Research in Economics (1999) **53**, 149–225
Article No. reec.1999.0187
Available online at http://www.idealibrary.com on IDEAL®

Recent results on belief, knowledge and the epistemic foundations of game theory

PIERPAOLO BATTIGALLI† and GIACOMO BONANNO‡

†*European University Institute, Fiesole (Florence) 50016, Italy*
‡*University of California, Davis, CA 95616-8578, U.S.A.*

Summary

We provide a self-contained, selective overview of the literature on the role of knowledge and beliefs in game theory. We focus on recent results on the epistemic foundations of solution concepts, including correlated equilibrium, rationalizability in dynamic games, forward and backward induction.© 1999 Academic Press

J.E.L. Classification: C72, D82.
Keywords: Games, belief, knowledge, interactive epistemology, solution concepts.

1. Introduction§

The objective of this paper is to provide a selective overview of the relatively recent literature on the role of beliefs in game theory, with particular focus on the foundations of solution concepts. In order to make the paper self-contained and accessible to the general reader, we begin in Section 2 with an overview of the "state space" representation of beliefs and knowledge, which has its roots in modal logic (cf. Chellas, 1984) and is a generalization of the "information partition" approach commonly used in information economics and game theory. Sections 3 and 4 are devoted to the study of what assumptions on the beliefs and reasoning of the players are implicit in various solution concepts. Section 3 focuses on strategic-form games, while Section 4 is devoted to extensive-form games. A complementary and in-depth discussion of many of the issues covered in this paper can be found in Dekel and Gul (1997).

† E-mail: battigal@iue.it
‡ E-mail: gfbonanno@ucdavis.edu
§ Prepared for the workshop on *Interactive Epistemology in Dynamic Games and Games of Incomplete Information*, Venice, June 21–22, 1998.

1090–9443/99/020149 + 77 $30.00/0

2. The semantic representation of knowledge and belief

2.1. REPRESENTING THE BELIEFS OF A SINGLE INDIVIDUAL

To represent the beliefs of an individual we start with a set of *states*, or possible worlds, Ω. Each state should be thought of as a complete description of the world. The subsets of Ω are called *events* and the set of all events is denoted by 2^{Ω}. A possibility correspondence $P : \Omega \to 2^{\Omega}$ associates with every state $\omega \in \Omega$ the set of states $P(\omega)$ that the individual considers possible at ω. The pair $\mathcal{F} = \langle \Omega, P \rangle$ is called a *belief frame*.†

From the possibility correspondence P a belief operator $B : 2^{\Omega} \to 2^{\Omega}$ is obtained as follows: $\forall E \subseteq \Omega, BE = \{\omega \in \Omega : P(\omega) \subseteq E\}$. BE can be interpreted as the event that (i.e., the set of states at which) individual i believes that event E has occurred.

REMARK 1: it is easily verified that the belief operator B satisfies the following properties:

Necessity: $B\Omega = \Omega$

Conjunction: $B\left(\bigcap_{j \in J} E_j\right) = \bigcap_{j \in J} BE_j$ where J is any index set

Monotonicity: if $E \subseteq F$ then $BE \subseteq BF$.

An operator $B : 2^{\Omega} \to 2^{\Omega}$ that satisfies Necessity, Conjunction and Monotonicity is called *normal*. Thus the operator that is obtained from a possibility correspondence is always normal. Instead of taking a possibility correspondence as primitive, one could start with a normal belief operator $B : 2^{\Omega} \to 2^{\Omega}$ and obtain from it a possibility correspondence as follows:

$$\forall \alpha \in \Omega, P(\alpha) = \{\omega \in \Omega : \alpha \in \neg B \neg \{\omega\}\}$$

(for every event $E \subseteq \Omega$, $\neg E$ denotes its complement in Ω). The two approaches are equivalent, in the sense the two mappings are one the inverse of the other.‡

† These structures are known in the modal logic and philosophy literature as *Kripke frames*. In this literature instead of a possibility correspondence $P : \Omega \to 2^{\Omega}$ it is more common to postulate an *accessibility relation* R on Ω. For $\alpha, \beta \in \Omega$, $\alpha R \beta$ reads "state β is accessible from state α". The two notions are equivalent. Given an accessibility relation R, the corresponding possibility correspondence is defined by: $\forall \alpha \in \Omega, P(\alpha) = \{\omega \in \Omega : \alpha R \omega\}$. Conversely, given a possibility correspondence P, the associated accessibility relation R is obtained as follows: $\forall \alpha, \beta \in \Omega$, $\alpha R \beta$ if and only if $\beta \in P(\alpha)$.

‡ Let $P : \Omega \to 2^{\Omega}$ be a possibility correspondence, $B : 2^{\Omega} \to 2^{\Omega}$ the associated belief operator ($\forall E \subseteq \Omega, BE = \left\{\omega \in \Omega : P(\omega) \subseteq E\right\}$) and $P' : \Omega \to 2^{\Omega}$ the possibility correspondence obtained from $B(\forall \alpha \in \Omega,\ P'(\alpha) = \{\omega \in \Omega : \alpha \in \neg B \neg \{\omega\}\})$. Then $P' = P$. Conversely, let B be a normal belief operator, P the possibility correspondence obtained from B and B' the belief operator obtained from P. Then $B = B'$.

Beliefs pertain to propositions. Events (that is, subsets of Ω) should be thought of as representing propositions. In order to establish the interpretation of events as propositions we need to introduce the notion of a model based on a frame.

We begin with a language with a modal operator $\Box$. The intended interpretation of $\Box\phi$ is "the individual believes that ϕ". The alphabet of the language consists of: (1) a finite or countable set Π of sentence letters (representing atomic propositions, such as "the earth is flat"), (2) the connectives $\neg$ (for "not"), $\vee$ (for "or"), and $\Box$, (3) the bracket symbols (and). The set Φ of formulae is obtained from the sentence letters by closing with respect to negation, disjunction and the operator $\Box$.† As is customary, we shall often omit the outermost brackets [e.g., we shall write $\phi \vee \psi$ instead of $(\phi \vee \psi)$] and use the following (metalinguistic) abbreviations: $\phi \wedge \psi$ for $\neg(\neg\phi \vee \neg\psi)$ (the symbol $\wedge$ stands for "and"), $\phi \rightarrow \psi$ for $(\neg\phi) \vee \psi$ (the symbol $\rightarrow$ stands for "if... then...") and $\phi \leftrightarrow \psi$ for $(\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$ (the symbol $\leftrightarrow$ stands for "if and only if").

Given a frame $\mathcal{F}$ one obtains a model $\mathcal{M}$ based on it by adding a function $f : \Pi \rightarrow 2^{\Omega}$ that associates with every sentence letter π the set of states at which π is true. For every formula $\phi \in \Phi$, the truth set of ϕ in $\mathcal{M}$, denoted by $\|\phi\|^{\mathcal{M}}$, is defined recursively as follows:

(1) If $\phi = (\pi)$ where π is a sentence letter, then $\|\phi\|^{\mathcal{M}} = f(\pi)$,

(2) $\|\neg\phi\|^{\mathcal{M}} = \neg\|\phi\|^{\mathcal{M}}$ (with slight abuse of notation, the symbol '$\neg$' is also used to denote complement: $\neg E = \Omega\backslash E$)

(3) $\|\phi \vee \psi\|^{\mathcal{M}} = \|\phi\|^{\mathcal{M}} \cup \|\psi\|^{\mathcal{M}}$,

(4) $\|\Box\phi\|^{\mathcal{M}} = \left\{\omega \in \Omega : P(\omega) \subseteq \|\phi\|^{\mathcal{M}}\right\}$.

If $\omega \in \|\phi\|^{\mathcal{M}}$ we say that ϕ is *true at state* ω in model $\mathcal{M}$. Thus according to (4), at state α the individual believes ϕ if and only if ϕ is true at every state that the individual considers possible at α, that is, if ϕ is true at every $\omega \in P(\alpha)$. If E is the truth set of some formula ϕ (that is, $E = \|\phi\|^{\mathcal{M}}$) and $B : 2^{\Omega} \rightarrow 2^{\Omega}$ is the belief operator, then BE is the truth set of the formula $\Box\phi$, that is, $BE = \|\Box\phi\|^{\mathcal{M}}$. Hence the interpretation of BE as the event that the individual believes E (or, more precisely, the proposition represented by event E). A formula ϕ is *valid in model* $\mathcal{M}$ if and only if it is true at every state, that is, if and only if $\|\phi\|^{\mathcal{M}} = \Omega$.

Properties of the possibility correspondence correspond to properties of beliefs, as explained in the following remark.

REMARK 2: fix a belief frame $\mathcal{F}$. Then (cf. Chellas, 1984: p. 164):
(1) *Non-empty valuedness* (or *seriality*) of the possibility correspondence P corresponds to consistency of beliefs, that is, the following are equivalent:

† Thus Φ is obtained recursively as follows: (i) for every sentence letter $\pi \in \Pi$, $(\pi) \in \Phi$, (ii) if $\phi, \psi \in \Phi$ then $(\neg\phi) \in \Phi$, $(\phi \vee \psi) \in \Phi$ and $(\Box\phi) \in \Phi$.

(i) $\forall \omega \in \Omega, P(\omega) \neq \emptyset$,
(ii) $\forall E \subseteq \Omega, BE \subseteq \neg B \neg E$,
(iii) for every model $\mathcal{M}$ based on $\mathcal{F}$ and for every formula ϕ, the formula $\Box\phi \to \neg\Box\neg\phi$ is valid in $\mathcal{M}$, that is, $\|\Box\phi \to \neg\Box\neg\phi\|^{\mathcal{M}} = \Omega$ (if the individual believes ϕ then she does not believe its negation).

(2) *Transitivity* of the possibility correspondence P corresponds to *positive introspection* of beliefs, that is, the following are equivalent:
(i) $\forall \alpha, \beta \in \Omega$, if $\beta \in P(\alpha)$ then $P(\beta) \subseteq P(\alpha)$,
(ii) $\forall E \subseteq \Omega, BE \subseteq BBE$,
(iii) for every model $\mathcal{M}$ based on $\mathcal{F}$ and for every formula ϕ, the formula $\Box\phi \to \Box\Box\phi$ is valid in $\mathcal{M}$ (if the individual believes ϕ then she believes that she believes ϕ).

(3) *Euclideanness* of the possibility correspondence P corresponds to negative introspection of beliefs, that is, the following are equivalent:
(i) $\forall \alpha, \beta \in \Omega$, if $\beta \in P(\alpha)$ then $P(\alpha) \subseteq P(\beta)$,
(ii) $\forall E \subseteq \Omega, \neg BE \subseteq B\neg BE$,
(iii) for every model $\mathcal{M}$ based on $\mathcal{F}$ and for every formula ϕ, the formula $\neg\Box\phi \to \Box\neg\Box\phi$ is valid in $\mathcal{M}$ (if the individual does not believe ϕ, then she believes that she does not believe ϕ).

The above three properties are usually taken as an expression of the notion of *rational belief*. A frame $\mathcal{F} = \langle \Omega, P \rangle$, where the possibility correspondence P satisfies seriality, transitivity and euclideanness is called a *KD45 frame*. From now on *we shall restrict attention to KD45 frames*.

REMARK 3: (graphical representation). We will make use of the following graphical representation of frames (and models). States are represented by points and for every two states α and β, $\beta \in P(\alpha)$ if and only if either (i) α and β are enclosed in the same cell (denoted by a rounded rectangle), or (ii) there is an arrow from α to the cell containing β, or (iii) there is an arrow from the cell containing α to the cell containing β.

For example, consider the following very simple frame: $\Omega = \{\alpha, \beta\}$, $P(\alpha) = P(\beta) = \{\beta\}$. Let $\mathcal{M}$ be the following model based on this frame: there is a single sentence letter π, representing the proposition "the earth is flat", which is true at β and false at α. This model is shown in Figure A according to the convention established in remark 3.

State α in this model represents a situation where, as a matter of fact, the earth is not flat but the individual incorrectly believes the earth to be flat. The possibility of incorrect beliefs is taken to be the distinguishing feature between knowledge and beliefs: only true facts can be known.

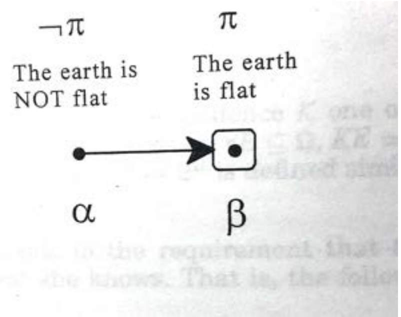


FIGURE A.

REMARK 4: fix a frame $\mathcal{F}$ and a state $\alpha \in \Omega$. Then (cf. Chellas, 1984: p. 164) *reflexivity* of the possibility correspondence P at α corresponds to correctness of beliefs at α, that is, the following are equivalent:†

(i) $\alpha \in P(\alpha)$,

(ii) $\forall E \subseteq \Omega$, if $\alpha \in BE$ then $\alpha \in E$ (equivalently, $\alpha \in \neg BE \cup E$),

(iii) for every model $\mathcal{M}$ based on $\mathcal{F}$ and for every formula ϕ, the formula $\Box\phi \to \phi$ is true at α, that is, $\alpha \in \|\Box\phi \to \phi\|^{\mathcal{M}}$ (if, at α, the individual believes ϕ then, at α, ϕ is indeed true).

If the possibility correspondence is reflexive at *every* state, then $\forall E \subseteq \Omega$, $BE \subseteq E$. This is called the *Truth Axiom*. When the Truth Axiom is imposed, one normally speaks of *knowledge* rather than belief and the corresponding frame is called an *S5 frame* and is characterized by the fact that the possibility correspondence gives rise to a partition of Ω.

REMARK 5: in a KD45 frame it is possible for an individual to believe something which is false (cf. Figure A). However, the individual always believes to have correct beliefs. This is a consequence of secondary reflexivity: a possibility correspondence is *secondary reflexive* if, $\forall \alpha, \beta \in \Omega$, if $\beta \in P(\alpha)$ then $\beta \in P(\beta)$. It can be shown that the following are equivalent.

(i) The possibility correspondence is secondary reflexive,

(ii) $\forall E \subseteq \Omega$, $B(\neg BE \cup E) = \Omega$,

(iii) For every model $\mathcal{M}$ and for every formula ϕ, the formula $\Box(\Box\phi \to \phi)$ is valid in $\mathcal{M}$ (the individual believes that if she believes ϕ then ϕ is true).

† The reader may have noticed a difference between the wording of remark 2 and that of remark 4. In remark 2 the properties of the possibility correspondence (seriality, transitivity and euclideanness) were assumed to hold globally, that is, at every state. Here we allow for the possibility that reflexivity may hold at some states but not others, that is, the property is treated as a local property. A global property is one that expresses the logic of rational belief. A local property, on the other hand, is one that is not required by the notion of rational belief.

Secondary reflexivity is implied by euclideanness. Thus in a KD45 frame the individual always believes to be correct in her beliefs.

2.2. BELIEFS BASED ON INFORMATION

It is often the case that individuals form their beliefs on the basis of information they receive. News that the Department of Justice is contemplating taking action against Microsoft conveys information about Microsoft's ability to release Windows 98 on the date originally announced. This information might lead some investors to believe that the release of Windows 98 will be delayed and that the price of Microsoft shares will drop. Others might form the contrary belief that Microsoft will not be deterred by the threat of legal action and that Windows 98 will be released on time, with no effect on the price of shares. In order to distinguish between information and beliefs, we consider a class of structures that represent the paradigm in the economics of information literature.† Let Ω be a set of states. The individual has a partition of the set Ω, representing her information or knowledge. We represent this partition by means of an information correspondence $\mathcal{K} : \Omega \to 2^{\Omega}$ that satisfies the following properties: reflexivity ($\forall\omega \in \Omega, \omega \in \mathcal{K}(\omega)$), transitivity ($\forall\alpha, \beta \in \Omega$, if $\beta \in \mathcal{K}(\alpha)$ then $\mathcal{K}(\beta) \subseteq \mathcal{K}(\alpha)$) and euclideanness ($\forall\alpha, \beta \in \Omega$, if $\beta \in \mathcal{K}(\alpha)$ then $\mathcal{K}(\alpha) \subseteq \mathcal{K}(\beta)$).‡ For every state $\omega \in \Omega$, $\mathcal{K}(\omega)$ is the set of states that, according to her information, the individual cannot rule out at ω. The individual's beliefs are represented, as before, by means of a belief correspondence $\mathcal{B} : \Omega \to 2^{\Omega}$ that satisfies seriality, transitivity and euclideanness. Furthermore, beliefs are based on information and depend only on it, in the following sense: $\forall\alpha, \beta \in \Omega$

$$\text{(R1) } \mathcal{B}(\alpha) \subseteq \mathcal{K}(\alpha)$$

$$\text{(R2) if } \beta \in \mathcal{K}(\alpha) \text{ then } \mathcal{B}(\beta) = \mathcal{B}(\alpha).$$

Whenever Ω is a non-empty set, $\mathcal{K} : \Omega \to 2^{\Omega}$ is reflexive, transitive and euclidean, $\mathcal{B} : \Omega \to 2^{\Omega}$ is serial, transitive and euclidean and together they satisfy (R1) and (R2), we call the structure $\langle\Omega, \mathcal{K}, \mathcal{B}\rangle$ a *KB-frame* ("KB" stands for "Knowledge and

† See, for example, Geanakoplos (1994: pp. 1456–1458). For the importance of the interaction between knowledge and belief in game theory see Dekel and Gul (1997).

‡ Transitivity is implied by the conjunction of reflexivity and euclideanness (see Chellas, 1984: p. 85). However, throughout the paper we shall allow for some redundancies when they add clarity to the exposition or make things look more familiar in the light of the existing literature.

Belief").† From the information correspondence $\mathcal{K}$ one obtains a knowledge operator $K : 2^{\Omega} \to 2^{\Omega}$ by setting, $\forall E \subseteq \Omega, KE = \{\omega \in \Omega : \mathcal{K}(\omega) \subseteq E\}$. The belief operator $B : 2^{\Omega} \to 2^{\Omega}$ is defined similarly (cf. Section 2.1).

REMARK 6: (R1) corresponds to the requirement that the individual always believe what she knows. That is, the following are equivalent.
(i) $\forall \omega \in \Omega$, $\mathcal{B}(\omega) \subseteq \mathcal{K}(\omega)$,
(ii) $\forall E \subseteq \Omega$, $KE \subseteq BE$.
Similarly, (R2) corresponds to the requirement that the individual know her own beliefs. That is, the following are equivalent:
(i) $\forall \alpha, \beta \in \Omega$, if $\beta \in \mathcal{K}(\alpha)$ then $\mathcal{B}(\beta) = \mathcal{B}(\alpha)$,
(ii) $\forall E \subseteq \Omega$, $BE \subseteq KBE$.‡

Note that in a KB frame it is still the case that the individual might have incorrect beliefs (although she is always correct in what she knows). Furthermore (cf. remark 5) she always believes her beliefs to be correct. However, it is not necessarily the case that she knows that her beliefs are correct. Consider the following extension of the example of Figure A: $\Omega = \{\alpha, \beta\}$, $\mathcal{B}(\alpha) = \mathcal{B}(\beta) = \{\beta\}$, $\mathcal{K}(\alpha) = \mathcal{K}(\beta) = \{\alpha, \beta\}$. This is shown in Figure B, where the (trivial) partition generated by $\mathcal{K}$ is denoted by a thick rectangle.

In this example the individual at state α incorrectly believes the earth to be flat and believes to be correct in her belief. However, given her information, she cannot rule out the possibility (represented by state α) that she is wrong in her belief that the earth is flat. It can be shown that if one imposes the requirement that the individual know to have correct beliefs, then belief and knowledge coincide. That is, the following are equivalent:§ (i) $\forall E \subseteq \Omega$, $K(\neg BE \cup E) = \Omega$ and (ii) $\forall E \subseteq \Omega$, $BE = KE$.

† In the economics of information literature beliefs are usually represented by a collection of probability measures, one for each cell of the information partition. In this case one can interpret $\mathcal{B}(\omega)$ as the support of the probability measure over the cell of the partition that contains state ω. It is easy to verify that, with this interpretation, the belief correspondence $\mathcal{B}$ indeed satisfies seriality, transitivity and euclideanness, as well as (R1) and (R2). See also Halpern (1991). Probabilistic beliefs are introduced in Section 3.1.

‡ As remarked before, belief and knowledge pertain to propositions and events should be thought of as representing propositions. In order to establish the interpretation of events as propositions we need to introduce a language with two operators: $\Box_B$ and $\Box_K$. The intended interpretation of $\Box_B\phi$ is "the individual believes ϕ" and the interpretation of $\Box_K\phi$ is "the individual knows ϕ". The notion of a model based on a frame is then developed as explained in the previous section. If $\mathcal{F}$ is a KB frame, then, for every model $\mathcal{M}$ based on it and for every formula ϕ, by (R1) the formula $\Box_K\phi \to \Box_B\phi$ is valid in $\mathcal{M}$ and by (R2) the formula $\Box_B\phi \to \Box_K\Box_B\phi$ is valid in $\mathcal{M}$.

§ Another way of obtaining the collapse of belief into knowledge is to assume that if the individual believes something then he believes that he knows it: $\forall E \subseteq \Omega$, $BE \subseteq BKE$ (cf. Lenzen, 1978).

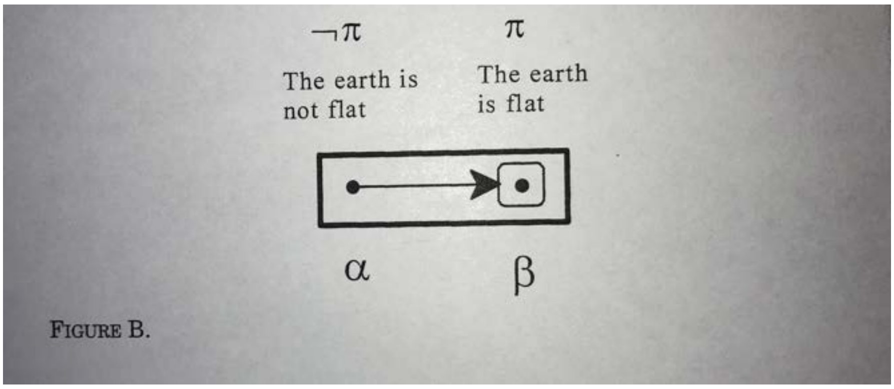


FIGURE B.

KB frames can be used to model change of beliefs over time in response to changes in information: between date t and date $t+1$ the individual might receive new information which might prompt her to revise the beliefs she held at time t. A basic principle in belief revision is the so called "principle of belief persistence", or "conservativity principle", which states that "When changing beliefs in response to new evidence, you should continue to believe as many of the old beliefs as possible" (Harman, 1986: p. 46). In particular, this means that if an individual gets new information, she has to accommodate it in her new belief set (the set of propositions she believes), and, if the new information is not inconsistent with the old belief set, then (i) the individual has to maintain all the beliefs she previously had and (ii) the change should be minimal in the sense that every proposition in the new belief set must be deducible from the union of the old belief set and the new information.

A "possible world" formalization of the principle easily comes to mind. The set of all the propositions that the individual believes corresponds to the set of states of the world that she considers possible and is a subset of the set of states that are not ruled out by the individual's information (or knowledge). The principle of belief persistence then requires that (1) if the individual considers a state possible and her new information does not exclude this state, then she continues to consider it possible, and (2) if the individual regards a particular state as impossible, then she should continue to regard it as impossible unless her new information excludes all the states that she previously regarded as possible. This is closely related to the well-known conditionalization rule to update probability measures. If an individual has probabilistic beliefs, the set of states that she considers possible is simply the support of her subjective probability measure. Let μ_0 be the probability measure representing the agents' beliefs *before* she receives information E and μ_n her subjective probability measure *after* she learns E. The "qualitative part" of the conditionalization rule states the following:

$$\text{if} \quad Supp(\mu_0) \cap E \neq \emptyset \text{ then } Supp(\mu_n) = Supp(\mu_0) \cap E$$

where $Supp(\mu)$ denotes the support of the probability measure μ.

In order to express this principle within the framework of KB frames, we need to index the belief and knowledge correspondences by time. Thus for every date t (where t is a natural number), we postulate a serial, transitive and euclidean belief correspondence $\mathcal{B}_t : \Omega \to 2^\Omega$ and a reflexive, transitive and euclidean knowledge correspondence $\mathcal{K}_t : \Omega \to 2^\Omega$. Furthermore, for every t, we impose (R1) and (R2), that is, $\forall\alpha, \beta \in \Omega$, $\mathcal{B}_t(\alpha) \subseteq \mathcal{K}_t(\alpha)$ and if $\beta \in \mathcal{K}_t(\alpha)$ then $\mathcal{B}_t(\alpha) = \mathcal{B}_t(\beta)$. Within our framework, the conditionalization rule can be stated as follows.

$$\forall t, \forall\omega, \text{ if } \mathcal{B}_t(\omega) \cap \mathcal{K}_{t+1}(\omega) \neq \emptyset \text{ then } \mathcal{B}_{t+1}(\omega) = \mathcal{B}_t(\omega) \cap \mathcal{K}_{t+1}(\omega). \tag{C}$$

The following modification of the example of Figure B is an illustration of a frame that satisfies condition (C): $\Omega = \{\alpha, \beta\}$, $\mathcal{B}_t(\alpha) = \mathcal{B}_t(\beta) = \{\beta\}$, $\mathcal{K}_t(\alpha) = \mathcal{K}_t(\beta) = \{\alpha, \beta\}$, $\mathcal{B}_{t+1}(\alpha) = \mathcal{K}_{t+1}(\alpha) = \{\alpha\}$, $\mathcal{B}_{t+1}(\beta) = \mathcal{K}_{t+1}(\beta) = \{\beta\}$. This is shown in Figure C. Here we have that $\mathcal{B}_t(\alpha) \cap \mathcal{K}_{t+1}(\alpha) = \emptyset$ hence condition (C) is trivially satisfied at α. On the other hand, $\mathcal{B}_t(\beta) = \mathcal{K}_{t+1}(\beta) = \mathcal{B}_{t+1}(\beta)$ and thus (C) is also satisfied at β. Note that at state α the individual (incorrectly) believes the earth to be flat and believes that he will continue to believe so in the future (although he cannot rule out, given his information, that in the future he will learn—and therefore believe—that the earth is not flat).

The following result is proved in Battigalli and Bonanno (1997*a*). Let $B_t : 2^\Omega \to 2^\Omega$ be the belief operator obtained from $\mathcal{B}_t$ and $K_t : 2^\Omega \to 2^\Omega$ be the belief operator obtained from $\mathcal{K}_t$.

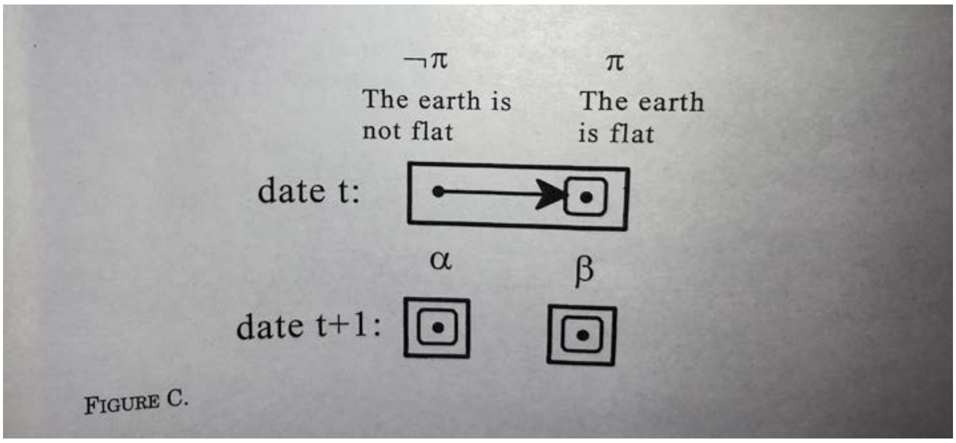


FIGURE C.

PROPOSITION 2.1: *fix an arbitrary time-indexed KB frame. Then the following are equivalent:*
(i) the frame satisfies (C)
(ii) $\forall E \subseteq \Omega$, $B_t E = B_t B_{t+1} E$.

Thus, according to the above proposition, the conservativity principle embodied in the conditionalization rule is equivalent to the requirement that if at any time the individual believes E then he must believe that he will continue to believe E in the future and, conversely, if he believes that in the future he will believe E then he must believe E now.†

2.3. INTERACTIVE BELIEFS AND THE NOTION OF COMMON BELIEF

Game theory models situations where individuals interact with other individuals. In such interactive situations it is important to model not only what a particular individual believes about the external world but also what she believes about other individuals, in particular about their beliefs. This is captured by the notion of interactive belief frame.

A *KD45 frame for interactive beliefs* is a tuple $\mathcal{F} = \langle N, \Omega, \{P_i\}_{i \in N} \rangle$, where $N = \{1, \ldots, n\}$ is a finite set of individuals, Ω is a set of states (or possible worlds) and for every individual $i \in N$, $P_i : \Omega \to 2^\Omega$ is i's possibility correspondence which satisfies seriality, transitivity and euclideanness. Individual i's belief operator $B_i : 2^\Omega \to 2^\Omega$ is defined as usual ($\forall E \subseteq \Omega$, $B_i E = \{\omega \in \Omega : P_i(\omega) \subseteq E\}$). The notion of model based on a frame is as explained before. The only modification consists in enlarging the language to include n modal operators $\Box_1, \Box_2, \ldots, \Box_n$, one for each individual. The intended interpretation of $\Box_i \phi$ is "individual i believes that ϕ". Thus if $\mathcal{F}$ is a frame, $E \subseteq \Omega$ an event and $\mathcal{M}$ a model based on $\mathcal{F}$ where E is the truth set of some formula ϕ (that is, $E = \|\phi\|^{\mathcal{M}}$), then $B_i E$ is the truth set of the formula $\Box_i \phi$, that is, $B_i E = \|\Box_i \phi\|^{\mathcal{M}}$. For example, consider the frame of Figure D and a model based on it where π is the atomic proposition "the universe is expanding", which is true at states τ and γ, that is, $\|\pi\| = \{\tau, \gamma\}$. Here the truth set of $\Box_1 \pi$ is $\{\tau\}$, while the truth set of $\Box_2 \pi$ is $\{\tau, \gamma\}$. It follows that the truth set of $\Box_1 \Box_2 \pi$ is $\{\tau\}$. State τ describes a world where in fact the universe is expanding and both individuals correctly believe that it is expanding; however, while individual 1 believes that individual 2

† For further characterizations of the conditionalization rule in terms of contractions and expansions of the individual's "belief set" (the set of formulae believed by her) brought about by changes in her knowledge set (the set of formulae known by her) see Battigalli and Bonanno (1997*a*). In that paper a syntactic analysis of belief revision is also given, together with a proof of soundness and completeness.

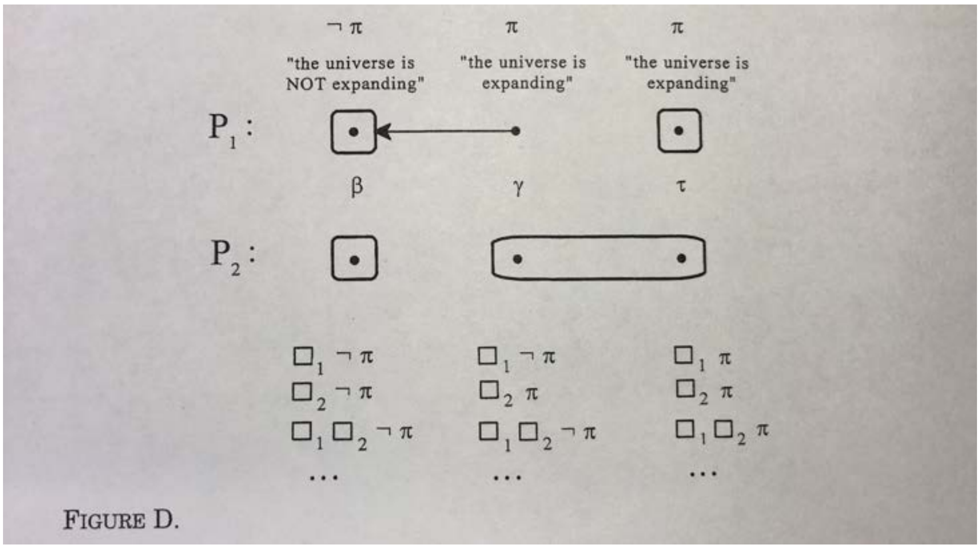

¬ π
"the universe is NOT expanding"
π
"the universe is expanding"
π
"the universe is expanding"
P1 :
β
γ
τ
P2 :
□1 ¬ π
□2 ¬ π
□1 □2 ¬ π
...
□1 ¬ π
□2 π
□1 □2 ¬ π
...
□1 π
□2 π
□1 □2 π
...

FIGURE D.

believes that the universe is expanding, individual 2 is uncertain as to whether 1 correctly believes that it is expanding or 1 incorrectly believes that it is not expanding ($\|\Box_1\neg\pi\| = \{\beta,\gamma\}$) and incorrectly attributes the same belief to individual 2 ($\|\Box_1\Box_2\neg\pi\| = \{\beta,\gamma\}$).

The common belief operator B_* is defined as follows. First, for every event $E \subseteq \Omega$, let $B_eE = \cap_{i\in N}B_iE$, that is, B_eE is the event that everybody believes E. For any operator B, define B^k, the kth iteration of B, as follows: for all E, $k \geq 1$, $B^0E = E$ and $B^kE = BB^{k-1}E$. The event that E is *commonly believed* is defined as the infinite intersection:

$$B_*E = B_eE \cap B_eB_eE \cap B_eB_eB_eE \cap \ldots = \bigcap_{k\geq 1} B_e^kE$$

Thus an event E is commonly believed if everybody believes it, everybody believes that everybody believes it, and so on, *ad infinitum*.

The corresponding *common possibility correspondence* $P_* : \Omega \to 2^\Omega$ is given by: for every $\alpha \in \Omega$, $P_*(\alpha) = \{\omega \in \Omega : \alpha \in \neg B_*\neg\{\omega\}\}$. P_* can be characterized† as the transitive closure of $\cup_{i\in N}P_i$, that is,

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

† See, for example, Bonanno (1996), Fagin *et al.* (1995), Halpern and Moses (1992), Lismont and Mongin (1994, 1995). These authors also show that the common belief operator can be alternatively defined by means of a finite list of axioms, rather than as an infinite conjunction.

In order to capture the notion of common belief in a model, one needs to extend the language by adding another operator $\Box_*$. If ϕ is a formula, the intended interpretation of $\Box_*\phi$ is "it is common belief that ϕ" and if $\mathcal{M}$ is a model where $E = \|\phi\|^{\mathcal{M}}$ then the truth set of $\Box_*\phi$ in $\mathcal{M}$ is given by $B_*E = \{\omega \in \Omega : P_*(\omega) \subseteq E\}$. For example, consider again the frame of Figure D. The common possibility correspondence is given by $P_*(\beta) = \{\beta\}$ and $P_*(\gamma) = P_*(\tau) = \{\beta, \gamma, \tau\}$. Figure E illustrates P_* for the model of Figure D (according to the convention established in remark 3) with the extended language that includes the common belief operator $\Box_*$. At state γ individual 1 wrongly believes that it is common belief that the universe is not expanding; hence, since $\gamma \in P_2(\tau)$, at state τ individual 2 considers it possible that individual 1 has such incorrect beliefs ($\neg\Box_2\neg\Box_1\Box_*\neg\pi$ is true at τ).

REMARK 7: note that, although P_* is always non-empty-valued (or serial) and transitive, in general it need not be euclidean, despite the fact that the individual possibility correspondences are (recall that P_* is euclidean if and only if B_* satisfies Negative Introspection: $\forall E \subseteq \Omega$, $\neg B_*E \subseteq B_*\neg B_*E$). For example, in the frame of Figure E, $\beta \in P_*(\tau) = \Omega$ but $P_*(\tau) \nsubseteq P_*(\beta) = \{\beta\}$. Let $E = \{\beta\}$. Then $\tau \in \neg B_*E$ but $\tau \notin B_*\neg B_*E$ since $\beta \in P_*(\tau)$ and $\beta \in B_*E$.

REMARK 8: it can be shown that in a KD45 interactive frame a proposition is commonly believed if and only if everybody believes that it is commonly believed: for every $E \subseteq \Omega$, $B_*E = \cap_{i\in N} B_i B_* E$.

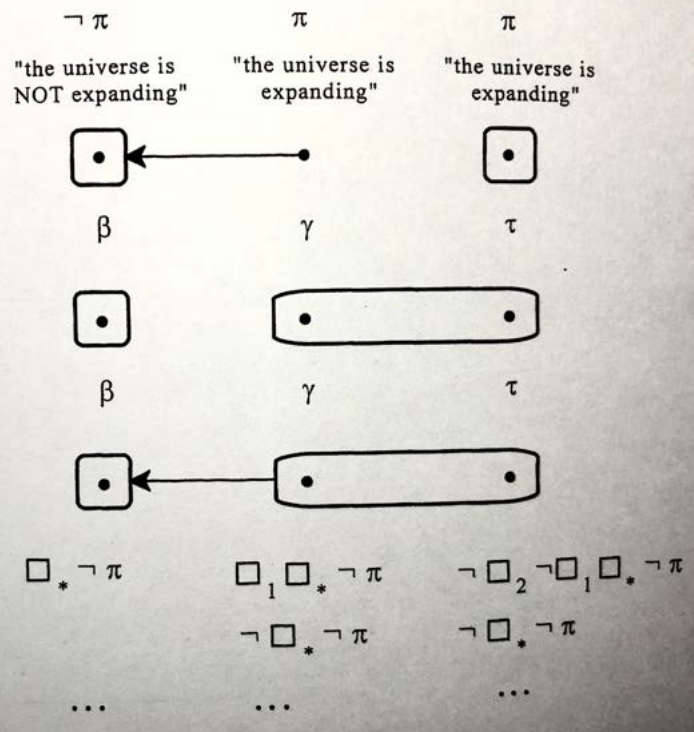

¬ π
π
π
"the universe is NOT expanding"
"the universe is expanding"
"the universe is expanding"
P₁ :
β
γ
τ
P₂ :
β
γ
τ
P* :
□* ¬ π
□₁□* ¬ π
¬ □₂ ¬□₁□* ¬ π
¬ □* ¬ π
¬ □* ¬ π
...
...
...

FIGURE E.

2.4. INTERSUBJECTIVE PROPERTIES OF BELIEFS

The structures that are most often used in the economics and computer science literature to discuss interactive beliefs/knowledge are partition structures.† Partition structures embody the S5 logic for individual beliefs, in particular the Truth Axiom, that is, the assumption that it is a necessary truth (true in all possible worlds of the model) that no one has any false beliefs. As Stalnaker (1994, 1996) points out, there is an important conceptual difference between a theory that builds S5 into the concept of belief (which—Stalnaker argues—is based on equivocating between knowledge and belief) and a theory that describes epistemic conditions under which knowledge and belief coincide, and then considers the consequences of assuming those conditions. In the latter, the Truth Axiom can be expressed locally (that is, as a property of the individuals' beliefs) as the condition that no one has any false beliefs and that it is common belief that no one has any false beliefs.

Let $\mathbf{T}_j \subseteq \Omega$ (for **Truth of *j*'s beliefs**) be the following event:‡

$$\mathbf{T}_j = \bigcap_{E \in 2^{\Omega}} (\neg B_j E \cup E)$$

Thus, for every $\alpha \in \Omega$, $\alpha \in \mathbf{T}_j$ if and only if individual j does not have any false beliefs at α (for every $E \subseteq \Omega$, if $\alpha \in B_j E$ then $\alpha \in E$).§ Let $\mathbf{T}$ (for **Truth**) be the event that no individual has any false beliefs:

$$\mathbf{T} = \bigcap_{j \in N} \mathbf{T}_j$$

For example, in the frame of Figure E, $\mathbf{T} = \{\beta, \tau\}$ and, therefore, $B_* \mathbf{T} = \{\beta\}$.

DEFINITION 2.2: *for every $\alpha \in \Omega$, the Truth Condition holds at α if and only if $\alpha \in \mathbf{T} \cap B_* \mathbf{T}$.*

The above definition is justified by the following observation. Given a frame $\langle N, \Omega, \{P_i\}_{i \in N} \rangle$, and a state $\tau \in \Omega$, define the *τ-reduced frame* as the frame $\langle N, \Omega', \{P'_i\}_{i \in N} \rangle$, where $\Omega' = P_*(\tau) \cup \{\tau\}$ and P'_i is the restriction of P_i to Ω'. Let B'_i be the corresponding belief operator of individual i and P'_* the corresponding common

† See, for example, Aumann (1976, 1987, 1989, 1995, 1996, 1998*a*, *b*), Geanakoplos (1992), Fagin *et al.* (1995).

‡ Throughout the paper, bold-face capital letters are used to denote events (sets of states) with a particular interpretation we want to emphasize. The letter used is meant to be suggestive of the interpretation.

§ Recall (cf. remark 4) that $\alpha \in \mathbf{T}_j$ if and only if $\alpha \in P_j(\alpha)$.

possibility correspondence. Then P'_* is the restriction of P_* to Ω' [in particular, $P'_*(\tau) = P_*(\tau)$] and, for every $E' \subseteq \Omega'$, $B'_iE' = B_iE' \cap \Omega'$. Fix a frame $\langle N, \Omega, \{P_i\}_{i\in N}\rangle$, and a state $\tau \in \Omega$ such that $\tau \in \mathbf{T} \cap B_*\mathbf{T}$. Then in the τ-reduced frame the following is true: $\forall i \in N$, $\forall E' \subseteq \Omega'$, $B'_iE' \subseteq E'$ (note, however, that in the original frame in general it is not true that $\forall i \in N$, $\forall E \subseteq \Omega$, $B_iE \subseteq E$: see Figure F(i)). Thus the τ-reduced frame is a partitional frame (unlike the original frame, in general). Figure F(ii) shows the τ-reduced frame corresponding to the frame of Figure F(i).

The intersubjective implications of the Truth Axiom ($\forall i \in N$, $\forall E \subseteq \Omega, B_iE \subseteq E$) are strong:

> The assumption that Alice believes (with probability one) that Bert believes (with probability one) that the cat ate the canary tells us nothing about what Alice believes about the cat and the canary themselves. But if we assume instead that Alice knows that Bert knows that the cat ate the canary, it follows, not only that the cat in fact ate the canary, but that Alice knows it, and therefore believes it as well (Stalnaker, 1996: p. 153).

This observation can be stated as a local property of beliefs, as follows. Given two individuals, i and j, and a state α, *i is like-minded with j* at α if and only if i shares all the beliefs that she attributes to j, that is, for every event E, if $\alpha \in B_iB_jE$ then $\alpha \in B_iE$. Let $\mathbf{L}_{ij}$ be the event that i **is like-minded with** j:

$$\mathbf{L}_{ij} = \bigcap_{E\subseteq\Omega} \left(\neg B_iB_jE \cup B_iE\right).$$

Let $\mathbf{L}$ be the event that every individual is **like-minded** with every other individual:

$$\mathbf{L} = \bigcap_{i\in N}\bigcap_{j\in N} \mathbf{L}_{ij}$$

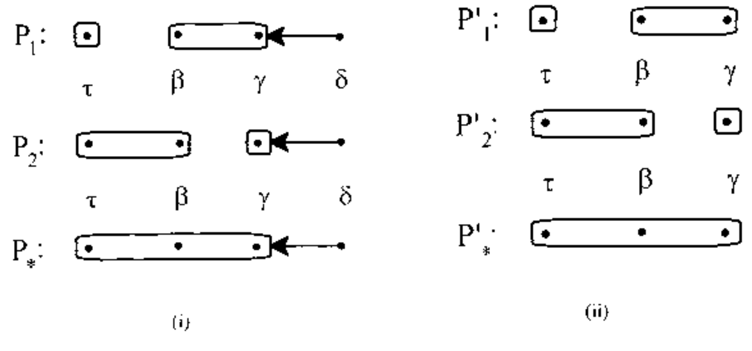


FIGURE F.

The following equivalence (proved in Bonanno & Nehring, 1998*a*) formalizes Stalnaker's observation and exhibits a converse to it. It is a straightforward consequence of secondary reflexivity.

PROPOSITION 2.3: $\mathbf{L}_{ij} = B_i\mathbf{T}_j$ *and, therefore,* $B_*\mathbf{L} = B_*\mathbf{T}$.

Thus i is like-minded with j if and only if i believes that j has correct beliefs; furthermore, common belief in like-mindedness is equivalent to common belief that no individual has any wrong beliefs. We call this latter property (represented by the event $B_*\mathbf{T}$) *common belief in no error*. It will be shown in Section 3 that the assumption of common belief in no error has important implications in the epistemic foundations of solution concepts in game theory (see, for example, Ben Porath, 1997; Stalnaker, 1994, 1996; Stuart, 1997).

A weaker property than common belief in no error is **Agreement**, defined as the common possibility of common belief in no error and denoted by **A**:

$$\mathbf{A} = \neg B_*\neg B_*\mathbf{T}$$

The term "Agreement" is justified by the fact that this property is equivalent to the impossibility of "agreeing to disagree" about qualitative belief indices (see Bonanno & Nehring, 1998*a*) and is thus a qualitative generalization of the notion of agreement introduced by Aumann (1976).

To gain further insight into the property of common belief in no error and the Truth Axiom we introduce two more properties that, together with Agreement, provide a decomposition of the Truth Axiom.

Let $\mathbf{T}_{CB}$ (for **Truth** *about* **common belief**) and $\mathbf{T}^*$ (for **Truth** *of* **common belief**) be the following events

$$\mathbf{T}_{CB} = \bigcap_{i \in N} \bigcap_{E \subseteq \Omega} (\neg B_i B_* E \cup B_* E)$$

$$\mathbf{T}^* = \bigcap_{E \subseteq \Omega} (\neg B_* E \cup E)$$

$\mathbf{T}_{CB}$ captures the notion that individuals are correct in their beliefs about what is commonly believed: $\alpha \in \mathbf{T}_{CB}$ if and only if, for every event E and individual i, if, at α, individual i believes that E is commonly believed, then, at α, E is indeed commonly believed (if $\alpha \in B_i B_* E$, then $\alpha \in B_* E$). On the other hand, $\alpha \in \mathbf{T}^*$ if and only if at α whatever is commonly believed is true (for every event E, if $\alpha \in B_* E$ then $\alpha \in E$).† Truth of common belief ($\mathbf{T}^*$) is a much weaker property than truth (or correctness) of

† It is straightforward that $a \in \mathbf{T}^*$ if and only if, $\alpha \in P_*(\alpha)$.

individual beliefs (**T**); in particular, every *KD*45 frame satisfies the property that $B_*\mathbf{T}^* = \Omega$, while in general $B_*\mathbf{T} \neq \Omega$. Thus, while typically $\mathbf{T} \cap B_*\mathbf{T} \neq \mathbf{T}$, it is always the case that $\mathbf{T}^* \cap B_*\mathbf{T}^* = \mathbf{T}^*$. In this sense $\mathbf{T}^*$ can be viewed as "truth with no intersubjective implications": the addition of $B_*\mathbf{T}^*$ to $\mathbf{T}^*$ does not yield a stronger property than $\mathbf{T}^*$.

The following proposition (proved in Bonanno & Nehring, 1998*a*) gives a decomposition of the Truth Axiom in terms of Agreement, Truth of common belief and (common belief in) Truth about common belief.

PROPOSITION 2.4: $\mathbf{T} \cap B_*\mathbf{T} = \mathbf{T}^* \cap B_*\mathbf{T}_{CB} \cap \mathbf{A}$.

REMARK 9: none of $\mathbf{T}^*$, $\mathbf{T}_{CB}$ and $B_*\mathbf{T}_{CB}$, either individually or in conjunction with the others, has any "agreement" implications. This can be seen from Figure G where $\mathbf{T}^* = \mathbf{T}_{CB} = B_*\mathbf{T}_{CB} = \Omega$ and yet at both τ and β the individuals agree to strongly disagree, in the sense that it is common belief that individual 2 believes E and individual 1 believes $\neg E$, where $E = \{\tau\} : B_*(B_1\neg E \cap B_2 E) = \Omega$. On the other hand, as remarked before, **A** is precisely the property that rules out disagreement.

As noted in remark 7, the common possibility correspondence P_* satisfies non-empty-valuedness and transitivity but not necessarily euclideanness. It follows that the common belief operator B_* satisfies consistency ($B_*E \subseteq \neg B_*\neg E$) and positive introspection ($B_*E \subseteq B_*B_*E$) but not necessarily negative introspection ($\neg B_*E \subseteq B_*\neg B_*E$). Thus Negative Introspection of common belief implies intersubjective restrictions on beliefs, which are uncovered in proposition 2.5 below.

Let (NI stands for "Negative Introspection")

$$\mathbf{NI}^* = \bigcap_{E \subseteq \Omega} (B_*E \cup B_*\neg B_*E)$$

Thus $\alpha \in \mathbf{NI}^*$ if and only if, for every event E, whenever at α it is not common belief that E, then, at α, it is common belief that

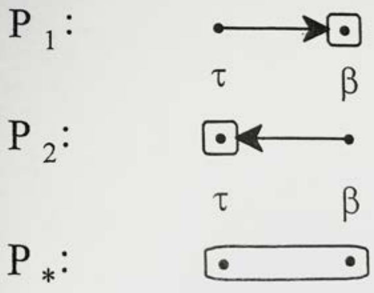


FIGURE G.

E is not commonly believed (if $\alpha \in \neg B_*E$ then $\alpha \in B_*\neg B_*E$). The following result is proved in Bonanno and Nehring (1998*b*).

PROPOSITION 2.5: $\mathbf{NI}^* = \mathbf{T}_{CB} \cap B_*\mathbf{T}_{CB}$.

Thus Negative Introspection of common belief is equivalent to Truth about common belief and common belief in it.† Since $\mathbf{NI}^*$ can be viewed as describing the "logic" of common belief, a global (or "axiomatic") version of proposition 2.5 is of some interest. It is provided in the following corollary.

COROLLARY 2.6: $\mathbf{NI}^* = \Omega$ *if and only if* $\mathbf{T}_{CB} = \Omega$.‡

It is clear from propositions 2.4 and 2.5 that $\mathbf{T}_{CB}$ captures an important intersubjective property of beliefs. It will be shown in the next section that $\mathbf{T}_{CB}$ can be interpreted as reflecting an intersubjective notion of caution.

2.5. KNOWLEDGE AND BELIEF AT THE INTERSUBJECTIVE LEVEL

In this section we extend the knowledge and belief frames of Section 2.2 to interactive situations. Integrated epistemic systems that jointly consider knowledge and belief have been studied in philosophy (Hintikka, 1962; Lentzen, 1978), artificial intelligence and computer science (Halpern, 1991; van der Hoek, 1993; van der Hoek & Meyer, 1995; Kraus & Lehmann, 1978), economics and game theory (Battigalli & Bonanno, 1997*a*; Dekel & Gul, 1997; Geanakoplos, 1994). The philosophy and artificial intelligence literature has dealt mainly with single-agent systems and the focus has been on the tendency of belief to collapse into knowledge as a result of plausible-looking axioms. In game theory a study of systems of knowledge and belief arises naturally in the context of extensive form games. In this section we focus on intersubjective properties of knowledge and beliefs and study their implications.

† One may wonder whether there is something qualitatively different about the truth of this very special type of beliefs. This question can be answered affirmatively, in that truth about common belief is necessary and sufficient for individuals' beliefs about common belief to coincide: we call this "Shared Worlds". (By comparison, having correct beliefs about what others believe, in general, does not imply sharing their beliefs.) Let **SW** be the following event: $\mathbf{SW} = \cap_{i\in N} \cap_{j\in N} \cap_{E\subseteq\Omega}(\neg B_iB_*E \cup B_jB_*E)$. **SW** captures the notion that individuals agree on what is commonly believed: $\alpha \in \mathbf{SW}$ if and only if, for every event E, whenever one individual believes that it is common belief that E, then every other individual believes that too. It can be shown that $\mathbf{SW} = \mathbf{T_{CB}}$.

‡ That is, the following are equivalent: (i) $\forall E \subseteq \Omega$, $\neg B_*E \subseteq B_*\neg B_*E$ and (ii) $\forall i \in N$, $\forall E \subseteq \Omega$, $B_iB_*E \subseteq B_*E$.

An *interactive KB-frame* is a tuple $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}, \{\mathcal{K}_i\}_{i\in N}\rangle$ where N is a set of individuals, Ω a set of states and, for every individual i, $\mathcal{B}_i : \Omega \to 2^\Omega$ is i's belief correspondence and $\mathcal{K}_i : \Omega \to 2^\Omega$ is i's information correspondence. $\mathcal{B}_i$ is assumed to be serial, transitive and euclidean while $\mathcal{K}_i$ is assumed to be reflexive, transitive and euclidean. Furthermore, together they satisfy the following properties: $\forall\alpha, \beta \in \Omega$, (R1) $\mathcal{B}_i(\alpha) \subseteq \mathcal{K}_i(\alpha)$ (knowledge implies belief) and (R2) if $\beta \in \mathcal{K}_i(\alpha)$ then $\mathcal{B}_i(\beta) = \mathcal{B}_i(\alpha)$ (knowledge of own beliefs). Let $B_i : 2^\Omega \to 2^\Omega$ and $K_i : 2^\Omega \to 2^\Omega$ be the associated belief and knowledge operators (respectively) of individual i. The common belief operator B_* is defined as in Section 2.3 and the common knowledge operator K_* is analogous: $K_eE = \cap_{i=1}^n K_iE$ is the event "everybody knows E" and $K_*E = \cap_{m\geq 1} K_e^m E$. Let $\mathcal{B}_* : \Omega \to 2^\Omega$ and $\mathcal{K}_* : \Omega \to 2^\Omega$ be the corresponding possibility correspondences:

$$\forall\alpha \in \Omega, \mathcal{B}_*(\alpha) = \{\omega \in \Omega : \alpha \in \neg B_*\neg\{\omega\}\},$$

$$\mathcal{K}_*(\alpha) = \{\omega \in \Omega : \alpha \in \neg K_*\neg\{\omega\}\}.$$

As explained before, $\mathcal{B}_*$ is the transitive closure of $\cup_{i\in N}\mathcal{B}_i$ and $\mathcal{K}_*$ is the transitive closure of $\cup_{i\in N}\mathcal{K}_i$.

As noted before (cf. remark 7) the common belief operator does not inherit all the properties of the individual belief operators, in particular it does not necessarily satisfy Negative Introspection. Having moved to a knowledge and belief framework, we now find a second property which is not reflected at the "common" level, namely property (R2). That is, whereas individuals always know what they believe, this is not necessarily so at the common level: it may be that the individuals don't commonly know what they commonly believe. This is illustrated in the following example.

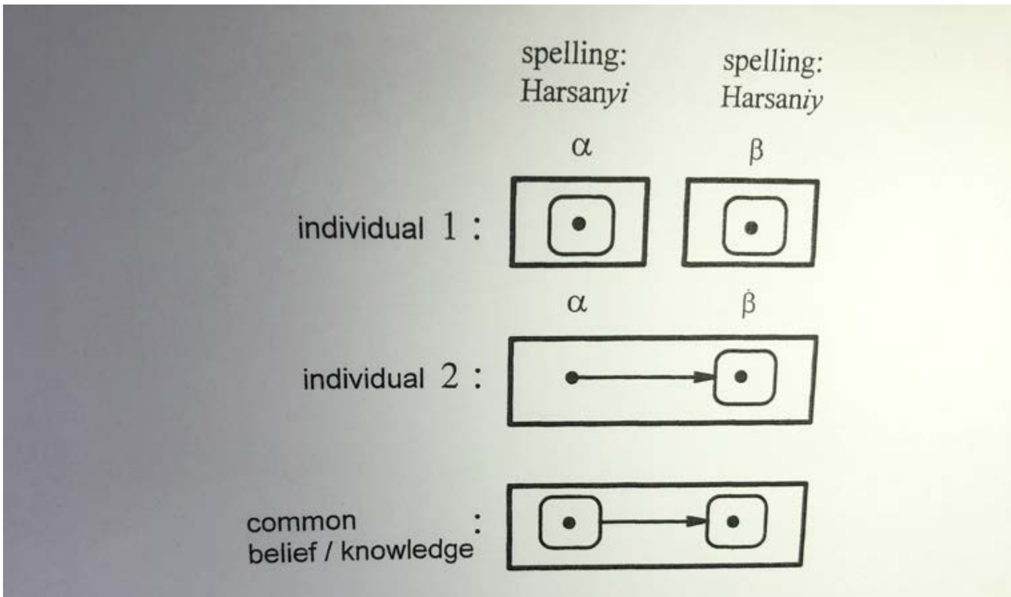


FIGURE H.

EXAMPLE 2.7: individual 1 is a game theorist who knows the correct spelling of his name (Harsan*yi*). Individual 2 mistakenly believes that the spelling is Harsan*iy*. She even believes this spelling to be common belief between them. These beliefs are represented by state α in Figure H, which represents the following frame: $\mathcal{B}_1(\alpha) = \mathcal{K}_1(\alpha) = \{\alpha\}$, $\mathcal{B}_1(\beta) = \mathcal{K}_1(\beta) = \{\beta\}$, $\mathcal{B}_2(\alpha) = \mathcal{B}_2(\beta) = \{\beta\}$, $\mathcal{K}_2(\alpha) = \mathcal{K}_2(\beta) = \{\alpha, \beta\}$. Thus $\mathcal{B}_*(\alpha) = \{\alpha, \beta\}$, $\mathcal{B}_*(\beta) = \{\beta\}$ and $\mathcal{K}_*(\alpha) = \mathcal{K}_*(\beta) = \{\alpha, \beta\}$. Let E be the event that represents the proposition "the spelling is Harsaniy", that is, $E = \{\beta\}$. Then, at state β, E is commonly believed, but it is not common knowledge that it is commonly believed (because individual 2's information set at β contains state α where E is not commonly believed). That is, $\beta \in B_*E$ but $\beta \notin K_*B_*E$.

The following events capture important intersubjective properties of beliefs and knowledge:

- *Common transparency:*

$$\mathbf{TRN}^* = \cap_{E\subseteq\Omega} (\neg B_*E \cup K_*B_*E).$$

- *Intersubjective caution:*

$$\mathbf{ICAU} = \cap_{i\in N} \cap_{E\subseteq\Omega} (\neg B_iB_*E \cup K_iB_*E).$$

- *Equivalence of common belief and common knowledge:*

$$\mathbf{EQU}^* = \cap_{E\subseteq\Omega} ((B_*E \cap K_*E) \cup (\neg B_*E \cap \neg K_*E)).$$

Thus $\omega \in \mathbf{TRN}^*$ if and only if, for every event E, if $\omega \in B_*E$ then $\omega \in K_*B_*E$; $\omega \in \mathbf{ICAU}$ if and only if, for every individual i and every event E, if $\omega \in B_iB_*E$ then $\omega \in K_iB_*E$; finally, $\omega \in \mathbf{EQU}^*$ if and only if, for every event E, $\omega \in B_*E$ if and only if $\omega \in K_*E$. $\mathbf{TRN}^*$ is the analogue, for common belief and knowledge, of property (R2) of individual beliefs/knowledge. **ICAU**, on the other hand, captures the following notion of intersubjective caution of beliefs. While, in general, an individual may simultaneously believe something and not know it (that is, he cannot rule out the possibility that he is wrong in his belief), for *common belief events* the individual's knowledge rules out the possibility that his beliefs might be wrong: if he *believes* that E is common belief then he also *knows* that E is common belief. $\mathbf{EQU}^*$ captures the property that common belief and common knowledge coincide.

The following result (proved in Bonanno & Nehring, 1998*c*) shows that the conjunction of common knowledge of intersubjective caution and Agreement (defined in the previous section) yields common belief in no error of beliefs ($B_*\mathbf{T}$, defined in the previous section) as well as common transparency.

PROPOSITION 2.8: *in an interactive KB-frame the following holds:*

$$\mathbf{A} \cap K_* \mathbf{ICAU} = B_* \mathbf{T} \cap \mathbf{TRN}^* \cap K_* \mathbf{TRN}^*$$

Thus proposition 2.8 gives an interpretation of common belief in no error as an expression of intersubjective caution. As shown in Section 3, common belief in no error is an important property in game theoretic reasoning.

The next result (also proved in Bonanno & Nehring, 1998*c*) shows that if one adds to common knowledge of intersubjective caution the hypothesis that it is common knowledge that only true facts are commonly believed, one obtains the collapse of common belief into common knowledge.

PROPOSITION 2.9: *in an interactive KB-frame the following holds:*

$$K_* \mathbf{ICAU} \cap K_* \mathbf{T}^* = \mathbf{EQU}^* \cap K_* \mathbf{EQU}^*$$

Further intersubjective properties of knowledge and beliefs are studied in Bonanno and Nehring (1998*c*).

3. Epistemic foundations of solution concepts: (A) strategic-form games

The objective of the literature on the epistemic foundations of game theory 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 and the next section we give an introduction to the general approach and review some of the main contributions.

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 programme 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 special 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. Section 2).

This section is devoted to the analysis of normal-form (or strategic-form) games, although some implications for extensive games are also discussed. First we need to introduce probabilistic beliefs.

3.1. BAYESIAN FRAMES

DEFINITION 3.1: *a finite interactive Bayesian frame (or Bayesian frame, for short) is a tuple* $\mathcal{B} = \langle N, \Omega, \{p_i\}_{i \in N} \rangle$, *where* $N = \{1, \ldots, n\}$ *is a finite set of individuals,* Ω *is a finite set of states and for every individual* $i \in N$, $p_i : \Omega \to \Delta(\Omega)$ *(where* $\Delta(\Omega)$ *denotes the set of probability distributions over* Ω*) is a function that specifies her probabilistic beliefs, satisfying the following property [we use the notation* $p_{i,\alpha}$ *rather than* $p_i(\alpha)$*]* : $\forall \alpha, \beta \in \Omega$,

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

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 $\|p_i = p_{i,\alpha}\| = \{\omega \in \Omega : p_{i,\omega} = p_{i,\alpha}\}$ the event that i's beliefs are given by $p_{i,\alpha} \in \Delta(\Omega)$. It is clear that the collection of subsets $\{\|p_i = p_{i,\omega}\| : \omega \in \Omega\}$ is a partition of Ω; it is often referred to as individual i's *type partition*.

Given a Bayesian frame $\mathcal{B}$, its qualitative frame (or frame, for short) is the tuple $\mathcal{Q} = \langle N, \Omega, \{P_i\}_{i \in N} \rangle$, where N and Ω are as in definition 3.1 and for every individual $i \in N$, $P_i : \Omega \to 2^{\Omega}$ is i's possibility correspondence, derived from i's probabilistic beliefs as follows:†

$$P_i(\alpha) = Supp(p_{i,\alpha})$$

Thus, for every $\alpha \in \Omega$, $P_i(\alpha)$ is the set of states that individual i considers possible (i.e., attaches positive probability to) at α. It follows from condition (1) of definition 3.1 that the possibility correspondence of every individual i is serial, transitive and euclidean. Thus the qualitative frame corresponding to a Bayesian frame coincides with the notion of interactive frame introduced in Section 2.3. Let $B_i : 2^{\Omega} \to 2^{\Omega}$ be the belief operator of individual i and $B_* : 2^{\Omega} \to 2^{\Omega}$ the corresponding common belief operator.

3.2. TYPE SPACES AND HIERARCHIES OF BELIEFS

Let us recapitulate some of the concepts introduced so far and elaborate on them. A *state of the world* should be thought of as a complete description of every relevant aspect of a situation. In

† If $\mu \in \Delta(\Omega)$, $Supp(\mu)$ denotes the support of μ, that is, the set of states that are assigned positive probability by μ.

an interactive epistemology framework this description has two components: a description of the *external* state (e.g., the weather or the actions of the individuals) and a description of the *epistemic* states of the individuals.

Fix an interactive beliefs frame $\mathcal{F} = \langle N, \Omega, \{P_i\}_{i \in N} \rangle$. In a modal logic approach we provide an interpretation of states of the world $\omega \in \Omega$ by introducing a set Π of primitive sentences about the external state and a function $f : \Pi \to 2^{\Omega}$ specifying the set of states at which a primitive sentence is true. Thus the set of sentences $\{\pi : \omega \in f(\pi)\}$ is the description of the external world at state ω (recall that if $\omega \notin f(\pi)$ then $\neg\pi$ is true at ω). A language Φ is constructed by introducing modal operators $\Box_i$ ("*i* believes that") and combining primitive sentences into formulae by means of logical connectives and modal operators. The possibility correspondences P_i $(i \in N)$ are used to specify which formulae involving epistemic operators are satisfied at a given state. Thus the set of formulae $\{\Box_i \phi : P_i(\omega) \subseteq \|\phi\|\}$ is a description of the epistemic state of individual i at ω. This set can be partitioned into formulae involving beliefs of different orders. *First-order beliefs* are individual *i*'s beliefs about the external world. *Second-order beliefs* are *i*'s beliefs about the external world *and* the first order beliefs of all the individuals *j*.† *nth-order beliefs* are *i*'s beliefs about (the external world and) the 1-to-$(n-1)$th-order beliefs of all the individuals *j*. Let Φ^0 represent the set of formulae which do not involve any epistemic operator. Then the *first-order beliefs set* of individual i at state ω is the set of formulae $\{\Box_i \phi : \phi \in \Phi^0, P_i(\omega) \subseteq \|\phi\|\}$. Let $\Phi^1 = \Phi^0 \cup \{\Box_j \phi : j \in N, \phi \in \Phi^0\}$. Then *i*'s *second-order beliefs set* at ω is the set of formulae $\{\Box_i \phi : \phi \in \Phi^1, P_i(\omega) \subseteq \|\phi\|\}$.‡ Individual *i*'s nth-order beliefs set can be constructed inductively.§

3.2.1. Bayesian frames, models and type spaces

The beliefs considered above are not probabilistic, but it should be intuitively clear that it makes sense to think of probabilistic beliefs of different orders. We introduced probabilistic beliefs in the

† We may also consider only *i*'s beliefs about other individuals' beliefs. The introspection properties take care of *i*'s beliefs about himself.

‡ Note that we define the second-order beliefs set at ω to be inclusive of the first-order beliefs set. Thus an element of this set is either a formula $\psi = \Box_i \phi$ with $\phi \in \Phi^0$ and $P_i(\omega) \subseteq \|\phi\|$, or a formula $\psi = \Box_i \Box_j \phi$ with $\phi \in \Phi^0$ and $P_j(\omega') \subseteq \|\phi\|$ for all $\omega' \in P_i(\omega)$. This is analogous to the definition of second-order probabilistic beliefs given below.

§ Let $\Phi^{n-1} = \Phi^{n-2} \cup \{\Box_j \phi : j \in N, \phi \in \Phi^{n-2}\}$.
Then the *nth-order beliefs set* of individual i at state ω is $\{\Box_i \phi : \phi \in \Phi^{n-1}, P_i(\omega) \subseteq \|\phi\|\}$.

previous section by means of Bayesian frames. By definition, frames do not provide any interpretation of the states of the world. Therefore, within a Bayesian frame, we cannot separate external states from epistemic states and we cannot describe beliefs of different orders. We might provide an interpretation using the modal logic approach. However, the formalization of probabilistic beliefs by means of modal logic turns out to be difficult.† We take an easier route. We assume that there is a well-defined set S (*finite* unless we explicitly say otherwise) of *external states*, whose interpretation we take for granted. Then we add to the Bayesian frame $\mathcal{B} = \langle N, \Omega, \{p_i\}_{i\in N}\rangle$ a function $\sigma : \Omega \to S$, where $\sigma(\omega)$ corresponds to the description of the external world at state ω. For example, if we have a finite set of primitive sentences $\Pi = \{\pi_1, \ldots, \pi_K\}$, we let $S = \{0, 1\}^{\Pi}$ (the set of functions from Π to $\{0, 1\}$), where $s(\pi_k) = 1$ ($s(\pi_k) = 0$) means that π_k is true (false) at external state s. Then the function σ corresponds to an interpretation function $f : \Pi \to 2^{\Omega}$, where $s = \sigma(\omega)$ if and only if $\omega \in f(\pi_k)$ for all π_k such that $s(\pi_k) = 1$. (Other examples of the specification of S are given in the following subsections: S can be the set of strategy profiles in a complete information game or the set of profiles of payoff-relevant states and strategies in a game with incomplete information.) The pair $\mathcal{M} = \langle \mathcal{B}, \sigma\rangle$ is an epistemic *model for* S.‡ The first-order (probabilistic) beliefs of individual i at state α in model $\mathcal{M}$ are given by the probability measure $\mu^1_{i,\alpha} \in \Delta(S)$ satisfying $\mu^1_{i,\alpha}(s) = \Sigma_{\omega\in\sigma^{-1}(s)} p_{i,\alpha}(\omega)$. Higher-order beliefs can be specified inductively. For example, i's second-order beliefs (beliefs about the external state and the other individuals' beliefs) at state α are given by the measure $\mu^2_{i,\alpha} \in \Delta\left(S \times [\Delta(S)]^{n-1}\right)$ satisfying

$$\mu^2_{i,\alpha}\left(s, \left(\mu^1_j\right)_{j\neq i}\right) = \sum_{\omega:\left(\sigma(\omega), \left(\mu^1_{j,\omega}\right)_{j\neq i}\right) = \left(s, \left(\mu^1_j\right)_{j\neq i}\right)} p_{i,\alpha}(\omega).$$

[Note that it is important that we specify second-order beliefs as joint beliefs about the external state *and* other individuals' (first order) beliefs. For example, if the external state describes each individuals' actions and we want to model the assumption that i believes that j is rational, we have to look for the set of states of the world where i assigns probability zero to every combination of actions and first-order beliefs for j that violate j's rationality.] Continuing this way, we can see that for any state

† We might introduce modal operators $\Box^p_i$ with the interpretation "i assigns probability at least p to. . .". But in order to have a countable language we should consider only rational values of p. See, for example, Fagin and Halpern (1994).

‡ Or an S-based belief space, in Mertens and Zamir's (1985) terminology.

of the world ω in a model for S there is a corresponding $(n+1)$-tuple $\left(\sigma(\omega), (\mu^1_{i,\omega}, \mu^2_{i,\omega}, \ldots)_{i\in N}\right)$ specifying the external state and an *infinite hierarchy of beliefs* for each individual i.

We now turn to a related, but somewhat more transparent representation of external and epistemic states due to Harsanyi (1967–68). The individuals in N are uncertain about the external state $s \in S$ and have beliefs about the external state and about each other's beliefs. The system of beliefs of individual i is determined by a parameter t_i, called the *epistemic type* of i, through a function θ_i which assigns to each t_i a probability measure on the set of external states *and* epistemic types of others. Epistemic models of this kind are called *type spaces*. We focus mainly on finite type spaces to avoid measure-theoretic technicalities.

DEFINITION 3.2: *let S be a finite set of external states. A finite type space for S is a tuple $\mathcal{T} = \langle N, S, \{T_i\}_{i\in N}, \{\theta_i\}_{i\in N}\rangle$ where $N = \{1, 2, \ldots, n\}$ is a finite set of individuals and for every individual $i \in N$, T_i is a finite set of types and $\theta_i : T_i \to \Delta(S \times T_{-i})$ is a function specifying the probabilistic beliefs of each type about the external states and the other individuals' types (where $T_{-i} = T_1 \times \cdots \times T_{i-1} \times T_{i+1} \times \cdots \times T_n$).*

A *state of the world* in a type space $\mathcal{T}$ for S is an $(n+1)$-tuple $(s, t_1, \ldots, t_n) \in S \times T_1 \times \cdots \times T_n$ specifying the external state and the epistemic types. Every epistemic type corresponds to an infinite hierarchy of beliefs. *First-order beliefs* are obtained in an obvious way: for every individual i and type $t \in T_i$, the first-order beliefs of t are given by the probability measure $\mu^1_{i,t} = mrg_S\, \theta_{i,t} \in \Delta(S)$, where mrg_S denotes "marginal distribution on S" and we write $\theta_{i,t}$ rather $\theta_i(t)$ for the probability measure assigned by type t of player i.† Once we have obtained the first-order-beliefs mappings $t \longmapsto \mu^1_{j,t}$ for all the individuals j, we can define the second order beliefs of any type t of any individual i : $\mu^2_{i,t} \in \Delta\left(S \times [\Delta(S)]^{n-1}\right)$ is the probability measure satisfying

$$\mu^2_{i,t}\left(s, \left(\mu^1_j\right)_{j\neq i}\right) = \theta_{i,t}\left(\left\{(s, t_{-i}) : \forall j \neq i, \mu^1_j = \mu^1_{j,t_j}\right\}\right).$$

Note that $mrg_S\, \mu^2_{i,t} = \mu^1_{i,t}$: since second-order beliefs (in our definition) are also beliefs about the external state, they must subsume first-order beliefs. Higher order beliefs for each type are constructed inductively. They satisfy an analogous "marginalization property" and assign probability zero to the lower order hierarchies of other individuals violating this property. Of course, the same properties are satisfied by the hierarchies of beliefs generated by a model for S.

† Thus $\mu^1_{i,t}(s) = \Sigma_{t_{-i}\in T_{-i}} \theta_{i,t}(s, t_{-i})$. Clearly, $\Sigma_{s\in S} \mu^1_{i,t}(s) = 1$.

Belief (certainty) operators can be defined in a quite straightforward way. We informally assume that every individual knows his epistemic type (cf. Sections 2.2 and 3.1). Therefore for any event $E \subseteq S \times T_1 \times \cdots \times T_n$ and type t_i of individual i the set of states in E that i does not rule out is $\left\{(s', t'_1, \ldots, t'_n) \in E : t'_i = t_i\right\}$. Let

$$E_{t_i} = \{(s', t'_{-i}) \in S \times T_{-i} : (s', t_i, t'_{-i}) \in E\}.$$

Type t_i *believes* (is certain of) E if $\theta_{i,t_i}(E_{t_i}) = 1$. Thus the event "player i believes E" is

$$B_i E = \left\{(s, t_i, t_{-i}) : \theta_{i,t_i}(E_{t_i}) = 1\right\}.$$

Mutual and common belief operators are then defined in the usual way. It can be verified that these belief operators satisfy all the properties of the corresponding operators defined for standard (serial, transitive, euclidean) frames.

As the following remark shows, a type space for S is essentially a model of S where the states of the world are explicitly decomposed into external and epistemic states. Furthermore, every model for S can be mapped into a corresponding type space.

REMARK 10: fix a finite set S of external states.

(1) For any model $\mathcal{M}$ for S, let $\mathcal{T}^{\mathcal{M}}$ denote the corresponding type space where, for each $i \in N$, $T_i \subseteq \Delta(\Omega)$ is the range of p_i and θ_i is such that, for all $\alpha \in \Omega$, $s \in S$, $(q_j)_{j \neq i} \in T_{-i} \subseteq [\Delta(\Omega)]^{n-1}$,

$$\theta_{i,p_{i,\alpha}}\left(s, (q_j)_{j\neq i}\right) = \sum_{\omega:\sigma(\omega)=s,\,(p_{j,\omega})_{j\neq i}=(q_j)_{j\neq i}} p_{i,\alpha}(\omega).$$

Then $\mathcal{M}$ and $\mathcal{T}^{\mathcal{M}}$ generate the same hierarchies of beliefs, that is, for all $\omega \in \Omega$, $i \in N$, $(\mu^1_{i,\omega}, \mu^2_{i,\omega}, \ldots) = (\mu^1_{i,p_{i,\omega}}, \mu^2_{i,p_{i,\omega}}, \ldots)$.

(2) For any type space $\mathcal{T}$ for S, let $\mathcal{M}^{\mathcal{T}}$ denote the corresponding model where $\Omega = S \times T_1 \times \ldots T_n$, $\sigma : \Omega \to S$ is the projection function and for each $i \in N$, p_i is such that, for all $(s, t_i, t_{-i}) \in \Omega$, $(s', t'_{-i}) \in S \times T_{-i}$,

$$p_{i,(s,t_i,t_{-i})}\left(s', t_i, t'_{-i}\right) = \theta_{i,t_i}\left(s', t'_{-i}\right).$$

Then $\mathcal{T}$ and $\mathcal{M}^{\mathcal{T}}$ generate the same hierarchies of beliefs, that is, for all $i \in N$, $\omega = (s, t_i, t_{-i}) \in \Omega$, $(\mu^1_{i,t_i}, \mu^2_{i,t_i}, \ldots) = (\mu^1_{i,\omega}, \mu^2_{i,\omega}, \ldots)$.

3.2.2. The universal type space

We have seen that for any given set S of external states we can use an epistemic model or a type space for S to provide consistent representations of the individuals' systems of beliefs.

In particular, every state of the world in the model (or type space) induces a consistent infinite hierarchy of beliefs. But not all the consistent hierarchies are generated in any particular finite model.† In fact, there are infinitely many conceivable first order beliefs, all the elements of $\Delta(S)$. More generally, even an infinite model need not represent all the consistent hierarchies of beliefs and, a priori, it is not even obvious that there exist infinite models representing all the consistent hierarchies. This is to be contrasted with what we can achieve using the modal logic approach to represent non-probabilistic beliefs. For any given system Σ of axioms and inference rules we can construct a *canonical model* which is precisely a model where every (epistemic or non-epistemic) formula in the language Φ which is consistent with Σ is satisfied at some state. A little more precisely, a state of the world in the canonical model is defined as a maximal set of formulae which are mutually consistent (do not generate contradictions) given the axioms and inference rules in Σ.

It turns out that if S is finite‡ we can provide a sort of probabilistic-beliefs-analog of the canonical model of modal logic, that is, a *universal* type space containing all the conceivable hierarchies of beliefs. We will argue that this is important to analyse the epistemic foundations of game theory (see Sections 3.3 and 4). Therefore we provide here a summary of this construction.

We want to define the set of all infinite hierarchies of beliefs satisfying the same consistency properties of the hierarchies that obtain at some state of a type space (or epistemic model) for S. For example, first order beliefs should be the marginals of second-order beliefs and third order beliefs should rule out hierarchies of beliefs of other individuals that do not satisfy this marginalization property.

Let us use the following notation:

- $X^0 = S$,
- $Z^1 = [\Delta(X^0)]^{n-1}, X^1 = X^0 \times Z^1$,
- $Z^2 = [\Delta(X^1)]^{n-1}, X^2 = X^1 \times Z^2$.

Given X^k we define

- $Z^{k+1} = [\Delta(X^k)]^{n-1}, X^{k+1} = X^k \times Z^{k+1}$.§

† Furthermore, the function σ in a model $\mathcal{M}$ for S need not be onto, which means that some conceivable external states are not realized at any state of the world in $\mathcal{M}$.

‡ Indeed if S is a "nice" topological space, e.g., a Polish, or a compact metric space.

§ It can be shown that each X^k is a measurable space and hence the set $\Delta(X^k)$ of all probability measures on X^k is well defined.

For any given individual i, Z^1 is the set of the conceivable combinations of first-order beliefs of all the other individuals. The set of his second-order beliefs is therefore $\Delta(S \times Z^1) = \Delta(X^1)$. His conceivable third-order beliefs are elements of $\Delta(X^2)$ that rule out any combination of first- and second-order beliefs which do not satisfy the marginalization property for at least one individual j. In general, the conceivable k-order beliefs ($k \geq 3$) of an individual i form only a subset of $\Delta(X^k)$, as some elements of $\Delta(X^k)$ do not rule out all the "inconceivable" hierarchies of length k for the other individuals. Furthermore, each conceivable hierarchy $(\mu^1, \ldots, \mu^k, \mu^{k+1}, \ldots)$ must satisfy the "marginalization property": $\mu^k = mrg_{X^{k-1}}\, \mu^{k+1}$. Thus the set of conceivable infinite hierarchies for any individual i is a subset $T_i^U \subset \Delta(X^0) \times \Delta(X^1) \times \cdots \times \Delta(X^k) \times \cdots$, which can be defined inductively as follows:[†]

- $Y_{-i}^1 = X^1$,
- $Y_{-i}^k = \{(s, ((\mu_j^1)_{j \neq i}, \ldots, (\mu_j^{k-1})_{j \neq i}, (\mu_j^k)_{j \neq i}) \in X^{k-1} \times Z^k :$
  $\forall j \neq i, mrg_{X^{k-2}}\, \mu_j^k = \mu_j^{k-1}, \mu_j^k(Y^{k-1}) = 1\}$,
- $T_i^U = \{(\mu_i^1, \mu_i^2, \ldots, \mu_i^k, \ldots) \in \Pi_{k=0}^{\infty} \Delta(X^k)$:
  $\forall k \geq 1, mrg_{X^{k-2}}\, \mu_i^k = \mu_i^{k-1}, \mu_i^k(Y_{-i}^{k-1}) = 1\}$.

The following fundamental result is due to Mertens and Zamir (1985).[‡] It shows that a vector

$$\left(s, \left((\mu_i^1, \mu_i^2, \ldots)\right)_{i \in N}\right) \in S \times \prod_{i \in N} T_i^U$$

is a complete and consistent description of the state of the world and that T_i^U can be interpreted as a set of epistemic types in a type space. Let $T_{-i}^U = \Pi_{j \neq i} T_j^U$.

PROPOSITION 3.3: *there is a "canonical homeomorphism" between T_i^U and $\Delta(S \times T_{-i}^U)$, that is, a bijective and bicontinuous*[§] *function $\theta_i^U : T_i^U \to \Delta(S \times T_{-i}^U)$ such that for all $t = (\mu_i^1, \mu_i^2, \ldots) \in T_i^U$ and for each integer $k \geq 1$,*

$$mrg_{X^{k-1}}\, \theta_{i,t}^U = \mu_i^k.$$

[†] All the individuals are symmetric in this construction because they all have beliefs about the same set of external states. But symmetry is not an important feature of the model.

[‡] See also Brandenburger and Dekel (1993), Heifetz and Samet (1996) and the references therein.

[§] It can be shown that $\Delta(X^k)$ $(k = 0, 1, \ldots)$, T_i^U and $\Delta(S \times T_{-i}^U)$ are complete, separable and metrizable with respect to the topology of weak convergence of measures. Continuity is defined with respect to this topology. See, for example, Brandenburger and Dekel (1993).

The proof of proposition 3.3 is beyond the scope of this paper, but the main idea of the proof is relatively simple and clarifies the whole construction. The main point should be familiar to those who know some theory of stochastic processes and Kolmogorov's extension theorem. Within space $S \times T^U_{-i}$ there are "finite-dimensional" events (measurable subsets) which are simply described by properties of the external state and the other individuals' beliefs up to order k. Let E^k denote such an event. For example, $E^0 = Y \times T^U_{-i}$ (where $Y \subset S$) is an event in $S \times T^U_{-i}$ that is only described by properties of the external state. All the beliefs of order $k+1$ or higher in hierarchy $t = (\mu^1_i, \mu^2_i, \ldots)$ assign a probability to event E^k (for beliefs of order $m = k+2, k+3, \ldots$ just take the marginal on X^k). *By consistency, all these probabilities must be the same*. For example, the probability assigned by hierarchy $t = (\mu^1_i, \mu^2_i, \ldots)$ to $E^0 = Y \times T^U_{-i}$ is $\mu^1_i(Y)$. Let $\theta^U_{i,t}(E^k)$ denote the probability assigned by t to E^k. Then we can take limits to obtain the probability of sets of the form $E = \cap_{k \geq 0} E^k$:

$$\theta^U_{i,t}(E) = \lim_{m \to \infty} \theta^U_{i,t}\left(\bigcap_{k=0}^{m} E^k\right).$$

Finally, we can define $\theta^U_{i,t}$ for all the (measurable) subsets of $S \times T^U_{-i}$ by (countable) additivity. Intuitively, this must be the "right" probability measure to assign to the infinite hierarchy of beliefs t. Now suppose we fix a probability measure $\mu_i \in \Delta(S \times T^U_{-i})$. Then we can derive an infinite hierarchy of beliefs $t = (\mu^1_i, \mu^2_i, \ldots)$ just by taking the marginal on each space X^k, $k = 0, 1, \ldots$, and it turns out that $\theta^U_i(t_i) = \mu_i$.

Proposition 3.3 means that we can think interchangeably of types as infinite hierarchies of beliefs and types as probability measures on the space of combinations of external states and other players' types. The first notion of type is *explicit*, because it relies on an iterative construction starting from first-order beliefs, a concept we already understood well. The second notion of type is *implicit*: its self-referential nature makes it possible to assign to every type an infinite hierarchy of beliefs, but this hierarchy is not explicitly given. This should ring a bell: we are back to type spaces. Indeed, proposition 3.3 shows that

$$\mathcal{T}^U = \left\langle N, S, \left\{T^U_i\right\}_{i \in N}, \left\{\theta^U_i\right\}_{i \in N}\right\rangle$$

is a type space. This particular type space, however, has two features: (1) it is infinite and infinite dimensional, (2) it is *universal* in the sense that it "contains" every type space (more precisely, it contains all the hierarchies of beliefs corresponding to the types of any type space $\mathcal{T}$for S). This is made formal by the following result:

PROPOSITION 3.4: *for every type space $\mathcal{T} = \left\langle N, S, \{T_i\}_{i\in N}, \{\theta_i\}_{i\in N}\right\rangle$ there is a unique n-tuple of functions $\varphi = (\varphi_i)$, $\varphi_i : T_i \to T_i^U$, such that for all $i \in N$, $t \in T_i$, $E \subset S \times T_{-i}^U$ (measurable)*

$$\theta_{i,t}\left(\left\{(s', t'_{-i}) \in S \times T_{-i} : (s, \varphi_{-i}(t_{-i})) \in E\right\}\right) = \theta^U_{i,\varphi_i(t)}(E)$$

where we write $\varphi_{-i}(t_{-i}) = (\varphi_j(t_j))_{j\neq i}$.

The meaning of proposition 3.4 is that there is one and only one way to assign to every type t of every player i in type space $\mathcal{T}$ a corresponding infinite hierarchy of beliefs (i.e., a type in $\mathcal{T}^U$) so that the same probability is assigned by t and $\varphi_i(t)$ to corresponding events. Note that the finiteness of $S \times T_{-i}$ implies that $\theta^U_{i,\varphi_i(t)}$ must be a probability measure with finite support. Indeed for each type $t \in T_i$ the hierarchy/type $\varphi_i(t) \in T_i^U$ "is certain" that his opponents' hierarchies of beliefs belong to the image set $\varphi_{-i}(T_{-i})$. Thus $\mathcal{T}$ corresponds to a finite, "belief-closed" subspace of $\mathcal{T}^U$.

The main step in the proof of proposition 3.4 is the inductive construction of the functions φ_i. These functions simply assign to each type $t \in T_i$ the infinite hierarchy of beliefs $(\mu^1_{i,t}, \mu^2_{i,t} \ldots)$ derived above.†

3.3. MODELS OF STRATEGIC-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 a finite set of strategies for player i and $u_i : S \to \Re$ (where $S = S_1 \times \cdots \times S_n$ and $\Re$ 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 of the interactive situation, in that it determines the choices that are available to the players and the preferences over strategy profiles, but does not specify the players' beliefs about each other or their actual choices. The notion of model of G provides a way of completing the description. Note that in the following definition it is implicitly assumed that G is a game with complete information (the payoff functions are common knowledge).

DEFINITION 3.5: *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 $\mathcal{B} = \left\langle N, \Omega, \{p_i\}_{i\in N}\right\rangle$, is a Bayesian frame (cf. definition 3.1) and, for every player i, $\sigma_i : \Omega \to S_i$ is a function that specifies for every state the choice made by player i at that state.*

† Note that the functions φ_i need not be injective because a type space may contain "duplicate" or "redundant" types.

For every state $\omega \in \Omega$, let $\sigma(\omega) = (\sigma_1(\omega), \ldots, \sigma_n(\omega)) \in S$ be the strategy profile played at ω and, for every player i, denote by $\sigma_{-i}(\omega) \in S_{-i}$ the $(n-1)$-tuple of strategies played by the players other than i. The association of a strategy profile with every state is what gives content to the beliefs of the players and allows the derivation of a profile of infinite hierarchies of beliefs for each state, as explained in Subsection 3.2. All the papers on the epistemic foundations of normal-form solution concepts considered here use these kind of epistemic models. Therefore we conform to this formalization. But it is worth stressing that all the results could be reformulated in terms of type spaces for S. We will argue in Section 4 that type spaces are particularly well suited for the epistemic analysis of extensive form games.

We first provide a precise definition of the event "player i is rational" within a model $\mathcal{M}$ of game G.

DEFINITION 3.6: *fix a strategy s_i and a probability measure $\mu \in \Delta(S_{-i})$. We say that s_i is a best response to μ—written $s_i \in r_i(\mu)$—if, for all $s_i' \in S_i$*

$$\sum_{s_{-i} \in S_{-i}} \left[u_i(s_i, s_{-i}) - u_i(s_i', s_{-i})\right] \mu(s_{-i}) \geq 0.$$

DEFINITION 3.7: *player i is rational at state $\alpha \in \Omega$ if her beliefs at α assign probability one to her choice at α and this choice is a best response to her (marginal) beliefs about the opponents' choices: let $s_{i,\alpha} = \sigma_i(\alpha)$, then*

(1) $P_i(\alpha) \subseteq \sigma_i^{-1}(s_{i,\alpha})$ and

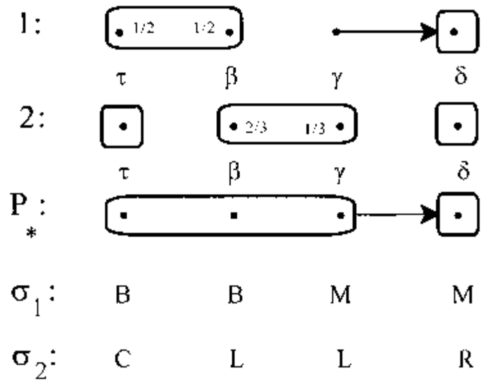
(2) $s_{i,\alpha} \in r_i(mrg_{S_{-i}} p_{i,\alpha})$.

Let $\mathbf{RAT}_i$ be the set of states where player i is rational and $\mathbf{RAT} = \cap_{i \in N} \mathbf{RAT}_i$ the event that all players are rational. Note that, by condition (1) of definition 3.7 if player i is rational, she is certain of being rational ($\mathbf{RAT}_i \subseteq B_i \mathbf{RAT}_i$), but the converse does not hold. However, many papers on the epistemic foundations of game theory adopt a stronger definition of model of a game assuming that condition (1) holds *globally* (and dropping (1) from the definition of rationality). In *these models*, if a player believes she is rational, she is indeed rational ($B_i \mathbf{RAT}_i \subseteq \mathbf{RAT}_i$). We prefer the more general formulation where (1) is assumed only locally as part of a player's rationality because this forces a more transparent formulation of results and because it is more appropriate for the analysis of extensive form games (see Section 4).

EXAMPLE 3.8: Figure I(ii) shows a model of the two-person game illustrated in Figure I(i). Here we have that $\mathbf{RAT}_1 = \{\tau, \beta\}$ and

		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

(i)



(ii)

FIGURE I.

$\mathbf{RAT}_2 = \Omega$; hence $\mathbf{RAT} = \{\tau, \beta\}$. Note also that $B_1\mathbf{RAT} = \{\tau, \beta\}$, $B_2\mathbf{RAT} = \{\tau\}$ and $B_*\mathbf{RAT} = \emptyset$.

3.4. RATIONALIZABILITY

The first solution concept we consider is rationalizability (Bernheim, 1984; Pearce, 1984), which is intended to capture the implications of rationality and common belief in it.

The hypothesis of rationality of all the players allows the elimination of all strategies that are never best responses. Furthermore, if every player is believed to be rational by everybody else, then no player should attach positive subjective probability to strategies of the other players that are never best responses. However, there might be strategies that are never best responses given such restrictions on beliefs. Then the hypothesis that everybody believes everybody else to be rational allows the elimination of such strategies too. This leads us to consider the following iterative elimination process:

- $\forall i \in N, S_i^0 = S_i, \forall k \geq 0, S^k = \Pi_{i \in N} S_i^k, S_{-i}^k = \Pi_{j \neq i} S_j^k$,
- $\forall i \in N$,

$$S_i^{k+1} = \left\{ s_i \in S_i^k : \exists \mu \in \Delta(S_{-i}), s_i \in r_i(\mu), \mu(S_{-i}^k) = 1 \right\},$$

- $S_i^{\infty} = \cap_{k \geq 1} S_i^k, S^{\infty} = \Pi_{i \in N} S_i^{\infty}$.†

Intuitively, this procedure ought to lead to the survival of all and only those strategies that are compatible with rationality and common belief in rationality. The surviving strategies are called *rationalizable*.‡ [Note that by finiteness of S there is some K such that $S^K = S^{\infty}$. In compact-continuous games S^{∞} is the (Hausdorff) limit of S^k as $k \to \infty$.]

By standard results in linear programming, a strategy of player i is never a best response if and only if it is strictly dominated, in the following sense. Recall that a probability distribution over S_i can be interpreted as a mixed strategy for player i. If $\nu_i \in \Delta(S_i)$ and $s_i \in S_i$, we denote by $\nu_i(s_i)$ the probability assigned to s_i by ν_i. A strategy $s_i \in S_i$ is *strictly dominated by* $\nu_i \in \Delta(S_i)$ on $\widehat{S}_{-i} \subseteq S_{-i}$ if, for all $s_{-i} \in \widehat{S}_{-i}$, $u_i(\nu_i, s_{-i}) > u_i(s_i, s_{-i})$, where $u_i(\nu_i, s_{-i}) = \Sigma_{x \in S_i} \nu_i(x) u_i(x, s_{-i})$. [For example, in the game of Figure J(i), strategy B of player 1 is strictly dominated by the mixture $(\frac{1}{2}A, \frac{1}{2}D)$.] Therefore we obtain the following alternative definition of rationalizable strategies:

PROPOSITION 3.9: *(Pearce, 1984) for all $k \geq 0$, $i \in N$,*

$$S_i^{k+1} = \{ s_i \in S_i^k : \forall \nu_i \in \Delta(S_i^k), s_i \textit{ is } \text{not} \textit{ strictly dominated by } \nu_i \textit{ on } S_{-i}^k \}.$$

For the game of Figure J(i), S^1, S^2 and $S^3 = S^{\infty}$ and the corresponding restricted games are shown in Figures J(ii)–(iv). In the game of Figure I(i), $S^{\infty} = \{(T, L), (T, C), (B, L), (B, C)\}$, 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.

† It is easily shown that, in the definition of S_i^{k+1}, on the one hand the restriction $s_i \in S_i^k$ can be eliminated (anyway, strategies cannot "come back"), on the other hand $s_i \in r_i(\mu)$ can be replaced by the *constrained* maximization condition $s_i \in \arg\max_{s_i' \in S_i^k} u_i(s_i', \mu)$. Thus a strategy surviving step k of the procedure need only be compared with other strategies surviving step k.

‡ This is the definition of "correlated" rationalizability. The definition first given by Pearce (1984) and Bernheim (1984) considered only best responses to uncorrelated beliefs.

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

(i) The game G

B is strictly dominated by (1/2 A, 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

(ii) 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

(iii) The game G^2

Now C is strictly dominated by (1/6 A, 5/6 D)

Player 2

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

(iv) The game $G^3 = G^\infty$

No strategy is strictly dominated

FIGURE J.

The following results provide an epistemic characterization of rationalizability. The first such characterization was explicitly provided by Tan and Werlang (1988) using a universal type space (cf. Section 4, which also provides a characterization of each subset S^k). The state space formulation used in propositions 3.10 and 3.11 is due to Stalnaker (1994) (see also Osborne & Rubinstein, 1994), but it was implicit in Brandenburger and Dekel (1987).†

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 I(ii), $\mathbf{S}^\infty = \{\tau, \beta\}$.

† In our terminology, Brandenburger and Dekel (1987) show that in a model of G where players are rational at every state, the players always play rationalizable strategies, and that there is a model of G where the players are rational at every state and every rationalizable profile is played at some state.

PROPOSITION 3.10: *let G be a game and $\mathcal{M}$ an arbitrary model of it. Then*

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

That is, if at a state all the players are rational and 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.

The converse of proposition 3.10 does not hold, the reason being that a model for G typically contains "too few" states of the world. First, the range of the strategy function σ may be smaller than S and even smaller than S^∞. Second, the set of infinite hierarchies of beliefs (epistemic types) corresponding to the states of the model may be too small to "rationalize" all the rationalizable strategies. A rather extreme example is provided by the following model of the game of Figure J: $\Omega = \{\tau\}$, $P_1(\tau) = P_2(\tau) = \{\tau\}$, $\sigma(\tau) = (A, a)$. Then $\tau \in \mathbf{S}^\infty \cap B_* \mathbf{S}^\infty$ but $\mathbf{RAT}_2 = \emptyset$ (and hence $B_* \mathbf{RAT} = \emptyset$), because player 2's belief that player 1 is playing A does not justify her choice of a. In general, that is, for any solution concept, there is always the possibility that at a state in a model the players make the "right" choices "accidentally" or "for the wrong reasons". The following theorem, however, explains the sense in which the notion of common belief in rationality can be thought of as equivalent to that of rationalizability.†

PROPOSITION 3.11: *let G be a game. Then there is a model $\mathcal{M}$ of G such that, for every $s \in S$, s is rationalizable if and only if there is an $\omega \in \Omega$ such that: (1) $\omega \in \mathbf{RAT} \cap B_* \mathbf{RAT}$, and (2) $\sigma(\omega) = s$.*

3.5. STRONG RATIONALIZABILITY

Note that propositions 3.10 and 3.11 are not based on any assumption of correctness of players' beliefs (cf. remark 4), that is, it is not assumed that if a player is certain of event E (i.e., attaches probability 1 to E) then E is indeed true. In particular, a player can be mistaken in ruling out some strategy choices of 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 & Nehring, 1996*b*). The following algorithm is similar to the iterative deletion of strictly

† The equivalence between rationalizability and common belief in rationality is made even more transparent within a universal type space, which—by definition—contains all the conceivable hierarchies of beliefs (cf. Tan & Werlang, 1988; and Section 4).

dominated strategies, but differs from the latter in that it requires the iterative deletion of *profiles* rather than strategies.

DEFINITION 3.12: *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 mixed strategy μ_i such that $u_i(\mu_i, x_{-i}) > u_i(x)$.] For every $k \geq 0$, define $S_s^k \subseteq S$ and $D_s^k \subseteq S$ as follows (the subscript s stands for "strong"): $S_s^0 = S$, D_s^k is the set of profiles that are inferior relative to S_s^k and $S_s^{k+1} = S_s^k \backslash D_s^k$. Let $S_s^\infty = \cap_{k=0}^{\infty} S_s^k$. The strategy profiles in S_s^∞ are called *strongly rationalizable*.

EXAMPLE 3.13: in the game of Figure K(i), the first step in the algorithm leads to the profiles shown in Figure K(ii) [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 K(iii) [now F is dominated by E and C is dominated by A] and the third and final step leads to the profiles shown in Figure K(iv) [now B is dominated by A]. Thus $S_s^\infty = \{(B, D), (C, D), (A, E), (A, F)\}$. Note that, on the other hand, every strategy profile is rationalizable, that is, $S^\infty = S$, since no player has any strictly dominated strategies.

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: $\mathbf{S}_s^\infty = \{\omega \in \Omega : \sigma(\omega) \in S_s^\infty\}$. Let $\mathbf{T}$ (Truth) be defined as in Section 2.4.

PROPOSITION 3.14: *(Stalnaker, 1994; see also Bonanno & Nehring, 1996b).*† *Let G be a game and $\mathcal{M}$ a model of it. Then*

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

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

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, no individual has false beliefs, then it is also true that the strategy profile actually played is strongly rationalizable.

† Stalnaker (1994: p. 63) incorrectly states the result as $B_*\mathbf{T} \cap B_*\mathbf{RAT} \subseteq \mathbf{S}_s^\infty$. Bonanno and Nehring (1996*b*) give a counterexample and prove the results as stated in proposition 3.14.

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

(i)

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

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

(ii)

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

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

(iii)

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

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

(iv)

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

FIGURE K.

To see that, in general, $B_*\mathbf{T} \cap B_*\mathbf{RAT} \not\subseteq \mathbf{S}_s^\infty$ consider the model of the game of Figure K(i) illustrated in Figure L. It is easy to check that $\mathbf{RAT} = \Omega$ (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, $B_*\mathbf{T} = \Omega$. However, while $\tau \in B_*\mathbf{T} \cap B_*\mathbf{RAT}$, $\sigma(\tau) = (A, D) \notin S_s^\infty$.

A partial converse to proposition 3.14 is given by the following result.

PROPOSITION 3.15: *let G be a game and* $s \in S_s^\infty$. *Then there is a model* $\mathcal{M}$ *of G such that: (1)* $\tau \in \mathbf{T} \cap B_*\mathbf{T} \cap B_*\mathbf{RAT}$, *and (2)* $\sigma(\tau) = s$.

The example of Figure K shows that strong rationalizability is considerably stronger than rationalizability. To stress this point, consider the extensive game of Figure M(i), whose normal form is shown in Figure M(ii).

For the normal form, $S^\infty = S$ (that is, all the strategy profiles are rationalizable), since no strategy of any player is strictly

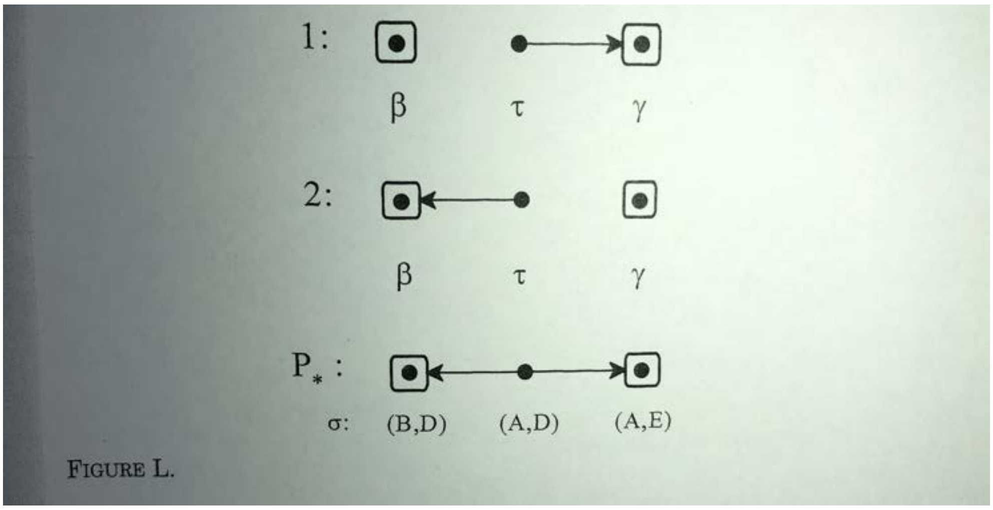
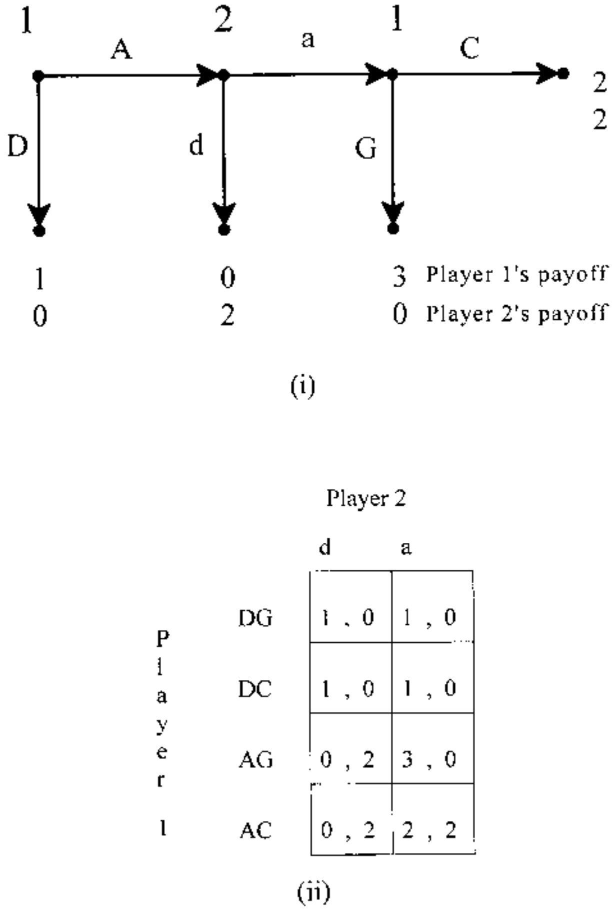


FIGURE L.



(i)

Player 2

		d	a
P l a y e r 1	DG	1 , 0	1 , 0
	DC	1 , 0	1 , 0
	AG	0 , 2	3 , 0
	AC	0 , 2	2 , 2

(ii)

FIGURE M.

dominated. Hence every outcome is compatible with common belief in rationality (in the sense of proposition 3.11). On the other hand, $S_s^\infty = \{(DG, d), (DG, a), (DC, d), (DC, a)\}$† 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.‡ The answer is negative, as the following example shows. Figure N(ii) shows a model of the normal form of the extensive game of Figure N(i). At state τ 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 $\tau \in \mathbf{T} \cap B_*\mathbf{T} \cap B_*\mathbf{RAT}$ (remember that $\mathbf{T} \cap B_*\mathbf{RAT} \subseteq \mathbf{RAT}$, in particular, player 1's choice of A is rational, given his belief that player 2 plays d and a with equal probability).

The extensive game of Figure N(i) 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 N(i) shows.§ Here there is a unique Nash equilibrium outcome, namely the payoff vector (7, 7, 7, 7). Yet in the model shown in Figure O(ii) at state τ the realized outcome is (2, 2, 2, 10) despite the fact that $\tau \in \mathbf{T} \cap B_*\mathbf{T} \cap B_*\mathbf{RAT}$.¶

3.6. CORRELATED EQUILIBRIUM

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

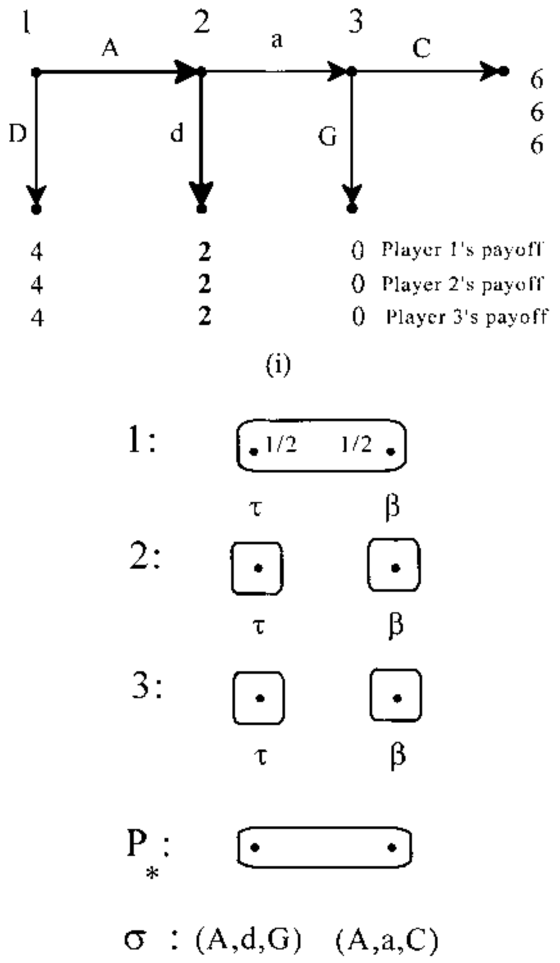
DEFINITION 3.16: *let G be a normal-form game. A correlated equilibrium distribution is a probability distribution p over the set S of strategy profiles such that, for every player i and every function* $d_i : S_i \to S_i$

† 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).

‡ Stalnaker (1994: p. 64, theorem 4) incorrectly makes this claim.

§ This example is due to Stalnaker (1996, pers. comm.).

¶ However, in perfect information games like the Centipede (see Section 4), which has a unique Nash equilibrium outcome in every subgame, strong rationalizability implies the Nash (and subgame perfect) equilibrium outcome (for a related result see Aumann, 1998*a*). Aumann and Brandenburger (1995) provide sufficient epistemic conditions for Nash equilibrium.



(i)

(ii)

FIGURE N.

$$\sum_{s\in S} u_i(s)p(s) \geq \sum_{s\in S} u_i(d_i(s_i), s_{-i})p(s) \tag{3.1}$$

EXAMPLE 3.17: consider the game of Figure P (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 (3.1) 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 (3.1)], d_U [defined by $d_U(x) = U$ for all x], d_D [defined by $d_D(x) = D$ for all x] and d_0 [defined by $d_0(U) = D$, $d_0(D) = U$]. With d_U the RHS of (3.1) is equal to $\frac{1}{2}5 + \frac{1}{2}0 = 2{\cdot}5$, with d_D it is equal to $\frac{1}{2}4 + \frac{1}{2}1 = 2{\cdot}5$, with d_0 it is equal to $\frac{1}{2}4 + \frac{1}{2}0 = 2$. Thus (3.1) is satisfied for player 1. Similar

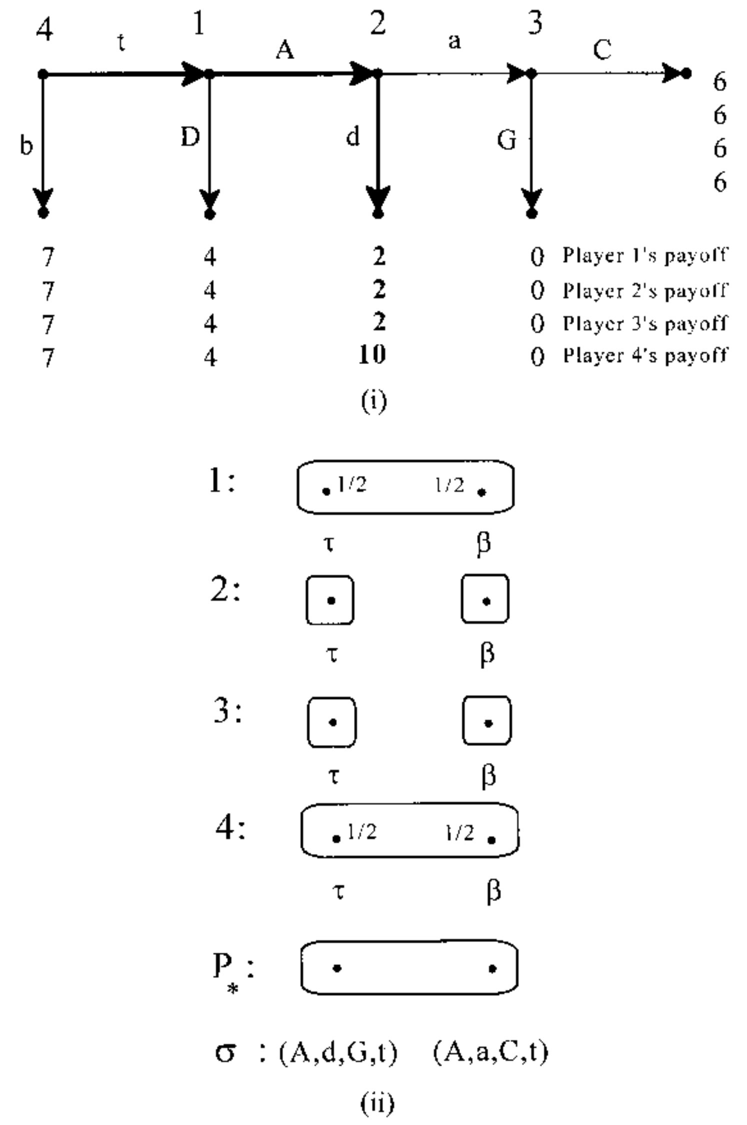
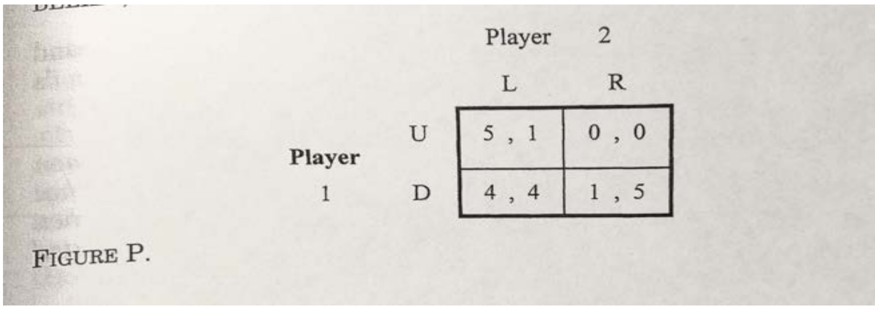


FIGURE O.

calculations show that (3.1) is also satisfied for player 2. Thus $p(U, L) = p(D, R) = \frac{1}{2}$ is a correlated equilibrium distribution.

Every Nash equilibrium is a correlated equilibrium.† Furthermore, every convex combination of Nash equilibria is also a correlated equilibrium. In a two-person zero-sum game all correlated

† For example, if s is a pure-strategy Nash equilibrium, take p such that $p(s) = 1$.



		Player 2	
		L	R
Player 1	U	5 , 1	0 , 0
	D	4 , 4	1 , 5

FIGURE P.

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 set of Nash equilibria.

One interpretation of the correlated equilibrium concept is that a correlated equilibrium distribution is the outcome of a Nash equilibrium of an expanded game with asymmetric information where each player privately observes a randomly generated, payoff-irrelevant signal before choosing her action. Correlation between the signals to different players induces (spurious) correlation between the players' actions. In other words, the players use a correlation device and a self-enforcing choice rule to co-ordinate their actions. But Aumann (1987) put forward another interpretation of correlated equilibrium "as an expression of Bayesian rationality". His interpretation relies on 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 ω. 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 $\sigma = (\sigma_1, \dots, \sigma_n)$. Player i is *rational at state* α if the strategy he chooses at α maximizes his expected utility calculated on the basis of his "posterior" beliefs $p_i(\cdot|H_i(\alpha))$:†

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

Note that here the states ω represent possible worlds, not the outcomes of a correlation device. According to this interpretation, there is *no ex ante stage* where the players contemplate which

† Defined by $p_i(\omega|H_i(\alpha)) = \frac{p^i(\omega)}{p^i(H_i(\alpha))}$ [where $p^i(H_i(\alpha)) = \Sigma_{x\in H_i(\alpha)} p^i(x)$] if $\omega \in H_i(\alpha)$ and $p_i(\omega|H_i(\alpha)) = 0$ otherwise.

signals they could receive and how they should react to them, and the "prior" p_i is simply a notational device to describe player i's beliefs at each possible world.

PROPOSITION 3.18: *(Aumann, 1987) if the players have a common prior p (i.e., if there is a probability measure p on Ω such that $p_1 = \ldots = p_n = p$) 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 of a game given in definition 3.5. The extra assumptions that Aumann introduces are: (1) that the possibility correspondences give rise to partitions and (2) that the "posterior" beliefs of the players are *Harsanyi consistent*, in the sense that they are derived from a common prior.† 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., 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, 1998; Dekel & Gul, 1997; Lipman, 1995) have pointed out that the meaning of a common prior in situations where there is no *ex ante* stage is highly problematic. This skepticism can be developed along the following lines. As shown in Section 3.2, the description of the "actual world" in terms of belief hierarchies generates a collection of "possible worlds" (combinations of external states and infinite hierarchies of beliefs), one of which is the actual world. This set of possible worlds, or states, gives rise to an epistemic model with type partitions for each individual. Thus—as Harsanyi (1967–68) noticed—there is a formal similarity between situations where the primitives are the individuals' belief hierarchies 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 n-tuple of infinite hierarchies of beliefs" (Gul, 1998: p. 924). As a result, notions such as that of a common prior, "seem to be

† Let us emphasize once again that the prior beliefs p^i of player i postulated by Aumann play no role: only the posterior beliefs $p_i(\cdot|H_i(\omega))$ are relevant. Indeed, given a model of a game according to definition 3.5, one can obtain a "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 $\{p_{i,\omega} : \omega \in \Omega\}$.

based on giving the artificially constructed states more meaning than they have" (Dekel & Gul, 1997: p. 115). Thus an essential step in providing a justification for correlated equilibrium in such situations 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 beliefs" (Dekel & Gul, 1997: p. 116).†

An interpretation of the desired kind of the common prior assumption in situation where there is no *ex ante* stage was provided recently (Bonanno & Nehring, 1996*a*; see also Feinberg, 1995; and Samet, 1996*b*, 1998) in terms of a generalized notion of absence of agreeing to disagree à la Aumann (1976), called consistency of expectations.

DEFINITION 3.19: *at state α there is Consistency of Expectations if there do* not *exist random variables $Y_i : \Omega \to \Re$ $(i \in N)$ such that:*

(1) $\forall \omega \in \Omega$, $\Sigma_{i \in N} Y_i(\omega) = 0$, 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 $\|E_i > 0\| = \{\omega \in \Omega : \Sigma_{\omega' \in \Omega} Y_i(\omega') p_{i,\omega}(\omega') > 0\}$.

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

DEFINITION 3.20: *for every $\mu \in \Delta(\Omega)$, let $\mathbf{HQC}_\mu$ (for Harsanyi Quasi Consistency with respect to the "prior" μ) be the following event: $\alpha \in \mathbf{HQC}_\mu$ if and only if*

(1) $\forall i \in N, \forall \omega, \omega' \in P_*(\alpha)$, if $\mu(\|p_i = p_{i,\omega}\|) > 0$ then $p_{i,\omega}(\omega') = \frac{\mu(\omega')}{\mu(\|p_i = p_{i,\omega}\|)}$ 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 $\|p_i = p_{i,\omega}\|$),‡ and

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

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

PROPOSITION 3.21:§ *at α, Consistency of Expectations is satisfied if and only if $\alpha \in \mathbf{HQC}$.*

Harsanyi Quasi Consistency may seem weaker than expected in that condition (2) of its definition only requires the derived

† For a defense of the common prior assumption see Aumann (1998*b*).

‡ Where, for every event E, $\mu(E) = \Sigma_{\omega \in E} \mu(\omega)$. Note that, for every $\omega \in \Omega$ and $i \in N$, $\omega \in \|p_i = p_{i,\omega}\|$. Thus $\mu(\omega) > 0$ implies $\mu(\|p_i = p_{i,\omega}\|) > 0$.

§ For a proof see Bonanno and Nehring (1996*a*). See also Feinberg (1995), Morris (1994) and Samet (1996*a*).

common prior to assign positive probability to some commonly possible state but allows the state representing the actual beliefs to be assigned zero "prior" probability. However, as illustrated in the example of Figure Q, Consistency of Expectations (and No Trade-type arguments) cannot deliver more.

In this example, at state τ individual 1 wrongly believes that it is common belief that the earth is flat, while individual 2 correctly believes that the earth is not flat and knows 1's incorrect beliefs. Expectation consistency is satisfied at τ (as well as at β). In fact, let Y_1 and Y_2 be random variables on $\{\tau, \beta\}$ 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 Consistency of Expectations is necessarily satisfied at τ. By proposition 3.21 there must be a μ such that $\tau \in \mathbf{HQC}_\mu$. 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 R(i) and R(ii) show a two-person zero-sum game with a unique correlated equilibrium (B,R), and an epistemic model of it.

In this example, at state τ (i) the players' beliefs satisfy Harsanyi Quasi Consistency ($\tau \in \mathbf{HQC}_\mu = \Omega$ 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 ($\tau \in \mathbf{T}$). Yet at τ the players play (T,L) which is not a correlated equilibrium strategy profile [no correlated equilibrium distribution assigns positive probability to (T,L)]. 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 state τ are "improbable" and constitute a

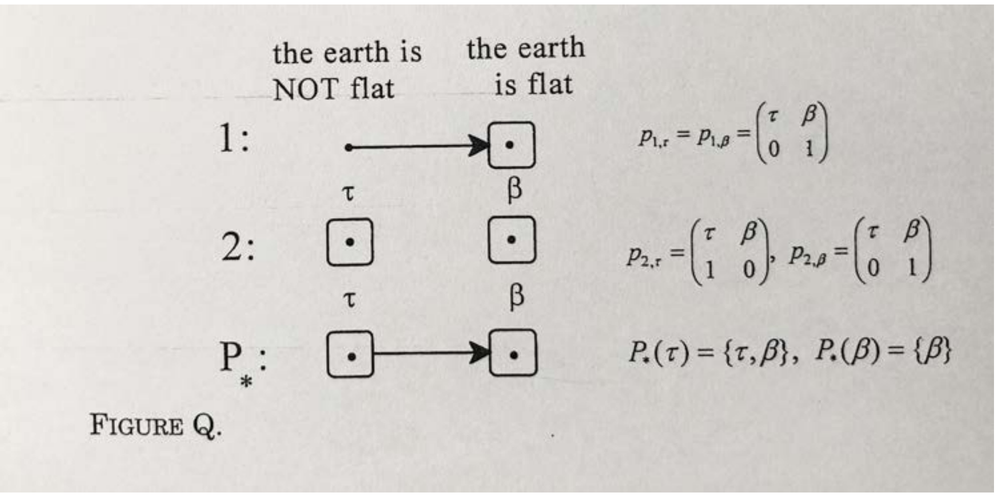
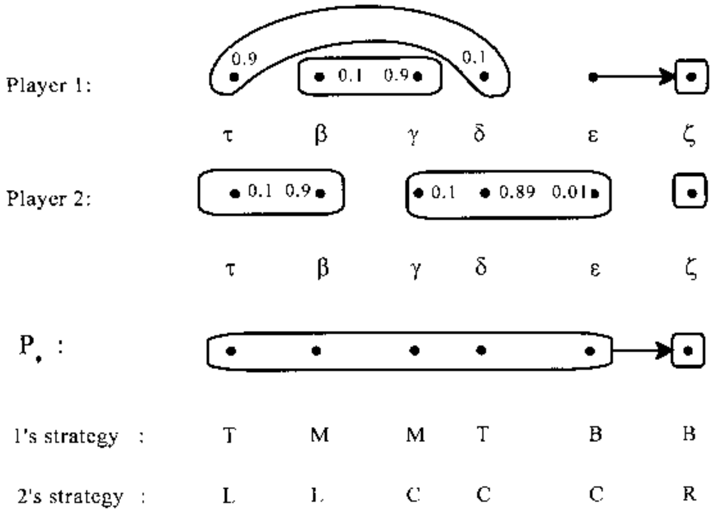

the earth is NOT flat
the earth is flat
1:
τ
β
2:
τ
β
P*:
$p_{1,\tau} = p_{1,\beta} = \begin{pmatrix} \tau & \beta \\ 0 & 1 \end{pmatrix}$
$p_{2,\tau} = \begin{pmatrix} \tau & \beta \\ 1 & 0 \end{pmatrix}$, $p_{2,\beta} = \begin{pmatrix} \tau & \beta \\ 0 & 1 \end{pmatrix}$
$P_*(\tau) = \{\tau, \beta\}$, $P_*(\beta) = \{\beta\}$

FIGURE Q.

		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

(i)



(ii)

FIGURE R.

null event. Indeed the actual beliefs of all players assign positive probability to τ.

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 falsely believes at ε that there is agreement that the true state is ζ. Hence $\mathbf{T}_{CB}$ is violated at ε, and therefore $B_*\mathbf{T}_{CB}$ at τ (the event $\mathbf{T}_{CB}$ was defined in Section 2.4).

Indeed—in the absence of collectively false beliefs—$B_*\mathbf{T}_{CB}$ 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 local character of our analysis, we call a strategy profile a *correlated equilibrium strategy profile* if it is played with positive probability in some correlated equilibrium (in the ordinary sense). Let **CE** be the event that (i.e., the set of states at which) a correlated equilibrium strategy profile is played.

PROPOSITION 3.22: *(Bonanno & Nehring, 1998b) fix an arbitrary finite normal-form game G and an arbitrary model of it. Then*

$$\mathbf{T}^* \cap B_*\mathbf{T}_{CB} \cap \mathbf{HQC} \cap B_*\mathbf{RAT} \subseteq \mathbf{CE}.$$

That is, if τ is a state where: (1) what is actually commonly believed is true and there is common belief in Truth about common belief, (2) Harsanyi Quasi Consistency of beliefs is satisfied and (3) there is common belief in rationality, then the strategy profile associated with τ (i.e., the strategy profile actually played) is a correlated equilibrium strategy profile. On the other hand, as the example of Figure R shows, if (2) and (3) are satisfied and instead of $\tau \in \mathbf{T}^* \cap B_*\mathbf{T}_{CB}$ one assumes $\tau \in \mathbf{T}$ then the strategy profile associated with τ need not be a correlated equilibrium.

REMARK 11: if the condition $\mathbf{T}^* \cap B_*\mathbf{T}_{CB}$ is weakened to $\mathbf{NI}^*$ (or, equivalently—cf. proposition 2.5—$\mathbf{T}_{CB} \cap B_*\mathbf{T}_{CB}$) then the conclusion is that it is common belief that a correlated equilibrium is played: $\mathbf{NI}^* \cap \mathbf{HQC} \cap B_*\mathbf{RAT} \subseteq B_*\mathbf{CE}$.

A converse to proposition 3.22 is given by the following result.

PROPOSITION 3.23: *let G be a game and $p \in \Delta(S)$ a correlated equilibrium distribution. Then there exists a model $\mathcal{M}$ of G, a probability measure $\mu \in \Delta(\Omega)$ and a state τ such that*

(1) $\tau \in \mathbf{T}^* \cap B_*\mathbf{T}_{CB} \cap \mathbf{HQC}_\mu \cap B_*\mathbf{RAT}$,

(2) the distribution over strategy profiles induced by μ restricted to $\{\tau\} \cup P_*(\tau)$ coincides with p and

(3) $\mu(\tau) > 0$ (so that the strategy profile actually played is in the support of p).

4. Epistemic foundations of solution concepts: (B) extensive-form games

The theory of extensive form (dynamic) games is more complex and more controversial than the theory of strategic form (static) games and until recently the epistemic foundations of extensive form solution concepts were not well understood. The fundamental reason of these difficulties is that a crucial ingredient of the theory is modelling how players would behave and what they would believe immediately after every (partial) history of play, including those that are inconsistent with the players' initial beliefs and/or with the theory. Game theorists agree that "static" solution concepts are too weak when applied to the strategic form of dynamic games, because they do not take into account that each player anticipates that her opponents would react rationally to whatever information they receive. But several "refinements" of strategic form solution concepts have been proposed without a clear understanding of their epistemic underpinnings. In fact, the static epistemic models presented in Section 3 do not have sufficient expressive power to represent the subtleties of the theory of dynamic games, because they cannot represent the conditional beliefs of the players and hence their counterfactual reasoning.†

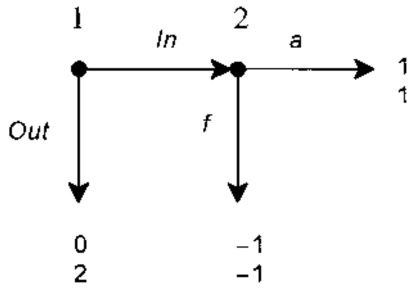
In the last few years our understanding of the foundations of the theory of extensive form games has been substantially improved by the development of adequate epistemic models, which make it possible to provide sufficient epistemic conditions and even full epistemic characterizations of some extensive form solution concepts. In this section we present some of these results within a common framework. Section 4.1 provides an informal discussion of the main issues using a few simple examples. The arguments presented here are formalized later. Section 4.2 introduces extensive form epistemic models. Since all the examples we are concerned with are multistage games with observed actions and (possibly) simultaneous moves in some stages (i.e., games with almost perfect information), we restrict our analysis to this class of games. This facilitates the discussion of interactive beliefs as the play unfolds. Section 4.3 presents the notion of conditional common belief in sequential rationality and a characterization of a weak notion of extensive form rationalizability. The methodology and concepts here are still quite similar to those developed for strategic form games. Section 4.4 features a radical departure from strategic form analysis in order to explore the epistemic foundations of solution concepts relying on a "forward induction" principle. Section 4.5 analyses epistemic independence and backward induction.

4.1. INTRODUCTORY EXAMPLES

4.1.1. The Entry game

The simplest example illustrating the differences between strategic form and extensive form analysis can be found in any recent textbook on game theory. The usual story is that there is a monopolistic market and player 1 is a potential entrant, while player 2 is the incumbent monopolist. Player 2 may fight

† However, static epistemic models can be used to provide an epistemic characterization of a weak extensive form refinement of the rationalizability solution concept (cf. Section 4.3).



(i)

		Player 2	
		a	f
Player	In	1 , 1	−1, −1
1	Out	0 , 2	0 , 2

(ii)

FIGURE S.

the entrant (f) or acquiesce (a) and split the market. The game is depicted in Figure S(i), while Figure S(ii) shows its strategic form.

The game has two Nash equilibria in pure strategies, (In, a) and (Out, f), but only the first one is "plausible". There is no disagreement about the right way to play this game: if the players understand the game (complete information) and are rational, and if player 1 believes that player 2 is rational, then they play (In, a), because the potential entrant anticipates that the incumbent would react optimally to entry. This is the so-called "backward induction" logic. Here we only want to emphasize that the standard argument used to deem (Out, f) "implausible" or "wrong" implicitly relies on the possibility of assigning a truth value to a subjunctive conditional. According to this argument, in equilibrium (Out, f) player 1 believes with positive probability that the statement "if I enter, player 2 fights" is true while he should be certain that the statement is false. But since player 1 stays out, the statement is counterfactual. If the statement were interpreted as the material implication "either I stay out, or player 2 fights" it would be true, because player 1 is actually staying out, and player 1 should be certain that the statement is true.

The backward induction logic used to solve the Entry game is uncontroversial in all two-stage games where (i) each active player has a dominant action in the second stage (which may depend on the outcome of the first stage) and (ii) anticipating a second-stage dominant choice yields a unique rational choice in the first stage. Stackelberg games and the twice repeated Prisoners' Dilemma are well-known examples. Intuitively, common belief—at the beginning of the game—in sequential rationality (conditional expected utility maximization) yields the backward induction solution. All the extensive form solution concepts considered in this section are consistent with initial common belief in sequential rationality and hence agree with the backward induction logic in such games. But in more complex games the equilibrium refinement capturing the backward induction logic, subgame perfection, becomes more problematic. On the one hand, there are backward-induction-solvable games with more than two stages where initial common belief in sequential rationality seems to be consistent with subgame imperfect outcomes and assuming common belief in sequential rationality at later stages is problematic. The finitely Repeated Prisoners' Dilemma, the Chainstore game and the Centipede game† (discussed below) are well-known examples. On the other hand, there are games, such as the Battle of the Sexes (BoS) with an Outside Option, where some subgame perfect equilibrium outcomes are inconsistent with the "forward induction" assumption that each player tries to "rationalize" the observed behaviour of her opponents.

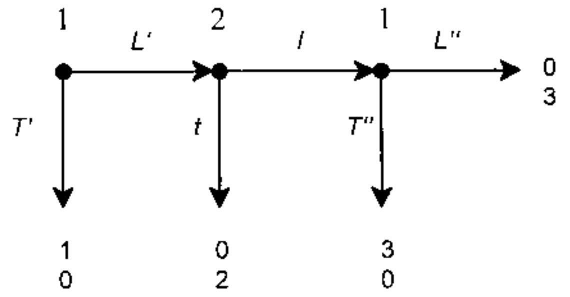
4.1.2. The Centipede

Figure T depicts a version of the Centipede game with its strategic form.† This is a multistage game of length K with alternating moves. At the kth stage of the game there are k dollars on the table, the active player can either take them and terminate the game or leave them on the table. In the second case one dollar is added on the table. The game is interrupted after K stages with the active player either taking K dollars or leaving them to the opponent.

Like the Repeated Prisoners' Dilemma, this game has a unique Nash equilibrium outcome and a unique subgame perfect equilibrium (backward induction) strategy profile: at each stage the active player is supposed to take the dollars. But there are other outcomes consistent with initial common belief in conditional

† The Chainstore game is a finite repetition of the Entry game where the incumbent sequentially faces different potential entrants in different markets (Selten, 1978). The Centipede game was first introduced by Rosenthal (1981) to discuss the "paradoxical" implications of backward induction in the Chainstore and other games.

† This is Reny's "Take-it-or-Leave-it" game (see, for example, Reny, 1985, 1995).



(i)

		Player 2	
		l	*t*
Player 1	*T'T''*	1 , 0	1 , 0
	T'L''	1 , 0	1 , 0
	L'T''	3 , 0	0 , 2
	L'L''	0 , 3	0 , 2

(ii)

FIGURE T.

expected utility maximization. Suppose that player 2 is initially certain that player 1 will apply the backward induction logic and take one dollar, but player 1 leaves it. Then player 2 will be surprised and may believe that player 1 is probably irrational and would leave him three dollars if given the opportunity. Given such beliefs, after action L', player 2 would rationally leave two dollars on the table. Suppose that player 1 is rational and assigns a sufficiently high probability to the event that player 2 is rational and would have such beliefs after L'. Then player 1 initially leaves one dollar (correctly) hoping to get three dollars later. In the situation just described, (0) each player is a conditional expected utility maximizer, (1) each player is initially certain of (0), (2) each player is initially certain of (1), and so on.

Can we obtain the backward induction outcome by assuming that there would be common belief in sequential rationality at later stages? No. Common belief in sequential rationality after action L' is impossible: if player 2 believes that player 1 is rational, he takes two dollars. If player 1 initially believes that player 2 is rational and that player 2 would believe that player 1 is rational after L', player 1 takes one dollar immediately. Player 1's beliefs about player 2 do not change after her own initial action. This leaves us with only two possibilities: either (a) player 1 is irrational and chooses L', or (b) player 1 is rational, chooses T', but would believe, if she chose L', that player 2 is rational and would believe that player 1 is rational. Since these two events are mutually exclusive, player 2 could not believe both after observing L'.

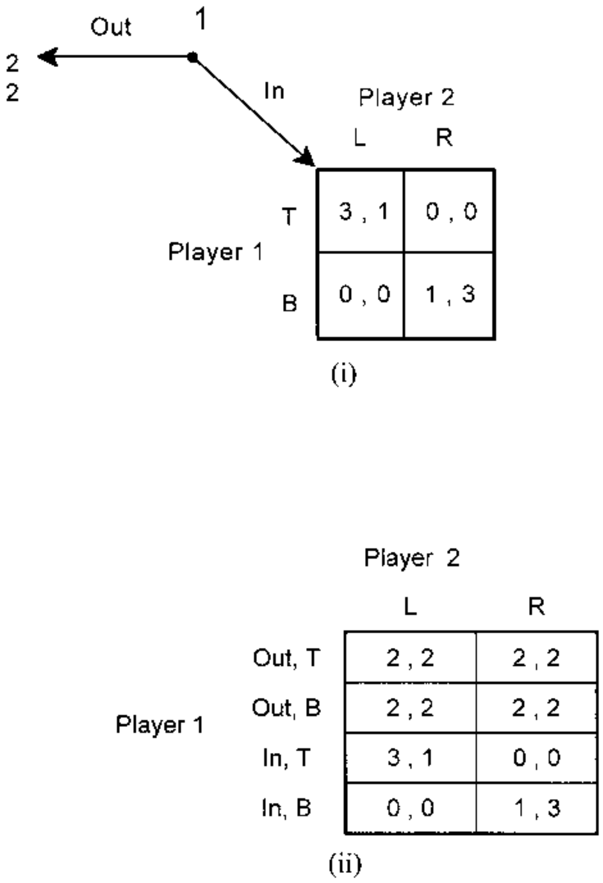
We conclude that, although the backward induction logic may seem as compelling as the rationalizability logic in static games, it cannot be justified by a straightforward extension of the epistemic assumptions characterizing normal form rationalizability.

4.1.3. The Battle of the Sexes (BoS) with an outside option

Consider the game depicted in Figure U. Player 1 (by convention a female) initially decides whether to play the BoS subgame (In) or take an outside option (Out) that yields an intermediate payoff. This game has two subgame perfect equilibrium outcomes, Out and $(In, (T, L))$,† but it is often argued that only the second is "reasonable". In fact, only the second equilibrium is consistent with the following assumptions: (0) all players are rational, (1) all players believe (0) whenever possible, (1) all players believe (0)&(1) whenever possible. By (0) player 1 does not play the strictly dominated strategy (In, B). On the other hand, strategy (In, T) can be rationalized by some beliefs. Therefore (0) and (1) imply that, if player 2 observed In, he would believe that player 1 is playing (In, T) and hence would respond with L.

Note that the same assumptions yield the backward induction outcome in the Centipede game of Figure T. In both cases the solution induced by these assumptions can be obtained by iteratively deleting weakly dominated strategies. In general, we may consider a longer list of assumptions where assumption (k + 1) is: all players believe (0)&(1)&. . . &(k) whenever possible. We will see that these assumptions are captured by a notion of extensive form rationalizability which is quite similar to the iterative (maximal) deletion of weakly dominated strategies. These assumptions correspond to a "forward induction" logic: each player always tries to "rationalize" the observed behaviour of her opponents, looking for its "most sophisticated" explanation.

† Out is supported by two equilibria in behavioural strategies: $[(Out, B), R]$ and $[(Out, \frac{3}{4}T + \frac{1}{4}R), (\frac{1}{4}L + \frac{3}{4}R)]$.



(i)

Player 1	Player 2: L	R
Out, T	2 , 2	2 , 2
Out, B	2 , 2	2 , 2
In, T	3 , 1	0 , 0
In, B	0 , 0	1 , 3

(ii)

FIGURE U.

4.2. EPISTEMIC MODELS FOR EXTENSIVE FORM GAMES

A finite *multistage game with observed actions* is a tuple

$$\Gamma = \langle N, \{A_i\}_{i \in N}, H, Z, \{u_i\}_{i \in N} \rangle$$

where $N = \{1, 2, \ldots, n\}$ is a set of players, A_i is a non-empty finite set of a priori feasible actions for player i, H is a non-empty finite set of *partial histories* of play, Z is a non-empty finite set of *complete (or terminal) histories* of play and $u_i : Z \to \Re$ is player i's payoff function. A history of length k is a finite sequence of action profiles

$$h = \left(\left(a_1^1, \ldots, a_n^1\right), \ldots, \left(a_1^k, \ldots, a_n^k\right) \right) \in A^k,$$

where $A = A_1 \times \cdots \times A_n$.† As usual, A_{-i} denotes the set of action profiles of all players other than i. For notational convenience the empty sequence—denoted by ϕ—is also regarded as a history preceding every other history and representing the beginning of the game. The set of (partial and complete) histories is naturally ordered by the relation "initial (proper) subhistory of", that is, $h = (a^1, \dots, a^k)$ *precedes* $h' = (b^1, \dots, b^m)$ if and only if $k < m$ and $(a^1, \dots, a^k) = (b^1, \dots, b^k)$. The set of histories with this partial order is a tree, with root given by the empty history ϕ.‡

For every sequence $h = (a^1, \dots, a^k) \in A^k$ and action profile $a \in A$ we denote the concatenation of h and a by $(h, a) = (a^1, \dots, a^k, a)$. The set of feasible actions for player i immediately after partial history $h \in H$ is

$$A_i(h) = \{a_i \in A_i : \exists a_{-i} \in A_{-i}, (h, (a_i, a_{-i})) \in H \cup Z\}.$$

The set of feasible action profiles immediately after $h \in H$ is $A_1(h) \times \cdots \times A_n(h)$. The set of (pure) strategies for player i is $S_i \subseteq (A_i)^H$, where $s_i \in S_i$ if and only if $s_i(h) \in A_i(h)$ for all $h \in H$.§ The complete history induced by a strategy profile $s \in S$ is denoted by $\zeta(s)$.¶ Thus $U_i = u_i \circ \zeta : S \to \Re$ is player i's strategic form payoff function. The strategic form of game Γ is $G_\Gamma = \langle N, \{S_i\}_{i \in N}, \{U_i\}_{i \in N} \rangle$.

For the sake of notational simplicity we are assuming that each player takes an action at each stage, but the assumption is completely innocuous, because the set of feasible actions of a player may be a singleton. Thus we are able to represent any combination of sequential and simultaneous moves. We say player i is *active* at $h \in H$ if $A_i(h)$ has at least two elements. Game Γ has *perfect information* if there is at most one active player at each history. Γ is *static*, or simultaneous, if $H = \{\phi\}$. Γ is *generic* if for all players $i \in N$ and terminal histories $z' \neq z''$, $u_i(z') \neq u_i(z'')$.

† See Fudenberg and Tirole (1991), Section 3.2.1 and Osborne and Rubinstein (1994), Sections 6.1.1 and 6.3.2.

‡ The sets of partial and complete histories have the following (quite obvious) properties: for all $h, h' \in \{\phi\} \cup (\cup_{k \geq 1} A^k)$, $a_i, b_i \in A_i$, $a_{-i}, b_{-i} \in A_{-i}$,

- if $h' \in H \cup Z$ and h precedes h', then $h \in H \backslash Z$,
- if $h \in H$, then h precedes some complete history $z \in Z$,
- if $(h, (a_i, a_{-i})) \in H \cup Z$ and $(h, (b_i, b_{-i}))) \in H \cup Z$, then $(h, (b_i, a_{-i})) \in H \cup Z$.

The first two properties imply that $H \cap Z = \emptyset$ and that the set $H \cup Z$ ordered by the natural precedence relation "initial subhistory of" is a tree with root ϕ where H is the set of non-terminal nodes and Z is the set of terminal nodes. The third property says that for every partial history the set of feasible action profiles is a Cartesian product of its projections on the action sets A_i, $i \in N$.

§ Recall that Y^X is the set of all functions with domain X and range Y.

¶ Thus, $\zeta(s) = (a^1, \dots, a^m)$ if and only if $s_i(a^1, \dots, a^k) = a_i^{k+1}$ for all $i \in N, k < m$.

[Note that to represent a game with observed actions and simultaneous moves in some stages with the standard graph-theoretic definition (e.g., Kreps & Wilson, 1982) we would have to eliminate inactive players from each stage and introduce (i) an artificial order among simultaneous moves and (ii) appropriate information sets. This would make the discussion of interactive knowledge and beliefs as the play unfolds more cumbersome and complex. In particular, we implicitly assume that a player receives information about past behaviour even when she is inactive, while the standard formulation of extensive form games represents only the information a player receives when she is active and implicitly assumes that no information is received otherwise. Thus the standard formulation does not allow a synchronous representation of interactive knowledge and beliefs.†]

Fix a multistage game with observed actions Γ. We may obtain an epistemic model of Γ simply by providing a model of (or type space for) its strategic form G_Γ. Such a model specifies the initial beliefs and the strategy of each player at each possible world. Although the formalism is the same as in Section 3, the most natural game-theoretic interpretation here is different. Since a strategy is a contingent plan of action, the model specifies, for each state of the world, the actions taken along the actual path but also the action that player i *would take* at each partial history h *if* history h occurred. In other words, we interpret the statement "if h occurred, i would take action a_i" as a subjunctive conditional. A strategy is a combination of subjunctive conditionals and we assume that the subjunctive conditional "if h occurred, i would take action a_i" is verified at state ω if and only if player i's strategy at ω is a function s_i such that $s_i(h) = a_i$, independently of whether history h occurs at state ω or not.‡

For the same reason, the most natural notion of rationality in an extensive form context is more demanding than in a normal form context. Intuitively, player i is rational if for every history h her continuation strategy at h maximizes her *conditional* expected payoff given h (see, for example, Kreps & Wilson, 1982). However, i's conditional beliefs given h are (implicitly) specified at a state ω of a model for G_Γ only if i's beliefs at ω assign positive probability to the event that h occurs. In this case, conditional beliefs given h can be derived via Bayes rule; otherwise, the model is silent about such beliefs. Thus it seems that in order to make sense of

† Battigalli and Bonanno (1997*b*) show how to extend the information structure of extensive form games (inheriting the perfect recall property) so that a player receives information at every node, including those owned by other players.

‡ Partial history h occurs at ω if the strategy profile at ω is s and h precedes $\zeta(s)$. Note that even the definition of rationality for static models implicitly relies on subjunctive conditionals, because a player compares the expected consequences of her actual decision with the consequences that (in her opinion) *would* occur *if* she chose a different action.

the notion of rationality in an extensive game framework we have to enrich the model by specifying, for each state of the world, a player's conditional beliefs for any given partial history h. In other words, a state of the world should describe, not only the action that player i would take at h, but also the conditional beliefs that i would have at h.

In order to describe conditional beliefs we first define the concept of conditional probability system (Rênyi, 1955; Myerson, 1986). Let Ω be a finite set and let $\mathcal{H}$ be a collection of non-empty subsets of Ω.† The set of all functions assigning to each element of $\mathcal{H}$ (a subset) a probability measure on Ω is $[\Delta(\Omega)]^{\mathcal{H}}$. For every such function $\mu \in [\Delta(\Omega)]^{\mathcal{H}}$, we write $\mu(\cdot|B)$ for the probability measure associated to subset $B \in \mathcal{H}$ and we interpret $\mu(E|B)$ as the probability of E given B. When convenient we write μ as a vector: $\mu = (\mu(\cdot|B))_{B\in\mathcal{H}}$. A *conditional probability system* on $\langle\Omega, \mathcal{H}\rangle$ is a function $\mu \in [\Delta(\Omega)]^{\mathcal{H}}$ such that for all $E \subseteq \Omega, B, C \in \mathcal{H}$,

(1) $\mu(B|B) = 1$,

(2) if $E \subseteq B \subseteq C$, then $\mu(E|C) = \mu(E|B)\mu(B|C)$.

Condition (1) is obvious. Condition (2) says that the usual rule to compute conditional probabilities applies whenever possible. The set of conditional probability systems on $\langle\Omega, \mathcal{H}\rangle$ is denoted $\Delta^{\mathcal{H}}(\Omega)$. Two special cases are worthy of attention. (i) Suppose that the set of states is a product $\Omega = S \times T$ and let $\mathcal{H}$ be a collection of non-empty subsets of S. Then we obtain a corresponding collection $\mathcal{H}_{S\times T}$ of subsets of Ω, that is,

$$\mathcal{H}_{S\times T} = \left\{B = S' \times T : S' \in \mathcal{H}\right\}.$$

The interpretation is that only the s-co-ordinate of the state (s, t) is (partially) observable and $\mathcal{H}_{S\times T}$ is a collection of potentially observable events. In this case, with a slight abuse of notation, we write $\Delta^{\mathcal{H}}(S \times T)$ for the set of conditional systems. (ii) If $\mathcal{H}$ is the collection of all non-empty subsets of Ω ($\mathcal{H} = 2^{\Omega}\setminus\{\emptyset\}$), then a conditional probability system on $\langle\Omega, \mathcal{H}\rangle$ is called *complete*. A complete conditional probability system represents a belief revision rule that can be applied for any information structure on Ω.‡ The set of complete conditional probability systems is denoted $\Delta^*(\Omega)$. We will focus mainly on case (i).

4.2.1. Type spaces for extensive form games

When the players reason about their best course of action at any point of the game they form beliefs about what their opponents

† $\mathcal{H}$ may, but need not be the family of events corresponding to the occurrence of partial histories in Γ.

‡ For references on belief revision see Gärdenfors (1988). See also Stalnaker (1996, 1998) and Brandenburger (1997).

would do and what they would believe immediately after some different histories of play, if such histories occurred. Similarly, as "external observers", or "theorists", we would like to be able to specify, for every possible world, what the players would do and what they would believe about each other's behaviour and beliefs at each partial history of the game. Extensive form type spaces are epistemic models with such expressive power.

Let S be the set of strategy profiles of game Γ and, for every $h \in H$, let $S(h)$ denote the set of profiles inducing, or "reaching," h, that is,

$$S(h) = \left\{ s \in S : h \text{ precedes } \zeta(s) \right\}.$$

Clearly, $S(h) = S_1(h) \times \cdots \times S_n(h)$, where $S_i(h)$ is the projection of $S(h)$ on S_i.† The collection of subsets

$$\mathcal{H} = \{S(h) : h \in H\}$$

represents a family of commonly observable events about players' behaviour. Note that $\mathcal{H}$ has a special structure, which reflects the tree-structure of $H : S = S(\phi) \in \mathcal{H}$ and for all S', $S'' \in \mathcal{H}$, either $S' \subseteq S''$, or $S' \supseteq S''$ or $S' \cap S'' = \emptyset$.

DEFINITION 4.1: *a finite type space for* Γ *is a tuple* $\mathcal{T} = \left\langle N, S, \mathcal{H}, \{T_i\}_{i \in N}, \{\theta_i\}_{i \in N} \right\rangle$ *where, for every* $i \in N$, T_i *is a finite set and* θ_i *is a function* $\theta_i : T_i \to \Delta^{\mathcal{H}}(S \times T_{-i})$ $(T_{-i} = \Pi_{j \neq i} T_j)$.‡

Note that definition 4.1 extends the notion of type space given in Section 3.2. The strategic form definition is obtained as a special case if Γ is a static game. Furthermore, since $S = S(\phi) \in \mathcal{H}$, the model specifies the actual *initial* beliefs of each player at the actual state: if i's type in a given state is $t \in T_i$, her initial beliefs at this state are given by the probability measure $\theta_{i,t}(\cdot | S \times T_{-i})$. Therefore an extensive form type space contains a type space in the usual sense and we can define the belief operators $B_i (i \in N)$ and the common belief operator B_* as before. But in this extended model we can do more: for every $i \in N$ and $h \in H$, we can define a belief operator $B_{i,h}$, where $B_{i,h}E$ is the event that player i would

† In games with imperfectly observed actions, perfect recall implies that, if I is an information set of player i, then $S(I) = S_i(I) \times S_{-i}(I)$, where $S(I)$ is the set of strategy profiles inducing a path through information set I.

‡ Infinite type spaces are similarly defined, but one has to take care of measure-theoretic issues. The extension of the concept to general extensive form games and dynamic games of incomplete information is straightforward.

Finite extensive form type spaces have been originally introduced by Ben Porath (1997) (the working paper version is dated 1992). The definition given here is due to Battigalli and Siniscalchi (1998*a*), who elaborate on the concept. In particular, they analyse type-morphisms between general (possibly) infinite spaces and construct a universal extensive form type space.

believe E immediately after history h, if h occurred. Since each history is commonly observed, it makes sense to define common belief operators $B_{*,h}(h \in H)$, where $B_{*,h}E$ is the event that, if h occurred, E would be commonly believed. Formally, for every event $E \subseteq S \times T_1 \times \cdots \times T_n$ and player $i \in N$ we define†

$$B_{i,h}E = \left\{(s, t, t_{-i}) : \theta_{i,t}(E_t|S(h) \times T_{-i}) = 1\right\},$$

$$B_{e,h}E = \bigcap_{i \in N} B_{i,h}E, B_{*,h}E = \bigcap_{k \geq 1} B^k_{e,h}E,$$

$$B_iE = B_{i,\phi}E, B_*E = B_{*,\phi}E.$$

It is easily checked that these belief operators have all the properties of the corresponding operators defined for Bayesian frames.‡ In particular, they satisfy consistency, conjunction, positive introspection and monotonicity, and the individual belief operators $B_{i,h}$ also satisfy negative introspection.

4.2.2. Hierarchies of conditional beliefs and universal type spaces

We know that a "static" epistemic model for a game with strategy space S implicitly specifies a profile of infinite hierarchies of beliefs for each state (see Section 3.2). Similarly, a type space for Γ implicitly specifies a profile of infinite hierarchies of *conditional* beliefs. The first-order conditional beliefs of type $t \in T_i$ are given by the conditional probability system

$$\mu^1_{i,t} = \left(mrg_S \theta_{i,t}(\cdot|S(h) \times T_{-i})\right)_{h \in H} \in \Delta^{\mathcal{H}}(S).$$

Given all the first-order mappings $t_j \longmapsto \mu^1_{j,t_j}(j \in N)$ we can define the second-order conditional beliefs of type $t \in T_i$: for all $h \in H$,

$$\mu^2_{i,t}\left((s, (\mu^1_j)_{j \neq i})|S(h) \times \left[\Delta^{\mathcal{H}}(S)\right]^{n-1}\right) =$$
$$\theta_{i,t}\left(\left\{(s, t_{-i}) : \forall j \neq i, \mu^1_{j,t_j} = \mu^1_j\right\}|S(h) \times T_{-i}\right).$$

Higher-order belief mappings can be obtained inductively. Thus, a hierarchy of conditional beliefs for a given player i specifies for

† Recall that, for each event E and type $t \in T_i$, E_t is event E from the point of view of type t, i.e., the set of elements of $S \times T_{-i}$ that are consistent with event E and type t (see Section 3.2).

‡ In fact, for each history $h \in H$, we can derive from $\mathcal{T}$ a type space $\mathcal{T}_h$ for $S(h)$: just take the belief functions $\theta_{i,h} : T_i \to \Delta(S(h) \times T_{-i})$, where $\theta_{i,h,t}(E) = \theta_{i,t}(E|S(h) \times T_{-i})$ for all $E \subseteq S(h) \times T_{-i}$. As we noticed in Section 3.2, a type space for $S(h)$ corresponds to a model of $S(h)$ in an obvious "beliefs preserving" way. The conditional belief operator $B_{i,h}$ corresponds to i's belief operator in this model.

each order k and each history h the beliefs that i would have at h about his opponents' contingent behaviour and lower-order conditional beliefs for any history h'.[†] As with "static" type spaces, at each state (s, t) the corresponding hierarchies of beliefs satisfy a natural "marginalization property" and assign zero probability to the states violating this property.

Any fixed type space contains only a small subset of the set of all conceivable hierarchies of conditional beliefs for game Γ. We have seen in the analysis of strategic form games how the "smallness", or incompleteness, of epistemic models prevents the formulation of simple "if and only if" characterizations of solution concepts like normal form rationalizability. The incompleteness of type spaces becomes more relevant in the context of extensive form games. In fact, if we want to formalize the assumption that a player would try to "rationalize" the observed actions of his opponents whenever possible (as in our discussion of the BoS with an Outside Option), we would like this player to be able to consider any conceivable profile of opponents' hierarchies of conditional beliefs. But in an incomplete type space this search for hierarchies that can rationalize some observed behaviour is artificially limited. Thus the possibility of constructing a universal type space is even more important for extensive form games. We present below the construction of a universal type space for Γ due to Battigalli and Siniscalchi (1998*a*), which is quite close to the standard construction of Section 3.2.[‡]

In analogy with the construction for the static case, we let $\Delta^{\mathcal{H}}(X^k)$ denote the set of all kth-order conditional beliefs (including the inconsistent ones) and Z^k denote the set of kth-order beliefs of a given player's opponents:

- $X^0 = S$,
- given X^{k-1} $(k \geq 1)$, $Z^k = [\Delta^{\mathcal{H}}(X^{k-1})]^{n-1}$, $X^k = X^{k-1} \times Z^k$.

[†] According to our definition of type space a player also has beliefs about her own behaviour. But we will assume that rational players are always certain of their behaviour.

[‡] In the construction we use a more general definition of conditional probability system on a possibly infinite space. Let Ω be a measurable space with a sigma-algebra of events $\mathcal{A}$ and let $\mathcal{H} \subseteq \mathcal{A}$. Since in our analysis the appropriate sigma-algebras are always understood, they are not explicit in our notation. A conditional probability system on $\langle \Omega, \mathcal{H} \rangle$ is a function $\mu \in \left[\Delta(\Omega)\right]^{\mathcal{H}}$ satisfying the same properties as in the finite case for all measurable subsets E and all $B, C \in \mathcal{H}$. We consider spaces of the form $\Omega = S \times T$ endowed with the information structure $\mathcal{H} \subseteq 2^S$, where S and $\mathcal{H}$ are derived from the finite game Γ, and T is some metrizable, separable and complete topological space endowed with the Borel sigma-algebra. With the usual slight abuse of notation we write $\Delta^{\mathcal{H}}(S \times T)$ for the set of conditional systems even if $\mathcal{H}$ is not a collection of subsets of $S \times T$. It can be shown that also $\Delta^{\mathcal{H}}(S \times T)$ is a metrizable, separable, complete topological space, given the (product) topology of weak convergence of measures.

Now we define the set $T_i^U \subset \Delta^{\mathcal{H}}(X^0) \times \Delta^{\mathcal{H}}(X^1) \times \cdots \times \Delta^{\mathcal{H}}(X^k) \times \cdots$ of "conceivable" hierarchies of beliefs, i.e., those which satisfy the appropriate consistency property:† for notational convenience we write $\mu_{i,h}^k$ for player i's beliefs conditional on the event corresponding to history h:

- $Y_{-i}^1 = X^1$,
- $Y_{-i}^k = \{(s, ((\mu_j^1)_{j \neq i}, \ldots, (\mu_j^{k-1})_{j \neq i}, (\mu_j^k)_{j \neq i}) \in X^{k-1} \times Z^k$: $\forall j \neq i, \forall h \in H, mrg_{X^{k-2}}\, \mu_{j,h}^k = \mu_{j,h}^{k-1}, \mu_{j,h}^k(Y^{k-1}) = 1\}$,
- $T_i^U = \{(\mu_i^1, \ldots, \mu_i^{k-1}, \mu_i^k, \ldots) \in \Pi_{k=0}^{\infty} \Delta^{\mathcal{H}}(X^k)$: $\forall k \geq 1, \forall h \in H, mrg_{X^{k-2}}\, \mu_{i,h}^k = \mu_{i,h}^{k-1}, \mu_{i,h}^k(Y_{-i}^{k-1}) = 1\}$.

The consistency property here says that the marginalization condition must be satisfied for every given history, and that all conditional beliefs must rule out other players' hierarchies violating this condition as well as the hierarchies not ruling out other players' hierarchies violating this condition, and so on. Let $T_{-i}^U = \Pi_{j \neq i} T_j^U$.

Proposition 4.2: *there is a "canonical homeomorphism" between T_i^U and $\Delta^{\mathcal{H}}(S \times T_{-i}^U)$, that is, a bijective and bicontinuous function $\theta_i^U : T_i^U \to \Delta^{\mathcal{H}}(S \times T_{-i}^U)$ such that for all $t = (\mu_i^1, \mu_i^2, \ldots) \in T_i^U$, $k \geq 1$ and $h \in H$,*

$$mrg_{X^{k-1}}\, \theta_{i,h,t}^U = \mu_{i,h}^k,$$

where $\theta_{i,h}^U$ is the h-co-ordinate function of θ_i^U.

In analogy with the static case, proposition 4.2 shows that

$$\mathcal{T}^U = \left\langle N, S, \mathcal{H}, \left\{T_i^U\right\}_{i \in N}, \left\{\theta_i^U\right\}_{i \in N} \right\rangle$$

is a type space for Γ. Type space $\mathcal{T}^U$ is universal in the following sense:

Proposition 4.3: *for every type space $\mathcal{T} = \langle N, S, \mathcal{H}, \{T_i\}_{i \in N}, \{\theta_i\}_{i \in N} \rangle$ there is a unique n-tuple of (measurable) functions $\varphi = (\varphi_i)_{i \in N}$, $\varphi_i : T_i \to T_i^U$, such that for all $i \in N$, $t \in T_i$, $E \subseteq S \times T_{-i}^U$ (measurable) and $h \in H$,*

$$\theta_{i,h,t}\left(\left\{(s', t'_{-i}) \in S \times T_{-i} : (s, \varphi_{-i}(t_{-i})) \in E\right\}\right) = \theta_{i,h,\varphi_i(t)}^U(E).$$

† Again, all the individuals are symmetric in this construction (as in the static case), but symmetry is a special feature due to symmetric information. In general extensive games each player has her own information structure $\mathcal{H}_i$ and this introduces an asymmetry in the construction.

4.2.3. State space models and belief revision

The following definition of a model for Γ is adapted from Stalnaker (1996)† and extends the notion of a model for a strategic form game.

DEFINITION 4.4: *a finite model for Γ is a tuple*

$$\left\langle N, \{S_i\}_{i\in N}, \{p_i\}_{i\in N}, \{\sigma_i\}_{i\in N}, \{q_i\}_{i\in N}\right\rangle$$

where $\left\langle N, \{S_i\}_{i\in N}, \{p_i\}_{i\in N}, \{\sigma_i\}_{i\in N}\right\rangle$ is a model for G_Γ with type partition $T_i = \left\{\|p_i = p_{i,\omega}\| : \omega \in \Omega\right\}$ (see definition 3.5) and, for each player i, $q_i : \Omega \to \cup_{Q\in T_i}\Delta^*(Q)$ is a function such that, for all $\alpha \in \Omega$, all $\omega \in \Omega$,

(1) $q_{i,\alpha} \in \Delta^*\left(\|p_i = p_{i,\alpha}\|\right)$ and $q_{i,\alpha}\left(\cdot|(\|p_i = p_{i,\alpha}\|)\right) = p_{i,\alpha}$,
(2) if $p_{i,\alpha} = p_{i,\omega}$, then $q_{i,\alpha} = q_{i,\omega}$,
(3) for each $h \in H$ there is some world $\omega \in \Omega$ such that $\sigma(\omega) \in S(h)$.

Recall that $\Delta^*(X) \subseteq \left[\Delta(X)\right]^{2^X\setminus\{\emptyset\}}$ is the set of *complete* conditional probability systems on a given set X. A complete CPS represents an individual's belief revision policy independently of the class of potentially observable events. Condition (1) of definition 4.4 says that for each state of the world α player i has a complete CPS $q_{i,\alpha}$ on the states where i's epistemic type is the same as in α and that the "initial beliefs" given by $q_{i,\alpha}$ coincide with $p_{i,\alpha}$ (actually, the p_i belief functions are redundant in this definition, but they facilitate the comparison with the standard definition of model for a game). By condition (2), the type partition fully expresses player i's epistemic attitudes, including her dispositions to revise her beliefs conditional on any possible event. Condition (3) says that each partial history obtains at some world. Representing epistemic types as complete CPSs allows us to ignore the information structure of Γ in the definition of the epistemic model, except for the "richness" condition (3).‡

An extensive form type space $\mathcal{T}$ can be derived from such a model $\mathcal{M}$ (cf. remark 10). Let $\mathbf{t}_j(\omega)$ denote the cell of player j's type partition containing world ω in $\mathcal{M}$. Then we can derive mappings $\theta_i : T_i \to \Delta^{\mathcal{H}}(S \times T_{-i})$ as follows:

$$\forall\alpha \in \Omega, \forall E \subseteq S \times T_{-i}, \forall h \in H, \theta_{i,\mathbf{t}_i(\alpha)}\left(E|S(h) \times T_{-i}\right) =$$
$$q_{i,\alpha}\left(\{\omega \in \mathbf{t}_i(\alpha) : (\sigma(\omega), \mathbf{t}_{-i}(\omega)) \in E\} \,|\, \{\omega \in \mathbf{t}_i(\alpha) : \sigma(\omega) \in S(h)\}\right).$$

† Stalnaker's original definition uses "epistemic priority" relations. It can be checked that there is a canonical bijection between the class of models à la Stalnaker and the class of models defined here.

‡ The stronger condition that the mapping σ be onto avoids any direct reference to the underlying extensive form. Stalnaker (1996) uses a notion of "perfect rationality" given by lexicographic expected utility maximization, which also exclusively relies on the strategic form. Perfect rationality implies conditional expected utility maximization at every relevant information set in the extensive form.

It is easily verified that we get essentially the same hierarchies of conditional beliefs at corresponding states of $\mathcal{T}$ and $\mathcal{M}$. We can also do the converse, i.e., obtain a model $\mathcal{M}$ for Γ from a (finite) type space $\mathcal{T}$ for Γ. But, except for trivial cases, there are several models $\mathcal{M}$ corresponding to $\mathcal{T}$, because in general there are many ways to complete a CPS on $(S, \mathcal{H})$.

4.3. SEQUENTIAL RATIONALITY AND COMMON BELIEF

Strategies have two interpretations in the present framework. A strategy for player i is (i) a component of a possible world describing a combination of subjunctive conditionals of the form "i would take action a_i at history $h \in H$", (ii) a plan of action which is part of i's beliefs about how the play will unfold and hence guides i's choice at any given history. If an individual rules out the possibility of mistakes in implementing her plan, she does not have to plan in advance for contingencies prevented by the plan itself. Yet a strategy s_i may specify behaviour at histories (more generally at information sets) whose occurrence is prevented by s_i. Thus a plan of action is typically a less complete description of a player's contingent behaviour than a strategy (see, for example, Rubinstein, 1991).† In this section we focus on the notion of sequential rationality of plans of action, but rather than working explicitly with plans of action, we take the equivalent and notationally simpler approach of checking the properties of a strategy s_i only at histories not prevented by s_i itself. Let $H(s_i)$ denote this set of histories, that is, $H_i(s_i) = \{h \in H : s_i \in S_i(h)\}$. Similarly, let $\mathcal{H}_{-i}(s_i) = \{S_{-i}(h) \subseteq S_{-i} : h \in H_i(s_i)\}$ denote the corresponding collection of subsets of opponents' strategy profiles. In Section 3 we assumed that a rational player is certain of her own strategy, which is a best response to her marginal beliefs about the opponents. Likewise, we assume here that a rational player is initially certain of her own strategy and, for each history consistent with it, would continue to be certain of her strategy and would optimize against her conditional beliefs about the opponents.

REMARK 12: fix $s_i \in S_i$ and $\mu \in \Delta^{\mathcal{H}}(S \times T_{-i})$. Suppose that

$$\forall h \in H_i(s_i), \mu(\{s_i\} \times S_{-i} \times T_{-i} | S(h) \times T_{-i}) = 1. \tag{4.1}$$

Then the vector of probability measures $\mu_{-i} = (mrg_{S_{-i}} \mu_{-i}(\cdot | S(h) \times T_{-i}))_{h \in H(s_i)}$ is a conditional probability system on $\langle S_{-i}, \mathcal{H}_{-i}(s_i) \rangle$, that is, $\mu_{-i} \in \Delta^{\mathcal{H}_{-i}(s_i)}(S_{-i})$.

† Formally, a *plan of action* is defined as a class of realization-equivalent strategies. The plan of action contained in s_i is given by the maximal set $[s_i] \subseteq S_i$ of strategies s'_i such that, for every $s_{-i} \in S_{-i}$, $\zeta(s_i, s_{-i}) = \zeta(s'_i, s_{-i})$. It turns out that $[s'_i] = [s_i]$ if and only if $s_i(h) = s'_i(h)$ for all h such that $s_i, s'_i \in S_i(h)$.

DEFINITION 4.5: *fix $s_i \in S_i$ and $\mu_{-i} \in \Delta^{\mathcal{H}_{-i}(s_i)}(S_{-i})$. We say that s_i is a (weakly) sequential best response to μ_{-i}—written $s_i \in r_i(\mu_{-i})$—if, for all $h \in H(s_i)$ and all $s'_i \in S_i(h)$,*

$$\sum_{s_{-i}} \left[U_i(s_i, s_{-i}) - U_i(s'_i, s_{-i})\right] \mu_{-i}(s_{-i}|S_{-i}(h)) \geq 0.$$

We say "weakly sequential" because expected payoff maximization is required only at histories consistent with the given strategy. This simply reflects that we are defining a notion of sequential rationality for plans of action. But we put "weakly" in parentheses because from now on we simply say "sequential".

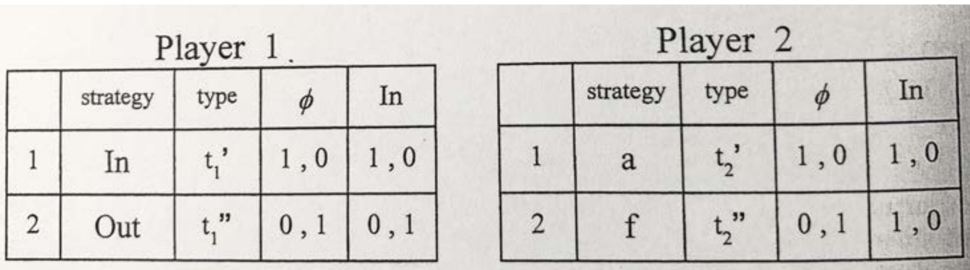
DEFINITION 4.6: *fix a type space $\mathcal{T}$ for Γ. Player i is sequentially rational at state $(s_i, s_{-i}, t_i, t_{-i})$ if*

(1) type t_i is certain of s_i whenever possible, that is, s_i and $\mu = \theta_{i,t_i}$ satisfy condition 4.1 above,

(2) s_i is a sequential best response to $(mrg_{S_{-i}}\, \theta_{i,t_i}(\cdot|S(h) \times T_{-i}))_{h \in H(s_i)}$ (the first-order conditional beliefs of type t_i about the opponents at histories consistent with s_i).

Let $\mathbf{SRAT}_i$ denote the set of states where player i is sequentially rational and let $\mathbf{SRAT} = \cap_{i \in N}\mathbf{SRAT}_i$ denote the event that all players are sequentially rational.

REMARK 13: (Tabular representation) when considering two-person games, we represent the essential features of finite type spaces as follows. First, we restrict our attention to states where condition 4.1 (certainty of one's own strategy) is satisfied and this is common belief. Then, for each player i we construct a matrix where each row fully specifies her behaviour and relevant beliefs. The first element is player i's strategy, the second element is her epistemic type label, the following elements are the marginal conditional distributions on $S_j \times T_j$ for every history $h \in H$ (i's beliefs about herself are implied by condition 4.1 at all histories consistent with the given strategy and arbitrary otherwise). For each $h \in H$, the kth element of the corresponding probability vector is the probability of row k in the matrix for player j. A state of the world is labelled by the ordered pair of indices of the corresponding rows in the matrices for player 1 and 2.



Player 1.

	strategy	type	ϕ	In
1	In	t_1'	1 , 0	1 , 0
2	Out	t_1''	0 , 1	0 , 1

Player 2

	strategy	type	ϕ	In
1	a	t_2'	1 , 0	1 , 0
2	f	t_2''	0 , 1	1 , 0

FIGURE V.

Example 4.7: the tables in Figure V represent the essential features of a type space for the Entry game (Figure S). Note that, in player 2's matrix, all the probability distributions of the column corresponding to history (In) assign probability one to the first row of player 1's matrix. This is implied by conditioning on the observed history. In fact, player 1's first row is the only one where she plays *In*. (Furthermore, in row 1, player 1 does not change her beliefs after playing *In* because she originally assigned probability one to this action. Her row-2 beliefs conditional on *In* are immaterial for our arguments.) Player 1 is (sequentially) rational at each state $\omega \in \{1, 2\} \times \{1, 2\}$. Player 2 is sequentially rational at $(1, 1)$ and $(2, 1)$. At state $(1, 1)$ there is common belief in sequential rationality.

Fix a partial history $h \in H$. We would like to characterize the strategies consistent with (history h, sequential rationality and) common belief in sequential rationality at h. The following iterative procedure is meant to capture this assumption (cf. Reny 1985, 1993, 1995):†

- $\forall i \in N$, $S^0_{i,h} = S_i(h)$, $\forall k \geq 0$, $S^k_h = \Pi_{i \in N} S^k_{i,h}$, $S^k_{-i,h} = \Pi_{j \neq i} S^k_{i,h}$,
- $\forall i \in N$,

$$S^{k+1}_{i,h} = \Big\{ s_i \in S_i(h) : \exists \mu_{-i} \in \Delta^{\mathcal{H}_{-i}(s_i)}(S_{-i}), s_i \in r_i(\mu_{-i}),$$
$$\mu(S^k_{-i,h} | S_{-i}(h)) = 1 \Big\},$$

- $S^\infty_h = \cap_{k \geq 1} S^k_h$.

† Reny's papers do not use an epistemic model. The epistemic foundations of Reny's work are provided in Battigalli and Siniscalchi (1998*a*). In particular, Reny (1993) defines a general class of procedures to characterize the set of strategies S^∞_F consistent with (sequential rationality and) common belief in (the opponent's) sequential rationality for a given set of histories $F \subseteq H$. Battigalli and Siniscalchi (1998*a*) show that, in two-person games, S^∞_F is indeed the projection on S of the following event (defined in $\mathcal{T}^U$, otherwise the projection is included in S^∞_F):

$$\mathbf{SRAT} \cap \left(\bigcap_{h \in F, i,j \in \{1,2\}, j \neq i} B_{i,h} \mathbf{SRAT}_j \right) \cap$$
$$\left(\bigcap_{h,g \in F, i,j \in \{1,2\}, j \neq i} B_{i,h} B_{j,g} \mathbf{SRAT}_i \right) \cap \ldots$$

Reny (1993) addresses the following problem: we could justify backward induction in generic perfect information games by assuming common belief in rationality for the class of all partial histories consistent with rationality except those where the active player has a dominant continuation strategy. Let $F(\Gamma)$ be this class for game Γ. Clearly this justification of backward induction is possible only if $S^\infty_{F(\Gamma)} \neq \emptyset$. It is easy to verify in simple examples that $S^\infty_{F(\Gamma)}$ may be empty (Centipede game) or not (Entry game). Reny shows that $S^\infty_{F(\Gamma)}$ is empty if and only if $F(\Gamma)$ is included in the set of backward-induction histories.

$S^1_{i,h}$ is the set of sequentially rational strategies consistent with h. The proposed interpretation of $S^2_{i,h}$ is the set of strategies consistent with h and with the event that player i is sequentially rational and would believe at h that everybody is sequentially rational. $S^{k+1}_{i,h}$ has a similar proposed interpretation. It is easily proved by induction that $S^{k+1}_h \subseteq S^k_h$. Since S is finite, there is some k^* such that $S^k_h = S^\infty_h$ for all $k \geq k^*$.

EXAMPLE 4.8: in the Centipede game depicted in Figure T, $S^k_{(L')} = \emptyset$ for all $k \geq 3$. In fact, taking into account that $L'L''$ is strictly dominated by T' which is the unique best response to t, we have $S^1_{(L')} = \{L'T''\} \times \{l, t\}$, $S^2_{(L')} = \{L'T''\} \times \{t\}$, $S^3_{(L')} = \emptyset$.

Note that the above procedure yields normal form rationalizability in static games. Furthermore, for $h = \phi$, the procedure is very similar to normal form rationalizability; the only difference is that "best response" is replaced by "sequential best response". The $\{S^k_\phi\}_{k\geq 1}$ procedure selects the backward induction solution in the Entry game and more generally in all "backward-induction-solvable" two-stage games, but not in more complex games. For example, the only strategy eliminated in the Centipede game of Figure T is (L', L'') and the only strategy eliminated in the BoS with an Outside Option (Figure U) is (In, B). We call this procedure *weak extensive form rationalizability* , where the adjective "weak" refers to the fact that the players do not necessarily try to "rationalize" their opponents behaviour after unexpected histories. A stronger notion of rationalizability featuring this "forward induction" principle is presented in Section 4.4.

Given the similarity to normal form rationalizability, it should not come as a surprise that weak extensive form rationalizability is related to a notion of iterated dominance. Recall that a strategy $s_i \in S_i$ is *weakly dominated* by a (pure or) mixed strategy $\mu_i \in \Delta(S_i)$ if and only if (a) $\forall s_{-i} \in S_{-i}$, $U_i(s_i, s_{-i}) \leq U_i(\mu_i, s_{-i})$ and (b) $\exists s^*_{-i} \in S_{-i}$, $U_i(s_i, s^*_{-i}) < U_i(\mu_i, s^*_{-i})$. Let $W_i \subseteq S_i$ denote the set weakly *undominated* pure strategies of player i's and let $W^p_i \supseteq W_i$ denote the set of player i's pure strategies not weakly dominated by other *pure* strategies. We say that a strategy $s_i \in S_i$ is *conditionally dominated at history* $h \in H$ if and only if $s_i \in S_i(h)$ and s_i is (strictly) dominated on the subset $S_{-i}(h)$ by some mixed strategy $\mu_i \in \Delta(S_i)$ with $Supp(\mu_i) \subseteq S_i(h)$. We let $C_i(h)$ denote the set of conditionally *undominated* strategies at h (note that $\neg S_i(h) \subseteq C_i(h)$). The following results relate conditional dominance, weak dominance and sequential rationality.†

† Results (1) and (2) hold for all finite extensive form games with perfect recall.

LEMMA 4.9: *(Battigalli, 1997; Ben Porath, 1997; Shimoji & Watson, 1998) fix a finite game with observed actions* Γ *and an arbitrary player i in* Γ.

(1) $W_i \subseteq S^1_{i,\phi} = \cap_{h \in H} C_i(h)$.

(2) If Γ *is generic,* $S^1_{i,\phi} \subseteq W^p_i$.

(3) If Γ *has perfect information,* $W_i = W^p_i$.

(4) If Γ *is generic and has perfect information,* $W_i = S^1_{i,\phi} = W^p_i$.

Let S^1W (S^1W^p) be the set of strategy profiles whose elements are not strictly dominated by a mixed strategy in the restriction of strategic game G_Γ to $W \subseteq S$ ($W^p \subseteq S$). Similarly, let S^kW (S^kW^p) be the set of strategy profiles whose elements are not strictly dominated by a mixed strategy in the restriction of G_Γ to $S^{k-1}W$ ($S^{k-1}W^p$).† The following result is a straightforward consequence of lemma 4.9 and—together with it—provides an "iterative dominance characterization" of weak extensive form rationalizability.

COROLLARY 4.10: *for all* $k \geq 1$,

(1) $S^kW \subseteq S^{k+1}_\phi$,

(2) if Γ *is generic,* $S^{k+1}_\phi \subseteq S^kW^p$,

(3) if Γ *is generic and has perfect information,* $S^kW = S^{k+1}_\emptyset = S^kW^p$.

EXAMPLE 4.11: using corollary 4.10 it is easy to check that the set of weakly extensive form rationalizable strategies in the Centipede game of Figure T is quite large: $S^\infty_\phi = S^1_\phi = \{T'T'', T'L'', L'T''\} \times \{l, t\}$. Similarly, for the BoS with an Outside Option, $S^\infty_\phi = S^1_\phi = \{(Out, T), (Out, B), (In, T)\} \times \{L, R\}$.

The following results show that the intended interpretation of the procedures $\{S^k_h\}_{k\geq 1}$ ($h \in H$) is indeed correct. In particular, they provide an epistemic characterization of weak extensive form rationalizability (cf. propositions 3.10 and 3.11).‡ For static

† The S^kW procedure has been first put forward by Dekel and Fudenberg (1990) to characterize the implications of iterated *weak* dominance that are robust to a "small amount" of incomplete information. Börgers (1994) shows that $S^\infty W$ is the set of strategies consistent with "almost common belief" (common belief with probability close to one) in rationality. Gul (1996) and Brandenburger (1992) provided other characterizations. On this see Dekel and Gul (1997).

‡ Weak extensive form rationalizability can be given an epistemic characterization using "static" epistemic models. Let $\mathcal{T}$ be a "static" type space for the strategic form game G_Γ. Say that i is *extended rational* at state $(s_i, s_{-i}, t_i, t_{-i})$ if there is some $\mu_{-i} \in \Delta^{\mathcal{H}_{-i}(s_i)}(S_{-i})$ such that $\mu_{-i}(\cdot|S_{-i}) = mrg_{S_{-i}}\theta_{i,t_i}$ and $s_i \in r_i(\mu_{-i})$. Then $s \in S^\infty_\phi$ if and only if there is some type space $\mathcal{T}$ for G_Γ and some state $(s, (t_i)_{i\in N})$ such that every i is extended rational and there is common belief in extended rationality at $(s, (t_i)_{i\in N})$.

games, we obtain a "type space" characterization of normal form rationalizability. The characterization of rationalizability due to Tan and Werlang (1988) is a special case of proposition 4.13 below. For any given type space $\mathcal{T}$ and history h, let $[h] := S(h) \times \Pi_{i \in N} T_i$ denote the event that h occurs. Recall that $B^0_{e,h}(E) = E$.

PROPOSITION 4.12: *(Ben Porath, 1997) let Γ be a finite game with observed actions and fix a type space $\mathcal{T}$ for Γ. Then, for all $h \in H$ and $k = 0, 1, 2 \ldots$*

(i) $\mathbf{SRAT} \cap (\cap^k_{j=0} B^j_{e,h}\mathbf{SRAT}) \cap [h] \subseteq S^{k+1}_h \times (\Pi_{i \in N} T_i)$ and

(ii) $\mathbf{SRAT} \cap B_{*,h}\mathbf{SRAT} \cap [h] \subseteq S^\infty_h \times (\Pi_{i \in N} T_i)$.

PROPOSITION 4.13: *(Battigalli & Siniscalchi, 1998a) let Γ be a finite game with observed actions and consider the universal type space $\mathcal{T}^U$ for Γ. Then, for all $s \in S$, $h \in H$ and $k = 0, 1, 2, \ldots,$*

(i) $s \in S^{k+1}_h$ if and only if there is a type profile $(t_i)_{i \in N} \in \Pi_{i \in N} T^U$ such that

$$\left(s, (t_i)_{i \in N}\right) \in \mathbf{SRAT} \cap \left(\bigcap_{j=0}^{k} B^j_{e,h}\mathbf{SRAT} \right) \cap [h],$$

(ii) $s \in S^\infty_h$ if and only if there is a type profile $(t_i)_{i \in N} \in \Pi_{i \in N} T^U$ such that

$$\left(s, (t_i)_{i \in N}\right) \in \mathbf{SRAT} \cap B_{*,h}\mathbf{SRAT} \cap [h].$$

(Ben Porath, 1997) Furthermore, there is a finite type space $\mathcal{T}$ for Γ satisfying the same properties.

EXAMPLE 4.14: proposition 4.13 implies that the set of strategy profiles consistent with initial common belief in sequential rationality in the Centipede game of Figure S is $\{T'T'', T'L'', L'T''\} \times \{l, t\}$ (cf. example 4.11). Figure W(i) shows a type space for this game. Since we are interested in plans of actions, we do not completely specify the strategy of player 1 when she chooses T'. We further simplify the tables by coalescing the columns corresponding to histories where player i has the same marginal distribution about player j. Figure W(ii) represents interactive beliefs at the beginning of the game with the usual graphical conventions (cf. remark 3). It can be checked that at state (2, 1) the players are sequentially rational and this is common belief at the beginning of the game. The outcome is (L', l, T''). This confirms our informal discussion of the Centipede game. State (1, 2) corresponds to the backward induction equilibrium. (This is the only state satisfying the Truth Condition.)

4.4 STRONG BELIEF AND FORWARD INDUCTION

As we mentioned in Section 4.1.3, we would like to formalize and characterize the following sequence of assumptions:

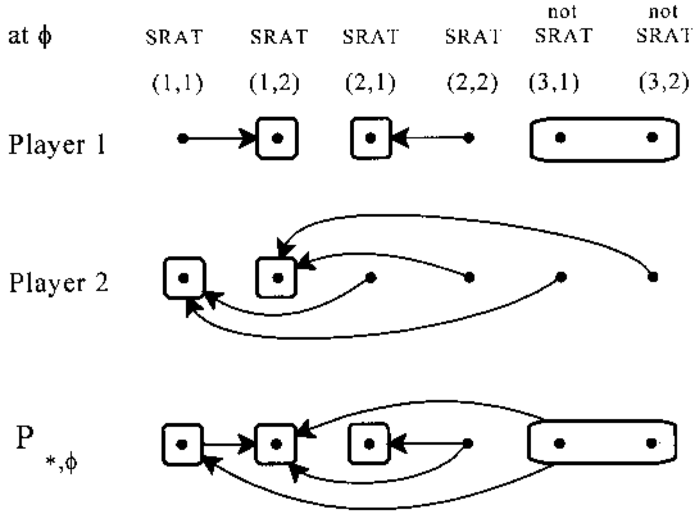
Player 1

	strategy	type	ϕ, L'	L'l
1	T'	t_1'	0 , 1	1 , 0
2	L'T"	t_1''	1 , 0	1 , 0
3	L'L"	t_1'''	p, 1−p	p, 1−p

Player 2

	strategy	type	ϕ	L',L'l
1	l	t_2'	1, 0, 0	0, 0, 1
2	t	t_2''	1, 0, 0	0, 1, 0

(i)

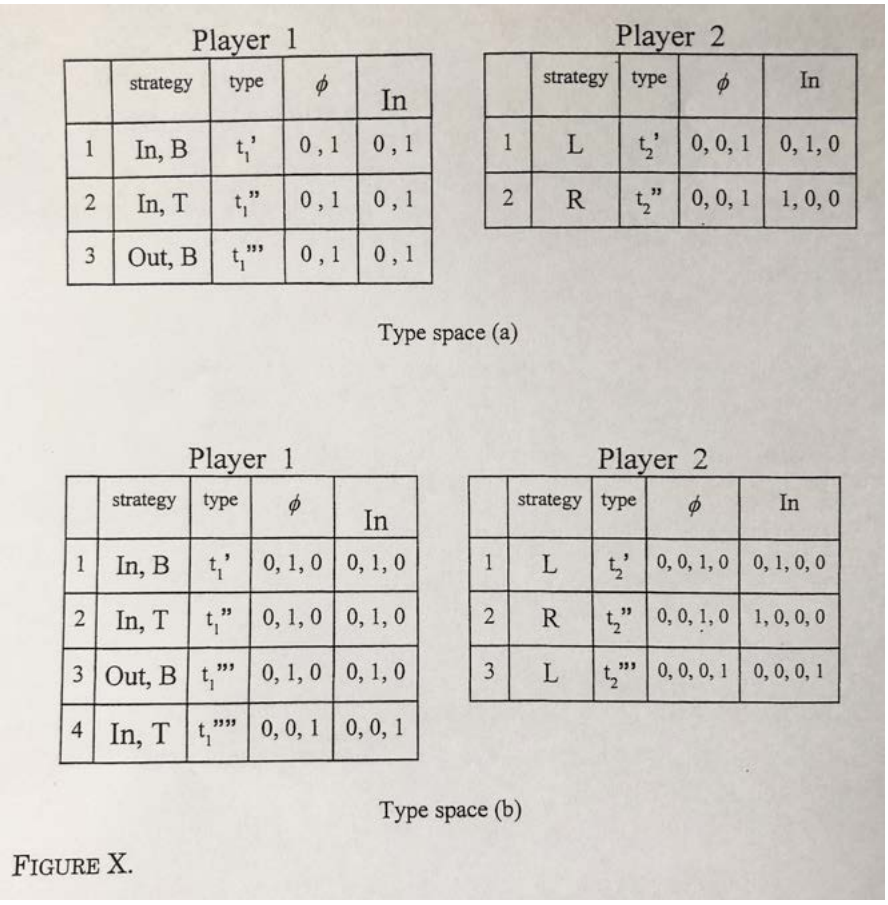


(ii)

FIGURE W.

(0) every player is sequentially rational,
(1) every player believes (0) whenever possible,
(2) every player believes (0)&(1) whenever possible,
. . .
(k+1) every player believes (0)&(1)&. . . &(k) whenever possible,
. . .

We have argued that assumptions (0), (1) and (2) eliminate the subgame perfect equilibrium outcome *Out* in the BoS with an Outside Option. But in order to appropriately formalize this argument we must be careful with the qualification "whenever possible". If we represent the players' interactive beliefs with a "small" type space, a player may find it impossible to rationalize his opponent's behaviour just because the space does not contain the conceivable epistemic types that rationalize such behaviour.



Player 1

	strategy	type	ϕ	In
1	In, B	t_1'	0 , 1	0 , 1
2	In, T	t_1''	0 , 1	0 , 1
3	Out, B	t_1'''	0 , 1	0 , 1

Player 2

	strategy	type	ϕ	In
1	L	t_2'	0, 0, 1	0, 1, 0
2	R	t_2''	0, 0, 1	1, 0, 0

Type space (a)

Player 1

	strategy	type	ϕ	In
1	In, B	t_1'	0, 1, 0	0, 1, 0
2	In, T	t_1''	0, 1, 0	0, 1, 0
3	Out, B	t_1'''	0, 1, 0	0, 1, 0
4	In, T	t_1''''	0, 0, 1	0, 0, 1

Player 2

	strategy	type	ϕ	In
1	L	t_2'	0, 0, 1, 0	0, 1, 0, 0
2	R	t_2''	0, 0, 1, 0	1, 0, 0, 0
3	L	t_2'''	0, 0, 0, 1	0, 0, 0, 1

Type space (b)

FIGURE X.

EXAMPLE 4.15: Figure X shows two type spaces for the BoS with an Outside Option. It can be checked that type space (a) can be embedded in type space (b), that is, the profiles of strategies and hierarchies of conditional beliefs corresponding to the states of space (a) are a subset of the profiles corresponding to states of space (b).† In space (a) there is no state consistent with the forward induction story, because player 2 is forced to believe, in the BoS subgame, that only irrational strategy/type pairs could have chosen *In*. On the contrary, in space (b) player 1's rational strategy/type pair $(In, T; t_1'''')$ chooses *In* and player 2's rational strategy/type pair (L, t_2''') rationalizes this move. State (4,3) of type space (b) is consistent with the forward induction story of Section 4.1.3.

Type space (b) in the above example is sufficiently rich to correctly represent the forward induction story, but in order to

† To be precise, we should also check the states inconsistent with assumption (1) of definition 4.5, but we are free to specify the beliefs at such states so that the claim holds.

provide a neat formalization it is better to work with a type space that contains all the conceivable epistemic types, that is, the universal type space for the given extensive form game. Here we present some concepts and results due to Battigalli and Siniscalchi (1997). Related ideas are discussed by Stalnaker (1998).

4.4.1. Extensive form rationalizability

We first introduce the solution procedure called "extensive form rationalizability" that is supposed to capture assumptions (0), (1), . . . ,(k),. . .

DEFINITION 4.16: *let $S_e^0 = S$. Assume that $S_e^1, \ldots, S_e^k$ have been defined. Then $s = (s_i)_{i \in N} \in S_e^{k+1}$ if and only if $s \in S_e^k$ and, for each player i, there exists some CPS $\mu_{-i} \in \Delta^{\mathcal{H}_{-i}(s_i)}(S_{-i})$ such that:*

(1) For each $h \in H(s_i)$, $S_{-i}(h) \cap S_{e,-i}^k \neq \emptyset \Rightarrow \mu_{-i}(S_{e,-i}^k | S_{-i}(h)) = 1$.

(2) $s_i \in r_i(\mu_{-i})$.

A strategy profile s is extensive form rationalizable if and only if $s \in \cap_{k>0} S_e^k$.

The preceding definition is very similar to that originally proposed by Pearce (1984) (see Battigalli, 1997). It can be checked that the only extensive form rationalizable profile in the BoS with an Outside Option is the "forward induction" equilibrium $[(In, T), L]$.

4.4.2. Strong (or robust) beliefs

Next we define the meaning of "believing event E whenever possible". This is captured by the notion of "strong" (or "robust") belief. † Fix the universal type space $\mathcal{T}^U$ for game Γ. Recall that an event in $\mathcal{T}^U$ is a (measurable) subset $E \subseteq S \times \Pi_{i \in N} T_i^U$.

DEFINITION 4.17: *for any event $E \neq \emptyset$, player i and type (hierarchy of conditional beliefs) $t \in T_i^U$ we say that type t strongly believes E (believes E whenever possible) if for all partial histories $h \in H$,*

$$E_t \cap (S(h) \times T_{-i}^U) \neq \emptyset \Rightarrow \theta_{i,t}(E_t | S(h) \times T_{-i}^U) = 1.$$

Let $B_i^s E$ denote the event that player i strongly believes E and let $B_e^s E$ denote the event that everybody strongly believes the (non-empty) event E, that is:

† "Robust belief" is the terminology used by Stalnaker (1998) for a similar concept.

- $B_i^s E := \{(s, t_i, t_{-i}) : \forall h \in H, E_{t_i} \cap (S(h) \times T_{-i}^U) \neq \emptyset \Rightarrow \theta_{i,t_i}(E_{t_i}|S(h) \times T_{-i}^U) = 1\}$,†
- $B_e^s E := \cap_{i \in N} B_i^s(E)$.

By inspecting the definition of strong belief, one notices that the event E itself determines the class of information sets $h \in H$ where player i's conditional beliefs are restricted. This simple observation has two important consequences. First, for arbitrary events E and F, we have $B_i^s(E \cap F) \supseteq B_i^s E \cap B_i^s F$, but *equality need not hold.* Clearly, every state of the world in $B_i^s E \cap B_i^s F$ is such that player i's belief at any information set $h \in H$ consistent with $E \cap F$ (i.e., such that $(E \cap F)_{t_i} \cap (S(h) \times T_{-i}^U) \neq \emptyset$) assigns probability one to both E and F, hence to $E \cap F$: thus, every such state is also an element of $B_i^s(E \cap F)$. However, the converse need not be true, because there might be an information set $h \in \mathcal{H}_i$ which is inconsistent with F but consistent with E (or vice versa): in this case, $B_i^s(E \cap F)$ places no restrictions on player i's beliefs at any such h, but clearly $B_i^s E \cap B_i^s F$ does. Thus, a given state of the world may be an element of the former set, but not of the latter.‡ As a result, one must be careful to interpret assumptions involving conjunctions of strong belief operators accurately.

Second, note that the argument above implies that, unlike standard epistemic operators, the strong belief operator B_i^s is *not* monotone (otherwise the inclusion relation $B_i^s(E \cap F) \supseteq B_i^s E \cap B_i^s F$ would necessarily hold as an equality).

Now we define an auxiliary operator that simplifies the representation of assumptions (0), (1), (2), ... For any event E, let

$$CE = E \cap B_e^s E$$

denote the set of states where E is true and everybody strongly believes E. We can define iterations of C in the usual way. In particular, we obtain the following identities:

$$\begin{aligned} C^0E &= E, \\ C^1E &= E \cap B_e^s E, \\ C^2E &= C\left(E \cap B_e^s E\right) = E \cap B_e^s E \cap B_e^s \left(E \cap B_e^s E\right), \\ &\dots \end{aligned}$$

It should be clear that the iterated application of the operator C yields a nested sequence of events which represent the informal assumptions (0), (1), (2), ... discussed above, that is, event C^k**SRAT** represents (0)&(1)&... &(k). We can actually be even more explicit:

† For $E = \emptyset$, let $B_i^s(E) = \emptyset$.

‡ Also note that there may be histories consistent with E and F but inconsistent with $E \cap F$. In this case $B_i^s(E) \cap B_i^s(F) = \emptyset$ because it is impossible to believe E and F at such histories.

REMARK 14: by inspection of the definitions above

$$C^m\mathbf{SRAT} = \bigcap_{i\in N}\left[\mathbf{SRAT}_i \bigcap \left(\cap_{k=0}^{m-1} B_i^s(C^k\mathbf{SRAT})\right)\right].$$

Therefore $(s, (t_i)_{i\in N}) \in C^m\mathbf{SRAT}$ if and only if, for each player i, s_i is a (weakly) sequential best response to the first-order beliefs of type t_i and t_i strongly believes $C^k\mathbf{SRAT}$ for all $k = 0, \ldots, n-1$.

4.4.3 Characterization

Having disposed of all the preliminaries, we are ready to state the main result:

PROPOSITION 4.18: *for every strategy profile $s \in S$ the following statements hold:*

(a) for all $k \geq 0$, $s \in S_e^{k+1}$ if and only if there exists a profile of infinite hierarchies of conditional beliefs $(t_i)_{i\in N} \in \Pi_{i\in N} T_i^U$ such that $(s, (t_i)_{i\in N}) \in C^k\mathbf{SRAT}$;

(b) $s \in \cap_{k=0}^{\infty} S_e^k$ if and only if there exists a profile of infinite hierarchies of beliefs $(t_i)_{i\in N} \in \Pi_{i\in N} T_i^U$ such that $(s, (t_i)_{i\in N}) \in \cap_{k=0}^{\infty} C^k\mathbf{SRAT}$.

4.4.4. Strong belief and backward induction

In the introductory discussion of Section 4.1 we noticed that assumptions (0), (1) and (2) yield the backward induction outcome in the Centipede game of Figure T. This observation can be generalized. In fact, extensive form rationalizability is generically equivalent to backward induction in games with perfect information (cf. Reny, 1992 and Battigalli, 1997), therefore we can derive from proposition 4.18 the following result.

PROPOSITION 4.19: *suppose that the given game Γ has perfect information and is generic. Then for every state $(s, (t_i)_{i\in N}) \in \cap_{k=0}^{\infty} C^k\mathbf{SRAT}$, $\zeta(s)$ (the complete history induced by s) is precisely the (unique) backward induction complete history.*

We emphasize that the joint assumptions (0)&(1)&(2)& ...do not imply that a player at a non-rationalizable partial history would play and/or expect the backward induction continuation. In certain games this is actually *inconsistent* with strong belief in sequential rationality. The following example (figure 3 in Reny, 1992), illustrates this point:

EXAMPLE 4.20: consider the game depicted in Figure Y. At the first iteration, $C^0\mathbf{SRAT} = \mathbf{SRAT}$ eliminates $A'D''$ for player 1 and

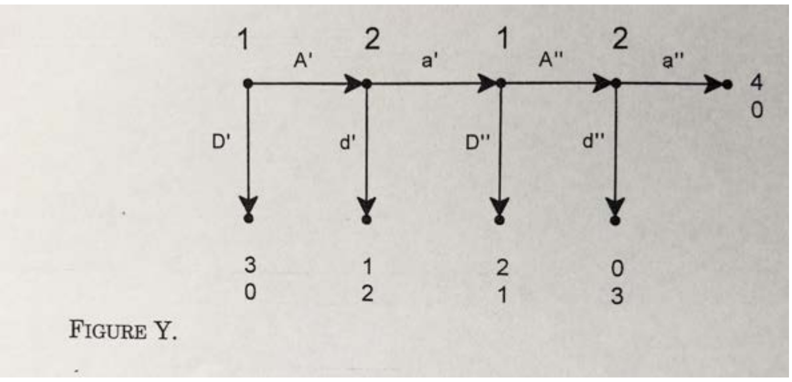


FIGURE Y.

$a'a''$ for player 2. Then, $C^1\mathbf{SRAT} = \mathbf{SRAT} \cap B_e^s\mathbf{SRAT}$ eliminates $A'A''$ and d', which yields the backward induction outcome. Hence, if player 2 were reached, he must conclude that it is not the case that everybody is sequentially rational and there is strong belief in sequential rationality. However, he can continue to (strongly) believe that everybody, in particular player 1, is sequentially rational. Of course, this implies that player 2 should expect player 1 to move Across at her second node. Event

$$C^2\mathbf{SRAT} = C^1\mathbf{SRAT} \cap B_e^s C^1\mathbf{SRAT} =$$
$$\mathbf{SRAT} \cap B_e^s\mathbf{SRAT} \cap B_e^s(\mathbf{SRAT} \cap B_e^s\mathbf{SRAT})$$

incorporates precisely this restriction. However, backward induction reasoning implies that player 2, upon being reached, should expect player 1 to move Down at her next node.

4.4.5. Rationalizability and conditional common belief in sequential rationality

We have seen that common belief in sequential rationality at a given partial history h may be impossible. Now, as a consequence of proposition 4.18, we can identify a set of histories consistent with common belief in sequential rationality: all the histories induced by extensive form rationalizable profiles.

PROPOSITION 4.21: *fix a partial history $h \in H$. If there is an extensive form rationalizable strategy profile s inducing h (i.e., $S_e^\infty \cap S(h) \neq \emptyset$), then there is a type space for Γ such that*

$$\mathbf{SRAT} \cap B_{*,h}\mathbf{SRAT} \cap [h] \neq \emptyset.$$

Note that the proposition provides only a sufficient condition. There are games with histories consistent with common belief in sequential rationality and yet unreachable by profiles of extensive form rationalizable strategies (Battigalli & Siniscalchi, 1997; see also Reny, 1985).

4.5. EPISTEMIC INDEPENDENCE AND BACKWARD INDUCTION

We say that player i's beliefs exhibit *epistemic independence* if information exclusively concerning opponent j does not affect i's beliefs about opponent k.† Here we relate the assumption of epistemic independence to the backward induction procedure. Consider the following modification of the Centipede game of Figure T: player 1 is split into two different players, 1' and 1", with identical payoffs. Suppose that (i) each player is sequentially rational and her beliefs exhibit epistemic independence, (ii) this is common belief at the beginning of the game. Then player 1' chooses T' and the other two players would choose their backward induction action if given the opportunity. In fact, since player 2 initially believes that player 1" is rational, she would not change this belief after L', a move of a different opponent. Therefore player 2 would anticipate T'' and—by sequential rationality—choose t after L'. Initial common belief in the event [rationality and epistemic independence] implies that player 1' anticipates choice t.

In order to formally define the epistemic independence assumption in the present setting we look at "marginal" CPSs on the sets of strategies and types of each player i. Let $\mathcal{H}_i = \{(S_i(h) \times T_i) : h \in H\}$ be the set of conditioning events concerning player i. The set of "marginal" CPSs on $\langle S_i \times T_i, \mathcal{H}_i \rangle$ is denoted by $\Delta^{\mathcal{H}_i}(S_i \times T_i)$. With a slight abuse of notation we also denote by $\Delta^{\mathcal{H}_i}(S_i)$ the set of marginal CPSs on the set of player i's strategies.

DEFINITION 4.22: *fix a type space $\mathcal{T}$ for Γ. We say that CPS $\mu \in \Delta^{\mathcal{H}}(S \times T_{-i})$ exhibits epistemic independence if there are marginal CPSs $\mu_i \in \Delta^{\mathcal{H}_i}(S_i)$, $\mu_j \in \Delta^{\mathcal{H}_j}(S_j \times T_j)(j \neq i)$ such that, for all $h \in H$, $\mu(\cdot|S(h) \times T_{-i})$ is the product measure obtained from $\mu_i(\cdot|S_i(h))$ and $(\mu_j(\cdot|S_j(h) \times T_j))_{j \neq i}$. Let* **I** *denote the set of states (s, t) such that θ_{i,t_i} exhibits epistemic independence for each $i \in N$.*

The epistemic independence assumption can be used to define modifications of the notions of weak extensive form rationalizability and extensive form rationalizability (see Battigalli & Siniscalchi, 1998*b*). Battigalli (1996) discusses the relationship between this notion of independence (for first-order beliefs) and

† The phrase "epistemic independence" is due to Stalnaker (1998). Aumann (1974), Stalnaker and others argue that *epistemic* independence is *not* a consequence of *causal* independence (the basic tenet that the plans and thought processes of different players cannot affect each other). This argument is now widely accepted.

consistency of assessments in the sense of Kreps and Wilson (1982). Here we simply consider the behavioural implications of epistemic assumptions concerning rationality and independence in a class of simple games. Recall that B_* denotes the "initial common belief" operator. Furthermore, let s^{BI} denote the (unique) *backward induction* strategy profile of a generic game with perfect information.

PROPOSITION 4.23: *(cf. Stalnaker, 1998; Battigalli & Siniscalchi, 1998b) let* Γ *be a finite and generic game with perfect information such that no player is active more than once along any play path (for all* $z \in Z$, $i \in N$, *there is at most one* h *preceding* z *such that* $A_i(h)$ *has at least two elements) and fix a type* $\mathcal{T}$ *space for* Γ.

(1) For all states $(s, t) \in B_*(\mathbf{I} \cap \mathbf{SRAT})$, *partial histories* $h, h' \in H$ *and players* $i, j \in N$, *if* h *precedes* h' *and* j *is active at* h', *then*

$$\theta_{i,t_i}\left(\{s_j^{BI}\} \times S_{-j} \times T_{-i} | S(h) \times T_{-i}\right) = 1.$$

(2) For all states $(s, t) \in (\mathbf{I} \cap \mathbf{SRAT}) \cap B_*(\mathbf{I} \cap \mathbf{SRAT})$, $s = s^{BI}$.

Part (1) of proposition 4.23 says that initial common belief in rationality and independence implies that first-order beliefs about future behaviour conform to backward induction. Part (2) is the self-explanatory consequence of this fact. The assumption that no player moves more than once in any play path is crucial. For example, it is easily checked that in the type space for the original Centipede represented in Figure W state (1,2) satisfies the epistemic assumptions of proposition 4.23, but violates backward induction.

Different epistemic foundations for the backward induction solution have been provided by Aumann (1995, 1996), Samet (1996*a*) and Balkenborg and Winter (1997). The epistemic models used in these papers differ substantially from the one used here because they do not represent Bayesian updating. This makes a formal comparison difficult. We refer to Stalnaker (1997) and Battigalli and Siniscalchi (1998*a*) for a discussion of these models. Aumann (1998*a*) provides a result for the Centipede that is related to our discussion of strong rationalizability in strategic form games in Subsection 3.5.

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. Hebrew University of Jerusalem, mimeo.

Aumann, R. (1995). Backward induction and common knowledge of rationality. *Games and Economic Behavior*, **8**, 6–19.

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

Aumann, R. (1998*a*). On the centipede game. *Games and Economic Behavior*, **23**, 97–105.

Aumann, R. (1998*b*). Reply to Gul. *Econometrica*, **66**, 929–938.

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

Balkenborg, D. & Winter, E. (1997). A necessary and sufficient epistemic condition for playing backward induction. *Journal of Mathematical Economics*, **27**, 325–345.

Battigalli, P. (1996). Strategic independence and perfect Bayesian equilibria, *Journal of Economic Theory*, **70**, 201–234.

Battigalli, P. (1997). On rationalizability in extensive games. *Journal of Economic Theory*, **74**, 40–61.

Battigalli, P. & Bonanno, G. (1997*a*). The logic of belief persistence. *Economics and Philosophy*, **13**, 39–59.

Battigalli, P. & Bonanno, G. (1997*b*). Synchronic information and common knowledge in extensive games. In M. Bacharach, L.A. Gerard-Varet, P. Mongin and H. Shin, Eds. *Epistemic Logic and the Theory of Games and Decisions*. Dordrecht: Kluwer.

Battigalli, P. & Siniscalchi, M. (1997). An epistemic characterization of extensive form rationalizability. Social Science Working Paper 1009, California Institute of Technology.

Battigalli, P. & Siniscalchi, M. (1998*a*). Hierarchies of conditional beliefs and interactive epistemology in dynamic games. Princeton University, mimeo.

Battigalli, P. & Siniscalchi, M. (1998*b*). Interactive beliefs, epistemic independence and strong rationalizability. Princeton University and European University Institute mimeo [forthcoming in *Research in Economics*].

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

Bernheim, D. (1984). Rationalizable strategic behavior. *Econometrica*, **52**, 1002–1028.

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

Bonanno, G. & Nehring, K. (1996*a*). How to make sense of the Common Prior Assumption under incomplete information. Working paper, University of California Davis. [Extended abstract In Itzhak Gilboa, Ed. *Theoretical Aspects of Rationality and Knowledge* (TARK 1998), San Francisco: Morgan Kaufman, pp. 147–160, 1998.]

Bonanno, G. & Nehring, K. (1996*b*). On Stalnaker's notion of strong rationalizability and Nash equilibrium in perfect information games. University of California Davis, memeo [forthcoming in *Theory and Decision*.]

Bonanno, G. & Nehring, K. (1998*a*). Assessing the Truth Axiom under incomplete information. *Mathematical Social Sciences*, **36**, 3–29.

Bonanno, G. & Nehring, K. (1998*b*). On the logic and role of negative introspection of common belief. *Mathematical Social Sciences*, **35**, 17–36.

Bonanno, G. & Nehring, K. (1998*c*). Intersubjective consistency of knowledge and belief. Working paper, University of California Davis.

Börgers, T. (1994). Weak dominance and approximate common knowledge. *Journal of Economic Theory*, **64**, 265–276.

Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In P. Dasgupta, D. Gale, O. Hart and E. Maskin, Eds. *Economic Analysis of Markets and Games*. Cambridge, MA: MIT Press.

Brandenburger, A. (1997). A logic of decision. Working Paper 98–039, Harvard Business School.

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

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

Chellas, B. (1984). *Modal Logic*. Cambridge, U.K.: Cambridge University Press.

Dekel, E. & Fudenberg, D. (1990). Rational behavior with payoff uncertainty. *Journal of Economic Theory*, **52**, 243–267.

Dekel, E. & Gul, F. (1997). Rationality and knowledge in game theory. In D. Kreps & K. Wallis, Eds. *Advances in Economics and Econometrics*. Cambridge, U.K.: Cambridge University Press.

Fagin, R. & Halpern, J. (1994). Reasoning about knowledge and probability. *Journal of the Association for Computing Machinery*, **41**, 340–367.

Fagin, R. Halpern, J. Moses, Y. & Vardi, M. (1995). *Reasoning about Knowledge*. Cambridge, MA: MIT Press.

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

Fudenberg, D. & Tirole, J. (1991). *Game Theory*. Cambridge, MA: MIT Press.

Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge, MA: MIT Press.

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

Geanakoplos, J. (1994). Common knowledge. In R. Aumann & S. Hart, Eds. *Handbook of Game Theory*, vol.2, pp. 1437–1496. Elsevier.

Gul, F. (1996). Rationality and coherent theories of strategic behavior. *Journal of Economic Theory*, **70**, 1–31.

Gul, F. (1998). A comment on Aumann's Bayesian view. *Econometrica*, **66**, 923–927.

Halpern, J. (1991). The relationship between knowledge, belief and certainty. *Annals of Mathematics and Artificial Intelligence*, **4**, 301–322.

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

Harman, G. (1986). *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Harsanyi, J. (1967–68). Games of incomplete information played by Bayesian players. Parts I, II, III. *Management Science*, **14**, 159–182, 320–334, 486–502.

Heifetz, A. & Samet, D. (1996). Topology-free typology of beliefs. Tel Aviv University, mimeo.

Hintikka, J. (1962). *Knowledge and Belief*. Cornell University Press.

Kreps, D. & Wilson, R. (1982). Sequential equilibria. *Econometrica*, **50**, 863–894.

Kraus, S. & Lehmann, D. (1988). Knowledge, belief and time. *Theoretical Computer Science*, **58**, 155–174.

Lentzen, W. (1978). Recent work in epistemic logic. *Acta Philosophica Fennica*, **30**, 1–220.

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

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

Lismont, L. & Mongin, P. (1995). Belief closure: a semantics for common knowledge for modal propositional logic. *Mathematical Social Sciences*, **30**, 127–153.

Mertens, J.F. & Zamir, S. (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.

Myerson, R. (1986). Multistage games with communication. *Econometrica*, **54**, 323–358.

Osborne, M. & Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, MA: MIT Press.

Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, **52**, 1029–1050.

Reny, P. (1985). Rationality, common knowledge and the theory of games. Department of Economics, Princeton University, mimeo.

Reny, P. (1992). Backward induction, normal form perfection and explicable equilibria. *Econometrica*, **60**, 626–649.

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

Reny, P. (1995). Rational behaviour in extensive form games. *Canadian Journal of Economics*, **28**, 1–16.

Rênyi, A. (1955). On a new axiomatic theory of probability. *Acta Mathematica Academiae Scientiarum Hungaricae*, **6**, 285–335.

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

Rubinstein, A. (1991). Comments on the interpretation of game theory. *Econometrica*, **59**, 909–904.

Samet, D. (1996*a*). Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, **17**, 230–251.

Samet, D. (1996*b*). Common priors and Markov chains. Tel Aviv University, mimeo.

Samet, D. (1998) Common priors and separation of convex sets. *Games and Economic Behavior*, **24**, 172–174.

Selten, R. (1978). The chainstore paradox. *Theory and Decision*, **9**, 127–159.

Shimoji, M. & Watson, J. (1998). Conditional dominance, rationalizability, and game forms. *Journal of Economic Theory*, **83**, 161–195.

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. (1998). Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, **36**, 31–56.

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

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

van der Hoek, W. (1993). Systems for knowledge and belief. *Journal of Logic and Computation*, **3**, 173–195.

van der Hoek, W. & Meyer, J.-J. Ch. (1995). *Epistemic Logic for Artificial Intelligence and Computer Science*. Cambridge, MA: Cambridge University Press.