EPISTEMIC FOUNDATIONS OF SOLUTION CONCEPT IN GAME THEORY: AN INTRODUCTION

GIACOMO BONANNO
KLAUS NEHRING

IN DB

Working Paper Series #97-21

Department of Economics
University of California
Davis, California 95616-8578

Epistemic Foundations of Solution Concepts in Game Theory: An Introduction

Giacomo Bonanno
and
Klaus Nehring

Working Paper Series No. 97-21
June, 1997

/

Paper prepared for an invited lecture at the Workshop on Bounded Rationality and Economic Modeling, Firenze, July 1 and 2, 1997 (Interuniversity Centre for Game Theory and Applications and Centre for Ethics, Law and Economics)

Abstract

We give an introduction to the literature on the epistemic foundations of solution concepts in game theory. **Only normal-form** games are considered. The solution concepts analyzed are **rationalizability,** strong rationalizability, correlated equilibrium and **Nash** equilibrium. The analysis is carried out locally in terms of properties of the belief hierarchies. Several examples are used throughout to illustrate definitions and concepts.

1. Introduction

The objective of the literature on the epistemic foundations of solution concepts in games is to determine what assumptions on the beliefs and reasoning of the players are implicit in various solution concepts. This is a recent line of inquiry in game theory and one that is gaining momentum. In this paper we give an introduction to the general approach and review some of the main contributions. We will provide a selective, rather than encompassing, survey. For a more ambitious and more comprehensive review of the issues dealt with in the literature on the epistemic foundations of game theory see Dekel and Gul (1997).

Why worry about the epistemic foundations of solution concepts? A common view is that results that relate epistemic conditions (such as common belief in rationality) to a particular solution concept help explain how introspection alone can lead players to act in accordance with it. The task of this research program is to identify for any game the strategies that might be chosen by rational and intelligent players who know the structure of the game and the preferences of their opponents and who recognize each other's rationality and knowledge.

Although several of the papers in the literature deal with the case of knowledge and common knowledge, we will take a more general point of view where the primitive concept is that of belief (and knowledge can be viewed as a particular form of belief: cf. Stalnaker, 1994, 1996).

The paper is devoted mainly to the analysis of normal-form (or strategic-form) games, although the implications for extensive games are sometimes discussed.[1] In Section 2 we discuss

[1] **There is, however, a very recent literature that deals with the epistemic foundations of solution concepts for extensive games, in particular backward induction in perfect-information games. *See*, for example. Aumann,**

Bayesian and qualitative frames and their properties and in Section 3 we use them to define the notion of a model for a **normal-form** game. In Section 4 we consider the notions of rationalizability and strong rationalizability, while Sections 5 and 6 are devoted to the epistemic foundations of correlated equilibrium and Nash equilibrium, respectively.

2. Bayesian and qualitative frames and their properties

DEFINITION 1. An interactive ***Bayesian frame*** (or Bayesian frame, for short)[2] is a tuple

$$\mathcal{B} = \langle N, \Omega, \tau, \{p_i\}_{i \in N} \rangle$$

where

- $N = \{1, ..., n\}$ is a finite set of individuals.
- Ω is a finite set of states (or possible worlds)[3]. The subsets of Ω are called events.
- $\tau \in 52$ is the "true" or "actual" state[4].
- for every individual $i \in N$, $p_i : 52 \to \Delta(\Omega)$ (where $\Delta(\Omega)$ denotes the set of probability distributions over Ω) is a function that specifies herprobabilistic beliefs, satisfying the following property [we use the notation $p_{i.\alpha}$ rather than $p_i(\alpha)$]: $\forall a, \beta \in \Omega$,

$$\text{if} \quad p_{i,\alpha}(\beta) > 0 \quad \text{then} \quad p_{i,\beta} = p_{i,\alpha} \tag{1}$$

(1995, 1996), Battigalli (1997), Battigalli and Siniscalchi (1997), Ben Porath (1997), Stalnaker (1996, 1997), Stuart (1997).

[2] For a similar definition see, for example, Aumann and Brandenburger (1995), Dekel and Gul (1997) and Stalnaker (1994, 1996).

[3] Finiteness of Ω is a common assumption in the literature (cf. Aumann, 1987, Aumann and Brandenburger, 1995, Dekel and Gul, 1997, Moms, 1994, Stalnaker, 1994, 1996).

Thus $p_{i,\alpha} \in \Delta(\Omega)$ is individual **i's** subjective probability distribution at state α and condition (1) says that every individual knows her own beliefs. We denote by $\|p_i = p_{i,\alpha}\|$ the event $\{\omega \in \Omega : p_{i,\omega} = p_{i,\alpha}\}$. It is clear that the set $\{\ \|p_i = p_{i,\omega}\| : \omega \in \Omega\ \}$ is a partition of Ω; it is often referred to as individual i's *type partition*.

DEFINITION 2. Given a Bayesian frame $\mathcal{B}$, its ***qualitative frame*** (or frame, for short) is the tuple $\mathcal{Q} = (N,\ \Omega,\ r,\ \{P_i\}_{i \in N})$ where N, Ω, and r are as in Definition 1 and

- for every individual $i \in N$, $P_i : \Omega \to 2^{\Omega} \backslash \varnothing$ is i's *possibility correspondence*, derived from i's probabilistic beliefs as follows: [5]

$$P_i(\alpha) := \text{supp}(p_{i,\alpha}).$$

Thus, for every $\alpha \in \Omega$, $P_i(\alpha)$ is the set of states that individual i considers possible at α.

REMARK 1. It follows from condition (1) of Definition 1 that the possibility correspondence of every individual **i** satisfies the following properties (whose interpretation is given in Footnote 7): $\forall\, \alpha, \beta \in \Omega$,

Transitivity:	if $\beta \in P_i(\alpha)$ then $P_i(\beta) \subseteq P_i(\alpha)$,
Euclideanness:	if $\beta \in P_i(\alpha)$ then $P_i(\alpha) \subseteq P_i(\beta)$.

REMARK 2 (Graphical representation). A non-empty-valued and transitive possibility correspondence $P : \Omega \to 2^{\Omega}$ can be uniquely represented (see Figures below) as an

[4] **We have included the true state in the definition of an interactive Bayesian model in order to stress the interpretation of the model as a representation of a particular profile of hierarchies of beliefs.**

[5] **If $\mu \in \Delta(\Omega)$, supp(μ) denotes the support of μ, that is, the set of states that are assigned positive probability by μ.**

asymmetric directed graph[6] whose vertex set consists of disjoint events (called *cells* and represented as rounded rectangles) and states, and each arrow goes from, or points to, either a cell or a state that does not belong to a cell. In such a directed graph, $\omega' \in P(\omega)$ if and only if either ω and ω' belong to the same cell or there is an arrow from ω, or the cell containing ω, to ω', or the cell containing ω'. Conversely, given a transitive directed graph in the above class such that each state either belongs to a cell or has an arrow out of it, there exists a unique non-empty-valued, transitive possibility correspondence which is represented by the directed graph. The possibility correspondence is euclidean if and only if all arrows connect states to cells and no state is connected by an arrow to more than one cell.
Finally, if – in addition – the possibility correspondence is reflexive, then one obtains a partition model where each state is contained in a cell and there are no arrows between cells.

Given a frame and an individual i, i's *belief* (or certainty) *operator* $B_i : 2^{\Omega} \to 2^{\Omega}$ is defined as follows: $\forall E \subseteq \Omega$, $B_i E = \{\omega \in \Omega : P_i(\omega) \subseteq E\}$. $B_i E$ can be interpreted as the event that (i.e. the set of states at which) individual i *believes for sure* that event *E* has occurred (i.e. attaches probability 1 to E).[7]

Notice that we have allowed for false beliefs by not assuming reflexivity of the possibility correspondences ($\forall \alpha \in \Omega, \alpha \in P_i(\alpha)$), which – as is well known (Chellas, 1984, p. 164) – is equivalent to the *Truth Axiom* (if the individual believes E then E is indeed true):

---

[6] **A directed graph is *asymmetric* if, whenever there is an arrow from vertex v to vertex v' then there is no arrow from v' to v.**

[7] **Thus Condition (1) of Definition 1 can be stated as follows: $\forall i \in N, \forall \alpha \in \Omega$, $\|p_i = p_{i,\alpha}\| = B_i \|p_i = p_{i,\alpha}\|$.**

$\forall E \subseteq \Omega,\ B_i E \subseteq E$[8].

The common belief operator B_* is defined as follows. First, for every $E \subseteq \Omega$, let $B_e E = \bigcap_{i \in N} B_i E$, that is, $B_e E$ is the event that everybody believes E. The event that E is commonly believed is defined as the infinite intersection:

$$B_* E = B_e E \cap B_e B_e E \cap B_e B_e B_e E \cap \ldots$$

The corresponding possibility correspondence P_* is then defined as follows: for every $\alpha \in \Omega$, $P_*(\alpha) = \{ \omega \in \Omega : \alpha \in \neg B_* \neg \{\omega\}\}$. It is well known that P_* can be characterized as the transitive ***closure*** of $\bigcup_{i \in N} P_i$, that is,

> $\forall \alpha, \beta \in \Omega$, $\beta \in P_*(\alpha)$ if and only if there is a sequence $\langle i_1, \ldots i_m \rangle$ in N and a sequence $\langle \eta_0, \eta_1, \ldots, \eta_m \rangle$ in Ω such that: (i) $\eta_0 = \alpha$. (ii) $\eta_m = \beta$ and (iii) for every $k = 0, \ldots, m-1$, $\eta_{k+1} \in P_{i_{k+1}}(\eta_k)$.

Note that, although P_* is always non-empty-valued and transitive, in general it need not be euclidean (despite the fact that the individual possibility correspondences are; recall that – cf. Footnote 7 – P_* is euclidean if and only if B_* satisfies Negative Introspection).

With reference to qualitative frames, we now define events that capture important properties of beliefs.

[8] It is well known (see Chellas, 1984, p. 164) that non-empty-valuedness of the possibility correspondence is equivalent to ***consistency*** of beliefs (an individual cannot simultaneously believe E and not E): $\forall E \subseteq \Omega$, $B_i E \subseteq \neg B_i \neg E$ (where, for every event F, $\neg F$ denotes the complement of F). Transitivity of the possibility correspondence is equivalent to ***positive introspection*** of beliefs (if the individual believes E then she believes that she believes E): $\forall E \subseteq \Omega$, $B_i E \subseteq B_i B_i E$. Finally, euclideanness of the possibility correspondence is equivalent to ***negative introspection*** of beliefs (if the individual does not believe E, then she believes that she does not believe E): $\forall E \subseteq \Omega$, $\neg B_i E \subseteq B_i \neg B_i E$.

Event	Corresponding property of beliefs
$\mathbf{T} = \bigcap_{i\in N} \bigcap_{E\in 2^\Omega} \neg(B_iE \cap \neg E)$	**No individual has false beliefs:** For every $a \in \Omega$, $a \in$ T if **and** only if no individual has any false beliefs at a (for every $i \in$ **N and** for every $\mathbf{E} \subseteq \mathbf{\Omega}$, if $\alpha \in \mathbf{B_i E}$ then $\alpha \in$ E)
$\mathbf{B_* T}$	**Common belief in no error:** For every $a \in$ St, $a \in$ B,T if and only if at a it is common belief that no individual has any false beliefs
$\mathbf{Q} = \neg B_* \neg B_* T$	**Quasi-coherence of beliefs:** For every $a \in \Omega$, $a \in \boldsymbol{Q}$ if and only if at a it is commonly possible that it is common belief that no individual has any false beliefs
$\mathbf{T}^* = \bigcap_{E\in 2^\Omega} \neg(B_*E \cap \neg E)$	**Truth of common belief:** $\alpha \in \mathbf{T}^*$ if and only if at a whatever is commonly believed is true (for every event E, if $\alpha \in$ B,E then $\alpha \in$ E)
$\mathbf{T_{CB}} = \bigcap_{i\in N} \bigcap_{E\in 2^\Omega} \neg(B_iB_*E \cap \neg B_*E)$	**Truth *about* common belief:** $\alpha \in \mathbf{T_{CB}}$ if and only if, for every event E and every individual i, if, at a, individual i believes that E is commonly believed, then, at a, E is indeed commonly believed (if $\alpha \in B_iB_*E$ then $\alpha \in B_*E$)
$\mathbf{NI} = \bigcap_{E\in 2^\Omega} (B_*E \cup B_*\neg B_*E)$	**Negative Introspection of common belief:** $a \in$ **NI** if and only if – for every event E – whenever at a it is not common belief that E, then, at a, it is common belief that E is not commonly believed (if $a \in \neg \mathbf{B_*E}$ then $a \in \mathbf{B_*\neg B_*E}$)

The following propositions establish **the** relationship between some of these properties.

PROPOSITION 1 (Bonanno and Nehring, 1997a). $\mathbf{NI} = \mathbf{T}_{CB} \cap B_* \mathbf{T}_{CB}$

PROPOSITION 2. (Bonanno and Nehring, 1997b). $T \cap B_*T = T^* \cap B_* \mathbf{T}_{CB} \cap \mathbf{Q}$.

3. Models of normal-form games

Throughout this paper we shall restrict attention to finite games. A finite normal-form or strategic-form game is a tuple $G = \langle N, \{S_i\}_{i\in N}, \{u_i\}_{i\in N}\rangle$ where $N = \{1, 2, \ldots, n\}$ is a set of ***players***, S_i is the set of strategies of player i and $u_i : S \to \mathbb{R}$ (where $S = S_1 \times \ldots \times S_n$ and $\mathbb{R}$ is the set of real numbers) is player i's von Neumann Morgenstern payoff (or utility) function. This (standard) definition of game represents only a partial description in that it determines the choices that are available to the players and the preferences that motivate the choices, but does not specify the players' beliefs about each other or their actual choices. The notion of model provides a way of completing the description.

DEFINITION 3. Fix a normal-form game G. A ***model*** of G is a pair $\mathcal{M} = \langle \mathcal{B}, \{\sigma_i\}_{i\in N}\rangle$ where $\mathcal{B} = \langle N, St, \tau, \{p_i\}_{i\in N}\rangle$ is a Bayesian frame and, for every player i, $\sigma_i : St \to S_i$ is a function that specifies for every state the choice made by player i at that state subject to the restriction that player i knows her own strategy:

$$\forall i \in N, \forall \alpha,\beta \in St, \text{ if } p_{i,\alpha} = p_{i,\beta} \text{ then } \sigma_i(\alpha) = \sigma_i(\beta).$$

For every state $\omega \in St$, let $\sigma(\omega) = (\sigma_1(\omega), \ldots, \sigma_n(\omega))$ be the strategy profile played at ω and, for every player i, denote by $\sigma_{-i}(\omega) = (\sigma_1(\omega),\ldots,\sigma_{i-1}(\omega),\sigma_{i+1}(\omega),\ldots,\sigma_n(\omega))$ the strategies played by the players other than i.

The addition of a strategy profile at every state is what gives content to the beliefs of the players.

DEFINITION 4. Player i is ***rational at state*** $a \in 52$ if her choice at a maximizes her expected utility, given her beliefs at a : for all $x \in S_i$,

$$\sum_{\omega \in P_i(\alpha)} u_i\left(s_i^{\alpha}, \sigma_{-i}(\omega)\right) p_{i,\alpha}(\omega) \geq \sum_{\omega \in P_i(\alpha)} u_i\left(x, \sigma_{-i}(\omega)\right) p_{i,\alpha}(\omega)$$

where $s_i^{\alpha} = \sigma_i(\alpha)$ (recall that i's own strategy is the same at every $o \in P_i(\alpha)$). Let $\mathbf{RAT}_i$ be the set of states where player **i** is rational and **RAT** = $\bigcap_{i \in N}$ $\mathbf{RAT}_i$ the event that all players are rational.

EXAMPLE 1. Figure 1b shows a model of the two-person game illustrated in Figure 1a. Here we have that **RAT,** = $\{\tau, \beta\}$ and **RAT,** = Ω ; hence **RAT** = $\{\tau, \beta\}$. Note also that B_1**RAT** = $\{\tau, \beta\}$, B_2**RAT** = $\{\tau\}$ and B_***RAT** = O.

Insert Figure 1

		Player 2		
		L	C	R
Player 1	T	4 , 6	3 , 2	8 , 0
	M	0 , 9	0 , 0	4 , 12
	B	8 , 3	2 , 4	0 , 0

Figure 1a

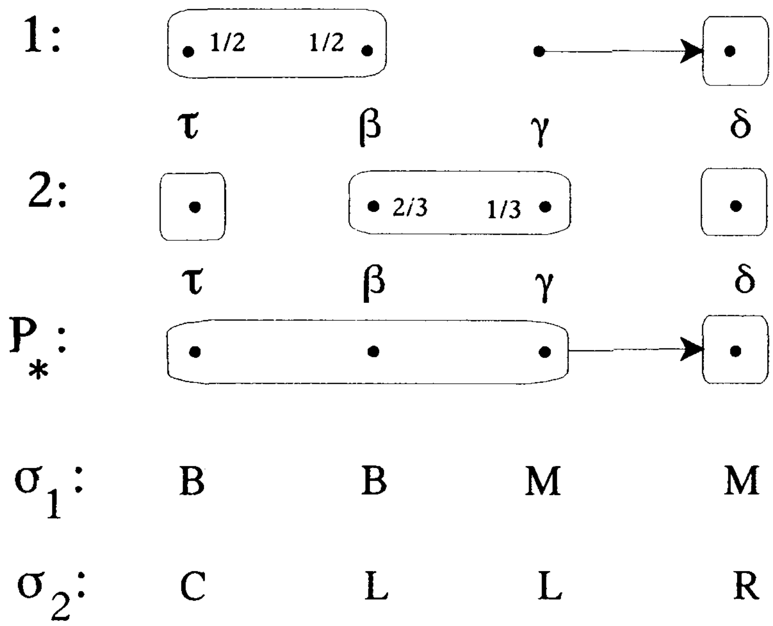


Figure 1b

REMARK 3. Note that, for every player i, $B_i\mathbf{RAT}_i$ = RAT,, that is, no player can have false beliefs about her own rationality and if she is rational then she knows it.[9] It follows that $\mathbf{B_*RAT} \subseteq$ RAT, that is, if it is common belief that all the players are rational, then they are indeed rational.[10] On the other hand, as Example 1 shows, in general RAT $\not\subseteq \mathbf{B_*RAT}$.

4. Rationalizability and Strong Rationalizability

The first solution concept we consider is rationalizability (Bernheim, 1984, Pearce, 1984).

[9] **If $\alpha \in B_i\mathbf{RAT}_i$ then $\beta \in$ RAT, for all $\beta \in I_i(\alpha)$. It follows that $\alpha \in \mathbf{RAT}_i$ since i's beliefs and strategy are the same at α as at any $\beta \in I_i(\alpha)$. The converse follows similarly.**

[10] **For all $i \in N$, $\mathbf{B_*RAT} \subseteq B_i\mathbf{RAT} \subseteq B_i\mathbf{RAT}_i = \mathbf{RAT}_i$.**

DEFINITION 5. For every player i, let $\Delta(S_i)$ be the set of probability distributions over S_i (the set of player i's mixed strategies). If $\mu_i \in \Delta(S_i)$ and $s_i \in S_i$, we denote by $\mu_i(s_i)$ the probability assigned to s_i by μ_i. A strategy $s_i \in S_i$ is ***strictly*** dominated by $\mu_i \in \Delta(S_i)$ if, for all $s_{-i} \in S_{-i}$, $u_i(\mu_i, s_{-i}) > u_i(s_i, s_{-i})$, where $u_i(\mu_i, s_{-i}) = \sum_{x \in S_i} \mu_i(x)\, u_i(x, s_{-i})$. [For example, in the game of Figure 2a, strategy B of player 1 is strictly dominated by the mixture $(\frac{1}{2}A, \frac{1}{2}D)$]. Given a game G, let G^1 be the game obtained by eliminating the strictly dominated strategies of every player; let G^2 be the game obtained from G^1 by eliminating the strictly dominated strategies of every player, etc. Let G^∞ be the game obtained from G after the iterative deletion of strictly dominated strategies and S^∞ the set of strategy profiles of G^∞. The profiles in S^∞ are called ***rationalizable***. For the game of Figure 2a, the games G^1, G^2 and $G^3 = G^\infty$ are shown in Figures 2b-d. In the game of Figure 1a, $S^\infty = \{(T,L), (T,C), (B,L), (B,C)\}$, since for Player 1 M is strictly dominated by T and – after deletion of M – for Player 2 R becomes strictly dominated by both L and C.

Insert Figure 2

Player 2

Player 1		a	b	c
	A	3 , 0	1 , 0	0 , 1
	B	1 , 1	0 , 2	1 , 1
	C	0 , 0	4 , 1	2 , 2
	D	0 , 3	1 , 0	3 , 2

Figure 2a

The game G: B is strictly dominated by $(\frac{1}{2}A, \frac{1}{2}D)$.

Player 2

Player 1		a	b	c
	A	3 , 0	1 , 0	0 , 1
	C	0 , 0	4 , 1	2 , 2
	D	0 , 3	1 , 0	3 , 2

Figure 2b

The game G^1: now b is strictly dominated by c.

Player 2

Player 1		a	c
	A	3 , 0	0 , 1
	C	0 , 0	2 , 2
	D	0 , 3	3 , 2

Figure 2c

The game G^2: now C is strictly dominated by $(\frac{1}{6}A, \frac{5}{6}D)$

Player 2

Player 1		a	c
	A	3 , 0	0 , 1
	D	0 , 3	3 , 2

Figure 2d

The game G^3: no strategy is strictly dominated; thus $G^3 = G^\infty$ and $S = \{(A,a), (A,c), (D,a), (D,c)\}$.

The following proposition was established by Bernheim (1984) and Pearce (1984) and proved more formally in an epistemic context by Brandenburger and Dekel (1987), Tan and Werlang (1988), Stalnaker (1994). Given a game G and a model $\mathcal{M}$ of it, with slight abuse of notation let $\mathbf{S}^\infty$ be the event that a strategy profile that survives iterated deletion of strictly dominated strategies is played: $\mathbf{S}^\infty = \{\omega \in \Omega : \sigma(\omega) \in S^\infty\}$. For example, in the model of Figure 1b, $\mathbf{S}^\infty = \{\tau, \beta\}$.

PROPOSITION 3. Let **G** be a game and $\mathcal{M}$ a model of it. Then

$$B_*\mathbf{RAT} \subseteq \mathbf{S}^\infty \cap B_*\mathbf{S}^\infty.$$

That is, if at a state there is common belief in rationality then the strategy profile played at that state is rationalizable and it is common belief that only rationalizable strategy profiles are played.

Proposition 3 is a consequence of the fact that a strategy $s_i \in S_i$ is a best response to some belief on (probability distribution over) S_{-i} if and only if it is not strictly dominated. Thus if $a \in B_*\mathbf{RAT}$ then $a \in$ RAT (since $B_*\mathbf{RAT} \subseteq$ RAT: see Remark 3) hence no player is choosing a strategy which is strictly dominated in G. Since, for every i, $B_*\mathbf{RAT} \subseteq B_i\mathbf{RAT}$, at a every player believes that no player has chosen a strictly dominated strategy in G. Hence no player is choosing a strategy which is strictly dominated in G^1, etc.

The converse of Proposition 3 does not hold. To see this, consider the following model of the game of Figure 2: $\Omega = \{\tau\}$, $P_1(\tau) = P_2(\tau) = \{\tau\}$, $\sigma(\tau) = (A,a)$. Then $\tau \in \mathbf{S}^\infty \cap B_*\mathbf{S}^\infty$ but $\mathbf{RAT}_2 = \emptyset$ and hence $B_*\mathbf{RAT} = \emptyset$. The following proposition gives a partial converse to Proposition 3 and shows that the notion of common belief in rationality is not stronger than the notion of rationalizability.

PROPOSITION 4. Let G be a game and $s \in S^\infty$. Then there is a model $\mathcal{M}$ of G such that: (1) $\tau \in B_*\mathbf{RAT}$, and (2) $\sigma(\tau) = s$.

In constructing the model of Proposition 4 one can take $\Omega = S^{\infty}$ and use the fact that for every strategy s_i of player i in game G^{∞} there is a probability distribution over the strategies of the opponents relative to which s_i is a best reply.

Propositions 3 and 4 are not based on any assumption of correctness of players' beliefs. Thus a player can be mistaken in the strategy choices and/or beliefs she attributes to the other players. A natural question to ask is whether ruling out incorrect beliefs further reduces the set of strategy profiles that can be played when there is common belief in rationality. The answer is affirmative, as Stalnaker (1994) shows (see also Bonanno and Nehring, 1996b). The following algorithm is similar to the iterative deletion of strictly dominated strategies, but differs from the latter in that it requires the iterative deletion of *profiles* rather than strategies.

DEFINITION 6. Given a normal-form game G, a strategy profile $x \in X \subseteq S$ is inferior relative to X if there exists a player i and a (possibly mixed) strategy μ_i of player i (whose support can be any subset of S_i, not necessarily the projection of X onto S_i) such that:

(1) $u_i(x) < u_i(\mu_i, x_{-i})$ and

(2) for all $s_{-i} \in S_{-i}$ such that $(x_i, s_{-i}) \in X$, $u_i(x_i, s_{-i}) \leq u_i(\mu_i, s_{-i})$.

[Thus if $X = S$ then x is inferior if and only if there is a player i for whom x_i is weakly dominated by some strategy s_i such that $u_i(s_i, x_{-i}) > u_i(x)$.] For every $k \geq 0$, define $S_s^k \subseteq S$ and $D_s^k \subseteq S$ as follows: $S^0 = S$, D_s^k is the set of profiles that are inferior relative to S_s^k and $S_s^{k+1} = S_s^k \setminus D_s^k$. Let $S_s^{\infty} = \bigcap_{k=1}^{\infty} S_s^k$. The strategy profiles in S_s^{∞} are called *strongly rationalizable.*

EXAMPLE 2. For the game illustrated in Figure 3, $S_s^{\infty} = \{(D,d), (D,a), (A,d)\}$. In fact, $(A, a) \notin S_s^{\infty}$ since it is inferior relative to S (for Player 1 A is weakly dominated by D and $u_1((A, a)) = 0 < u_1((D, a)) = 1$). Note that, on the other hand, $S^{\infty} = S$ (that is, all strategy profiles are rationalizable) since no player has any strictly dominated strategies.

Insert Figure 3

		Player 2 d	a
Player	D	1 , 1	1 , 1
1	A	1 , 1	0 , 0

Figure 3

EXAMPLE 3. In the game of Figure 4a, the first step in the algorithm leads to the profiles shown in Figure 4b [for Player 2 D is weakly dominated by E and for Player 1 C is weakly dominated by B], the second step leads to the profiles shown in Figure 4c [now F is dominated by E and C is dominated by A] and the third and final step to the profiles shown in Figure 4d [now B is dominated by A]. Thus $S_s^\infty = \{(B,D), (C,D), (A,E), (A,F)\}$. Note again that that $S^\infty = S$ since no player has any strictly dominated strategies.

Insert Figure 4

		Player 2 D	E	F
Player	A	2 , 0	2 , 2	0 , 2
1	B	2 , 2	1 , 2	5 , 1
	C	2 , 0	1 , 0	1 , 5

Figure 4a

$S_s^0 = S, D_s^0 = \{(A, D), (C, F)\}$

		Player 2		
		D	E	F
Player 1	A		2 , 2	0 , 2
	B	2 , 2	1 , 2	5 , 1
	C	2 , 0	1 , 0	

Figure 4b

$S_s^1 = \{(A, E), (A, F), (B,D), (B, E), (B, F), (C, D), (C, E)\}$

$D_s^1 = \{(C, E), (B, F)\}$

		Player 2		
		D	E	F
Player 1	A		2 , 2	0 , 2
	B	2 , 2	1 , 2	
	C	2 , 0		

Figure 4c

$S_s^2 = \{(A, E), (A, F), (B,D), (B, E), (C, D)\}$, $D_s^2 = \{(B, E)\}$.

		Player 2		
		D	E	F
Player 1	A		2 , 2	0 , 2
	B	2 , 2		
	C	2 , 0		

Figure 4d

$S_s^3 = S_s^\infty \ \{(A, E), (A, F), (B,D), (C, D)\}$, $D_s^3 = \emptyset$.

Given a game G and a model $\mathcal{M}$ of it, with slight abuse of notation let $\mathbf{S}_s^\infty$ be the event that a strongly rationalizable strategy profile is played: $\mathbf{S}_s^\infty = \{w \in \Omega : \sigma(\omega) \in S_s^\infty\}$.

PROPOSITION 5 (Stalnaker, 1994; see also Bonanno and **Nehring, 1996b).**[11] Let G be a game and $\mathcal{M}$ a model of it. Then

$$\text{(1)} \qquad \mathbf{B_*T} \cap \mathbf{B_*RAT} \subseteq \mathbf{B_*S_s^\infty} \quad \text{and}$$

$$\text{(2)} \qquad \mathbf{T^*} \cap \mathbf{B_*T} \cap \mathbf{B_*RAT} \subseteq \mathbf{S_s^\infty} \cap \mathbf{B_*S_s^\infty}.$$

That is, if there is common belief in no error and common belief in rationality, then it is common belief that only strongly rationalizable profiles are played. If, furthermore, Truth of common belief also holds, then it is also true that the strategy profile actually played is strongly rationalizable.

Proof. Fix an arbitrary $\alpha \in \mathbf{B_*T} \cap \mathbf{B_*RAT}$. (1) For every $\omega \in \mathbf{P}_*(\alpha)$ define $j(\omega)$ as follows: $j(\omega) = \infty$ if $\sigma(\omega) \in S_s^\infty$ and $j(\omega) = k \in \mathbb{N}$ (where $\mathbb{N}$ is the set of non-negative integers) if $\sigma(\omega) \in S_s^k$ and $\sigma(\omega) \notin S_s^{k+1}$. Clearly $j(\omega)$ is well defined, since $\sigma(\omega) \in S_s^0$ for all $\omega \in \Omega$. Let $\bar{k}$ be the minimum of $\{j(\omega)\}_{\omega \in \mathbf{P}_*(\alpha)}$. Suppose that $\mathbf{P}_*(\alpha) \not\subseteq \{\omega \in \Omega : \sigma(\omega) \in S_s^\infty\}$. Then $\bar{k} < \infty$. Let $\bar{\omega} \in \mathbf{P}_*(\alpha)$ be such that $j(\bar{\omega}) = \bar{k}$. Then $\sigma(\bar{\omega}) \in D_s^{\bar{k}}$, that is, $\sigma(\bar{\omega})$ is inferior relative to $S_s^{\bar{k}}$. Thus there is a player i and a (possibly mixed) strategy μ_i of player i such that:

$$u_i(\mu_i, s_{-i}) \geq u_i(\sigma_i(\bar{\omega}), s_{-i}) \text{ for all } s_{-i} \in S_{-i} \text{ such that } (\sigma_i(\bar{\omega}), s_{-i}) \in S_s^{\bar{k}}, \text{ and} \qquad (1)$$

$$u_i(\mu_i, S_{-i}(\bar{\omega})) > u_i(\sigma(\bar{\omega})). \qquad (2)$$

Since $\bar{\omega} \in \mathbf{P}_*(\alpha)$ and $\alpha \in \mathbf{B_*T}$, $\bar{\omega} \in P_i(\bar{\omega})$, which implies that $p_{i,\bar{\omega}}(\bar{\omega}) > 0$. By definition of $\mathbf{P}_*$, $P_i(\bar{\omega}) \subseteq \mathbf{P}_*(\bar{\omega})$. By transitivity of $\mathbf{P}_*$, since $\bar{\omega} \in \mathbf{P}_*(\alpha)$, $\mathbf{P}_*(\bar{\omega}) \subseteq \mathbf{P}_*(\alpha)$. By definition of $\bar{k}$,

---

[11] **Stalnaker (1994, p. 63) incorrectly states the result as $\mathbf{B_*T} \cap \mathbf{B_*RAT} \subseteq \mathbf{S_s^\infty}$. Bonanno and Nehring (1996b) give a counterexample and prove the results as stated in Proposition** 5.

$P_*(\alpha) \subseteq \{\omega \in \Omega : \sigma(\omega) \in S_s^{\bar{k}}\}$. Hence $P_i(\varpi) \subseteq \{\omega \in \Omega : \sigma(\omega) \in S_s^{\bar{k}}\}$. It follows from this and (1)and (2) that

$$\sum_{y \in P_i(\varpi)} p_{i,\varpi}(y) u_i(\sigma(y)) < \sum_{y \in P_i(\varpi)} p_{i,\varpi}(y) u_i(\mu_i, \sigma_{-i}(y)),$$

[recall that, for all $y \in P_i(\varpi)$, $\sigma_i(y) = \sigma_i(\varpi)$] that is, player i is not rational at ϖ. Hence $\tau \notin B_*RAT_i$, yielding a contradiction. Part (2) follows directly from (1) and the definition of $T^\bullet$. ■

To see that, in general, $\mathbf{B_*T} \cap \mathbf{B_*RAT} \nsubseteq \mathbf{S_s^\infty}$ consider the model of the game of Figure 3 illustrated in Figure 5. It is easy to check that **RAT** = 52 (indeed, for $x \in \{\beta, \gamma\}$, $\sigma(x)$ is a Nash equilibrium). Hence at τ (indeed at every state) it is common belief that all the players are rational Furthermore there is common belief (at τ, indeed at every state) that no player has false beliefs, that is, **B,T** = 52. However, while $\tau \in \mathbf{B_*T} \cap \mathbf{B_*RAT}$, $\sigma(\tau) = (A,a) \notin S_s^\infty$.

Insert Figure 5

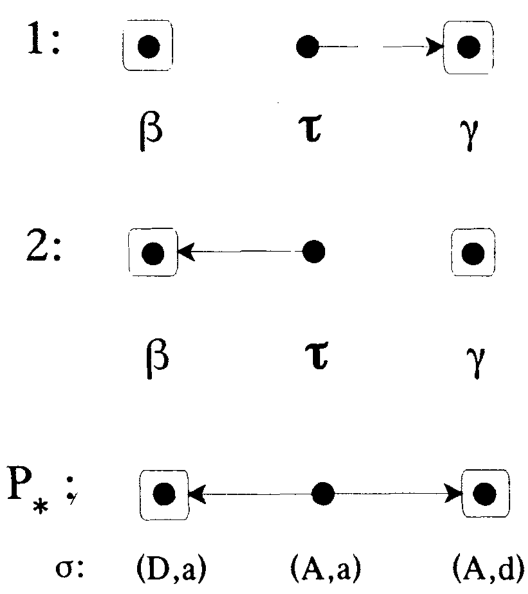


Figure 5

A partial converse to Proposition 5 is given by the following result.

PROPOSITION 6. Let G be a game and $s \in S_s^\infty$. Then there is a model $\mathcal{M}$ of G such tha:: (1) $\tau \in T \cap B_*T \cap B_*RAT$, and (2) $\sigma(\tau) = s$.

The example of Figure 4 shows that strong rationalizability is considerably stronger than rationalizability. To stress this point, consider the extensive game of Figure 6a, whose normal form is shown in Figure 6b.

Insert Figure 6

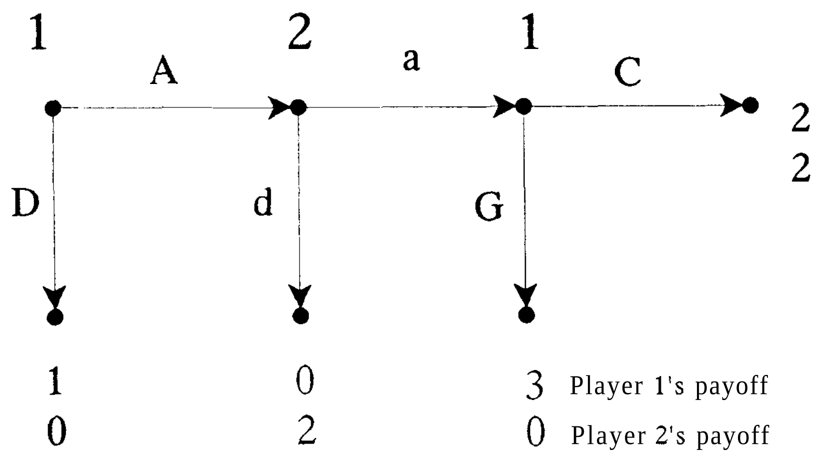


Figure 6a

		Player 2	
		d	a
Player 1	DG	1 , 0	1 , 0
	DC	1 , 0	1 , 0
	AG	0 , 2	3 , 0
	AC	0 , 2	2 , 2

Figure 6b

For the normal form, $S^{\infty} = S$ (that is, ali the strategy profiles are rationalizable), since no strategy of any player is strictly dominated. Hence *every* outcome is compatible with common belief in rationality. On the other hand, $S_s^{\infty} = \{(DG, d), (DG, a), (DC, d), (DC, a)\}$[12] and all the strategy profiles in S_s^{∞} give rise to the Nash equilibrium outcome, namely the payoff vector (1,0).

One might wonder whether the above example can be generalized to the claim that in the normal form of an extensive game with perfect information strong rationalizability implies the play of a Nash equilibrium outcome.[13] The answer is negative, as the following example

[12] In the first round (AG, a) and (AC,a) are eliminated [the first because d weakly dominates a, the second because AG weakly dominates AC]; in the second round (AG, d) and (AC, d) are eliminated (because 1's strategy is dominated by DG).

[13] Stalnaker (1994,p. 64, Theorem 4) incorrectly makes this claim.

shows.[14] Figure 7b shows a model of the normal form of the extensive game of Figure 7a. At the true state the players choose (A, d, G), which is not a Nash equilibrium; furthermore, there is no Nash equilibrium that gives rise to the outcome (2,2,2). Note that $\tau \in T \cap B_* T \cap RAT \cap B_* RAT$ (in particular, Player 1's choice of A is rational, given his belief that Player 2 plays d and a with equal probability).

Insert Figure 7

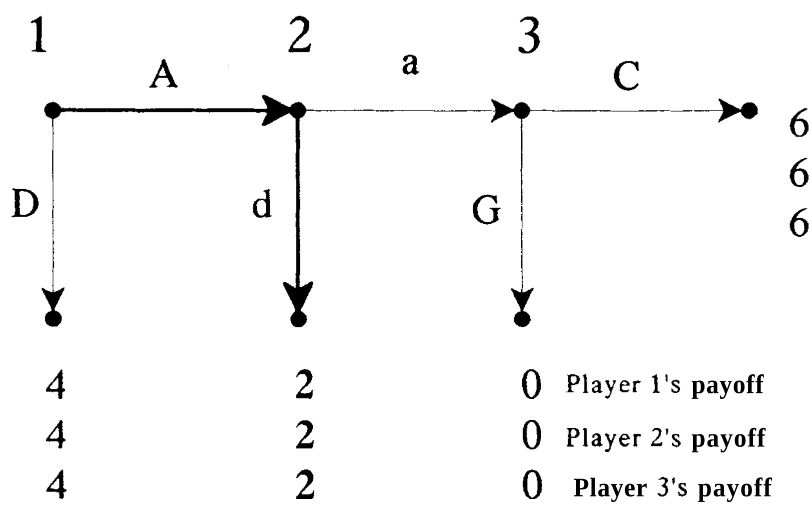


Figure 7a

[14] This example is due to Battigalli (1996, private communication). For a similar example *see* Bonanno and Nehring (1996b).

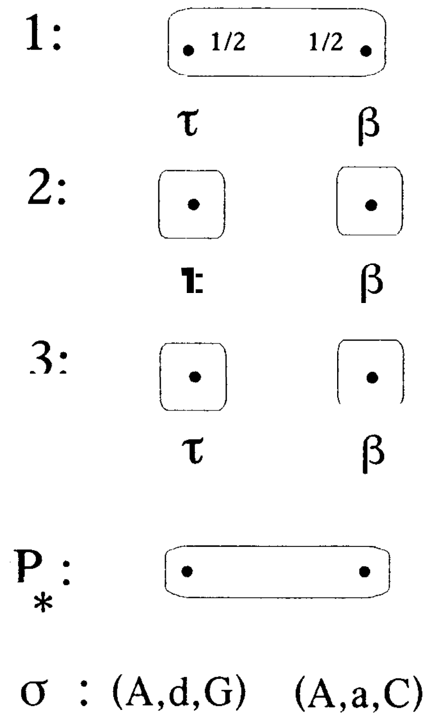


Figure 7b

The extensive game of Figure 7a has several Nash equilibria and more than one Nash equilibrium outcome. Does strong rationalizability imply Nash equilibrium outcome if there is a unique such outcome? Once again, the answer is negative as the following modification of the game of Figure 7a shows.[15] Here there is a unique Nash equilibrium outcome, namely the payoff vector (7, 7, 7, 7). Yet in the model shown in Figure 8b at τ the realized outcome is (2, 2, 2, 10) despite the fact that $\tau \in T \cap B_*T \cap RAT \cap B_*RAT$.

Insert Figure 8

[15] This example is due to Stalnaker (1996, private communication).

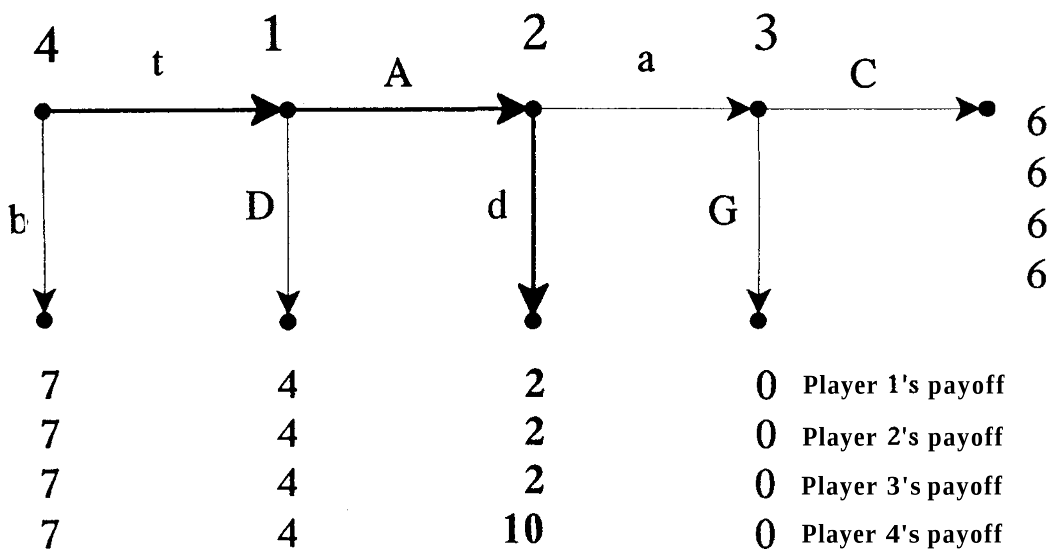


Figure 8a

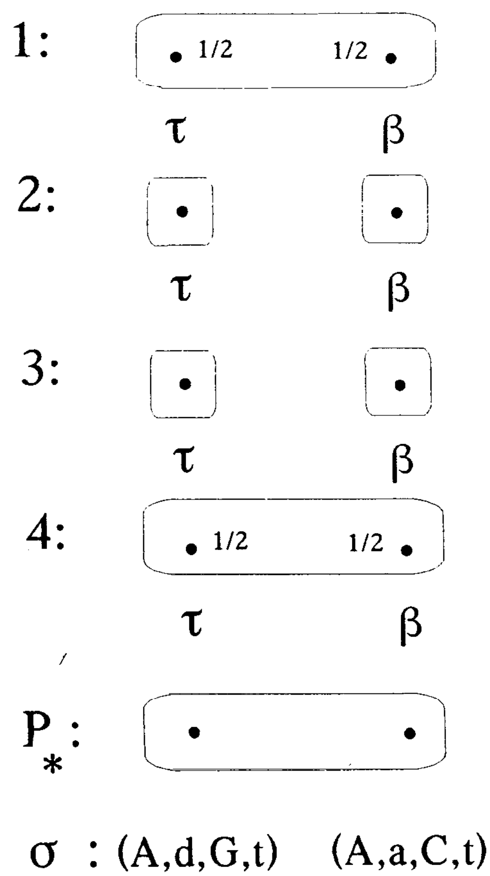


Figure 8b

5. Correlated equilibrium

We now turn to the notion of correlated equilibrium which was introduced by Aumann (1974, 1987).

DEFINITION 7. Let *G* be a normal-form game. ***A correlated equilibrium distribution*** is a probability distribution p over the set S of strategy profiles such that, for every player **i** and every function $d_i : S_i \to S_i$

$$\sum_{s \in S} u_i(s)\, p(s) \geq \sum_{s \in S} u_i(d_i(s_i), s_{-i})\, p(s) \tag{2}$$

EXAMPLE 4. Consider the game of Figure 9 (discussed by Aumann, 1974) and the following distribution: $p(U,L) = p(D,R) = \frac{1}{2}$. Consider Player 1. The left-hand side of (2) is equal to $\frac{1}{2}\,5 + \frac{1}{2}\,1 = 3$. The possible functions $d : \{U, L\} \to \{U, L\}$ are the identity function id (which gives the LHS of (2)), d_U (defined by $d_U(x) = U$ for all x), d_D (defined by $d_D(x) = D$ for all x) and d_o (defined by $d_o(U) = D$, $d_o(D) = U$). With d_U the RHS of (2) is equal to $\frac{1}{2}\,5 + \frac{1}{2}\,0 = 2.5$, with d_D it is equal to $\frac{1}{2}\,4 + \frac{1}{2}\,1 = 2.5$. Thus (2) is satisfied for Player 1. Similar calculations show that (2) is also satisfied for Player 2. Thus $p(U,L) = p(D,R) = \frac{1}{2}$ is a correlated equilibrium distribution.

It is easy to see that every Nash equilibrium is also a correlated equilibrium.[16] Furthermore, every convex combination of Nash equilibria is also a correlated equilibrium. In a two-person zero-sum game all correlated equilibria are convex combinations of pairs of optimal (maxmin and minmax) strategies. Thus if a two-person zero-sum game has a unique pure-strategy Nash equilibrium s then s is the unique correlated equilibrium point. However, in general, there are correlated equilibria that are outside the convex hull of the Nash equilibria.

[16] **For example, if s is a pure-strategy Nash equilibrium, take p such that p(s) = 1.**

Aumann (1987) proved the following result. Let Ω be a set of states; for every player i let $\mathcal{H}_i$ be a partition of Ω and denote by $H_i(\omega)$ the element of the partition that contains state w. Let $p^i \in \Delta(\Omega)$ be individual i's "prior" such that $p^i(H_i) > 0$ for all $H_i \in \mathcal{H}_i$. Let $\sigma_i : \Omega \to S_i$ be a function that specifies i's choice of strategy at every state, satisfying the property that if $\omega' \in H_i(\omega)$ then $\sigma_i(\omega') = \sigma_i(\omega)$, that is, player i knows his own strategy. Let $a = (a,, \ldots, \sigma_n)$. Player i is rational at state a if the strategy he chooses at α maximizes his expected utility calculated on the basis of his *posterior* beliefs $p_i(\cdot \mid H_i(\alpha))$ defined by $p_i(\omega \mid H_i(\alpha)) = \dfrac{p^i(\omega)}{p^i(H_i(\alpha))} = \dfrac{p^i(\omega)}{\sum_{x \in H_i(\alpha)} p^i(x)}$ if $\omega \in H_i(\alpha)$ and $p_i(\omega \mid H_i(\alpha)) = 0$ if $\omega \notin H_i(\alpha)$:

$$\sum_{\omega \in \Omega} u_i(\sigma(\omega))\, p_i(\omega \mid H_i(\alpha)) \geq \sum_{\omega \in \Omega} u_i(x, \sigma_{-i}(\omega))\, p_i(\omega \mid H_i(\alpha)) \qquad \forall x \in S_i$$

PROPOSITION 7 (Aumann, 1987). If the players have a common prior (i.e. if there is a probability measure p on Ω such that $p, = \ldots = p_n = p$) and each player is rational at every state, then the probability distribution induced by p on S is a correlated equilibrium distribution.

It is clear that the structure considered by Aumann is just a special case of the notion of model given in Definition 3. The extra assumptions that Aumann introduces are: (1) that the possibility correspondences give rise to partitions and (2) that the beliefs of the players are Harsanyi consistent, in the sense that they can be derived from a common prior.[17] An interesting question is therefore whether Aumann's theorem can be generalized to the case where the possibility correspondences are non-partitional (i.e. where some players might have false beliefs). In order to do so one first needs to have a local definition of Harsanyi consistency (i.e.

---

[17] Note that the prior beliefs p^i of player i postulated by Aumann play no role: only the posterior beliefs $p_i(\cdot \mid H_i(\omega))$ are relevant. Indeed, given a model of a game according to Definition 3, one can obtain a (local) "prior" for player i by taking any convex combination of the different beliefs (types) of that player, that is, a prior of player i is any point in the convex hull of $\{p_{i,\omega} : w \in P_*(\tau)\}$.

of the existence of a common prior). However, obtaining a local formulation of the notion of a common prior is only part of the difficulty. Recent contributions (Gul, 1996, Dekel and Gul, 1997, Lipman, 1995) have pointed out that the ***meaning*** of a common prior in situations of incomplete information is highly problematic. This skepticism can be developed along the following lines. As Mertens and Zamir (1985) showed in their classic paper, the description of the "actual world" in terms of belief hierarchies generates a collection of "possible worlds", one of which is the actual world. This set of possible worlds, or states, gives rise to a formal similarity between situations of incomplete information and those of asymmetric information (where there is an ex ante stage at which the individuals have identical information and subsequently update their beliefs in response to private signals). However, while a state in the latter represents a real contingency, in the former it is "a fictitious construct, used to clarify our understanding of the real world" (Lipman, 1995, p.2), "a notational device for representing the profile of infinite hierarchies of beliefs" (Gul, 1996, p. 3). As a result, notions such as that of a common prior, "seem to be based on giving the artificially constructed states more meaning than they have" (Dekel and Gul, 1997, p.115). Thus an essential step in providing a justification for correlated equilibrium under incomplete information is to provide an interpretation of the common prior based on "assumptions that do not refer to the constructed state space, but rather are assumed to hold in the true state", that is, assumptions "that only use the artificially constructed states the way they originated – namely as elements in a hierarchy of belief" (Dekel and Gul, 1997, p.116).

An interpretation of the desired kind of the common prior assumption under incomplete information was provided recently (Bonanno and Nehring, 1996a; see also Feinberg, 1995) in terms of a generalized notion of absence of agreeing to disagree a la Aumann (1976), called consistency of expectations.

DEFINITION 8. At state α there is Consistency of Expectations if there do not exist random variables $Y_i : \Omega \to \mathbb{R}$ ($i \in N$) such that: (1) $\forall \omega \in P$, $\sum_{i \in N} Y_i(\omega) = 0$, and (2) at α it is common belief that, for every individual i, i's subjective expectation of Y_i is positive, that is, $\alpha \in B_*(\|E_1 > 0\| \cap \ldots \cap \|E_n > 0\|)$, where $\|E_i > 0\| = \{\omega \in \Omega : \sum_{\omega' \in \Omega} Y_i(\omega') p_{i,\omega}(\omega') > 0\}$.

Consistency of Expectations turns out to be equivalent to a particular local version of the Common Prior Assumption defined as follows.

DEFINITION 9. For every $\mu \in \Delta(\Omega)$, let $\mathbf{HQC}_\mu$ (for Harsanyi Quasi Consistency with respect to the "prior" y) be the following event: $\alpha \in \mathbf{HQC}_\mu$ if and only if

(I) $\forall i \in N$, $\forall \omega, \omega' \in P_*(\alpha)$, if $\mu(\|p_i = p_{i,\omega}\|) > 0$ then $p_{i,\omega}(\omega') = \dfrac{\mu(\omega')}{\mu(\|p_i = p_{i,\omega}\|)}$ if $\omega' \in \|p_i = p_{i,\omega}\|$ and $p_{i,\omega}(\omega') = 0$ otherwise (that is, $p_{i,\omega}$ is obtained from μ by conditioning on $\|p_i = p_{i,\omega}\|$)[18], and

(2) $\mu(P_*(\alpha)) > 0$.

If $\alpha \in \mathbf{HQC}_\mu$, μ is a *local common prior at* α. Furthermore, let $HQC = \bigcup_{\mu \in \Delta(\Omega)} \mathbf{HQC}_\mu$.

PROPOSITION 8.[19] At α Consistency of Expectations is satisfied if and only if $\alpha \in$ HQC.

The above proposition shows that HQC is the natural way of expressing Harsanyi consistency locally.

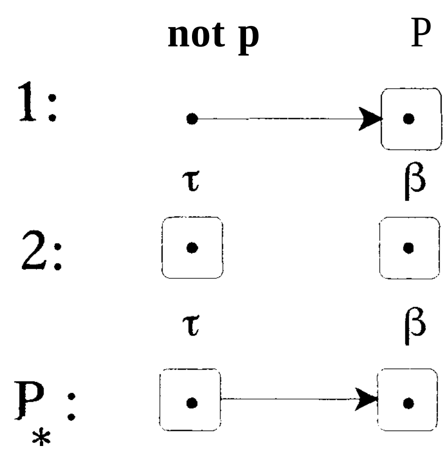
---

[18] **Where, for every event E, $\mu(E) = \sum_{\omega \in E} \mu(\omega)$. Note that, for every $\omega \in \Omega$ and $i \in N$, $w \in \|p_i = p_{i,\omega}\|$. Thus $\mu(\omega) > 0$ implies $\mu(\|p_i = p_{i,\omega}\|) > 0$.**

[19] **For a proof see Bonanno and Nehring (1996a). This result is a local version of Morris's (1994) characterization of no trade under asymmetric information. See also Feinberg (1995) and Samet (1996).**

Harsanyi Quasi Consistency may seem weaker than expected in that condition (2) of its definition only requires the derived common prior to assign positive probability to some commonly possible state but allows the true state to be assigned zero "prior" probability. As illustrated in the example of Figure 9, Agreement and No Trade-type arguments cannot deliver more.

Insert Figure 9



$$p_{1,\tau} = p_{1,\beta} = \begin{pmatrix} \tau & \beta \\ 0 & 1 \end{pmatrix},$$

$$p_{2,\tau} = \begin{pmatrix} \tau & \beta \\ 1 & 0 \end{pmatrix}, \; p_{2,\beta} = \begin{pmatrix} \tau & \beta \\ 0 & 1 \end{pmatrix}$$

$$P_*(\tau) = \{\tau, \beta\}, P_*(\beta) = \{\beta\}$$

Figure 9

In this example, at the true state individual 1 wrongly believes that it is common belief that p, while individual 2 correctly believes that not p is the case and knows 1's incorrect beliefs. Expectation consistency is satisfied at the true state (as well as at (3). In fact, let Y_1 and Y_2 be random variables on $\{r, \beta\}$ such that $Y_2 = -Y_1$ and suppose that $\tau \in B_*\|E_1 > 0\|$, that is, at τ it is common belief that individual 1's expectation of Y_1 is positive. Then $Y_1(\beta) > 0$, hence $Y_2(\beta) < 0$. Thus $\beta \notin \|E_2 > 0\|$, that is, at β individual 2's expectation of Y_2 cannot be positive. Since $\beta \in P_*(\tau)$, it follows that $\tau \notin B_*\|E_2 > 0\|$. Thus Agreement is necessarily satisfied at τ. By Proposition 7 there must be a μ such that $\tau \in HQC_\mu$. Indeed such a local common prior is given by $\mu(\beta) = 1$.

Is Harsanyi Quasi Consistency an adequate epistemic basis for correlated equilibrium? Perhaps not too surprisingly in view of the previous example, Harsanyi Quasi Consistency is insufficient by *itself,* as demonstrated by the following example. Figures 10a and 10b show a two-person zero-sum game with a unique correlated equilibrium (B,R), and an epistemic model of it.

Insert Figure 10

		Player 2 L	C	R
Player 1	T	10, −10	−10, 10	−7 , 7
	M	−10, 10	10, −10	−7 , 7
	B	7 , −7	7 , −7	0 , 0

Figure 10a

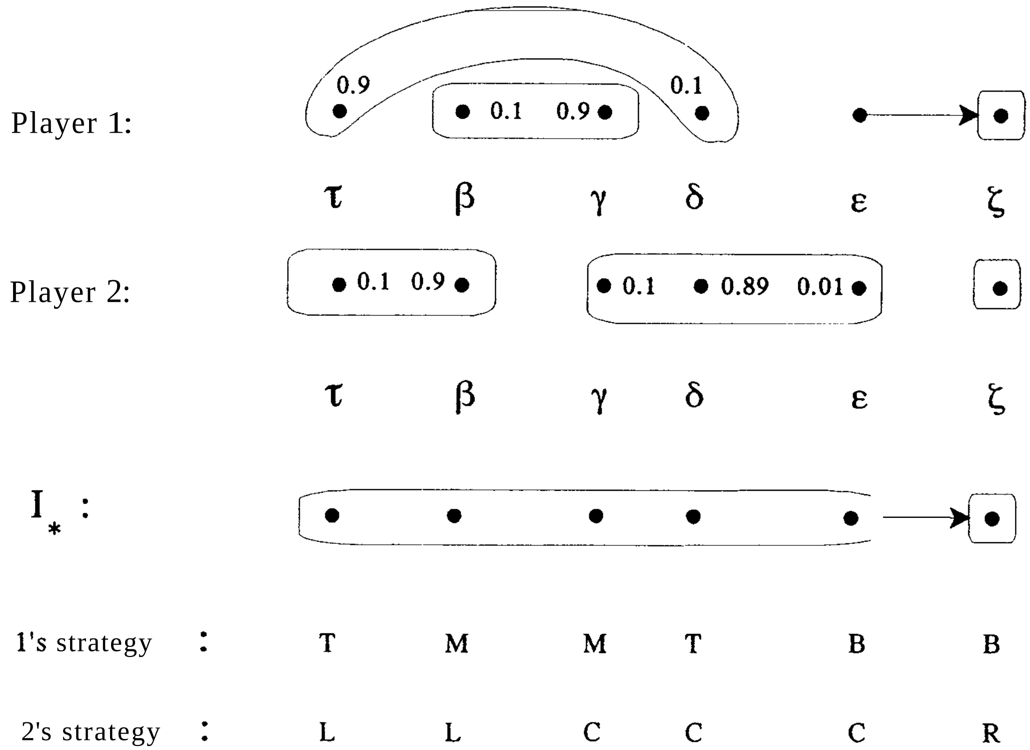


Figure 10b

In this example, at τ (i) the players' beliefs satisfy Harsanyi Quasi Consistency ($\tau \in \mathrm{HQC}_{\mu} = \Omega$ where $\mu(\zeta) = 1$), (ii) there is common belief in rationality ($P_{*}(\tau) = \Omega$ and at every state each player's strategy is optimal given her beliefs) and (iii) no individual has any false beliefs. Yet at τ the players play (T,L) which is not a correlated equilibrium.

Note that in the above example, although the derived common prior assigns zero probability to τ, there is no sense in which the belief hierarchies described by the true state are "improbable" and constitute a null event. Indeed the actual beliefs of all players assign positive probability to τ.

The above example is in fact quite general. By a straightforward generalization of its construction any profile of correlated rationalizable strategies – where one strategy is a unique best response to some distribution over correlated rationalizable strategies of the other players –

can be realized at the true state τ of a Bayesian frame where $\tau \in$ **HQC** (and no individual has false beliefs).

What seems to go wrong in the example is that, while Player 2 believes Player 1 to be wrong at ε, this does not show up as disagreement – and hence as a violation of Harsanyi Quasi Consistency – since Player 1 falseiy believes at ε that there is agreement that the true state is ζ. Hence $\mathbf{T}_{CB}$ is violated at ε, and therefore $\mathbf{B}_*\mathbf{T}_{CB}$ at $\mathbf{t}$.

Indeed – in the absence of false beliefs at the true state – $\mathbf{B}_*\mathbf{T}_{CB}$ is exactly what needs to be added to HQC to ensure the play of a correlated equilibrium strategy-profile, as the following theorem shows.

To take account of the incomplete information context, we call a strategy profile a ***correlated equilibrium*** if it is played with positive probability in some correlated equilibrium (in the ordinary sense).

PROPOSITION 9 (Bonanno and Nehring, 1997). Fix an arbitrary finite normal-form game G and an arbitrary model of G such that:

(1)	$\tau \in \mathbf{T} \cap B_*\mathbf{T}_{CB}$	(the actual beliefs of the players are correct and there is common belief in Truth about common belief),
(2)	$\tau \in B_*\mathbf{RAT}$,	(there is common belief in rationality)
(3)	$\tau \in$ HQC	(Harsanyi Quasi Consistency of beliefs, that is, Agreement, is satisfied).

Then the strategy profile associated with τ (i.e. the strategy profile actually played) is a correlated equilibrium.

On the other hand, as the example of Figure 10 shows, if (2) and (3) are satisfied and (1) is weakened to $\tau \in \mathrm{T}$ then the strategy profile associated with τ need not be a correlated equilibrium.

REMARK 4. If condition (1) is weakened to $\tau \in \mathrm{NI}$ (or, equivalently – cf. Proposition 1 – $\tau \in \mathbf{T}_{CB} \cap \mathbf{B}_*\mathbf{T}_{CB}$) then the conclusion is that $\tau \in \mathbf{B}_*\mathbf{CE}$, where CE is the event that a correlated equilibrium is played; that is, at the true state it is common belief that a correlated equilibrium is played.

Thus one sees that once the rather mild-looking property of Negative Introspection of common belief is satisfied, HQC is re-instated with the proper strength.

A converse to Proposition 9 is given by the following result.

PROPOSITION 10. Let G be a game and $p \in \Delta(S)$ a correlated equilibrium distribution. Then there exists a $\mu \in \Delta(\Omega)$ and model $\mathcal{M}$ of G such that (1) $\tau \in \mathrm{T} \cap \mathbf{B}_*\mathbf{T}_{CB} \cap \mathrm{HQC}_\mu \cap \mathrm{B}_*\mathbf{RAT}$, (2) the distribution over strategy profiles induced by μ restricted to $\{\tau\} \cup \mathrm{P}_*(\tau)$ coincides with p and (3) $\mu(\tau) > 0$ (so that the strategy profile actually played is in the support of p).

6. Nash equilibrium

We conclude by examining the epistemic foundations of Nash equilibrium, which (together with its refinements) is without doubt the solution concept most used in applications. The above examples (e.g. Figure 7) show that common belief in rationality, even in the presence of Truth and common belief in Truth, is not sufficient to guarantee the play of a Nash equilibrium. There are further difficulties, however, due to the fact that some Nash equilibria involve mixed strategies. The models we have considered are models of particular ways a game is played, and a particular ***pure*** strategy profile will always be realized at the true state of the model and, indeed, at every state. The notion of model (cf. Definition 3) incorporates the assumption that each player knows the strategy he actually plays. One could easily weaken this assumption by allowing players to delegate their choice of strategy to a random device. However, as Aumann (1987, p.15) observes,

> "In the traditional view of strategy randomization, the players use a randomizing device, such as a coin flip, to decide on their actions. This view has always had difficulties. Practically speaking, the idea that serious people would base important decisions on the flip of a coin is difficult to accept. Conceptually, too, there are problems. The reason a player must randomize in equilibrium is only to keep others from deviating; for himself, randomizing is unnecessary."

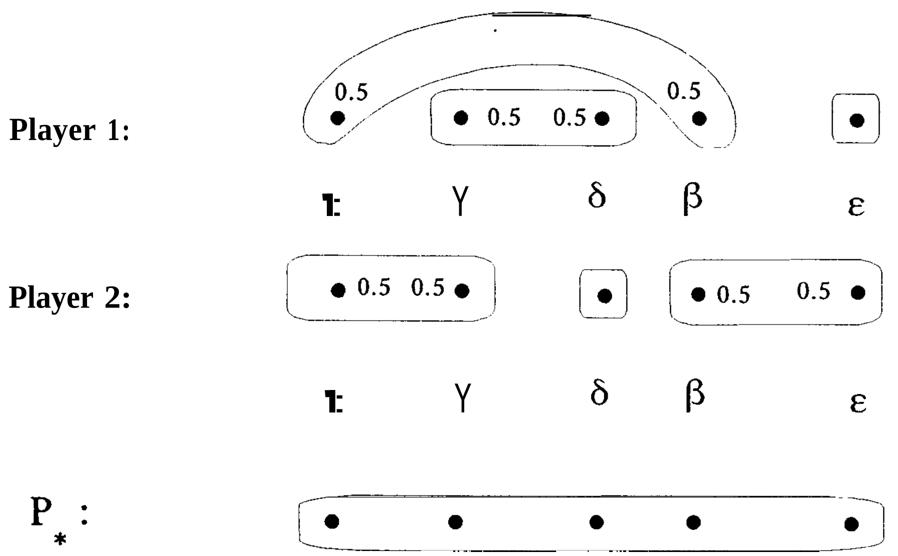
Elaborating on an idea of Harsanyi (1973), Aumann's suggestion was to view a mixed strategy of player **i** not as an actual choice by player **i** but as an expression of the uncertainty in the *other* players' mind concerning the choice made by i.

DEFINITION 10. Given a model of a game, we can extract a *conjecture* of player i, defined as a function $\chi_i : \Omega \to \Delta(S_{-i})$ that associates with every state α the probability distribution over S_{-i} induced by player i's beliefs at α. For example, consider the zero-sum matching penny game of Figure 11a and the model of Figure 11b (taken from Stalnaker, 1994, p. 59). The functions χ_1 and χ_2 are shown in Figure 11b. At every state except ε, Player 1 believes that Player 2 is choosing h an t with equal probability. At every state except δ, Player 2 believes that Player 1 is choosing H an T with equal probability. Note that at the true state τ the *conjectures* of the players form a mixed strategy Nash equilibrium. Note that the mixed strategy of Player 2 represents in fact the belief of Player 1 and vice versa. Note also that at τ common belief in rationality fails; in fact, $\mathbf{RAT_1} = \Omega$ and $\mathbf{RAT_2} = \{\tau, \gamma, \beta, \varepsilon\}$ so that $\mathbf{RAT} = \{\tau, \gamma, \beta, \varepsilon\}$ and $B_* \mathbf{RAT} = \emptyset$.

Insert Figure 11

		Player 2	
		h	t
Player 1	H	1 , −1	−1 , 1
	T	−1 , 1	1 , −1

Figure 11a



σ :	(H,h)	(T,h)	(T,t)	(H,t)	(T,t)
χ_1 :	$(\frac{1}{2}h, \frac{1}{2}t)$	$(\frac{1}{2}h, \frac{1}{2}t)$	$(\frac{1}{2}h, \frac{1}{2}t)$	$(\frac{1}{2}h, \frac{1}{2}t)$	t
χ_2 :	$(\frac{1}{2}H, \frac{1}{2}T)$	$(\frac{1}{2}H, \frac{1}{2}T)$	T	$(\frac{1}{2}H, \frac{1}{2}T)$	$(\frac{1}{2}H, \frac{1}{2}T)$

Figure 11b

The above example generalizes. Given a probability distribution $p \in \Delta(S_2)$ denote by $\|\chi_1 = p\|$ the event that Player 1 has conjecture p: $\|\chi_1 = p\| = \{\omega \in \Omega : \chi_1(\omega) = p\}$. Similarly, for $q \in \Delta(S_1)$ let $\|\chi_2 = q\| = \{\omega \in \Omega : \chi_2(\omega) = q\}$.

PROPOSITION 11 (Aumann and Brandenburger, 1995). Let *G* be a two-person normal-form game and $\mathcal{M}$ a model of it. Let $p \in \Delta(S_2)$ and $q \in \Delta(S_1)$. Then for every $\alpha \in T \cap B_1 RAT \cap B_2 RAT \cap B_1\|\chi_2 = q\| \cap B_2\|\chi_1 = q\|$, the pair $(\chi_1(\alpha), \chi_2(\alpha))$ is a Nash equilibrium of G.

When the number of players is greater than 2, complications arise due to the fact that the conjecture of player **i** is not a mixed strategy of another player, but a probability distribution on (n−1)-tuples of strategies of all the other players. However, **i's** conjecture does induce a mixed strategy for each player $j \neq i$ (the marginal on S_j of **i's** overall conjecture). However, different players other than **j** may have different conjectures about j. Since j's component of the putative equilibrium is meant to represent the conjectures of the other players (other than j), and these may differ across j's opponents, it is not clear how j's component should be defined. Aumann and Brandenburger (1995) however show that if the players have a common prior, their rationality is is mutually known and their conjectures are commonly known then for each player j, all the other players i agree on the same conjecture χ_j about j; and the resulting profile $(\chi_1, \ldots, \chi_n)$ is a Nash equilibrium. The authors also show, through a series of examples, that the conditions stated are "tight", in the sense that if any one of them is not met then the claim is no longer true.

7. Conclusion

The aim of this paper has been to introduce the approach and some of the main results of the recent literature on the epistemic foundations of game theory. Not all the contributions were reviewed. In particular, we left out those papers that deal with extensive-form solutions concepts. Recent papers have examined the foundations of backward induction in perfect information games (Aumann, 1995, 1996, Ben Porath, 1997, Stalnaker, 1996, 1997, Stuart 1997) and of extensive form rationalizability (Battigalli 1997, Battigalli and Siniscalchi, 1997). Several issues arise in this context, namely whether or not ex ante rationality is sufficient, whether an explicit analysis of counterfactuals is required, etc. A careful review of this literature would require a paper as long as this one.

References

Aumann, R. **(1974),** Subjectivity and correlation in randomized strategies, ***Journal*** of ***Mathematical*** Economics, **1, 67-96.**

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

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

Aumann, R. **(1989),** Notes on interactive epistemology, mimeo, Hebrew University of Jerusalem.

Aumann, R. **(1995),** Backward induction and common knowledge of rationality, Games ***and*** Economic Behavior, **88, 6-19.**

Aumann, R. **(1996a),** Deriving backward induction in the centipede game without assuming rationality at **unreached** vertices, mimeo, The Hebrew University, Jerusalem.

Aumann, R. **(1996b),** Reply to **Binmore,** Games ***and*** Economic Behavior, 17, 138-146.

Aumann, R. and A. Brandenburger **(1995),** Epistemic conditions for Nash equilibrium, Econometrica, 63, 1161-1180.

Battigalli, P. **(1997),** Hierarchies of conditional beliefs and interactive epistemology in dynamic games, mimeo, Princeton University.

Battigalli, P. and M. Siniscalchi **(1997),** An epistemic characterization of extensive form rationalizability, mimeo, Princeton University.

Ben-Porath, E. **(1997),** Rationality, Nash equilibrium and backward induction in perfect information games, Review of Economic Studies, 64, 23-46.

Bernheim, D. **(1984),** Rationalizable strategic behavior, ***Econometrica*,** 52, 1007-28.

Bonanno, G. **(1996),** On the logic of common belief, Mathematical Logic Quarterly, 42, 305-311.

Bonanno, G. and K. Nehring **(1996a),** How to make sense of the Common Prior Assumption under incomplete information, Working paper, University of California Davis.

Bonanno, G. and K. Nehring **(1996b),** On Stalnaker's notion of strong rationalizability and Nash equilibrium in perfect information games, mimeo, University of California Davis.

Bonanno, G. and K. Nehring **(1997a),** On the logic and role of negative introspection of common belief, Working Paper, University of California Davis.

Bonanno, G. and K. Nehring **(1997b),** Assessing the Truth Axiom under incomplete information, Working paper, University of California Davis.

Brandenburger, A. and E. Dekel **(1987),** Rationalizability and correlated equilibria, ***Econometrica*,** 55,1391-1402.

Brandenburger, A. and E. Dekel **(1993), Hierarchies** of beliefs and common knowledge, ***Journal*** of Economic Theory, 59,189-198.

Chellas, B. **(1984),** Modal logic, Cambridge University Press, Cambridge.

Dekel, E. and F. Gul (1997), Rationality and knowledge in game theory, in Kreps D. M. and K. F. Wallis (eds.), Advances in Economics and *Econometrics*, vol. 1, Cambridge University Press.

Fagin, R., J. Halpern, Y. Moses and M. Vardi (1995), Reasoning about knowledge, MIT Press, Cambridge.

Feinberg, Y. (1995), A converse to the Agreement Theorem, Discussion Paper # 83, Center for Rationality and Interactive Decision Theory, Jerusalem.

Geanakoplos, J. (1989), Game theory without partitions and applications to speculation and consensus, Cowles Foundation Discussion Paper 914, Yale University.

Geanakoplos, J. (1992), Common knowledge, *Journal* of Economic Perspectives, 6, 53-82.

Geanakoplos, J. (1994), Common knowledge, in R. J. Aumann and S. Hart (Eds), Handbook of Game Theory, Vol. 2, Elsevier.

Gul, F. (1996), A comment on Aumann's Bayesian view, mimeo, Northwestern University [forthcoming in *Econometrica*].

Halpern, J. and Y. Moses (1992), A guide to completeness and complexity for modal logics of knowledge and belief, *Artificial* intelligence, 54, 319-379.

Harsanyi, J. (1967-68), Games with incomplete information played by "Bayesian players", Parts I-III, Management Science, 8, 159-182, 320-334,486-502.

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

Lipman, B. (1995), Approximately common priors, mimeo, University of Western Ontario.

Lismont, L. and P. Mongin (1994), On the logic of common belief and common knowledge, Theory and Decision, 37, 75-106.

Mertens, J-F. and S. Zamir (1985), Formulation of Bayesian analysis for games with incomplete information, *International* Journal of Game Theory, 14, 1-29.

Morris, S. (1994), Trade with heterogeneous prior beliefs and asymmetric information, *Econometrica*, 62,1327-1347.

Pearce, D. (1984), Rationalizable strategic behavior and the problem of perfection, , *Econometrica*, 52, 1029-50.

Samet, D. (1996), Common priors and separation of convex sets, mimeo, Tel Aviv University.

Stalnaker, R. (1994), On the evaluation of solution concepts, Theory and Decision, 37, 49-74.

Stalnaker, R. (1996), Knowledge, belief and counterfactual reasoning in games, Economics and Philosophy, 12, 133-163.

Stalnaker, R. (1997), Belief revision in games: forward and backward induction, forthcoming in Mathematical Social Sciences.

Stuart, H. (1997), Common belief of rationality in the finitely repeated Prisoners' Dilemma, Games and Economic Behavior, 19, 133-143.

Tan, T. and S. Werlang (1988), The Bayesian foundation of solution concepts of games, *Journal* of Economic Theory, 45,370-391.