game G. The following definitions will facilitate the subsequent analysis of learning in normal form games.

[14] Milgrom and Roberts [17] have independently developed a model similar to this one that emphasizes serially undominated strategies in games with possibly infinite sets of (pure) strategies.

[15] I suspect that stochastic versions of the results presented below would hold if players knew sufficiently rich samples of the past outcomes. Proving this, however, would require a substantially more complicated model and arguments.

Definition 12: *For $t \geq 1$, a t-period history $h^t \in H^t$ is a t-tuple of actions (i.e., pure strategy) profiles. Hence, $H^t := A^t := (\prod_{i=1}^n A_i)^t$. For $t' \leq t$, $h^t(t')$ will denote the entry in coordinate t' of h^t. The first t' entries of h^t will be denoted $h^t(-t') \in H^{t'}$. The pair $(\hat{h}^t, h^T) \in H^{t+T}$ is the $t+T$ period history with $\hat{h}^t$ as its first t and h^T as its last T entries. The history of actions for player i associated with h^t are denoted by h_i^t, $h_i^t(t')$ and $h_i^t(-t')$. For any t-period history, $P_e(a_i, h^t)$ denotes the empirical frequency of the action $a_i \in A_i$ in the history h^t. That is, $P_e(a_i, h^t)$ is the cardinality of the set $\{t' \leq t \mid h_i^t(t') = a_i\}$.*

Definition 13: *A learning model $L = (b_1, b_2, \ldots, b_n)$ is an n-tuple of correspondence $b_i : \bigcup_{t=1}^{\infty} H^t \to S_{-i}$.*

Thus, any b_i specifies the set of conjectures for a player i after every history h^t. Note that a learning model specifies only the rules as to how players form conjectures and not how they behave. I will assume that players always choose some best response to their conjectures. Since I am concerned only with finite games, the result below can be generalized to the case in which players choose strategies with payoff within ϵ of a best response, provided ϵ is small enough.

Clearly, a learning model L and the requirement that players choose best responses to their conjectures restricts the set of histories that can be observed. I will also allow for the possibility that some initial history is already in place before the learning model L is adopted. The role of the initial history is to capture the possibility that, at the early stages of the learning process, players might choose actions somewhat randomly and, hence, histories which are not consistent with any notion of rationality may precede the formal learning stage.

Definition 14: *The history h^T is consistent with the learning model $L = (b_1, b_2, \ldots, b_n)$ given the initial history $\hat{h}^t$, if for all $t' = 0, 1, \ldots, T-1$ and for all i,*

$$h_i^T(t'+1) \in B_i\big(b_i\big(\hat{h}^t, h^T(-t')\big)\big) \quad \textit{where} \quad \big(\hat{h}^t, h^T(-0)\big) = \hat{h}^t.$$

In any context where the same game is played repeatedly, it is logically conceivable that outcomes observed in the past will have no impact on behavior in the future. The

tenuousness of the relationship between past and future play becomes more apparent in learning models, since such models typically require that players ignore the effect of their current actions on the future behavior (i.e., repeated game effects) of their opponents. Nevertheless, it is possible and perhaps even plausible that the past will have a bearing on the future. Learning models analyze this possibility by imposing explicit restrictions on the belief formation process. The naive learning model of this section will be characterized by the requirement that players assign low probabilities to actions which have been observed infrequently in the past. This requirement will be called δ-minimal history dependence.

Definition 15: *For $\delta \in (0,1)$, a conjecture s_{-i} satisfies δ-minimal history dependence* (MHD$_\delta$)*, given history h^t, if for all $j \neq i$, $\pi_j(s_{-i})(a_j) \leq \delta$ whenever $P_e(a_j, h^t) \leq \delta/2$. A learning model $L = (L_1, L_2, \ldots, L_n)$ satisfies* MHD$_\delta$ *if, for every i, t, h^t and $b_i \in L_i$, $b_i(h^t)$ satisfies* MHD$_\delta$*. L satisfies cautious* MHD$_\delta$ *if, in addition to satisfying* MHD$_\delta$*, $b_i(h^t) \in \operatorname{Int} S_{-i}$ for all i, t, and h^t.*

Note that the fictitious play algorithm is MHD$_\delta$ for all δ (see Samuelson [25]). Similarly, the so-called Bayesian learning models in which players assume that they are faced with a stationary distribution of behavior from their opponents would satisfy MHD$_\delta$, provided we start the process off with a suitably long initial history.

I will be concerned with the case in which δ is small. Hence, MHD$_\delta$ is indeed the requirement that infrequent actions are assigned low probabilities. The following proposition establishes that for δ small, MHD$_\delta$ eventually leads to rationalizability and cautious MHD$_\delta$ eventually leads to perfect τ-rationalizability.

Proposition 9: *There exists $\delta^* \in (0,1)$ such that, for every $t = 1, 2, \ldots$, there exists some $T^* \geq 1$ with the following property: for all $\hat{h}^t \in H^t$, L satisfying (cautious)* MHD$_\delta$*, $\delta \in (0, \delta^*)$, and h^T consistent with L given $\hat{h}^t$, $T \geq T^*$ implies $h_i^T(T)$ is a (perfect τ-) rationalizable action for all i.*

Proposition 9 states roughly that, if δ is small and T is large, any action observed at time T will be rationalizable. Conspicuously absent from the statement of the proposition are statements of the form "all rationalizable actions will eventually be played." Also note that MHD-learning is a model of learning in which rational agents do not contemplate the

rationality of their opponents. Thus, Proposition 9 observes that rational but somewhat naive agents will, if they pay some attention to history, end up in a situation in which rationality is common knowledge (i.e., only rationalizable strategies will be played).

In contrast, the learning model presented by Sanchirico [27] provides assumptions under which a history generated by a learning model will necessarily converge to behavior associated with a minimal τ-theory among τ-theories in which rationality is common knowledge. That is, a τ-theory of the form $\tau = (\rho, \rho)$ such that $\hat{\tau} = (\hat{\rho}, \hat{\rho})$ is a τ-theory, $\rho = (R_i)_{i=1}^n$, $\hat{\rho} = (\hat{R}_i)_{i=1}^n$, and $R_i \cap \hat{R}_i \neq \emptyset$ for some i implies $\hat{\rho} \succeq \rho$. Sanchirico's result shows how a plausible learning model could lead to the type of restrictions (or refinements) studied in this paper.

5. Conclusion

This paper is an attempt at reconciling many ideas of the refinements literature with the well-known criticisms of Nash equilibrium and its refinements. The key concepts are coherence and exogenous restrictions on beliefs. Coherence is the statement that the predicted behavior should imply the assumed exogenous restrictions on beliefs (as opposed to being merely consistent with these beliefs) and that no belief over rational actions should be ruled out. These two ideas are used in conjunction to suggest an alternative analysis of many problematic elements of game theory, such as iterative dominance, backward induction and invariance. In particular, I have attempted to argue that many of the paradoxes of game theory result from incorporating apparently plausible (exogenous) restrictions on beliefs into the Rationality Hypothesis.

Much of the paper deals with the issue of how type 1 and type 2 restrictions can lead to more precise predictions than what would be implied by rationalizability. This is not to say that every environment will entail an abundance of such factors that will lead to the most restrictive or most favored (such as backward induction) τ-theories. Nor would I wish to suggest that the primary focus of research in game theory should be to articulate and understand such restrictions. My objective is to simply argue that these factors could conceivably be incorporated into a theory of rational strategic behavior, and that the best way of doing this is by abandoning the preconception that all such factors will boil down to finding a single principle of rationality.

The entire approach of this paper can be incorporated into the lexicographic probability model of Blume, Brandenburger and Dekel [6]. The key point is that the first order probabilities in the lexicon would denote beliefs regarding the actions of rational players, while the higher order probabilities would denote the beliefs about the behavior of irrational players. Proposition 0 establishes that this is equivalent to assuming that ϵ, the probability of irrationality, is small.

The issue of communication has been omitted entirely. The literature on cheap talk or pregame communication [see, for example, Farrell [11] and Rabin [21] has in common with this paper the objective of combining exogenous restrictions with the Rationality Hypothesis. Yet these models violate what I have called coherence. For the time being, I can offer no model of communication. However, it does not appear too implausible that a reasonable model of communication can be developed within the framework of τ-theories. A plausible model of communication which showed that communication always leads to a certain subclass of type 1 restrictions[16] would provide some support for my claim that a better understanding of the main concerns of game theory requires distinguishing between exogenous restrictions on beliefs and implied restrictions on rational behavior. Providing such a model is, however, beyond the scope of the current paper.

The preceding analysis has been restricted to the case of complete information. The extension of the kind of analysis outlined in this paper to the problem of asymmetric information is left for future work.

[16] In a similar sense, the work of Sanchirico [27] can be said to show that learning leads to a subclass of type 1 restrictions.

References

1. R. Aumann, Correlated equilibrium as an expression of Bayesian Rationality, *Econometrica* 55 (1987),1–18.

2. K. Basu, On the non-existence of a rationality definition for extensive games, *International Journal of Game Theory* 19 (1990), 33–44.

3. K. Basu, and J.W. Weibull, Strategy subsets closed under rational behavior, Discussion Paper #62, John M. Olin Program for the Study of Economic Organization and Public Policy (1992), Princeton University.

4. E. Ben-Porath, Rationality in extensive form games, mimeo (1992), Northwestern University.

5. D. Bernheim, Rationalizable strategic behavior, *Econometrica* 52 (1984), 1007–1028.

6. L. Blume, A. Brandenburger and E. Dekel, Lexicographic probabilities and equilibrium refinements, *Econometrica* 59 (1991), 81–98.

7. G. Bonanno, The logic of rational play in games of perfect information, *Economics and Philosophy* 7 (1991), 37–65.

8. T. Börgers, Weak dominance and approximate common knowledge of rationality, mimeo (1990), Universität Basel.

9. T. Börgers and L. Samuelson, Cautious utility maximization and iterated weak dominance, *International Journal of Game Theory* 21 (1992), 13–25.

10. E. Dekel and D. Fudenberg, Rational behavior with payoff uncertainty," *Journal of Economic Theory* 52 (1992), 243–67.

11. J. Farrell, Meaning and credibility in cheap talk games, *Games and Economic Behavior* 5 (1993), 514–31.

12. J. C. Harsanyi, Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points," *International Journal of Game Theory* 2 (1973), 1–23.

13. E. Kalai, and D. Samet, Persistent equilibria in strategic games, *International Journal of Game Theory* 13 (1984), 129–144.

14. E. Kohlberg, and J. F. Mertens, On the strategic stability of equilibria, *Econometrica* 54 (1986), 1003–1037.

15. D. M. Kreps, P. Milgrom, D. J. Roberts, and R. Wilson, Rational cooperation in the repeated prisoner's dilemma, *Journal of Economic Theory* 27 (1982), 245–252.

16. D. M. Kreps, and R. Wilson, Sequential Equilibria, *Econometrica* 50 (1982), 863–894.

17. P. Milgrom, and D. J. Roberts, Adaptive and sophisticated learning in normal form games, *Games and Economic Behavior* 3 (1991), 82–100.

18. D. Pearce, Ex ante equilibrium: strategic behavior and the problem of perfection, working paper (1982), Princeton University.

19. D. Pearce, Rationalizable strategic behavior and the problem of perfection, *Econometrica* 52 (1984), 1008–1050.

20. M. J. Rabin, Incorporating behavioral assumptions into game theory, (1994), 69–87, in J. Friedman (ed.) "Problems of Coordination in Economic Activity", Dordrecht, Netherlands: Kluwer Academic Publishers.

21. M. J. Rabin, A model of pre-game communication," *Journal of Economic Theory* 61 (1994), 370–91.

22. P. Reny, Backward induction, normal form perfection and explicable equilibria, *Econometrica* 60 (1992), 627–649.

23. P. Reny, Common belief and the theory of games with perfect information, *Journal of Economic Theory* 59 (1993), 257–274.

24. R. W. Rosenthal, Games of perfect information, predatory pricing and the chain-store paradox, *Journal of Economic Theory* 25 (1981), 92–100.

25. L. Samuelson, Evolutionary foundations of solution concepts for finite, two-player, normal-form games, (1988), in M. Vardi (ed.), "Theoretical Aspects of Reasoning About Knowledge", Morgan Kaufmann, Los Altos, California.

26. L. Samuelson, Dominated strategies and common knowledge, *Games and Economic Behavior* 4 (1991), 284–313.

27. C. Sanchirico, Strategic intent and the salience of past play: a probabilistic model of learning in games, mimeo (1993), Department of Economics, Yale University.

28. R. Selten, Re-examination of the perfectness concept for equilibrium in extensive games, *International Journal of Game Theory* 4 (1975), 22–25.

6. Appendix

Many of the proofs of the propositions rely on similar arguments. In order to avoid repetition, I will present certain key steps as lemmas.

Lemma 0: *Let $\hat{S}_{-i} \subset S_{-i}$ be an arbitrary set of conjectures. Then $B_i(\hat{S}_{-i})$ is closed.*

Proof: Suppose $s_i^m \in B_i(\hat{S}_{-i})$ is a sequence converging to s_i. There must exist $\overline{m}$ such that for all $m \geq \overline{m}$ and $a_i \in A_i$, $s_i(a_i) > 0$ implies $s_i^m(a_i) > 0$. Then the linearity of U_i implies $s_i \in B_i(s_{-i}^m)$ whenever $s_i^m \in B_i(s_{-i}^m)$. Hence $s_i \in B_i(\hat{S}_{-i})$. ∎

Lemma 1: *For any G, $s_i \in S_i$ is strictly dominated if and only if it is not a best response to some conjecture $s_{-i} \in S_{-i}$; furthermore, s_i is (weakly) dominated if and only if it is not a best response to some conjecture s_{-i} such that $\pi_j(s_{-i}) \in \operatorname{Int} S_j \ \forall j \neq i$.*

Proof: For two-person games, Lemma 1 is proved in Pearce [19] (Lemmas 3 and 4 in the appendix). Since S_{-i}'s allow for correlated conjecture, the n-person case follows from the same arguments. ∎

Proof of Proposition 0: Part (i) is straightforward. To prove part (ii), observe that $R_i^{k+1} \subset R_i^k$. Moreover, if the set of pure strategies in R_i^{k+1} and R_1^k is the same for all i, then $R_i^{k+2} = R_i^{k+1}$ for all i. Hence, the existence of the desired k^* follows from the finiteness of the game G. From part (i) and $\rho^0 \succeq \hat{\rho}$, we have, by induction, $\mathbf{B}^\epsilon(\rho^{k-1}, \rho') = \rho^k \succeq \hat{\rho} = \mathbf{B}^\epsilon(\hat{\rho}, \rho')$ for all k and, in particular, for $k = k^*$. This concludes the proof of part (ii).

To prove part (iii), for $X_i \subset S_i$ let $\mathrm{E}X_i = \{s_i \in X_i \mid s_i$ places the same probability on each element of its support, for any $X_i \subset S_i\}$. Let $X_i^\epsilon = \mathbf{B}_i^\epsilon(\tau)$ for $\tau = (\rho, \rho')$ and $\rho = (R_i)_{i=1}^n$. By the linearity of U^i, $\mathrm{E}X_i^\epsilon = \mathrm{E}X_i^{\epsilon'}$ iff $X_i^\epsilon = X_i^{\epsilon'}$. Since $X_i^{\epsilon'} \subset X_i^\epsilon$ whenever $\epsilon' \leq \epsilon$, it follows from the finiteness of $\mathrm{E}X_i^\epsilon$ that for some $\overline{\epsilon} > 0$, $\mathrm{E}X_i^\epsilon = \mathrm{E}X_i^{\overline{\epsilon}}$ for all $\epsilon < \overline{\epsilon}$. Hence, for $\overline{\epsilon}$ sufficiently small $\mathbf{B}_i^\epsilon(\tau) = \mathbf{B}_i^{\overline{\epsilon}}(\tau)$ for all i and $\epsilon < \overline{\epsilon}$. To prove that $\overline{\epsilon}$ can be chosen so as to satisfy the final assertion of part (iii), note that, since each R_i is closed, $C_{-i}(\tau)$ is closed. Let $Y(x)$ denote the set of all conjectures to which the pure strategy x is a best response. Since U^i is continuous, $Y(x)_{x \in A_i}$ is a (finite) collection of closed sets. It follows that for any $y \in S_{-i}$, we can find an open set θ_y that contains y such that $\theta_y \cap Y(x) = \emptyset$ for all $x \notin B_i(y)$. Since the collection $(\theta_y)_{y \in S_{-i}}$ is an open cover of the compact set $C_{-i}(\tau)$, it has a finite subcover $\theta = \bigcup \theta_y$. Thus, the sets $S_{-i} \backslash \theta$ and $C_{-i}(\tau)$ are disjoint compact sets. Assume $S_{-i} \backslash \theta \neq \emptyset$. Let $\delta = d\big(S_{-i} \backslash \theta, C_{-i}(\tau)\big) = \min\big\{||y' - y''|| \mid y' \in S_{-i} \backslash \theta, \ y'' \in C_{-i}(\tau)\big\}$. It follows that $d\big(\hat{y}, C_{-i}(\tau)\big) < \delta$ implies $\hat{y} \in \theta_y$ for some θ_y of the finite subcover θ. But for $\overline{\epsilon}$ small enough, the set of all $(\overline{\epsilon} - \tau)$-allowable conjectures is within δ of $C_{-i}(\tau)$ (note that this is true even if $S_{-i} \backslash \theta = \emptyset$). Hence, any best response to such a conjecture $\hat{y} \in \theta_y$ is also a best response to y, which yields the desired conclusion. ∎

Proof of Proposition 1: Let $\tilde{\rho} = \mathbf{B}\big(\rho, (S_i)_{i=1}^n\big)$ and $\hat{\rho} = \mathbf{B}(\rho, \rho)$. By part (i) of Proposition 0, $\tilde{\rho} \succeq \hat{\rho}$. By Lemma 0, $\rho = (R_i)_{i=1}^n$ and either (ρ, ρ) or $\big(\rho, (S_i)_{i=1}^n\big)$ is a τ-theory implies each R_i is closed. Hence, by part (iii) of Proposition 0, $\hat{\rho} \succeq \tilde{\rho}$. Thus, $\tilde{\rho} = \hat{\rho}$, which establishes the desired result. ∎

Proof of Proposition 2: It is well-known that $\mathbf{B}(\tau^*) = \rho^*$ (the set of all best responses to the set of all rationalizable conjectures is the set of rationalizable strategies). Hence, ρ^* is a τ-theory. By Lemma 0, $\rho = (R_i)_{i=1}^n$ implies each R_i is closed. Hence by Proposition 0, we have $\mathbf{B}(\rho, \rho) = \mathbf{B}\big(\rho, (S_i)_{i=1}^n\big) \succeq \mathbf{B}(\rho, \rho) = \rho$, so it suffices to show that $\rho^* \succeq \mathbf{B}(\rho, \rho)$. Let $\rho(t) = \big(R(t)\big)_{i=1}^n$ for $t = 1, 2, \ldots$, be the collection of sets of strategies used in the definition of rationalizability (Definition 1). Note that $R_i(0) = S_i$ for all i; hence $\rho(0) \succeq \rho$. Then by induction and part (i) of Proposition 0,

$$\mathbf{B}\big(\rho(t), \rho(t)\big) \succeq \mathbf{B}(\rho, \rho)$$

for all t. This yields the desired conclusion. ∎

Proof of Proposition 3: Let $S_i^u = \overline{A}_i^u$ and $\mathbf{B}_i^u$ denote the mapping $\mathbf{B}_i$ for the game G^u (hence, $\mathbf{B}_i^u(\cdot) \subset S_i^u$). Pick $s_i \in \mathbf{B}_i\big(\rho^p, (\operatorname{Int} S_j)_{j=1}^n\big)$. Since s_i is a best response to an interior conjecture, it follows that $s_i \in S_i^u$. Note that by Lemmas 0 and 1 R_i^p is closed. Then by applying parts (i) and (iii) of Proposition 0 we have

$$s_i \in \mathbf{B}_i\big(\rho^p, (\operatorname{Int} S_j)_{j=1}^n\big) \subset S_i^u \cap \mathbf{B}_i\big(\rho^p, (S_j)_{j=1}^n\big) \subset S_i^u \cap \mathbf{B}_i(\rho^p, \rho^p).$$

By Proposition 1 and part (i) of Proposition 0,

$$S_i^u \cap \mathbf{B}_i(\rho^p, \rho^p) \subset S_i^u \cap \mathbf{B}_i\big(\tau^*(G^u)\big) = \mathbf{B}_i^u\big(\tau^*(G^u)\big) = R_i^*(G^u).$$

The last equality follows from applying Proposition 2 to the game G^u. The equality that precedes it follows from the fact that s_i is a best response in G, and since s_i is a strategy in the game G^u, all best responses in G^u are best responses in G. Hence, $s_i \in R_i^p$.

Next assume $s_i \in R_i^p = R_i^*(G^u) \cap R_i^u$. Then $s_i \in \mathbf{B}_i^u\big(\tau^*(G^u)\big)$ (by Proposition 2) and since only dominated strategies are removed to obtain G^u, we have $\mathbf{B}_i^u\big(\tau^*(G^u)\big) \subset \mathbf{B}_i\big(\tau^*(G^u)\big)$ (that is, $\mathbf{B}_i^u(\hat{\tau}) = \mathbf{B}_i(\hat{\tau}) \cap S_i^u$ for all $\hat{\tau}$). Hence, $s_i \in \mathbf{B}_i\big(\tau^*(G^u)\big)$ and $s_i \in R_i^u$. Then by Lemma 1, $s_i \in \mathbf{B}_i(\hat{s}_{-i})$ for $\hat{s}_{-i}$ such that $\pi_j(s_i) \in \operatorname{Int} S_j$ for all $j \neq i$ and $s_i \in \mathbf{B}_i(s_{-i})$ for $s_{-i} \in C_{-i}\big(\tau^*(G^u)\big) = C_{-i}(\rho^p, \rho^p)$. Then $s_i \in \mathbf{B}_i\big(\lambda s_{-i} + (1-\lambda)\hat{s}_{-i}\big)$ for all $\lambda \in [0, 1]$. Hence, $s_i \in \mathbf{B}_i^\epsilon\big(\rho^p, (\operatorname{Int} S_j)_{j=1}^n\big)$ for all ϵ. That is, $s_i \in \mathbf{B}_i\big(\rho^p, (\operatorname{Int} S_j)_{j=1}^n\big)$.

Proof of Proposition 4: Let $\rho = (R_i)_{i=1}^n$. From Lemma 0 and Proposition 0, we have

(1) $\mathbf{B}(\rho, \rho) = \mathbf{B}\big(\rho, (S_i)_{i=1}^n\big) \succeq \mathbf{B}(\rho, \rho) = \rho.$

Obviously,

(2) $s_i \in \mathbf{B}_i(\tau)$ for some perfect τ-theory implies $s_i \in S_i^u$.

In proving Proposition 3, it was noted that since only dominated strategies are removed in obtaining G^u, we have

(3) $\mathbf{B}_i^u(\hat{\tau}) = \mathbf{B}_i(\hat{\tau}) \cap S_i^u$ for all $\hat{\tau} \in \Upsilon$.

Finally in proving Proposition 2, it was established that

(4) $\mathbf{B}(\rho, \rho) \succeq \rho$ implies $\rho^* \succeq \rho$.

Suppose $s_i \in R_i$. Then by (1), we have $s_i \in \mathbf{B}_i(\rho, \rho)$ and by (2) we have $s_i \in S_i^u$. Therefore $s_i \in \mathbf{B}_i^u(\rho, \rho) = \mathbf{B}_i(\rho, \rho) \cap S_i^u$ by (3). Thus, applying (4) to game G^u, we get $s_i \in R_i^*(G^u)$. Hence $s_i \in R_i^*(G^u) \cap R_i^u = R_i^p$ as desired. ∎

Proof of Proposition 5: Let $\tau_k = (A_i^d, \Sigma_i^k)_{i=1}^n$ where $\Sigma_i^k = \{s_i \in \operatorname{Int} S_i \mid a_i' \in A_i(t)$ and $a_i \in A_i(t-1) \backslash A_i(t)$ implies $s_i(a_i') > k s_i(a_i)\}$. Let $X^k = X_1^k \times X_2^k = B_1(\tau_k) \times B_2(\tau_k)$. It follows from Lemma 0 that X^k is compact. From part (i) of Proposition 0, it follows that $X^{k+1} \subset X^k$. Hence, $\cap X^k = (\cap X_1^k) \times (\cap X_2^k) \neq \emptyset$. Moreover, since the set EX_i^k as defined in the proof of part (iii) of Proposition 0 is finite and $EX_i^k = EX_i^{k'}$ implies $X_i^k = X_i^{k'}$, it follows that there exists some K such that $X^k = X^K$ for all $k \geq K$. Let $R_i^d = B_i(\tau_K)$ and $\rho^d = (R_i^d)$ and $\rho = (\Sigma_i^K)_{i=1}^n$. Next I will prove that $R_i^d \cap A_i = A_i^d$ and hence $C_{-i}(\tau_K) = C_{-i}(\rho^d, \rho)$ and $\rho^d = B(\tau_K) = B(\rho^d, \rho)$ as desired.

Suppose $a_i \notin A_i^d$. Then there exists $s_i' \in \operatorname{Int} \overline{A_i(t)}$ such that $U_i(s_i', a_j) \geq U_i(a_i, a_j)$ for all $a_j \in A_j(t)$ and $\epsilon = U_i(s_i', a_j') - U_i(a_i, a_j') > 0$ for some $a_j \in A_j(t')$. Furthermore, in any stage of the iterative removal algorithm, it can not be the case that all strategies dominating a given strategy are dominated. Hence, we can without loss of generality assume $s_i' \in \operatorname{Int} \overline{A_i(t')}$.

$$\begin{aligned} U_i(s_i', s_j) - U_i(a_i, s_j) &= \sum_{a_j \in A_j} s_j(a_j)\big[U_i(s_i', a_j) - U_i(a_i, a_j)\big] \\ &\geq s_j(a_i')\big[U_i(s_i', a_j') - U_i(a_i, a_j')\big] \\ &\quad + \sum_{a_j \in A_j \backslash A_j(t)} s_j(a_j)\big[U_i(s_i', a_j) - U_i(a_i, a_j)\big] \\ &\geq \epsilon s_j(a_i') - m\overline{\epsilon}p \end{aligned}$$

where m is the cardinality of A_j, $\overline{\epsilon}$ is the maximum of $\big|U_i(s_i', a_j) - U_j(a_i, a_j)\big|$ for all $a_j \in A_j$, and $p = \max_{a_j \in A_j \backslash A_j(t)} s_j(a_j)$. But $\epsilon s_i(a_i') - m\overline{\epsilon}p \geq p[\epsilon k\overline{\epsilon} - m\overline{\epsilon}]$ whenever $s_j \in C_{-i}(\tau_k)$ and hence by choosing $k > \frac{m}{\epsilon}$, we can guarantee that $U_i(s_i', s_j) > U_i(a_i, s_j)$ for all $s_j \in C_{-j}(\tau_k)$. Hence $a_i \notin R_i^d$.

Next assume that $a_i \in A_i^d$. Then there exists by Lemma 1 a collection (s_j^t), for $t = 0, 1, \ldots$, such that $s_j^t \in \operatorname{Int} \overline{A_j(t)}$ and $a_i \in B_i(s_j^t)$ for all t. Thus by linearity, $a_i \in B_i(\sum_t \alpha^t s_j^t)$ for all $(\alpha^t)_{t=1}^{\overline{t}}$ such that $\sum \alpha^t = 1$ and $\alpha^t > 0$. Clearly we can choose $(\alpha^t)_{t=1}^{\overline{t}}$ so that $\sum_t \alpha^t s_j^t \in C_{-i}(\tau_K)$ so that $a_i \in B_i(\sum_t \alpha^t s_j^t)$ implies $a_i \in R_i^d$ as desired.

Finally, suppose $\hat{\tau} = (\hat{\rho}, \rho)$ is a τ-theory. Let $\rho(t) = \big(A_i(t)\big)_{i=1}^n$ for $t = 0, 1, \ldots$. Clearly, $\rho(0) \succeq \hat{\rho}$. If $\rho(t) \succeq \hat{\rho}$, then by part (i) of Proposition 0, $B_i\big(\rho(t), \rho\big) \succeq \hat{\rho}$. Since G is a finite game, there exists $\overline{t}$ such that $A_i^d = A_i(t)$ for all i. Hence, $\rho^d \succeq \hat{\rho}$ which proves that $\tau^d = (\rho^d, \rho)$ has no type 1 restrictions. ∎

Proof of Proposition 6: Let $S_i^x = \overline{A}_i^x$ and $\mathbf{B}_i^{G^x}$ denote the mapping B_i^x for the game G^x. In proving Proposition 3, we noted that $B_i^u(\hat{\tau}) = B_i(\hat{\tau}) \cap S_i^u$ for any $\hat{\tau}$. But if $s_i \in S_i^u$, then s_i is a best response to some conjecture s_{-i} that reaches every information set (Lemma 1), and hence $s_i \in R_i^x \subset S_i^x$. Thus, $\emptyset \neq \mathbf{B}_i(\hat{\tau}) \cap S_i^u \subset \mathbf{B}_i(\hat{\tau}) \cap S_i^x$ and $\mathbf{B}_i^{G^x}(\hat{\tau}) = \mathbf{B}_i(\hat{\tau}) \cap S_i^x$ for all $\hat{\tau}$. Let $s_i \in \mathbf{B}_i^x(\tau^e) = \mathbf{B}_i(\tau^e) \cap R_i^x$. Hence by part (i) of Proposition 0, $s_i \in \mathbf{B}_i\big(\tau^*(G^x)\big)$ and by Proposition 2, $s_i \in R_i^*(G^x)$; and since $s_i \in R_i^x$ by assumption we have $s_i \in R_i^e$. Let $s_i \in R_i^e = R_i^*(G^x) \cap R_i^x$. Then $s_i \in \mathbf{B}_i^{G^x}\big(\tau^*(G^x)\big) \cap R_i^x$. But as noted above $\mathbf{B}_i^{G^x}\big(\tau^x(G^x)\big) = \mathbf{B}_i\big(\tau^*(G^x)\big) \cap S_i^x$ so $s_i \in \mathbf{B}_i\big(\tau^*(G^x)\big) \cap R_i^x = \mathbf{B}_i^x\big(\tau(G^x)\big)$ as desired.

Now $\rho = \mathbf{B}^x(\rho, \rho')$ implies $\rho \succeq (R_i^x)_{i=1}^n$ and $\mathbf{B}(\rho, \rho) \succeq \rho$ by Propositions 0 and 1. But since $\mathbf{B}_i^{G^x}(\rho, \rho) = \mathbf{B}_i(\rho, \rho) \cap S_i^x$ for all i, we have $\mathbf{B}^{G^x}(\rho, \rho) \succeq \rho$. By part (4) of Proposition 4, this implies $\rho^*(G^x) \succeq \rho$. Hence, $\rho^e \succeq \rho$. ∎

Proof of Proposition 7: Follows from the definition of a perfect τ-theory and the observation that conjectures in a perfect τ-theory reach every information set.

Proof of Proposition 8: Consider the agent normal form Γ_a of Γ. Let b_i^0 be the behavioral strategy for player i that places equal probability $1 - \epsilon$ on the unique backward induction action at every information set and $\frac{\epsilon}{k-1}$ on each of the remaining actions, where k is the total number of actions at the information set. Let s_i^0 be the strategy in Γ for player i that is equivalent to b_i^0. For ϵ small enough, the only best response to s_i^0 by player j is the backward induction strategy (note that given our definition of a strategy in extensive forms, this best response is a unique strategy). Thus, if a_1, a_2 are the two backward induction strategies for Γ, then $\tau = \Big(\big(\{a_1\}, \{a_2\}\big), \big(\{s_1^0\}, \{s_2^0\}\big)\Big)$ is the desired extensive form τ-theory. ∎

Proof of Proposition 9: By Proposition 0 there exists $\overline{\epsilon}$ such that

$$\mathbf{B}_i^{\overline{\epsilon}}(\tau^t) = \mathbf{B}_i^{\overline{\epsilon}}(\hat{\tau}^t) = R_i(t+1) \text{ for all } t = 0, 1, 2, \ldots \overline{t}$$

where $R_i(t)$'s are the sets used to define R_i^* and $\overline{t}$ is the last iteration of the definition (i.e., $R_i(t) = R_i^*$ for all $t \geq \overline{t}$), $\tau^t = \big(R_i(t), S_i\big)_{i=1}^n$, and $\hat{\tau}^t = \big(R_i(t), R_i(t)\big)_{i=1}^n$. Set $\delta = \frac{\overline{\epsilon}}{K}$ where K is the total number of pure strategies. After any initial history, no strategy $s_i \notin \mathbf{B}_i(\tau^0)$ will be played. This implies that, by some finite time v, the relative frequency $a_i \notin R_i(1)$ will fall below $\frac{\delta}{2}$ so that MHD_δ will require that $b_i(h^v)$ is a $(\overline{\epsilon} - \tau^1)$-allowable conjecture. Hence, $s_i \in B_i\big(b_i(h^v)\big)$ will imply $s_i \in R_i(\hat{\tau}^1)$ for all i, which in finite time will imply that the relative frequency of strategies $a_i \notin R_i(2)$ will fall below $\frac{\delta}{2}$, etc., so that in finite time MHD_δ will imply that $s_i \in B_i\big(b_i(h^v)\big)$ only if $s_i \in R_i^*$. The cautious case requires a straightforward adjustment of the above argument. ∎