Strong Belief and Forward Induction Reasoning*

Pierpaolo Battigalli
Istituto di Economia Politica, Università Bocconi,
20136 Milano, Italy; pierpaolo.battigalli@uni-bocconi.it

Marciano Siniscalchi
Department of Economics, Princeton University,
Princeton, NJ 08544-1021; marciano@princeton.edu

September 2001

RUNNING TITLE: Strong Belief and Forward Induction

CORRESPONDING AUTHOR: Marciano Siniscalchi
Economics Department, Princeton University
Princeton, NJ 08544-1021
email: marciano@princeton.edu
phone (609) 258-4024; fax (609) 258-6419

*This is a significantly expanded and revised version of Battigalli and Siniscalchi [6]. We thank Drew Fudenberg for helpful comments. We also benefited from conversations with Adam Brandenburger, Eddie Dekel, Faruk Gul, and seminar participants at Barcelona, Caltech, Harvard, Michigan, MIT, Northestern, Princeton, Stanford, Tel Aviv, Tilburg, Toulouse, Torino and UCL. The usual disclaimer applies. Financial support from the European University Institute and Bocconi University (Battigalli) and IGIER–Bocconi University, Kellogg Graduate School of Management, and NSF Grant N. 9911490 (Siniscalchi) is gratefully acknowledged.

Abstract

We provide a unified epistemic analysis of some forward-induction solution concepts in games with complete and incomplete information. We suggest that forward induction reasoning may be usefully interpreted as a set of assumptions governing the players' belief revision processes, and define a notion of strong belief to formalize these assumptions. Building on the notion of strong belief, we provide an epistemic characterization of extensive-form rationalizability and the intuitive criterion, as well as sufficient epistemic conditions for the backward induction outcome in generic games with perfect information.

Journal of Economic Literature Classification Numbers: C72, D82.

KEYWORDS: Conditional Belief, Strong Belief, Forward Induction, Rationalizability, Intuitive Criterion.

1 Introduction

Forward-induction reasoning[1] is motivated by the assumption that unanticipated strategic events, including deviations from a putative equilibrium path, result from purposeful choices. Thus, if a player observes an unexpected move, she should *revise her beliefs* so as to reflect its likely purpose.

However, in order to divine the purpose of unexpected moves, a player must formulate assumptions about her opponents' rationality and strategic reasoning. This paper focuses on these assumptions and emphasizes their rôle in guiding the players' belief revision process, and hence their behavior (cf. Stalnaker [31, 32]). In particular, we adopt a model of interactive conditional beliefs based on Battigalli and Siniscalchi [7] and propose a formal analysis of forward-induction reasoning whose centerpiece is the notion of "strong belief."

We say that a player *strongly believes* event E if she believes that E is true at the beginning of the game, *and continues to do so as long as E is not falsified by the evidence.*[2] In other words, E serves as a "working hypothesis."

The notion of strong belief allows us to provide a unified epistemic analysis of different versions of forward induction, listed below in an order that, loosely speaking, reflects the complexity of the corresponding assumptions about beliefs:

- In its simplest form, forward-induction reasoning involves the assumption that, upon observing an unexpected (but undominated) move of her opponent, a player maintains the "working hypothesis" that the latter is rational (for example, see [24], pp 110-111). Strong belief in the rationality of opponents captures precisely this type of argument.

- In the context of signalling games, we show that strong belief in rationality and in a candidate equilibrium path justifies the deletion of equilibrium-dominated messages for each sender type. This leads to an epistemic characterization of the *intuitive criterion* of Cho and Kreps [15].

[1] To the best of our knowledge the earliest example of forward induction reasoning is due to Elon Kohlberg. See Van Damme [33] and Kohlberg [21] for excellent surveys and references on forward induction equilibria. Non-equilibrium solution concepts featuring forward induction are put forward and/or analyzed by Asheim and Dufwenberg [1], Battigalli [4, 5], Pearce [25] and Reny [27].

[2] In a different formal setting, Stalnaker [32] independently introduced the notion of "robust belief," which captures a similar intuition.

- *Extensive-form rationalizability* (Battigalli [4, 5], Pearce [25]) is based on the informal assumption that a player interprets unexpected moves of her opponents in a manner consistent with the highest possible "degree of strategic sophistication." Using the notion of strong belief, we formalize this assumption in the context of our epistemic model, and obtain an epistemic characterization of extensive-form rationalizability.

  Since extensive-form rationalizability induces the backward-induction outcome in generic perfect information games (cf. [5] and [27]), our analysis additionally provides sufficient epistemic conditions for *backward induction.*

The above results are meant to illustrate "typical" applications of strong belief to the analysis of forward-induction reasoning. Thus, we have restricted our attention to (relatively) well-known solution concepts and examples. However, as we suggest in Section 6, different assumptions involving rationality and strong belief may be used to obtain characterizations of other known extensive-form solution concepts that embody notions of forward induction—as well as to derive new solution concepts which may be more germane to specific applications.

On a similar note, we *do not* wish to suggest that any of the solution concepts analyzed in this paper should be regarded as embodying the "right" notion of forward induction. Rather, we suggest that the notion of strong belief allows to uncover and make explicit certain assumptions about the belief revision processes associated with different versions of forward-induction reasoning.

Finally, normal-form solution concepts such as strategic stability (cf. Kohlberg and Mertens [22]) and iterated weak dominance also typically select outcomes consistent with versions of forward-induction reasoning. When this is the case, normal-form analysis may be viewed as providing an alternative rationale for "forward induction *outcomes*". However, normal-form analysis is unlikely to shed light on the aspect of forward induction *reasoning* we emphasize, namely the players' belief revision process.

Following Battigalli and Siniscalchi [7], in the model of interactive beliefs adopted here, a state comprises a specification of the strategy and *epistemic type* of each player. Every epistemic type corresponds to a conditional probability system over opponents' strategies and types—hence, implicitly, to an infinite hierarchy of conditional beliefs on opponents' actions and conditional beliefs.

As we argue in Section 3, the analysis of the behavioral implications of forward induction is considerably simplified by focusing on *belief-complete* models. Loosely

speaking, in such models, *every* conceivable hierarchy of conditional beliefs a player may hold about her opponents is represented by an epistemic type.

We have already mentioned some of the key references on forward induction. Further comments on the related literature are deferred to the discussion section.

The remainder of the paper is organized as follows. The framework is introduced in Section 2. Section 3 provides the formal definition of strong belief and illustrates its features by means of an example. Section 4 provides a characterization of extensive-form rationalizability. Section 5 contains our characterization of the intuitive criterion. Section 6 discusses some modelling choices and possible extensions of the analysis, and comments on the related literature. All proofs are contained in the Appendix.

2 The Framework

This Section introduces most of the required game-theoretic notation, and summarizes the features of type spaces that will be relevant to our analysis. Further details may be found in Battigalli and Siniscalchi [7].

2.1 Multistage Games

We focus on dynamic, finite games, and allow for the possibility that payoff functions may not be commonly known. This may reflect imperfect knowledge of the opponents' preferences, or of the link between actions and payoff relevant consequences. Therefore, in general we allow for incomplete information.

In order to keep notation at a minimum, our analysis shall deal with multistage games with *observable actions*,[3] although our framework and techniques can be adapted to deal with general extensive-form games (see Section 6 for further details).

We shall be interested in the following primitive objects: a set $I = \{1, \ldots, |I|\}$ of players, a finite collection $\mathcal{H}$ of *(non-terminal) histories*,[4] including the *empty history* ϕ, a finite collection of *terminal histories* $\mathcal{Z}$, and, for each player $i \in I$, a

[3]For a complete definition see Fudenberg and Tirole [18], §3.3, §8.2.3 or Osborne and Rubinstein [24], §6.3.2, §12.3 (note that [24] uses "perfect information" to refer to all games with observable actions, including those featuring simultaneous moves).

[4]Histories are sequences of consecutive action profiles.

finite collection Θ_i of *payoff types* and a payoff function $u_i : \mathcal{Z} \times \Theta \to \mathbb{R}$, where $\Theta = \Theta_1 \times ... \times \Theta_I$. Each element $\theta_i \in \Theta_i$ represents Player i's private information about the unknown payoff-relevant aspects of the game. If the set Θ contains only one element, we say that the game has *complete information*.

As the game progresses, each player is informed of the history that has just occurred. However, a player is never informed of her opponents' payoff types. The set of feasible actions for Player i may depend on previous history, but not on his private information θ_i, and it is denoted $A_i(h)$. Player i is *active* at $h \in \mathcal{H}$ if $A_i(h)$ contains more than one element. There are simultaneous moves at h if at least two players are active at h. If there is only one active player at each $h \in \mathcal{H}$, we say that the game has *perfect information*.

Moreover, we shall make use of certain derived objects. First, for every $i \in I$, we shall denote by S_i the set of *strategies* available to Player i (where a strategy is defined as a function $s_i : \mathcal{H} \to \bigcup_{h \in \mathcal{H}} A_i(h)$ such that $s_i(h) \in A_i(h)$ for all h).[5] In keeping with standard game-theoretic notation, we let $S = \prod_{i \in I} S_i$ and $S_{-i} = \prod_{j \neq i} S_j$.

For any $h \in \mathcal{H} \cup \mathcal{Z}$, $S(h)$ denotes the set of strategy profiles which induce history h; its projections on S_i and S_{-i} are denoted by $S_i(h)$ and $S_{-i}(h)$, respectively. The correspondence $S(\cdot)$ provides a convenient strategic-form representation of the information structure.

We denote by $\Sigma_i = S_i \times \Theta_i$ the set of *strategy-payoff type pairs* for Player i and let $\Sigma = \prod_{i \in I} \Sigma_i$ and $\Sigma_{-i} = \prod_{j \neq i} \Sigma_i$.

Using this notation, we can define a strategic-form payoff function $U_i : \Sigma_i \times \Sigma_{-i} \to \mathbb{R}$ in the usual way: for all $z \in \mathcal{Z}$, $(s_i, \theta_i) \in \Sigma_i$ and $(s_{-i}, \theta_{-i}) \in \Sigma_{-i}$, if $(s_i, s_{-i}) \in S(z)$, then $U_i(s_i, \theta_i, s_{-i}, \theta_{-i}) = u_i(z, (\theta_j)_{j \in I})$.

Finally, for every strategy s_i, we let $\mathcal{H}(s_i) = \{h \in \mathcal{H} : s_i \in S_i(h)\}$ denote the collection of histories consistent with s_i.

Note that the structure $(\mathcal{H}, \mathcal{Z}, I, (\Theta_i, u_i)_{i \in I})$ is not a game with incomplete information in the sense of Harsanyi [20], because it contains no description of the possible interactive beliefs about payoff types. Such description will be provided in the following subsections within a richer framework encompassing interactive beliefs conditional on (non-terminal) histories.

[5] A strategy of player i at an hypothetical stage of the game where she does not yet know her payoff type would be a map with domain $\Theta_i \times \mathcal{H}$. Here, however, we are not assuming that such an "ex ante stage" exists. Therefore, we take the point of view of a player who knows her payoff type. This simplifies the analysis and emphasizes the incomplete-information interpretation of our framework.

2.2 Conditional Beliefs and Type Spaces

As the game progresses, players update and/or revise their conjectures in light of newly acquired information. In order to account for this process, we represent beliefs by means of *conditional probability systems* (see Rênyi [28]).

Fix a player $i \in I$. For a given measure space $(X_i, \mathcal{X}_i)$, consider a non-empty collection $\mathcal{B}_i \subseteq \mathcal{X}_i$ of events such that $\emptyset \notin \mathcal{B}_i$. The interpretation is that Player i is uncertain about the "true" element $x \in X_i$, and $\mathcal{B}_i$ is a collection of observable events – or "relevant hypotheses" – concerning x.

Definition 1 *A* conditional probability system *(or CPS) on* $(X_i, \mathcal{X}_i, \mathcal{B}_i)$ *is a mapping* $\mu(\cdot|\cdot) : \mathcal{X}_i \times \mathcal{B}_i \to [0,1]$ *such that, for all* $B, C \in \mathcal{B}_i$ *and* $A \in \mathcal{X}_i$*, (1)* $\mu(B|B) = 1$*, (2)* $\mu(\cdot|B)$ *is a probability measure on* $(X_i, \mathcal{X}_i)$*, and (3)* $A \subseteq B \subseteq C$ *implies* $\mu(A|B)\mu(B|C) = \mu(A|C)$*.*

We assume that X_i is a topological space, and it is understood that $\mathcal{X}_i$ is the Borel sigma algebra on X_i. Therefore we often omit to mention $\mathcal{X}_i$ explicitly and we refer only to X_i and $\mathcal{B}_i$. The set of probability measures on X_i is denoted by $\Delta(X_i)$. The set of conditional probability systems on $(X_i, \mathcal{B}_i)$ can be regarded as a subset of $[\Delta(X_i)]^{\mathcal{B}_i}$ and is denoted by $\Delta^{\mathcal{B}_i}(X_i)$. $\Delta(X_i)$ is endowed with the topology of weak convergence of measures and $[\Delta(X_i)]^{\mathcal{B}_i}$ is endowed with the product topology.

Throughout this paper, we shall be interested solely in "relevant hypotheses" corresponding to the event that a certain partial history has occurred. Thus, Player i's *first-order* (conditional) *beliefs* about her opponents' behavior and payoff types may be represented by taking $X_i = \Sigma_{-i}$ and $\mathcal{B}_i = \{B \subseteq \Sigma_{-i} : B = S_{-i}(h) \times \Theta_{-i} \text{ for some } h \in \mathcal{H}\}$. We denote the collection of CPSs on $(\Sigma_{-i}, \mathcal{B}_i)$ thus defined by $\Delta^{\mathcal{H}}(\Sigma_{-i})$. Since S_{-i} and $\mathcal{H}$ are finite, $\Delta^{\mathcal{H}}(\Sigma_{-i})$ is easily seen to be a closed subset of the Euclidean $|\mathcal{H}| \cdot |\Sigma_{-i}|$-dimensional space.

To represent Player i's *higher-order beliefs*, we introduce the notion of an extensive-form type space. The conditional beliefs of each player j are parametrized by her *epistemic type* $t_j \in T_j$, where T_j is a compact topological space. A state of the world is an array $\omega = (\omega_j)_{j \in I} = (s_j, \theta_j, t_j)_{j \in I}$ of strategies, payoff types and epistemic types. We consider a set of "possible worlds" $\Omega = \prod_{j \in I} \Omega_j \subseteq \prod_{j \in I}(\Sigma_j \times T_j)$, where every combination $(s_j, \theta_j)_{j \in I} \in \Sigma$ occurs at some state. Player i has conditional beliefs about the strategies, payoff types and epistemic types of her opponents. Therefore the structure $(X_i, \mathcal{B}_i)$ is specified as follows:

$X_i = \prod_{j \neq i} \Omega_j = \Omega_{-i}$ and

$$\mathcal{B}_i = \{B \in \mathcal{X}_i \ : \ B = \{(s_{-i}, \theta_{-i}, t_{-i}) \in \Omega_{-i} : s_{-i} \in S_{-i}(h)\} \text{ for some } h \in \mathcal{H}\}.$$

The set of CPSs on $(\Omega_{-i}, \mathcal{B}_i)$ will be denoted by $\Delta^{\mathcal{H}}(\Omega_{-i})$.

Definition 2 *(cf. Ben Porath [9]) A* type space *on $(\mathcal{H}, S(\cdot), \Theta, I)$ is a tuple $\mathcal{T} = (\mathcal{H}, S(\cdot), \Theta, I, (\Omega_i, T_i, g_i)_{i \in I})$ such that, for every $i \in I$, T_i is a compact topological space and*

1. *Ω_i is a closed subset of $\Sigma_i \times T_i$ such that $\text{proj}_{\Sigma_i} \Omega_i = \Sigma_i$;*
2. *$g_i = (g_{i,h})_{h \in \mathcal{H}} : T_i \to \Delta^{\mathcal{H}}(\Omega_{-i})$ is a continuous mapping.*[6]
   For any $i \in I$, $g_{i,h}(t_i)$ denotes the beliefs of epistemic type t_i conditional on h.[7]

Thus, at any "possible world" $\omega = (s_i, \theta_i, t_i)_{i \in I} \in \Omega$, we specify each player i's *dispositions to act* (her strategy s_i) and *dispositions to believe* (her system of conditional probabilities $g_i(t_i) = (g_{i,h}(t_i))_{h \in \mathcal{H}}$), together with her *payoff type*. These dispositions also include what a player *would* do and think at histories that are inconsistent with ω (history h is inconsistent with, or counterfactual at, $\omega = (s, \theta, t)$ if $s \notin S(h)$).[8] We call "event" any Borel subset of Ω.

Notably absent in our definition of a type space is the description of the beliefs of a player about herself. We omit such beliefs because the epistemic assumptions appearing in our results only involve beliefs about the opponents. Thus, beliefs about oneself do not play an explicit role. But our analysis is consistent with the standard assumption that a player knows her beliefs and assigns probability one to the strategy she intends to carry out.

Type spaces encode a collection of infinite hierarchies of CPSs for each player. It is natural to ask whether there exists a type space which encodes *all* "conceivable" hierarchical beliefs. Mertens and Zamir [23] and Brandenburger and

[6] It would make sense to assume that g_i is injective, but this is immaterial for our arguments. Continuity of the mapping g_i means that the index set T_i inherits the topological structure of the beliefs set $\Delta^{\mathcal{H}}(\Omega_{-i})$. For more on this see [23] and [7].

[7] Once we have specified a type space $\mathcal{T}$, we may derive a Bayesian game *à la* Harsanyi by taking, for each epistemic type t_i, the the initial marginal beliefs over $\Theta_{-i} \times T_{-i}$. Of course, Harsanyi-consistency (i.e. the possibility to derive beliefs at each state from a common prior) is satisfied only in special cases.

[8] We comment on the role of actions at certain counterfactual histories before Definition 4.

Dekel [12] answered this question in the affirmative when beliefs are represented by probability measures on a compact or Polish space; Battigalli and Siniscalchi [7] provide a counterpart of these results in the present "dynamic" setting where beliefs are represented by CPSs.

Consider the following definition.

Definition 3 *A* belief-complete type space *on* $(\mathcal{H}, S(\cdot), \Theta, I)$ *is a type space* $\mathcal{T} = (\mathcal{H}, S(\cdot), \Theta, I, (\Omega_i, T_i, g_i)_{i \in I})$ *such that, for every* $i \in I$, $\Omega_i = \Sigma_i \times T_i$ *and the function* g_i *maps* T_i onto $\Delta^{\mathcal{H}}(\prod_{j \neq i} \Sigma_j \times T_j)$.[9]

It is shown in [7] that a belief-complete type space may always be constructed (for finite games and also for "well-behaved" infinite games) by taking the sets of epistemic types to be the collection of *all* possible hierarchies of conditional probability systems that satisfy certain intuitive coherency conditions. Also, every type space may be viewed as a belief-closed subspace of the space of infinite hierarchies of conditional beliefs.[10] Finally, since we assume that the set of external states Σ is finite and hence compact, the sets T_i $(i \in I)$ of epistemic types in the belief-complete type space thus constructed are compact topological spaces, as assumed in our definition.

2.3 Sequential Rationality

Our basic behavioral assumption is that each Player i chooses and carries out a strategy $s_i \in S_i$ that is optimal, given her payoff type θ_i and her beliefs, conditional upon any history consistent with s_i. This does not impose restrictions on the actions specified at histories that cannot obtain if Player i follows the strategy s_i. Thus, we use a sequential best response property which applies to plans of action[11] as well as strategies (see, for example, [29] and [27]).[12]

[9] We use "complete" in the same sense as Brandenburger [11], who shows (in a different framework) that a (belief-) complete, filter-theoretic type space does not exists (see also [13]). Of course, this notion of completeness is not to be confused with the topological one.

[10] [7] uses a slightly different definition of type space. But all the arguments in [7] can be easily adapted to the present framework.

[11] Intuitively, a plan of action for player i is silent about which actions would be taken by i if i did not follow that plan. Formally, a *plan of action* is a class of realization-equivalent strategies. In generic extensive games, a plan of action is a strategy of the reduced normal form.

[12] Hence, our analysis could be carried out in a more parsimonious (but less conventional) formal setup, wherein each player's behavior at a state is described by a plan of action.

Definition 4 *Fix a CPS $\mu_i \in \Delta^{\mathcal{H}}(\Sigma_{-i})$. A strategy $s_i \in S_i$ is a* sequential best reply *to μ_i for payoff type $\theta_i \in \Theta_i$ if and only if, for every $h \in \mathcal{H}(s_i)$ and every $s_i' \in S_i(h)$,*

$$\sum_{(s_{-i},\theta_{-i})\in\Sigma_{-i}} [U_i(s_i,\theta_i,s_{-i},\theta_{-i}) - U_i(s_i',\theta_i,s_{-i},\theta_{-i})]\mu_i(\{(s_{-i},\theta_{-i})\}|S_{-i}(h)\times\Theta_{-i}) \geq 0$$

For any CPS $\mu_i \in \Delta^{\mathcal{H}}(\Sigma_{-i})$, let $r_i(\mu_i)$ denote the set of pairs $(s_i,\theta_i) \in \Sigma_i$ such that s_i is a sequential best reply to μ_i for θ_i.

It can be shown by standard arguments that (a) for all (θ_i,μ_i) the set of sequential best replies to μ_i for θ_i is nonempty (i.e. $\text{proj}_{\Theta_i} r_i(\mu_i) = \Theta_i$ for all μ_i) and (b) r_i is an upper-hemicontinuous correspondence.

It is convenient to introduce the following additional notation. Fix a type space $\mathcal{T}$. For every player $i \in I$, let $f_i = (f_{i,h})_{h\in\mathcal{H}} : T_i \to [\Delta(\Sigma_{-i})]^{\mathcal{H}}$ denote her first-order belief mapping; that is, for all $t_i \in T_i$ and $h \in \mathcal{H}$,

$$f_{i,h}(t_i) = \text{marg}_{\Sigma_{-i}} g_{i,h}(t_i).$$

It is easy to see that $f_i(t_i) \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ for every $t_i \in T_i$; also, f_i is continuous.

We say that Player i is *rational* at a state ω in $\mathcal{T}$ if and only if

$$\omega \in R_i = \{(s,\theta,t) \in \Omega \ : \ (s_i,\theta_i) \in r_i(f_i(t_i))\}$$

(Note that R_i is closed because the correspondence $r_i \circ f_i$ is upper hemicontinuous. Hence R_i is an event.) We shall also refer to the events $R = \bigcap_{i\in I} R_i$ ("every player is rational") and $R_{-i} = \bigcap_{j\neq i} R_j$ ("every opponent of Player i is rational").

2.4 Conditional Belief Operators

The next building block is the epistemic notion of (conditional) *probability-one belief*, or (conditional) *certainty*. Recall that an epistemic type encodes the beliefs a player would hold, should any one of the possible histories occur. This allows us to formalize statements such as, "Player i would be certain that Player j is rational, were she to observe history h."

Let $\mathcal{A}_i$ denote the sigma-algebra of events $E \subseteq \Omega$ such that $E = \Omega_i \times \text{proj}_{\Omega_{-i}} E$. $\mathcal{A}_i$ is the collection of events concerning Player i. The collection of events concerning the opponents of Player i, $\mathcal{A}_{-i}$, is similarly defined.

The *conditional (probability-one) belief operator* for player $i \in I$ given history $h \in \mathcal{H}$ is a map $\mathrm{B}_{i,h} : \mathcal{A}_{-i} \to \mathcal{A}_i$ defined by[13]

$$\forall E \in \mathcal{A}_{-i}, \quad \mathrm{B}_{i,h}(E) = \{(s, \theta, t) \in \Omega \ : \ g_{i,h}(t_i)(\mathrm{proj}_{\Omega_{-i}} E) = 1\}.$$

For any $E \in \mathcal{A}_{-i}$, $\mathrm{B}_{i,h}(E)$ corresponds to the statement "Player i would be certain that her opponents' strategies, payoff and epistemic types are consistent with E, were she to observe history h."

For each player i and history $h \in \mathcal{H}$, the operator $\mathrm{B}_{i,h} : \mathcal{A}_{-i} \to \mathcal{A}_i$ satisfies the standard properties[14] of falsifiable beliefs (see, for example, Chapter 3 of Fagin *et al* [16]); in particular, it satisfies

- *Conjunction*: For all events $E, F \in \mathcal{A}_{-i}$, $\mathrm{B}_{i,h}(E \cap F) = \mathrm{B}_{i,h}(E) \cap \mathrm{B}_{i,h}(F)$;
- *Monotonicity*: For all events $E, F \in \mathcal{A}_{-i}$: $E \subseteq F$ implies $\mathrm{B}_{i,h}(E) \subseteq \mathrm{B}_{i,h}(F)$.

Finally, we shall often be interested in formalizing assumptions such as "Every player believes that her opponents are rational." In order to simplify notation, we introduce an auxiliary "mutual belief" operator. For any Borel subset $E \subseteq \Omega$ such that $E = \prod_{i \in I} \mathrm{proj}_{\Omega_i} E$, and for any history $h \in \mathcal{H}$, let

$$\mathrm{B}_h(E) = \bigcap_{i \in I} \mathrm{B}_{i,h}(\Omega_i \times \mathrm{proj}_{\Omega_{-i}} E).$$

For instance, if $I = \{1, 2\}$ and $E = R$, then $R = R_1 \cap R_2$ and $R_i = \Omega_{-i} \times \mathrm{proj}_{\Omega_i} R$ for $i \in I$; thus, $\mathrm{B}_h(R) = \mathrm{B}_{1,h}(R_2) \cap \mathrm{B}_{2,h}(R_1)$.

3 Strong Belief

With the basic framework and notation in place, we now turn to the main focus of this paper, the notion of strong belief. This section provides the basic definition, and illustrates the key features of strong beliefs by means of a simple example. Finally, it draws a first connection with forward-induction reasoning.

[13] For any $E \in \mathcal{A}_{-i}$, $\mathrm{B}_{i,h}(E)$ is closed, hence measurable; this follows from the continuity of $g_{i,h}$, via an application of the *portmanteau* theorem. Clearly, $\mathrm{B}_{i,h}(E) \in \mathcal{A}_i$.

[14] It is easy to extend the definition of $\mathrm{B}_{i,h}$ to *all* Borel subsets of Ω in a manner consistent with *all* properties of falsifiable beliefs: see e.g. [7]. However, this extension requires additional notation, and is irrelevant for our analysis.

3.1 Definition

We say that Player i *strongly believes* that an event $E \neq \emptyset$ is true (i.e. adopts E as a "working hypothesis") if and only if she is certain of E at all histories consistent with E.[15] Formally, for any type space $\mathcal{T}$, define the operator $\mathrm{SB}_i : \mathcal{A}_{-i} \to \mathcal{A}_i$ by $\mathrm{SB}_i(\emptyset) = \emptyset$ and

$$\mathrm{SB}_i(E) = \bigcap_{h \in \mathcal{H}:\ E \cap [h] \neq \emptyset} \mathrm{B}_{i,h}(E)$$

for all events $E \in \mathcal{A}_{-i} \backslash \{\emptyset\}$, where $[h] := \prod_{j \in I} S_j(h) \times \Theta_j \times T_j$ is the event "history h occurs."[16]

As in Subsection 2.4, it is convenient to define an auxiliary "mutual strong belief" operator. For any Borel subset $E \subseteq \Omega$ such that $E = \prod_{i \in I} \mathrm{proj}_{\Omega_i} E$, and for any history $h \in \mathcal{H}$, let

$$\mathrm{SB}(E) = \bigcap_{i \in I} \mathrm{SB}_i(\Omega_i \times \mathrm{proj}_{\Omega_{-i}} E).$$

As in the case of conditional belief, if $I = \{1, 2\}$ and $E = R$, then $\mathrm{SB}(R) = \mathrm{SB}_1(R_2) \cap \mathrm{SB}_2(R_1)$.

3.2 Belief and Strong Belief

The features of strong beliefs are best illustrated by means of a comparison with conditional beliefs. Note first that $\mathrm{SB}_i(E) \subseteq \mathrm{B}_{i,\phi}(E)$ for all $E \in \mathcal{A}_{-i}$; that is, strong belief implies initial certainty. More generally, $\mathrm{SB}_i(E) \subseteq \mathrm{B}_{i,h}(E)$ for all $E \in \mathcal{A}_{-i}$ and $h \in \mathcal{H}$ such that $[h] \cap E \neq \emptyset$.

Unlike conditional belief, strong belief does not satisfy conjunction and monotonicity. To illustrate this point (as well as others later on), we refer to a well-known game-theoretic example: the Battle of the Sexes with an outside option. The game is depicted in Figure 1.

Table I describes a type space for the game under consideration; we shall denote it by $\mathcal{T}$. Since there is complete information (each set Θ_i is a singleton), we simply omit payoff types.

[15] An analogous notion (called "absolutely robust belief") was independently put forth by Stalnaker [32].

[16] For any partial description p of the world, such as a history, a strategy, a player's beliefs, we let $[p]$ denote the set of states of the world satisfying p.

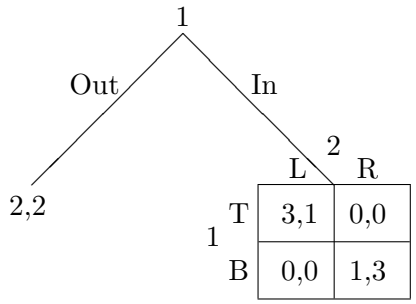

Figure 1: The Battle of the Sexes with an Outside Option

n_1	ω_1	$g_{1,\phi}(t_1)$	$g_{1,(\text{In})}(t_1)$
1	(InB, t_1^1)	0,1,0	0,1,0
2	(InT, t_1^1)	0,1,0	0,1,0
3	(OutB, t_1^1)	0,1,0	0,1,0
4	(OutT, t_1^1)	0,1,0	0,1,0
5	(InT, t_1^2)	0,0,1	0,0,1

n_2	ω_2	$g_{2,\phi}(t_2)$	$g_{2,(\text{In})}(t_2)$
1	(L, t_2^1)	0,1,0,0,0	0,1,0,0,0
2	(R, t_2^2)	0,0,1,0,0	1,0,0,0,0
3	(L, t_2^3)	0,0,0,0,1	0,0,0,0,1

Table I: The Type Space $\mathcal{T}$

The table specifies the sets $T_1 = \{t_1^1, t_1^2\}$ and $T_2 = \{t_2^1, t_2^2, t_2^3\}$ of epistemic types, the sets Ω_1, Ω_2 and $\Omega = \Omega_1 \times \Omega_2$, and the maps $g_i : T_i \to \Delta^{\mathcal{H}}(\Omega_{-i})$, as required by our definitions. Note that $\text{proj}_S\Omega = S$. It will be notationally convenient to denote pairs $\omega_i = (s_i, t_i)$ by $\omega_i^{n_i}$, where n_i is the corresponding line number in the relevant table; thus, $\omega_1^5 = (\text{InT}, t_1^2)$.

We keep using square brackets to denote events corresponding to histories or strategies. For example, in type space $\mathcal{T}$, the event "Player 1 chooses Out at ϕ" is $[\text{Out}] = \{\omega_1^3, \omega_1^4\} \times \Omega_2$. Also, the notation $[\mathbf{s}_i = s_i]$ corresponds to the event that Player i adopts strategy s_i; for instance, the event "Player 2 would choose R if the subgame were reached" is $[\mathbf{s}_2 = \text{R}] = \Omega_1 \times \{\omega_2^2\}$.

Player 2 might entertain one or both of the following initial hypotheses, corresponding to events in $\mathcal{T}$:

- "Player 1 is rational": $R_1 = \{\omega_1^3, \omega_1^4, \omega_1^5\} \times \Omega_2$.
- "Player 1 initially believes that Player 2 would play R after observing In": $\text{B}_{1,\phi}([\mathbf{s}_2 = \text{R}]) = (\Omega_1 \setminus \{\omega_1^5\}) \times \Omega_2$.

The two hypotheses *jointly* imply that Player 1 chooses Out:

$$R_1 \cap \mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}]) = \{\omega_1^3, \omega_1^4\} \times \Omega_2 = [\mathrm{Out}] = \Omega \backslash [\mathrm{In}].$$

However, R_1 and $\mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])$ are *individually* consistent with Player 1 choosing In: $R_1 \cap [\mathrm{In}] = \{\omega_1^5\} \times \Omega_2 \neq \emptyset$ and $\mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}]) \cap [\mathrm{In}] = \{\omega_1^1, \omega_1^2\} \times \Omega_2 \neq \emptyset$. Therefore $\mathrm{SB}_2(R_1) \subseteq \mathrm{B}_{2,(\mathrm{In})}(R_1)$ and $\mathrm{SB}_2\left(\mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])\right) \subseteq \mathrm{B}_{2,(\mathrm{In})}\left(\mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])\right)$.

It follows that Player 2 cannot strongly believe R_1 and strongly believe $\mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])$ in the same state, or else she would hold contradictory beliefs after In:

$$\mathrm{SB}_2(R_1) \cap \mathrm{SB}_2(\mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])) \subseteq \mathrm{B}_{2,(\mathrm{In})}([\mathrm{Out}]) = \emptyset.$$

On the other hand, Player 2 can strongly believe the *joint* hypothesis $R_1 \cap \mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])$. In particular, at ω_2^2 Player 2 initially believes $R_1 \cap \mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])$ and would *give up his belief in Player 1's rationality after In*:

$$\mathrm{SB}_2\left(R_1 \cap \mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])\right) = \Omega_1 \times \{\omega_2^2\} \neq \emptyset.$$

This shows that SB_2 does not satisfy conjuction. A similar argument shows that strong belief does not satisfy monotonicity: for instance, $R_1 \cap \mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}]) \subseteq R_1$, but $\mathrm{SB}_2(R_1 \cap \mathrm{B}_{1,\phi}([\mathbf{s}_2 = \mathrm{R}])) \not\subset \mathrm{SB}_2(R_1) = \Omega_1 \times \{\omega_2^3\}$.

3.3 Strong Belief and Forward Induction

Affinities between intuitions about forward induction and the notion of strong belief emerge clearly in the analysis of the Battle of the Sexes with an outside option.

The usual "forward-induction analysis" of this game runs as follows. Observe first that the strategy profile (OutB, R) is a subgame-perfect equilibrium. It is sustained by Player 2's implicit threat to play R in the simultaneous-moves subgame, were Player 1 to deviate and choose In at the initial history.

According to forward-induction reasoning, this threat is not credible: InB is strictly dominated for Player 1, so *if in the subgame Player 2 believes that Player 1 is rational*, he should not expect her to follow In with R. On the other hand, the subgame-perfect equilibrium (InT, L) passes the forward-induction test.

The key step in this argument is the italicized statement about Player 2's beliefs. First note that at state (ω_1^3, ω_2^2), which corresponds to the subgame-perfect equilibrium (OutB, R), the players are rational and there is initial common certainty of the opponent's rationality, that is,

$$(\omega_1^3, \omega_2^2) \in R_i \cap \mathrm{B}_{i,\phi}(R_{-i}) \cap \mathrm{B}_{i,\phi}(\mathrm{B}_{-i,\phi}(R_i)) \cap \mathrm{B}_{i,\phi}(\mathrm{B}_{-i,\phi}(\mathrm{B}_{i,\phi}(R_{-i}))) \cap \ldots \quad \text{for } i = 1, 2.$$

But, as noted above, $(\omega_1^3, \omega_2^2) \notin \mathrm{B}_{2,(\mathrm{In})}(R_1)$.

On the other hand, forward-induction reasoning suggests that Player 2's conditional beliefs following the unexpected move In should still be consistent with Player 1's rationality. To capture this intuition, assume that Player 2 strongly believes in Player 1's rationality; this leads to an epistemic characterization of the forward-induction solution. Note that $\mathrm{SB}_2(R_1) = \Omega_1 \times \{\omega_2^3\}$ and $R_2 = \Omega$; thus,

$$R_1 \cap R_2 \cap \mathrm{SB}_2(R_1) = \{(\omega_1^{n_1}, \omega_2^3) \ : \ n_1 = 3, 4, 5\}$$

If we now add the further assumption that Player 1 is initially certain that Player 2 is rational and strongly believes that Player 1 is rational, we obtain

$$R_1 \cap R_2 \cap \mathrm{SB}_2(R_1) \cap \mathrm{B}_{1,\phi}\left(R_2 \cap \mathrm{SB}_2(R_1)\right) = \{(\omega_1^5, \omega_2^3)\}$$

i.e. we identify the strategy profile (InT,L).

3.4 The Pitfalls of Incomplete Type Spaces

We wish to point out an important consequence of the fact that strong belief fails monotonicity and conjunction: *analyzing an extensive-form game in the framework of an incomplete type space introduces implicit and potentially undesirable restrictions on forward-induction reasoning.*

Consider for instance the game in Figure 1, together with the type space $\mathcal{T}'$ described in Table II.

n_1	ω_1	$g_{1,\phi}(t_1)$	$g_{1,(\mathrm{In})}(t_1)$
1	(InB, t_1^1)	0,1	0,1
2	(InT, t_1^1)	0,1	0,1
3	(OutB, t_1^1)	0,1	0,1
4	(OutT, t_1^1)	0,1	0,1

n_2	ω_2	$g_{2,\phi}(t_2)$	$g_{2,(\mathrm{In})}(t_2)$
1	(L, t_2^1)	0,1,0,0	0,1,0,0
2	(R, t_2^2)	0,0,1,0	1,0,0,0

Table II: The Type Space $\mathcal{T}'$

$\mathcal{T}'$ is a belief-closed subspace of $\mathcal{T}$. Indeed $\Omega' \subset \Omega$ and every state $\omega \in \Omega'$ corresponds to the same profile of strategies and hierarchies of CPSs in $\mathcal{T}$ and $\mathcal{T}'$. To emphasize that events and belief operators are defined within the latter type space, we write R_i', $\mathrm{SB}_i'(\cdot)$ and so forth.

The type space $\mathcal{T}'$ incorporates the assumption that Player 1, if rational, never chooses In, and that Player 2 strongly believes this. Ω' incorporates other

restrictions as well: for instance, at any state $\omega' \in \Omega'$ there is common certainty conditional on both ϕ and (In) that either Player 1 is rational or she chooses In.

Intuitively, these assumptions break the forward-induction argument: if Player 2 observes that the simultaneous-moves game is reached, he *must* conclude that Player 1 is irrational, and hence may be planning to choose B. But then Player 2 may rationally respond with R.

Formally, observe first that $R'_1 = \{\omega_1^3, \omega_1^4\} \times \Omega'_2$. Next, note that $\mathrm{SB}'_2(R'_1) = \Omega'_1 \times \{\omega_2^2\}$: since there is no state in the type space $\mathcal{T}'$ consistent both with Player 1's rationality and with the event that the subgame is reached, the assumption that Player 2 strongly believes that Player 1 is rational puts no constraint on Player 2's beliefs after (In). On the other hand, Player 2 must initially believe that Player 1 is rational, which singles out type t_2^2. It is then easy to see that

$$R'_1 \cap R'_2 \cap \mathrm{SB}'_2(R'_1) \cap \mathrm{B}'_{1,\phi}\left(R'_2 \cap \mathrm{SB}'_2(R'_1)\right) = \{(\omega_1^3, \omega_2^2), (\omega_1^4, \omega_2^2)\},$$

where both (ω_1^3, ω_2^2) and (ω_1^4, ω_2^2) yield outcome Out: by *restricting* the type space, we make Out consistent with forward induction!

To relate this to the properties of strong belief, note that $R'_1 = R_1 \cap \Omega'$; therefore

$$\mathrm{SB}'_2(R'_1) = \mathrm{SB}_2(R_1 \cap \Omega') = \Omega_1 \times \{\omega_2^2\} \neq \mathrm{SB}_2(R_1) \cap \mathrm{SB}_2(\Omega') = \emptyset$$

(a failure of conjunction) and

$$R'_1 \cap \mathrm{SB}'_2(R'_1) = (R_1 \cap \Omega') \cap \mathrm{SB}_2(R_1 \cap \Omega') \not\subseteq R_1 \cap \mathrm{SB}_2(R_1)$$

(a failure of monotonicity).

In general, our epistemic assumptions reflecting forward-induction reasoning interact with the restrictions on beliefs implicit in the belief-incomplete type space $\mathcal{T}'$. The violations of conjunction and monotonicity exhibited here mirror this interaction.

The type space $\mathcal{T}'$ is not "rich enough" to capture the intuitive forward induction argument in this example. In general, we need to ensure that our epistemic analysis of forward induction is not biased by extraneous (and perhaps non-transparent) restrictions on the players' hierarchical beliefs. Since any belief-incomplete type space incorporates such restrictions, adopting a belief-complete type space is the simplest way to avoid potential biases.[17]

[17]For more on this see Section 6.

4 Iterated Strong Beliefs and Rationalizability

We now turn to the implications of *iterated strong beliefs* about the players' rationality. The main result of this section is an epistemic characterization of extensive-form rationalizability. We also present related results on backward induction and the relationship between extensive-form rationalizability and common certainty of rationality at a given history.

Throughout this section, we adopt a standard notation for the n-fold composition of operators. Fix a map $\mathcal{O} : \mathcal{A} \to \mathcal{A}$; then, for any event $E \in \mathcal{A}$, let $\mathcal{O}^0(E) = E$ and, for $n \geq 1$, let $\mathcal{O}^n(E) = \mathcal{O}(\mathcal{O}^{n-1}(E))$. We begin with a caveat on the subtle issues one has to deal with when defining iterations involving the strong belief operator.

4.1 A Caveat on Iterated Strong Beliefs

The epistemic analysis of static games with complete information shows that a strategy profile s survives $n+1$ steps of iterated (maximal) deletion of dominated strategies if and only if it is consistent with mutual certainty of rationality of order n, i.e. if and only if there exists a profile of epistemic types t (in some type space) such that $(s,t) \in \bigcap_{m=0}^{n} \mathrm{B}^m(R)$, where B is the mutual certainty operator. Similar results involving mutual belief in rationality at a specific history can be proved for dynamic games with complete or incomplete information (see [9] and [7]).

A formal analogy with such results might suggest considering assumptions of the form $\bigcap_{m=0}^{n} \mathrm{SB}^m(R)$. However, consider the event

$$\bigcap_{m=0}^{2} \mathrm{SB}^m(R) = \bigcap_{i \in I} \left(R_i \cap \mathrm{SB}_i(R_{-i}) \cap \mathrm{SB}_i \left(\bigcap_{j \neq i} \mathrm{SB}_j(R_{-j}) \right) \right).$$

The key observation is that, although the $\mathrm{SB}(R)$, $\mathrm{SB}(\mathrm{SB}(R))$ and $\mathrm{SB}(R \cap \mathrm{SB}(R))$ are nonempty in any belief-complete model, it may still be the case that $\mathrm{SB}(R) \cap \mathrm{SB}(\mathrm{SB}(R)) = \emptyset$. Thus, one may have $\bigcap_{m=0}^{2} \mathrm{SB}^m(R) = \emptyset$. This is an instance of the general observation that the strong belief operator need not satisfy the conjunction property (see Section 3).

The game in Figure 2 offers an example. It can be checked that (in a belief-complete model) $\mathrm{proj}_S R = (S_1 \setminus \{\mathrm{A_1D_2}\}) \times (S_2 \setminus \{\mathrm{a_1a_2}\})$ and $\mathrm{proj}_S R \cap \mathrm{SB}(R) = \{\mathrm{D_1D_2}, \mathrm{D_1A_2}\} \times \{\mathrm{a_1d_2}\}$.[18] Although history $(\mathrm{A_1})$ is consistent with R_1 and with

[18]Formally, both equalities follow from Proposition 6. The intuition is that $\mathrm{A_1D_2}$ is strictly

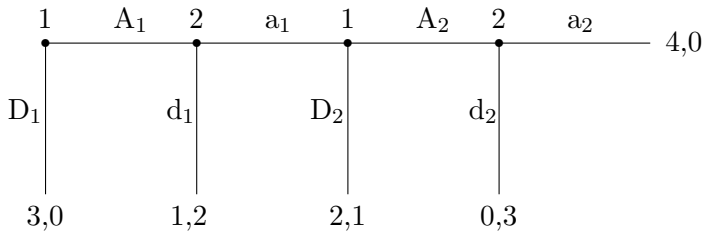

Figure 2: A perfect-information game (Reny [27])

the assumption $\mathrm{SB}_1(R_2)$ (which, by itself, has no behavioral implications), it is clearly inconsistent with $R_1 \cap \mathrm{SB}_1(R_2)$; thus, Player 2 cannot assign probability one to both R_1 and $\mathrm{SB}_1(R_2)$ conditional upon observing A_1, which implies that $\mathrm{SB}(R) \cap \mathrm{SB}(\mathrm{SB}(R)) = \emptyset$.

4.2 Strong Belief and the Best-Rationalization Principle

The *best-rationalization principle* (Battigalli [4]) requires that players' beliefs conditional upon observing a history $h \in \mathcal{H}$ be consistent with the highest degree of "strategic sophistication" of their opponents.

Our analysis clarifies what is meant by "strategic sophistication" in terms of interactive beliefs. Moreover, it illustrates how iterated strong beliefs may be employed to formulate assumptions about the players' belief revision policy—in this case, to ensure that they attribute the highest degree of strategic sophistication to their opponents at each history.

We introduce an auxiliary operator that allows us to express complex events concerning interactive strong beliefs in a compact way. Define

$$\mathrm{CSB}(E) = E \cap \mathrm{SB}(E)$$

for any event $E \subseteq \Omega$ such that $E = \prod_{i \in I} \mathrm{proj}_{\Omega_i} E$. Thus, $\mathrm{CSB}(E)$ is the set of states where each player i strongly believes E_{-i}, and such beliefs happen to be *correct*. Hence the notation CSB. Also let $\mathrm{CSB}^\infty(E) = \bigcap_{n \geq 0} \mathrm{CSB}^n(E)$.

Note that operator CSB inherits the non-monotonicity of the strong belief operator. However, since CSB satisfies the Truth axiom ($\mathrm{CSB}(E) \subseteq E$), the iterations of CSB give rise to a weakly decreasing sequence of events. Therefore

dominated for Player 1, and $\mathrm{a}_1\mathrm{a}_2$ is not sequentially rational; the further assumption that players strongly believe that these strategies will not be chosen eliminates $\mathrm{A}_1\mathrm{A}_2$, $\mathrm{d}_1\mathrm{d}_2$ and $\mathrm{d}_1\mathrm{a}_2$.

$\bigcap_{m=0}^{n} \mathrm{CSB}^m(E) \neq \emptyset$ whenever $\mathrm{CSB}^m(E) \neq \emptyset$ for each $m = 0, ..., n$, and the difficulties described in Subsection 4.1 do not arise.

For every $n \geq 0$, we associate the event $\mathrm{CSB}^n(R)$ with n-th order strategic sophistication:

A minimally sophisticated player is simply rational: $\mathrm{CSB}^0(R) = R$.

A first-order strategically sophisticated player is rational, and also maintains whenever possible the hypothesis that her opponents are rational: $\mathrm{CSB}^1(R) = R \cap \mathrm{SB}(R)$.

More interestingly, a second-order strategically sophisticated player is rational, and maintains the hypothesis that her opponents are first-order strategically sophisticated until the latter is contradicted by the evidence. However, when this happens, *she switches to the assumption that her opponents are simply rational*, and maintains this hypothesis until it, too, is contradicted. Formally, this corresponds to the event

$$\mathrm{CSB}^2(R) = R \cap \mathrm{SB}(R) \cap \mathrm{SB}(\mathrm{CSB}^1(R)) = \bigcap_{i \in I} R_i \cap \mathrm{SB}_i(R_{-i}) \cap \mathrm{SB}_i \left(R_{-i} \cap \bigcap_{j \neq i} \mathrm{SB}_j(R_{-j}) \right).$$

In the game of Figure 2, at any state $\omega \in \mathrm{CSB}^2(R)$ Player 2 believes *at the initial node* that Player 1 is rational *and* that Player 1 strongly believes that her opponent is rational. However, as soon as Player 2 observes A_1, he abandons the assumption $\mathrm{SB}_1(R_2)$ but retains the assumption R_1.

More generally, for every $n \geq 0$,

$$\mathrm{CSB}^n(R) = R \cap \bigcap_{m=0}^{n-1} \mathrm{SB}(\mathrm{CSB}^m(R)); \quad \text{also,} \quad \mathrm{CSB}^\infty(R) = R \cap \bigcap_{n \geq 0} \mathrm{SB}(\mathrm{CSB}^n(R))$$

which may now be seen to capture the intuition behind the best-rationalization principle.

The main result of this section states that rationality and the best-rationalization principle completely characterize extensive-form rationalizability (Pearce [25] and Battigalli [5]). The following is an extension of this solution procedure to the present incomplete-information framework.

Definition 5 *Consider the following procedure.*

(Step 0) *For every $i \in I$, let $\Sigma_i^0 = \Sigma_i$. Also, let $\Sigma_{-i}^0 = \prod_{j \neq i} \Sigma_i^0$ and $\Sigma^0 = \prod_{i \in I} \Sigma_i^0$.*

(Step $n > 0$) *For every $i \in I$, and for every $s_i \in S_i$ and $\theta_i \in \Theta_i$, let $(s_i, \theta_i) \in \Sigma_i^n$ if and only if $(s_i, \theta_i) \in \Sigma_i^{n-1}$ and there exists a CPS $\mu \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ such that*

1. *$(s_i, \theta_i) \in r_i(\mu)$;*
2. *$\forall h \in \mathcal{H}$, if $\Sigma_{-i}^{n-1} \cap [S_{-i}(h) \times \Theta_{-i}] \neq \emptyset$, then $\mu(\Sigma_{-i}^{n-1} | S_{-i}(h) \times \Theta_{-i}) = 1$.*

Also let $\Sigma_{-i}^n = \prod_{j \neq i} \Sigma_i^n$ and $\Sigma^n = \prod_{i \in I} \Sigma_i^n$.

Finally, let $\Sigma^\infty = \bigcap_{k \geq 0} \Sigma^n$. The profiles of strategies and payoff types in Σ^∞ are said to be extensive-form rationalizable.

In the Battle of the Sexes with an outside option $\Sigma^\infty = \Sigma^3 = \{(\mathrm{InT}, \mathrm{L})\}$, while in the game of Figure 2, $\Sigma^\infty = \Sigma^2 = \{\mathrm{D_1D_2}, \mathrm{D_1A_2}\} \times \{\mathrm{a_1d_2}\}$.

Proposition 6 *For any belief-complete type space, (i) $\Sigma^{n+1} = \mathrm{proj}_\Sigma \mathrm{CSB}^n(R)$ for all $n \geq 0$, and (ii) $\Sigma^\infty = \mathrm{proj}_\Sigma \mathrm{CSB}^\infty(R)$.*

We emphasize that Proposition 6 should *not* be viewed as providing unqualified support to extensive-form rationalizability. Rather, it is intended to clarify the epistemic assumptions underlying this solution concept, and hence enable potential users to judge whether or not these assumptions are appropriate, or plausible, in a specific situation.

To illustrate this point, consider the game form in Figure 2, and replace the payoff vectors (3,0), (1,2), (2,1), (0,3) and (4,0) with $(1,0)$, $(0,2)$, $(3,1)$, $(2,4)$ and $(5,3)$ respectively; the resulting game is a four-legged Centipede (cf. [24], pp 106-7). The extensive-form rationalizability solution $\Sigma^4 = \{\mathrm{D_1D_2}, \mathrm{D_1A_2}\} \times \{\mathrm{d_1d_2}, \mathrm{d_1a_2}\}$ is obtained in four steps, and corresponds to the backward-induction solution in terms of outcome and conditional first-order beliefs. Proposition 6 implies that $(s_1, s_2) \in \Sigma^4$ if and only if there are epistemic types t_1 and t_2 such that, for $i = 1, 2$

$$\begin{aligned}(s_i, t_i) \quad \in \quad & \mathrm{proj}_{\Omega_i} \mathrm{CSB}^3(R) = \\ & R_i \cap \mathrm{SB}_i(R_{-i}) \\ & \cap \mathrm{SB}_i\left(R_{-i} \cap \mathrm{SB}_{-i}(R_i)\right) \\ & \cap \mathrm{SB}_i\left(R_{-i} \cap \mathrm{SB}_{-i}(R_i) \cap \mathrm{SB}_{-i}\left(R_i \cap \mathrm{SB}_{-i}(R_i)\right)\right).\end{aligned}$$

The events R_1, $R_1 \cap \mathrm{SB}_1(R_2)$ and $R_1 \cap \mathrm{SB}_1(R_2) \cap \mathrm{SB}_1(R_2 \cap \mathrm{SB}_2(R_1))$ correspond to increasing degrees of strategic sophistication of Player 1. The first entails

no restrictions on the latter's behavior; the second rules out the strategy A_1A_2, because R_2 rules out a_1a_2; finally, the third forces the choice of D_1 at the initial node, because $R_2 \cap SB_2(R_1)$ rules out a_1d_2.

Now take the point of view of Player 2. At the initial node, he attributes the highest degree of strategic sophistication to Player 1, and expects her to choose D_1. If, however, Player 1 were to choose A_1 at the initial node, Player 2 would attribute her the second-highest degree of strategic sophistication, in accordance with the best-rationalization principle. This implies that Player 2 would expect Player 1 to choose D_2 at the third node; hence, Player 2 would best-respond by choosing d_1. Anticipating this, Player 1 will choose D_1 at the initial node.

The example illustrates how iterations of the CSB operator formalize the best-rationalization principle. However, it also illustrates that the latter embodies rather strong assumptions about the players' belief-revision policies.

We can also clarify the connection between extensive-form rationalizability and common certainty of rationality at a given history h: if a history h is consistent with extensive-form rationalizability, then it is also consistent with rationality and common certainty of rationality.

Proposition 7 *For all histories $h \in \mathcal{H}$, in any belief-complete type space, $\Sigma^\infty \cap [S(h) \times \Theta] \neq \emptyset$ implies $[h] \cap \bigcap_{n \geq 0} B_h^n(R) \neq \emptyset$.*

Note that Proposition 7 only provides a *sufficient* condition. Reny [26] provides an example where a non-extensive-form-rationalizable history is consistent with common certainity of rationality.

4.2.1 Strong Belief and Backward Induction

Battigalli [5] shows that, in generic games with perfect and complete information, extensive-form rationalizability is outcome-equivalent to backward induction (for a related result, see Reny [27]). Note that, since Σ is finite and $\Sigma^{n+1} \subseteq \Sigma^n$, there is some $N \geq 0$ such that $\Sigma^\infty = \Sigma^N$. Hence, Proposition 6 also provides a set of *sufficient* epistemic conditions for the backward induction outcome:

Proposition 8 *Suppose the game under consideration has complete and perfect information and no player is indifferent among payoffs at different terminal nodes. Then there exists an integer $N \geq 0$ such that for any belief-complete type space, any strategy profile $s \in \text{proj}_S \text{CSB}^N(R)$ induces the unique backward-induction outcome.*

Our results provide an explicit set of assumptions about the players' beliefs revision processes leading to backward-induction play. But it should be noted that these assumptions do *not* imply that a player at a non-rationalizable history/node would play and/or expect the backward-induction continuation. Indeed, in certain games this is actually *inconsistent* with the forward-induction logic of the best-rationalization principle (cf. Reny [27]). For example, in the game of Figure 2, backward-induction reasoning implies that Player 2, upon being reached, should expect Player 1 to choose D_2 at her next node; as we noted above, our assumptions imply that Player 2 rules out D_2, because A_1D_2 is strictly dominated by D_1D_2 for Player 1, whereas A_1A_2 may at least be justified by the "unsophisticated" belief that Player 2 will irrationally play a_1a_2.

5 Strong Belief and the Intuitive Criterion

The strong belief operator may also be used to analyze equilibrium refinements motivated by forward-induction considerations. As an example, in this section we provide an epistemic characterization of the Intuitive Criterion (Cho and Kreps [15]).

Consider a (finite) signaling game: Player 1 (the *Sender*) is active at the first stage and Player 2 (the *Receiver*) is active at the second stage; the payoff type of Player 1 is unknown, while the payoff type of Player 2 is known, thus, we may write $\Theta_1 = \Theta$. For the sake of simplicity, we assume that the set of feasible actions of the Sender does not depend on her payoff-type and that the set of feasible actions for the Receiver does not depend on the Sender's action.[19] Table III summarizes our notation for signalling games.

The actions of the Sender will be referred to as messages or signals; those of the Receiver will also be called responses.

In this framework, an external state is given by a tuple $\sigma = (m, \theta, s_2) \in M \times \Theta \times A^M$ and a state of the world is a tuple (σ, t_1, t_2) where t_1 and t_2 are—respectively—the epistemic types of the Sender and Receiver in a (belief-complete) type space based on $\Sigma = \Sigma_1 \times \Sigma_2$ and $\mathcal{H}$. We say that outcome ζ is π_0-feasible if there is a behavioral profile (π_1, π_2) such that (π_0, π_1, π_2) generates ζ. With a slight abuse of notation we denote the marginal and conditional probabilities derived from ζ as follows: $\zeta(\theta)$, $\zeta(m)$, $\zeta(m, a)$, $\zeta(m|\theta)$, $\zeta(m, a|\theta)$, $\zeta(\theta|m)$, $\zeta(a|m)$.

[19] The first assumption is already part of the (relatively) general framework adopted here. Removing these assumptions is straightforward but requires a more complex notation.

Object	**Notation**	**Remarks**
Payoff-Types for Player 1	$\theta \in \Theta = \Theta_1$	
Actions, Behavioral Strategies	$m \in M = A_1, \quad \pi_1(\cdot) \in [\Delta(M)]^{\Theta}$ $a \in A = A_2, \quad \pi_2(\cdot\|\cdot) \in [\Delta(A)]^{M}$	$S_1 = M$, $\Sigma_1 = M \times \Theta$ $\Sigma_2 = S_2 = A^M$
Histories	$\mathcal{H} = \{\phi\} \cup M$	
Player 2's prior about θ	$\pi_0 \in \Delta^0(\Theta)$	$\pi_0(\theta) > 0$ for all $\theta \in \Theta$.
Player 2's belief system	$\nu(\cdot\|\cdot) \in [\Delta(\Theta)]^M$	
Outcome or outcome distribution	$\zeta \in \Delta(\Theta \times M \times A)$	

Table III: Notation for Signalling Games.

Note that if ζ is π_0-feasible $\zeta(m|\theta)$ and $\zeta(m, a|\theta)$ are always well defined, because $\zeta(\theta) = \pi_0(\theta) > 0$ for all θ; moreover, $\zeta(a|m, \theta) = \zeta(a|m)$.

Definition 9 *A π_0-feasible outcome ζ is a* self-confirming-equilibrium outcome *if there is a $|\Theta|$ -tuple of behavioral strategies $\left(\pi_2^\theta\right)_{\theta\in\Theta}$ (where $\pi_2^\theta \in [\Delta(A)]^M$) such that, for all $\theta \in \Theta$, $m \in M$, $a \in A$,*
(1) if $\zeta(m|\theta) > 0$, then $m \in \arg\max_{m'} \sum_{a'} \pi_2^\theta(a'|m')u_1(\theta, m', a')$,
(2) if $\zeta(m, a) > 0$, then $a \in \arg\max_{a'} \sum_{\theta'} \zeta(\theta'|m)u_2(\theta', m, a')$,
(3) if $\zeta(m) > 0$, then $\pi_2^\theta(a|m) = \zeta(a|m)$.

Our definition of self-confirming-equilibrium outcome agrees with the definition of self-confirming equilibrium with unitary beliefs put forward by Fudenberg and Levine [17], if each incarnation θ of the Sender is regarded as an individual player selected by chance with probability $\pi_0(\theta)$. The behavioral strategy π_2^θ is to be interpreted as a conjecture of incarnation θ of the Sender about the Receiver. Clearly, every sequential-equilibrium outcome is also a self-confirming-equilibrium outcome. But the converse does not hold, because in a self-confirming-equilibrium outcome the (randomized) choices of different types may be justified by different conjectures about Player 2, and actions following off-equilibrium messages need not be optimal. Cho and Kreps [15] put forward the Intuitive Criterion as a test for sequential-equilibrium outcomes, but clearly the same criterion can be naturally be applied to self-confirming-equilibrium outcomes (cf. Kohlberg [21], p 23, footnote 17).

For any π_0-feasible outcome ζ, we let $u_1^\zeta(\theta) = \sum_{m,a} \zeta(m, a|\theta)u_1(\theta, m, a)$ denote

the expected payoff for type θ. For any subset of types $\emptyset \neq \Theta' \subseteq \Theta$ and message m, $BR_2(\Theta', m)$ is the set of best responses to beliefs concentrated on Θ' given message m. Consider the following procedure.

Definition 10 *[Intuitive Criterion] Fix a self-confirming-equilibrium outcome ζ and a message $m \in M$ such that $\zeta(m) = 0$. Let*

$$\overline{\Theta}(m;\zeta) = \left\{ \theta \in \Theta : u_1^{\zeta}(\theta) > \max_{a \in BR_2(\Theta, m)} u_1(\theta, m, a) \right\}$$

$$A(m;\zeta) = \begin{cases} BR_2(\Theta \setminus \overline{\Theta}(m;\zeta), m) & \overline{\Theta}(m;\zeta) \neq \emptyset \\ BR_2(\Theta, m) & \overline{\Theta}(m;\zeta) = \emptyset \end{cases}$$

Outcome ζ satisfies the Intuitive Criterion *(IC) if and only if, for every message $m \in M$ with $\zeta(m) = 0$ and every payoff-type $\theta \in \Theta$, there exists an action $a \in A(m;\zeta)$ such that $u_1(\theta, m, a) \leq u_1^{\zeta}(\theta)$.*

Informally, a candidate outcome *fails* the Intuitive Criterion if a Sender's type may deviate to an off-equilibrium message and expect to obtain a higher payoff than she receives according to ζ, provided that the Receiver applies forward induction whenever he observes an unexpected message (cf. [15]).

Our objective is to clarify the epistemic assumptions leading to this criterion. Cho and Kreps [15] argue that

> "the Intuitive Criterion relies heavily on the common knowledge of the fixed candidate equilibrium outcome and, in particular, attaches a very specific meaning (a conscious attempt to break that equilibrium) to defections from the supposed equilibrium."

Thus, the equilibrium path plays a different role than the specification of off-equilibrium-path behavior and beliefs. To anticipate, our characterization states that outcome ζ satisfies the Intuitive Criterion if and only if the assumption that the Receiver's beliefs "agree" with ζ is consistent with certain assumptions involving initial belief and strong belief.

Say that Player i's beliefs *agree with outcome* ζ at state $(\sigma_i, t_i, \omega_{-i})$ if $f_{i,\phi}(t_i)$ (the initial first-order beliefs of t_i) yields the same (conditional) probabilities as ζ. In particular, the event "the Sender's beliefs agree with ζ" is

$$\begin{array}{rcl} [\zeta]_1 & = & \{(\sigma_1, t_1, \omega_2) \in \Omega : \\ \forall m & \in & M, \forall a \in A, \zeta(m) > 0 \Rightarrow f_{1,\phi}(t_1)\left(\{s_2 : s_2(m) = a\}\right) = \zeta(a|m)\}. \end{array}$$

Similarly, the event "the Receiver's beliefs agree with ζ" is

$$[\zeta]_2 = \{(\omega_1, s_2, t_2) \in \Omega : \forall (\theta, m) \in \Sigma_1, f_{2,\phi}(t_2)(\{(\theta, m)\}) = \zeta(\theta, m)\}.$$

Part (1) of the following proposition is a preliminary step of some independent interest, similar in spirit to Theorem A in Aumann and Brandenburger [3].

Part (2) provides two alternative (but closely related) characterizations of the Intuitive Criterion.

Define the following events:

$$\begin{aligned} IC_1^\zeta &\equiv R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2) \\ IC_2^\zeta &\equiv R_2 \cap [\zeta]_2 \cap \mathrm{SB}_2(IC_1^\zeta) \\ IC^\zeta &\equiv R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(IC_2^\zeta) \end{aligned}$$

Proposition 11 *Fix a π_0-feasible outcome ζ.*
(1) If $\bigcap_{i=1,2} [\zeta]_i \cap \mathrm{B}_{i,\phi}\left(R_{-i} \cap [\zeta]_{-i}\right) \neq \emptyset$ in some type space, then ζ is a self-confirming-equilibrium outcome.
(2) The following statements about outcome ζ are equivalent:
(a) For any belief-complete type space, $\{(\theta, m) : \zeta(\theta, m) > 0\} \subseteq \mathrm{proj}_{\theta \times M}\, IC^\zeta$;
(b) For any belief-complete type space, $[\zeta]_2 \cap \mathrm{B}_{2,\phi}(IC^\zeta) \neq \emptyset$;
(c) ζ is a self-confirming-equilibrium outcome satisfying the Intuitive Criterion.

We provide some intuition for the characterization of the Intuitive Criterion via the event IC^ζ. First, strong belief in the event $IC_1^\zeta = R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2)$ captures the forward-induction assumption that, upon observing an off-equilibrium message $m \in M$, the Receiver's beliefs are concentrated on types for which m is not equilibrium-dominated, given that the Sender does not expect the Receiver to choose conditionally dominated actions. Second, at each state $\omega \in IC^\zeta$, the Sender is rational, and her initial beliefs agree with ζ; moreover, she expects the Receiver to play a best response to a belief consistent with equilibrium domination if she chooses to "deviate" from ζ. Then characterization (2.a) follows: ζ passes the Intuitive Criterion if and only if, for every (θ, m) with $\zeta(\theta, m) > 0$, there exists a state $\omega \in IC^\zeta$ in which the Sender's type is θ and she sends message m. (2.b) is essentially a restatement of (2.a) in terms of the initial beliefs of the Receiver.

6 Discussion

6.1 Extensions

More general information structures and characterizations. In order to focus on the properties and applications of strong belief, we have confined our analysis of extensive-form rationalizability to the simplified setting of games with observable actions. However, Proposition 6 immediately extends to general extensive games. Moreover, the result can be generalized in order to incorporate "exogenous restrictions" on players' first order beliefs in the sense that, for any event F about the players' first order beliefs, one can obtain the behavioral implications of the epistemic assumptions $\text{CSB}^n(R \cap F)$ $(n = 1, 2, ...)$ with a modification of the extensive-form rationalizability procedure. We refer the interested reader to the previous version of this paper [6] and to [8].

Other results on refinements. Similarly, the techniques employed in our analysis of the Intuitive Criterion may be adapted to study other refinements for signalling games. For example, equilibrium dominance is characterized as follows. Let $ED_2^\zeta \equiv R_2 \cap [\zeta]_2 \cap \text{SB}_2(R_1 \cap [\zeta]_1)$ and $ED^\zeta = R_1 \cap [\zeta]_1 \cap \text{B}_{1,\phi}(ED_2^\zeta)$. Then ζ is a self-confirming equilibrium outcome satisfying the test of equilibrium dominance if and only if $\{(\theta, m) : \zeta(\theta, m) > 0\} \subseteq \text{proj}_{\theta \times M}\ ED^\zeta$, and the latter holds if and only if $[\zeta]_2 \cap \text{B}_{2,\phi}(ED^\zeta) \neq \emptyset$.

In [8], we build on a generalization of our Proposition 6 to provide an epistemic characterization of the Iterated Intuitive Criterion. We conjecture that other forward-induction refinements for signalling games and more general incomplete incomplete information games may also be analyzed using a combination of the techniques presented in Sections 4 and 5.

Beliefs about oneself. In the analysis of static games, it is standard to assume that a player knows her epistemic type and strategy (of course, in an incomplete information setting she also knows her payoff type). We could adapt this assumption to our analysis of dynamic games in several ways. In [6] and [7] we assume that players know their epistemic types and, if rational, they assign probability one to their plan of action at each history consistent with it. Versions of our results can be proved for these extended epistemic models.

6.2 Related Literature

Extensive–Form Type Spaces. Finite (hence incomplete) extensive-form type spaces are introduced in Ben Porath [9] to characterize common certainty of rationality at the beginning of a perfect information game. Battigalli and Siniscalchi [7] provide a general analysis of (finite and infinite) type spaces for extensive-form games and show the existence of a belief-complete type space, a building block of our analysis.

Belief Revision. Belief revision (mostly in a single-person setting) has been studied extensively in the philosophy literature. See, e.g., Gärdenfors [19] and references therein.

In that literature, the following alternative framework for belief revision is often employed. First, one fixes a logically closed set of propositions, called a *belief set*, that an individual accepts as true. Then, for each proposition that the individual may subsequently learn to be true, one considers a corresponding logically closed belief set, representing the individual's new epistemic state. Belief revision is defined via axioms relating prior and posterior belief sets.

A similar construction and a characterization of strong belief in terms of belief sets can be carried out in the present setting, with the proviso that each player can only learn about the occurrence of a history $h \in \mathcal{H}$.

Belief Revision in Games and Forward Induction. Our paper is related to work by Stalnaker and Board. Stalnaker [31] puts forward a normal-form, finite epistemic model, which can also be used to analyze extensive-form reasoning. This model is used by Stalnaker [32] to provide a brief discussion of forward induction and by Board [10] to characterize some extensive-form solution concepts, including extensive-form rationalizability. Some comments on the relationship between our work and Stalnaker's are warranted.

From a substantive viewpoint, Stalnaker [32] independently proposes a notion of "absolutely robust belief" that is analogous to our strong belief. He employs this notion to sketch a characterization of the following procedure: perform two rounds of elimination of weakly dominated strategies, followed by iterated strict dominance. In some simple games (such as the one we analyze in Section 3), this procedure singles out the forward-induction outcome.

Our analysis employs the notion of strong belief to analyze *different* solution concepts, i.e. extensive-form rationalizability and the Intuitive Criterion.[20] In this respect, Stalnaker's [32] result complements ours.

[20]Stalnaker's iterative procedure is clearly unrelated to the Intuitive Criterion. It also differs

From a technical standpoint, the main difference between our type spaces and Stalnaker's epistemic model is that, for each state, our model specifies beliefs conditional on *observable* events only, while Stalnaker's model specifies beliefs conditional on *every* event, including unobservable events concerning the beliefs of the players. This prevents the construction of belief-complete models by standard methods.[21] Stalnaker and Board are thus forced to qualify their characterization results with the proviso that the incomplete model at hand is "sufficiently rich" to allow for forward-induction reasoning in the game under consideration (see the discussion in Section 3.4). This is made precise by Board [10]. However, this notion of "richness" depends crucially on the *payoffs* of the game, as well as on the specific *solution concept* one wishes to characterize. Finally, characterizing the notion of richness in any given context is somewhat cumbersome. Adopting belief-complete type spaces makes it possible to avoid these complications altogether.

Normal-Form Solution Concepts and Forward-Induction Outcomes. The present paper focuses on explicit forward-induction *reasoning* in extensive games. However, forward-induction *strategies* are also selected by appropriate normal-form solution concepts. Epistemic characterizations of two such solution concepts are discussed below.

It is well-known that iterated weak dominance supports forward-induction outcomes in several games (see, e.g., [24], section 6.6). Brandenburger and Keisler [14] provide an epistemic characterization of this solution concept. In their model, players' types correspond to lexicographic sequences (a generalization of lexicographic probability systems that allows for an uncountable state space) over the set of opponents' strategies and types. We note that, as in our paper, a crucial ingredient in the analysis is the existence of belief-complete type spaces of this

---

from extensive-form rationalizability: for instance, the latter selects the forward-induction outcome in the game Burning Money, whereas Stalnaker's procedure does not. Also observe that, in a large class of games, extensive-form rationalizability is equivalent to iterated weak dominance (Battigalli [5]).

As our analysis indicates, extensive-form rationalizability is based on the assumption that, at any history, a player's beliefs about her opponents are consistent with the highest *degree of strategic sophistication* compatible with observed game play. Stalnaker's procedure is based on the simpler assumption that players believe that their opponents are rational, whenever this is compatible with observed game play.

The notion of "degrees of strategic sophistication" does not play any rôle in Stalnaker's analysis. On the other hand, it is central to our characterization of extensive-form rationalizability. Section 4.1 indicates that formalizing this notion requires some care.

[21] We are not aware of any proof of existence of belief-complete models à la Stalnaker.

nature, which Brandenburger and Keisler also establish.

In the context of a finite, normal-form epistemic model, Asheim and Dufwenberg [1] propose a notion of "full admissible consistency" of a player's preferences with the game being played and with the preferences of his opponent. Correspondingly, they define a solution concept, "full permissibility", which selects the forward-induction outcome in games such as the Battle of the Sexes with an outside option. Their main result shows that common "certain belief" of full admissible consistency characterizes fully permissible sets of strategies. The authors provide a thorough discussion of the differences and similarities between standard forward-induction arguments and full permissibility, as well as between the latter and iterated weak dominance: see [1], section 5.1.

Backward Induction. Aumann [2] originated a literature where partitional epistemic models are used to provide sufficient conditions for the backward-induction strategies, or path, in generic perfect-information games (see Section 6 in [7] for a discussion of this literature). We emphasize that, as was noted in Section 4, Proposition 8 only provides sufficient conditions for the backward-induction *path.*

(Iterated) Intuitive Criteron. Sobel *et al.* ([30], Proposition 2) relate the iterated intuitive criterion (IIC) to extensive-form rationalizability in a modified signalling game. Thus, their result concerns the equivalence of certain iterative deletion procedures. The epistemic characterization of the IIC we provide in [8] partially builds on their work. Christian Ewerhart (private communication) also provides a non-epistemic analysis of the Intuitive Criterion, whereby players' "assumptions" about each other are represented by sets of strategy profiles. The suggested interpretation of these "assumptions" is reminiscent of the events appearing in Proposition 11.

7 Appendix: Proofs

Lemma 12 *Fix a map $\tau_{-i} : \Sigma_{-i} \to T_{-i}$. Also, fix a first-order CPS $\delta \in \Delta^{\mathcal{H}}(\Sigma_{-i})$. Then there exists an epistemic type $t_i \in T_i$ such that, for each $h \in \mathcal{H}$, $g_{i,h}(t_i)$ has finite support and*

$$g_{i,h}(t_i)\left((\sigma_{-i}, \tau_{-i}(\sigma_{-i}))\right) = \delta(\sigma_{-i}|\Sigma_{-i}(h))$$

for all $\sigma_{-i} \in \Sigma_{-i}$.

Proof. Define a *candidate* CPS μ on $\Sigma_{-i} \times T_{-i}$ by setting

$$\mu\left(\left\{(\sigma_{-i}, \tau_{-i}(\sigma_{-i}))\right\} | \Sigma_{-i}(h) \times T_{-i}\right) = \delta(\sigma_{-i} | \Sigma_{-i}(h))$$

for every $h \in \mathcal{H}$, and extending the assignments by additivity. Properties (1) and (2) in Definition 1 follow immediately from the observation that the map $\sigma_{-i} \mapsto (\sigma_{-i}, \tau_{-i}(\sigma_{-i}))$ yields an embedding of $\bigcup_{h \in \mathcal{H}} \operatorname{supp}\left[\delta(.|\Sigma_{-i}(h))\right] \subseteq \Sigma_{-i}$ (a finite set) in $\Sigma_{-i} \times T_{-i}$, so that, for every $h \in \mathcal{H}$, $\mu(.|\Sigma_{-i}(h) \times T_{-i})$ is indeed a probability measure on $\Sigma_{-i} \times T_{-i}$. By the same argument, μ must also satisfy Property (3), i.e. it must be a CPS; of course, each $\mu(.|\Sigma_{-i}(h) \times T_{-i})$ has finite support by construction. Since g_i is onto, there exists a type $t_i \in T_i$ such that $g_i(t_i) = \mu$. By construction, t_i satisfies the property stated in the Lemma. ■

Lemma 13 *For every $i \in I$ and $n \geq 0$, $\Sigma_i^{n+1} \neq \emptyset$; furthermore $\sigma_i \in \Sigma_i^{n+1}$ if and only if there exists a CPS $\delta \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ such that $\sigma_i \in r_i(\delta)$ and*

$$\forall m = 0, \ldots, n-1,\ \forall h \in \mathcal{H}: \quad \Sigma_{-i}^{m+1} \cap \Sigma_{-i}(h) \neq \emptyset \Rightarrow \delta(\Sigma_{-i}^{m+1} | \Sigma_{-i}(h)) = 1 \quad (1)$$

Proof: omitted (for a similar result see [5], Theorem 1 and Corollary 1).

Proof of Proposition 6

(i) We proceed by induction. Part of the argument consists in showing that, for every $n \geq 0$ and $i \in I$, we can associate to each σ_i an epistemic type $t_i = \tau_i^{n+1}(\sigma_i)$ in such a way that $(\sigma_i, \tau_i^{n+1}(\sigma_i))_{i \in I} \in \mathrm{CSB}^n(R)$ whenever $(\sigma_i)_{i \in I} \in \Sigma^{n+1}$.

(Step 0.) Fix $(\sigma, t) \in \mathrm{CSB}^0(R) = R$. Then by definition $\sigma_i \in r_i(f_i(t_i))$ for every $i \in I$, which implies that $\sigma \in \Sigma^1$.

Conversely, for each $i \in I$ and $\sigma_i \in \Sigma_i$, pick $\tau_i^0(\sigma_i) \in T_i$ arbitrarily. Now fix $\sigma \in \Sigma^1$, and for each player $i \in I$, let $\delta \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ be such that $\sigma_i \in r_i(\delta)$. Now Lemma 12 yields a type $\tau_i^1(\sigma_i) \in T_i$ such that $g_{i,h}(\tau_i^1(\sigma_i))(\{(\sigma'_j, \tau_j^0(\sigma_j))_{j \neq i}\}) = \delta(\sigma'_{-i} | \Sigma_{-i}(h))$ for every $\sigma'_{-i} \in \Sigma_{-i}$, and hence $f_i(\tau_i^1(\sigma_i)) = \delta$. Thus, $(\sigma_i, \tau_i^1(\sigma_i))_{i \in I} \in R$.

Finally, for each $i \in I$, we complete the definition of the function $\tau_i^1(\cdot)$ by letting $\tau_i^1(\sigma_i) = \tau_i^0(\sigma_i)$ for $\sigma_i \in \Sigma_i \setminus \Sigma_i^1$.

(Step $n > 0$.) Now assume that Part (i) has been shown to hold for $m = 0, \ldots, n-1$, and that, for each such m, we have defined functions $\tau_i^{m+1} : \Sigma_i \to T_i$

such that $(\sigma_i, \tau_i^{m+1}(\sigma_i))_{i\in I} \in \text{CSB}^m(R)$ whenever $\sigma \in \Sigma^{m+1}$. Finally, let the functions $\tau_i^0(\cdot)$ be defined as above. We will prove that (a) $(\sigma, t) \in \text{CSB}^n(R)$ implies $\sigma \in \Sigma^{n+1}$ and (b) we can construct functions $\tau_i^{n+1} : \Sigma_i \to T_i$ such that $(\sigma_i, \tau_i^{n+1}(\sigma_i))_{i\in I} \in \text{CSB}^n(R)$ whenever $\sigma \in \Sigma^{n+1}$.

Note that, for every $n \geq 1$,

$$\text{CSB}^n(R) = R \cap \bigcap_{i\in I} \left\{ \bigcap_{m=0}^{n-1} \text{SB}_i(\Omega_i \times [\text{proj}_{\Omega_{-i}} \text{CSB}^m(R)]) \right\}. \tag{2}$$

Also note that, for any $i \in I$, $h \in \mathcal{H}$ and event E such that $E = \Omega_i \times \text{proj}_{\Omega_{-i}} E$ (that is, $E \in \mathcal{A}_{-i}$),

$$E \cap (\Sigma(h) \times T) \neq \emptyset \quad \Leftrightarrow \quad [\text{proj}_{\Sigma_{-i}} E] \cap \Sigma_{-i}(h) \neq \emptyset. \tag{3}$$

(a) Now consider $(\sigma, t) \in \text{CSB}^n(R)$ and fix $i \in I$. Let $\delta = f_i(t_i)$ be the first-order belief of type t_i. Eq. (2) yields $\sigma_i \in r_i(\delta)$. By the induction hypothesis, $\text{proj}_{\Sigma_{-i}} \text{CSB}^m(R) = \Sigma_{-i}^{m+1}$ for every $m = 0, ..., n-1$. Thus, (3) implies that, for every $m = 0, \ldots, n-1$ and $h \in \mathcal{H}$, $\Sigma_{-i}^{m+1} \cap \Sigma_{-i}(h) = [\text{proj}_{\Sigma_{-i}} \text{CSB}^m(R)] \cap \Sigma_{-i}(h) \neq \emptyset$ if and only if $[\Omega_i \times \text{proj}_{\Omega_{-i}} \text{CSB}^m(R)] \cap (\Sigma(h) \times T) \neq \emptyset$. The definition of strong belief and (2) imply that, whenever the latter condition is satisfied, $g_{i,h}(t_i)(\text{proj}_{\Omega_{-i}} \text{CSB}^m(R)) = \delta(\Sigma_{-i}^{m+1} | \Sigma_{-i}(h)) = 1$. Therefore δ satisfies (1), and Lemma 13 implies that $\sigma_i \in \Sigma_i^{n+1}$.

(b) Define

$$m_{-i}(\sigma'_{-i}) = \max\{m = 0, \ldots, n \ : \ \sigma'_{-i} \in \Sigma_{-i}^m\}$$

for every $i \in I$ and $\sigma'_{-i} \in \Sigma_{-i}$ (recall that $\Sigma_{-i}^0 = \Sigma_{-i}$, so $m_{-i}(\cdot)$ is well-defined). Now consider $\sigma \in \Sigma^{n+1}$ and fix a player $i \in I$. By Lemma 13, we can find a CPS $\delta \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ satisfying (1). By (3) and the induction hypothesis, for $h \in \mathcal{H}$ and $m = 0, \ldots, n-1$, $[\Omega_i \times \text{proj}_{\Omega_{-i}} \text{CSB}^m(R)] \cap (\Sigma(h) \times T) \neq \emptyset$ if and only if $\Sigma_{-i}^{m+1} \cap \Sigma_{-i}(h) \neq \emptyset$. Moreover, by (1), if the latter inequality holds, then $\delta(\Sigma_{-i}^{m+1} | \Sigma_{-i}(h)) = 1$.

Define $\tau_{-i} : \Sigma_{-i} \to T_{-i}$ by letting

$$\tau_{-i}(\sigma'_{-i}) = (\tau_j^{m_{-i}(\sigma'_{-i})}(\sigma'_j))_{j\neq i} \quad \forall \sigma'_{-i} \in \Sigma_{-i};$$

Lemma 12 now yields a type $\tau_i^{n+1}(\sigma_i) \in T_i$ such that

$$g_{i,h}(\tau_i^{n+1}(\sigma_i))(\{(\sigma'_j, \tau_j^{m_{-i}(\sigma'_{-i})}(\sigma'_j))_{j\neq i}\}) = \delta(\sigma'_{-i} | \Sigma_{-i}(h))$$

for all $h \in \mathcal{H}$ and $\sigma'_{-i} \in \Sigma_{-i}$. Now note that, for $m = 0, \ldots, n-1$,

$$\sigma'_{-i} \in \Sigma_{-i}^{m+1} \quad \Rightarrow \quad (\sigma'_j, \tau_j^{m_{-i}(\sigma'_{-i})}(\sigma'_j))_{j \neq i} \in \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^m(R)$$

because, (1) $m_{-i}(\sigma'_{-i}) \geq m+1$ if $\sigma'_{-i} \in \Sigma_{-i}^{m+1}$; (2) if $m_{-i}(\sigma'_{-i}) \geq 1$ then, by the induction hypothesis,

$$(\sigma'_j, \tau_j^{m_{-i}(\sigma'_j)}(\sigma'_j))_{j \neq i} \in \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^{m_{-i}(\sigma'_{-i})-1}(R);$$

and finally (3) the sets $\{\mathrm{CSB}^m(R)\}_{m \geq 0}$ are monotonically decreasing. But then

$$g_{i,h}(\tau_i^{n+1}(\sigma_i))(\mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^m(R)) = 1$$

for any $m = 0 \ldots n-1$ and $h \in \mathcal{H}$ such that $[\Omega_i \times \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^m(R)] \cap (\Sigma(h) \times T) \neq \emptyset$, because, by the preceding argument, supp $\delta(\cdot|\Sigma_{-i}(h)) \subseteq \Sigma_{-i}^{m+1}$ at any such history.

Moreover, by construction $f_i(\tau_i^{n+1}(\sigma_i)) = \delta$, so $\sigma_i \in r_i(f_i(\tau_i^{n+1}(\sigma_i)))$.

Repeating the argument for every $i \in I$ yields a profile of types $(\tau_i^{n+1}(\sigma_i))_{i \in I}$ which, by (2), satisfies $(\sigma_i, \tau_i^{n+1}(\sigma_i))_{i \in I} \in \mathrm{CSB}^n(R)$. This shows how to define the functions τ_i^{n+1} on Σ_i^{n+1}. To complete the induction step, for each $i \in I$ we now extend τ_i^{n+1} to the whole Σ_i by letting $\tau_i^{n+1}(\sigma'_i) = \tau_i^n(\sigma'_i)$ for every $\sigma'_i \in \Sigma_i \setminus \Sigma_i^{n+1}$. This conludes the proof of part (i).

(ii) By Lemma 13, $\Sigma^n \neq \emptyset$ for every $n \geq 0$. Thus, Part (i) implies that $\mathrm{CSB}^n(R) \neq \emptyset$ for every $n \geq 0$. Then $\mathrm{CSB}^\infty(R)$ is nonempty, because T is compact by assumption and the nested, nonempty closed sets $\{\mathrm{CSB}^n(R)\}_{n \geq 0}$ form a family with the finite intersection property.

Now suppose $(\sigma, t) \in \mathrm{CSB}^\infty(R)$. Since, by Part (i), $\Sigma^{n+1} = \mathrm{proj}_\Sigma \mathrm{CSB}^n(R)$ for any $n \geq 0$, we conclude that $\sigma \in \Sigma^n$ for every $n \geq 1$; so $\sigma \in \bigcap_{n \geq 1} \Sigma^n = \Sigma^\infty$. Hence $\mathrm{proj}_\Sigma \mathrm{CSB}^\infty(R) \subseteq \Sigma^\infty$.

Next, let N be the smallest integer such that $\Sigma^N = \Sigma^\infty$ (which must exist because Σ is finite). Pick any $\sigma \in \Sigma^N = \Sigma^\infty$ and consider the sequence of sets $M(m, \sigma) = \mathrm{CSB}^{(N-1)+m}(R) \cap (\{\sigma\} \times T)$, $m \geq 0$ (let $M(0, \sigma) = \{\sigma\} \times T$ if $N = 0$). Each set $M(m, \sigma)$ is nonempty and closed; also, the sequence of sets $M(m, \sigma)$ is decreasing, and hence has the finite intersection property. Then there is some profile of epistemic types t such that $(\sigma, t) \neq \bigcap_{m \geq 0} M(m, \sigma) \subseteq \mathrm{CSB}^\infty(R)$. It follows that $\Sigma^\infty \subseteq \mathrm{proj}_\Sigma \mathrm{CSB}^\infty(R)$. ■

Proof of Proposition 7

Proof. We claim that, for all $n \geq 0$ and $h \in \mathcal{H}$, $\Sigma^\infty \cap \Sigma(h) \neq \emptyset$ implies $\text{CSB}^n(R) \subseteq \text{B}_h^n(R)$; the assertion of the Proposition then follows immediately.

By definition, $\text{CSB}^0(R) = \text{B}_h^0(R) = R$, so the claim is true for $n = 0$. Assume it is true for some $n \geq 0$. Recall that $\text{CSB}^{n+1}(R) = \text{CSB}^n(R) \cap \text{SB}(\text{CSB}^n(R))$. Suppose that $\Sigma^\infty \cap \Sigma(h) \neq \emptyset$. By Theorem 6, this implies $\text{CSB}^n(R) \cap [h] \neq \emptyset$. Then, by definition of strong belief, $\text{SB}(\text{CSB}^n(R)) \subseteq \text{B}_h(\text{CSB}^n(R))$. By the induction hypothesis, $\text{CSB}^n(R) \subseteq \text{B}_h^n(R)$. By monotonicity of B_h, $\text{B}_h(\text{CSB}^n(R)) \subseteq \text{B}_h(\text{B}_h^n(R)) = \text{B}_h^{n+1}(R)$. Therefore, we conclude that $\Sigma^\infty \cap \Sigma(h) \neq \emptyset$ implies $\text{CSB}^{n+1}(R) \subseteq \text{SB}(\text{CSB}^n(R)) \subseteq \text{B}_h(\text{CSB}^n(R)) \subseteq \text{B}_h^{n+1}(R)$. ■

Proof of Proposition 11

Preliminaries. Note first that we can identify $\Delta(S_2)$ with $\Delta^{\mathcal{H}}(S_2)$, because $S_2(m) = S_2(\phi) = S_2$ for all m. To simplify the notation, for any CPS $\nu \in \Delta^{\mathcal{H}}(\Sigma_1)$, we write $\nu(\theta, m) = \nu((\theta, m)|\Sigma_1)$ and $\nu(\theta|m) = \nu((\theta, m)|\Theta \times \{m\})$. Let $\mu \in \Delta(S_2)$, $\nu \in \Delta^{\mathcal{H}}(\Sigma_1)$, $\pi_2 \in [\Delta(A)]^M$, $\zeta \in \Delta(\Theta \times M \times A)$.

We say that: (i) μ *agrees with* ζ if $\forall m, \forall a$, $\zeta(m) > 0$ implies $\zeta(a|m) = \mu(\{s_2 : s_2(m) = a\})$; (ii) ν *agrees with* ζ if $\forall \theta, \forall m$, $\nu(\theta, m) = \zeta(\theta, m)$; (iii) π_2 *agrees with* ζ if $\forall m, \forall a$, $\zeta(m) > 0$ implies $\pi(a|m) = \zeta(a|m)$; (iv) μ is the *mixed representation* of π_2, and π_2 is the *behavioral representation* of μ, if $\forall s_2$, $\mu(s_2) = \prod_{m \in M} \pi_2(s_2(m)|m)$; (v) m is a best reply for θ to π_2, written $(\theta, m) \in r_1(\pi_2)$, if $(\theta, m) \in r_1(\mu)$ and μ is the mixed representation of π_2.

We often need to associate an epistemic type to every payoff type-message pair or, respectively, strategy. When this is the case, we denote the epistemic type by $\tau_1(\theta, m)$ and $\tau_2(s_2)$ (cf. Lemma 12).

We prove (2) first. The argument involves six claims. For some of them we only sketch the proof.

Claim 1. For every $s_2 \in S_2$, there exists $t_2 \in T_2$ such that $(s_2, t_2) \in \text{proj}_{\Omega_2} R_2$ if and only if $s_2(m) \in BR_2(\Theta, m)$ for all $m \in M$.

Proof (sketch): It is easily checked that $\exists \nu \in \Delta^{\mathcal{H}}(\Sigma_1)$ such that $s_2 \in r_2(\nu)$ if and only if $\forall m$, $s_2(m) \in BR_2(\Theta, m)$. By Lemma 12, completeness of the type space implies the claim. ■

Recall that $IC_1^\zeta \equiv R_1 \cap [\zeta]_1 \cap \text{B}_{1,\phi}(R_2)$.

Claim 2. For every $(\theta, m) \in \Sigma_1$, there exists $t_1 \in T_1$ such that $(\theta, m, t_1) \in \text{proj}_{\Omega_1} IC_1^\zeta$ if and only if there exists $\pi_2^{\theta,m} \in [\Delta(A)]^M$ such that m is a best reply for θ to $\pi_2^{\theta,m}$, $\pi_2^{\theta,m}$ agrees with ζ and

$$\forall a \in A, \forall m' \in M,\ \pi_2^{\theta,m}(a|m') > 0 \Rightarrow a \in BR_2(\Theta, m'). \tag{4}$$

Proof: Fix $(\theta, m) \in \Sigma_1$ and suppose there exists t_1 as above. Let $\pi_2^{\theta,m}$ be the behavioral representation of $f_{1,\phi}(t_1)$. Then $(\theta, m) \in r_1(\pi_2^{\theta,m})$ and $\pi_2^{\theta,m}$ agrees with ζ. To see that Equation (4) is also satisfied, observe first that $\pi_2^{\theta,m}(a|m') > 0$ implies that, for some $s_2 \in \text{supp } f_{1,\phi}(t_1)$, $s_2(m') = a$. By assumption, supp $f_{1,\phi}(t_1) \subseteq \text{proj}_{S_2} R_2$, so there is some t_2 such that $(s_2, t_2) \in \text{proj}_{\Omega_2} R_2$; by Claim 1, this implies $a \in BR_2(\Theta, m')$.

Conversely, suppose that $(\theta, m) \in r_1(\pi_2^{\theta,m})$, $\pi_2^{\theta,m}$ agrees with ζ and (4) holds. Let $\mu \in \Delta(S_2)$ be the mixed representation of $\pi_2^{\theta,m}$. Then $(\theta, m) \in r_1(\mu)$ and μ agrees with ζ. Moreover, for every $s_2 \in S_2$ such that $\mu(s_2) > 0$, and for every $m' \in M$, $s_2(m') \in BR_2(\Theta, m')$; thus, by Claim 1, there exists an epistemic type $\tau_2(s_2)$ such that $(s_2, \tau_2(s_2)) \in \text{proj}_{\Omega_2} R_2$. By Lemma 12, there exists an epistemic type $t_1 \in T_1$ such that

$$\forall s_2 \in \text{supp } \mu, \quad g_{1,\phi}(t_1)(\{(s_2, \tau_2(s_2))\}) = \mu(s_2).$$

Thus, $g_{1,\phi}(t_1)(\text{proj}_{\Omega_2} R_2) = 1$ and therefore $(\theta, m, t_1) \in \text{proj}_{\Omega_1} IC_1^\zeta$. ■

Claim 3a. $[\zeta]_2 \cap \text{B}_{2,\phi}(IC_1^\zeta) \neq \emptyset$ implies that ζ satisfies

$$\forall \theta, \forall m,\ \zeta(\theta, m) > 0 \Rightarrow \sum_{a \in A} u_1(\theta, m, a)\zeta(a|m) = u_1^\zeta(\theta) \tag{5}$$

$$\forall \theta, \forall m,\ \zeta(\theta, m) = 0,\ \zeta(m) > 0 \Rightarrow u_1^\zeta(\theta) \geq \sum_{a \in A} u_1(\theta, m, a)\zeta(a|m) \tag{6}$$

$$\forall m,\ \zeta(m) = 0 \Rightarrow \forall \theta, \exists a \in BR_2(\Theta, m) :\ u_1(\theta, m, a) \leq u_1^\zeta(\theta) \tag{7}$$

$$\forall m, \forall a,\ \zeta(m, a) > 0 \Longrightarrow a \in BR_2(\Theta, m) \tag{8}$$

Proof: The assumption clearly implies $\{(\theta, m) : \zeta(\theta, m) > 0\} \subset \text{proj}_{\Sigma_1} R_1 \cap [\zeta]_1 \cap \text{B}_{1,\phi}(R_2)$. Thus, for any $(\theta, m) \in \Sigma_1$ with $\zeta(\theta, m) > 0$, there exists an epistemic type $\tau_1(\theta, m) \in T_1$ such that $(\theta, m, \tau_1(\theta, m)) \in R_1 \cap [\zeta]_1 \cap \text{B}_{1,\phi}(R_2)$, and hence, by Claim 2, a corresponding behavioral strategy $\pi_2^{\theta,m}$ that agrees with ζ, and satisfies $(\theta, m) \in r_1(\pi_2^{\theta,m})$ and (4).

To see that (5) holds, consider θ, m, m' with $\zeta(\theta, m) > 0$ and $\zeta(\theta, m') > 0$ and the corresponding behavioral strategies $\pi_2^{\theta,m}$, $\pi_2^{\theta,m'}$. Since $(\theta, m) \in r_1(\pi_2^{\theta,m})$ and $(\theta, m') \in r_1(\pi_2^{\theta,m'})$, $\sum_a u_1(\theta, m, a)\pi_2^{\theta,m}(a|m) \geq \sum_a u_1(\theta, m', a)\pi_2^{\theta,m}(a|m')$ and $\sum_a u_1(\theta, m', a)\pi_2^{\theta,m'}(a|m') \geq \sum_a u_1(\theta, m, a)\pi_2^{\theta,m'}(a|m)$. By agreement of $\pi_2^{\theta,m}$and $\pi_2^{\theta,m'}$ with ζ this implies $\sum_a u_1(\theta, m, a)\zeta(a|m) = \sum_a u_1(\theta, m', a)\zeta(a|m')$. Since $u_1^\zeta(\theta) = \sum_{m:\zeta(\theta,m)>0} \zeta(m|\theta) \sum_a u_1(\theta, m, a)\zeta(a|m)$, we obtain (5).

To see that (6) and (7) hold, consider θ, m, m' such that $\zeta(\theta, m') > 0$ and $\zeta(\theta, m) = 0$. Then (5) and $(\theta, m') \in r_1(\pi_2^{\theta,m'})$ imply $\sum_a u_1(\theta, m', a)\pi_2^{\theta,m'}(a|m') = u_1^\zeta(\theta) \geq \sum_a u_1(\theta, m, a)\pi_2^{\theta,m'}(a|m)$. If $\zeta(m) > 0$, then (6) follows from agreement of $\pi_2^{\theta,m'}$ with ζ. If $\zeta(m) = 0$, then (4) (swapping the roles of m and m') and $(\theta, m') \in r_1(\pi_2^{\theta,m\prime})$ imply that there must be $a \in \text{supp } \pi_2^{\theta,m'}(\cdot|m) \subset BR_2(\Theta, m)$ such that $u_1(\theta, m, a) \leq u_1^\zeta(\theta)$, as claimed in (6).

To see that (8) holds, note that the assumption implies $[\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2)) \neq \emptyset$. ■

Recall that $IC_2^\zeta \equiv R_2 \cap [\zeta]_2 \cap \mathrm{SB}_2(IC_1^\zeta)$.

Claim 3b. For every $s_2 \in S_2$, there exists $t_2 \in T_2$ such that $(s_2, t_2) \in \text{proj}_{\Omega_2} IC_2^\zeta$ if and only if Equations (5), (6), (7) and (8) hold and, moreover, there exists $\nu \in \Delta^{\mathcal{H}}(\Sigma_1)$ such that $s_2 \in r_2(\nu)$, ν agrees with ζ and

$$\forall m \in M, \ \left(\zeta(m) = 0, \ \overline{\Theta}(m; \zeta) \neq \Theta\right) \Rightarrow \nu(\overline{\Theta}(m; \zeta)|m) = 0. \tag{9}$$

Proof: (Only if). Fix $s_2 \in S_2$. If $t_2 \in T_2$ as above can be found, then in particular $[\zeta_2] \cap \mathrm{B}_{2,\phi}(IC_1^\zeta) \neq \emptyset$, so Claim 3a implies that (5), (6), (7) and (8) hold. Let $\nu = f_2(t_2)$. Then $s_2 \in r_2(\nu)$ and ν agrees with ζ. To show that (9) also holds, consider $m \in M$ such that $\zeta(m) = 0$ and $\overline{\Theta}(m; \zeta) \neq \Theta$. We prove that $\nu(\overline{\Theta}(m; \zeta)|m) = 0$.

First, we claim that $g_{2,m}(t_2)(\text{proj}_{\Omega_1} R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2)) = 1$. To see this, note that, since $\overline{\Theta}(m; \zeta) \neq \Theta$, there exists $\theta^* \in \Theta$ and $a^* \in BR_2(\Theta, m)$ such that $u_1(\theta^*, m, a^*) \geq u_1^\zeta(\theta^*)$. Construct a behavioral strategy $\pi_2^{\theta^*,m}$ as follows: (i) $\pi_2^{\theta^*,m}$ agrees with ζ; (ii) for all $m' \in M$ such that $\zeta(m') = 0$ and $m' \neq m$, let $\pi_2^{\theta^*,m}(a(\theta^*, m')|m') = 1$, where $a(\theta^*, m')$ is as in Equation (7); and finally let $\pi_2^{\theta^*,m}(a^*|m) = 1$. By construction, $(\theta^*, m) \in r_1(\pi_2^{\theta^*,m})$ and $\pi_2^{\theta^*,m}$ satisfies (4). Thus, by Claim 2, there exists $t_1 \in T_1$ such that $(\theta^*, m, t_1) \in \text{proj}_{\Omega_1} R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2)$. Hence, $R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2) \cap [m] \neq \emptyset$. Since $(s_2, t_2) \in \text{proj}_{\Omega_2} \mathrm{SB}_2(R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2))$, the claim follows.

Now let θ be such that $f_{2,m}(t_2)(\theta, m) = \nu(\theta|m) > 0$. By the previous result, there exists $t_1 \in T_1$ such that $(\theta, m, t_1) \in \text{proj}_{\Omega_1} R_1 \cap [\zeta]_1 \cap \mathrm{B}_{1,\phi}(R_2)$. Let $\pi_2^{\theta,m}$ be the

behavioral representation of $f_{1,\phi}(t_1)$. Then $\pi_2^{\theta,m}$ agrees with ζ, $(\theta, m) \in r_1(\pi_2^{\theta,m})$, (4) holds (see Claim 2) and $\sum_{a\in A} u_1(\theta, m, a)\pi_2^{\theta,m}(a|m) \geq u_1^{\zeta}(\theta)$ (by (5)). Thus, $\theta \notin \overline{\Theta}(m;\zeta)$ and Equation (9) must hold.

(If). Suppose that $\nu \in \Delta^{\mathcal{H}}(\Sigma_1)$ agrees with ζ, $s_2 \in r_2(\nu)$ and Equations (5), (6), (7), (8) and (9) are satisfied. For all (θ, m) such that $\zeta(\theta, m) > 0$, let $\pi_2^{\theta,m}$ the behavioral strategy that agrees with ζ and satisfies $\pi_2^{\theta,m}(a(\theta, m')|m') = 1$ for all m' with $\zeta(m') = 0$, where the action $a(\theta, m')$ is as in Equation (7).

Equations (5), (6) and (7) ensure that $(\theta, m) \in r_1(\pi_2^{\theta,m})$. Moreover, $\pi_2^{\theta,m}$ satisfies Equation (4), so by Claim 2 we can find an epistemic type $\tau_1(\theta, m) \in T_1$ such that $(\theta, m, \tau_1(\theta, m)) \in \text{proj}_{\Omega_1} IC_1^{\zeta}$. Thus, any pair in the support of $\nu(\cdot|\Sigma_1)$ can be matched with a suitable epistemic type. Hence, by Bayes' Rule, the same is true of points (θ, m) in the support of $\nu(\cdot|m)$ for all messages m with $\zeta(m) > 0$.

Next, we consider pairs (θ, m) such that $\nu(\theta|m) > 0$ and $\zeta(m) = 0$. If $\overline{\Theta}(m;\zeta) = \Theta$, choose an epistemic type $\tau_1(\theta, m)$ arbitrarily. Otherwise, by Equation 9 and the definition of $\overline{\Theta}(m;\zeta)$, there exists $a^* \in BR_2(\Theta, m)$ such that $u_1^{\zeta}(\theta) \leq u_1(\theta, m, a^*)$. Let $\pi_2^{\theta,m}$ be the behavioral strategy that agrees with ζ and satisfies $\pi_2^{\theta,m}(a^*|m) = 1$ and $\pi_2^{\theta,m}(a(\theta, m')|m') = 1$ for all $m' \neq m$ with $\zeta(m') = 0$. Again, $(\theta, m) \in r_1(\pi_2^{\theta,m})$ and Equation (4) holds, so Claim 2 yields an epistemic type $\tau_1(\theta, m)$ such that $(\theta, m, \tau_1(\theta, m)) \in \text{proj}_{\Omega_1} IC_1^{\zeta}$.

By Lemma 12, there exists $t_2 \in T_2$ such that, for all $(\theta, m) \in \Sigma_1$ and $h \in \mathcal{H} = \{\emptyset\} \cup M$,

$$g_{2,\phi}(t_2)(\{(\theta, m, \tau_1(\theta, m))\}) = \nu(\theta, m),\ g_{2,\phi}(t_2)(\{(\theta, m, t_1(\theta, m))\}) = \nu(\theta|m).$$

Since $f_2(t_2) = \nu$, $(s_2, t_2) \in \text{proj}_{\Omega_2} R_2 \cap [\zeta]_2$; the above construction also ensures that $g_{2,\phi}(t_2)(\text{proj}_{\Omega_1} IC_1^{\zeta}) = 1$, and similarly $g_{2,m}(t_2)(\text{proj}_{\Omega_1} IC_1^{\zeta}) = 1$ for all $m \in M$ with $\zeta(m) > 0$.

Now consider $m \in M$ such that $\zeta(m) = 0$ and $IC_1^{\zeta} \cap [m] \neq \emptyset$. Then, by Claim 2, for some $\theta \in \Theta$ and $\pi_2^{\theta,m}$, $(\theta, m) \in r_1(\pi_2^{\theta,m})$, $\pi_2^{\theta,m}$ agrees with ζ and (4) holds. Therefore, since $\zeta(m) = 0$, there must exist $a^* \in BR_2(\Theta, m)$ such that $u_1^{\zeta}(\theta) \leq u_1(\theta, m, a^*)$; thus, $\theta \notin \overline{\Theta}(m;\zeta)$. The preceding construction now ensures that $g_{2,m}(t_2)(\text{proj}_{\Omega_1} IC_1^{\zeta}) = 1$, as needed. ■

Recall that $IC^{\zeta} \equiv R_1 \cap [\zeta]_1 \cap \text{B}_{1,\phi}(IC_2^{\zeta})$.

Claim 4. $\{(\theta, m) : \zeta(\theta, m) > 0\} \subset \text{proj}_{\Sigma_1} IC^{\zeta}$ if and only if ζ is a self-confirming equilibrium that satisfies the Intuitive Criterion.

(Only if): The assumption implies that $IC^{\zeta} \neq \emptyset$; hence, $IC_2^{\zeta} \neq \emptyset$ and $[\zeta]_2 \cap \text{B}_{2,\phi}(IC_1^{\zeta}) \neq \emptyset$. Claim 3a then implies that Equations (5) and (6) hold.

Moreover, for every (θ, m) with $\zeta(\theta, m) > 0$, there exists an epistemic type $\tau_1(\theta, m) \in T_1$ such that $(\theta, m, \tau_1(\theta, m)) \in \text{proj}_{\Omega_1} IC^\zeta = R_1 \cap [\zeta]_1 \cap \text{B}_{1,\phi}(IC_2^\zeta)$. For any such pair (θ, m), let $\pi_2^{\theta,m}$ be the behavioral representation of $f_{1,\phi}(\tau_1(\theta, m))$; then $\pi_2^{\theta,m}$ agrees with ζ, and m is a best reply for θ to $\pi_2^{\theta,m}$. In particular, for all m' with $\zeta(m') = 0$, there exists an action $a(m', \theta)$ such that $\pi_2^{\theta,m}(a(m', \theta)|m') > 0$ and $u_1^\zeta(\theta) \geq u_1(\theta, m', a(m', \theta))$. Furthermore, we claim that (a) for all $m' \in M$ such that $\zeta(m') = 0$, $a(m', \theta) \in A(\zeta; m')$; and (b) for all $(m', a) \in M \times A$ such that $\zeta(m', a') > 0$, $a \in \arg\max_{a'} \sum_{\theta'} \zeta(\theta'|m') u_2(\theta', m', a)$.

To see this, note first that, for all $(m', a) \in M \times A$, $\pi_2^{\theta,m}(a|m') > 0$ and $g_{1,\phi}(\tau_1(\theta, m))(\text{proj}_{\Omega_2} IC_2^\zeta) = 1$ imply that there exists $s_2 \in \text{supp } f_{1,\phi}(\tau_1(\theta, m))$ and $t_2 \in T_2$ such that $s_2(m') = a$ and $(s_2, t_2) \in IC_2^\zeta$. Claim 3b then implies that, for some CPS $\nu \in \Delta^{\mathcal{H}}(\Sigma_1)$ that agrees with ζ, $s_2 \in r_2(\nu)$ and Equation (9) holds. Thus, in particular, $s_2(m') \in \arg\max_a \sum_{\theta'} \nu(\theta'|m') u_2(\theta', m', a)$. If $\zeta(m') = 0$, Equation (9) implies that $\nu(\overline{\Theta}(m'; \zeta)|m') = 0$ whenever $\overline{\Theta}(m'; \zeta) \neq \Theta$, so (a) holds. If $\zeta(m', a) > 0$, then (b) follows because ν agrees with ζ.

To see that ζ is a self-confirming equilibrium, for every $\theta \in \Theta$, let π_2^θ be a behavioral strategy that agrees with ζ and such that $\pi_2^\theta(a(m, \theta)|m) = 1$ for all $m \in M$ with $\zeta(m) = 0$. By the preceding observations, the profile $\{\pi_2^\theta\}_{\theta \in \Theta}$ satisfies Conditions (1) and (3) in Definition 9. Moreover, (b) above shows that Condition (2) is also satisfied.

Finally, for all m, θ with $\zeta(\theta, m) = 0$, (a) above shows that the actions $a(m, \theta)$ satisfy Definition 10, so ζ passes the Intuitive Criterion.

(If): Suppose now ζ is a self-confirming equilibrium that passes the Intuitive Criterion. By Definition 10, for every $\theta \in \Theta$ and $m \in M$ with $\zeta(m) = 0$, there exists $a(\theta, m) \in A(m; \zeta)$ such that $u_1^\zeta(\theta) \geq u_1(\theta, m, a(\theta, m))$, and a belief $\nu^{\theta,m} \in \Delta(\Theta)$ such that $a(\theta, m) \in \arg\max_a \sum_{\theta'} u_2(\theta', m, a) \nu^{\theta,m}(\theta')$, and $\nu^{\theta,m}(\overline{\Theta}(m; \zeta)) = 0$ if $\overline{\Theta}(m; \zeta) \neq \Theta$.

Fix $\theta \in \Theta$ and define a behavioral strategy π_2^θ as follows: π_2^θ agrees with ζ, and $\pi_2^\theta(a(\theta, m)|m) = 1$ for all $m \in M$ with $\zeta(m) = 0$. Observe that, by the choice of actions $a(\theta, \cdot)$ and Conditions (1) and (3) in Definition 9, $\zeta(\theta, m) > 0$ implies $(\theta, m) \in r_1(\pi_2^\theta)$.

Correspondingly, define a belief system ν^θ as follows: ν^θ agrees with ζ, and $\nu^\theta(\theta'|m) = \nu^{\theta,m}(\theta')$ for all $m \in M$ with $\zeta(m) = 0$.

By construction, ν^θ satisfies Equation (9). Moreover, let μ^θ be the mixed representation of π_2^θ; for every strategy $s_2 \in \text{supp } \mu^\theta$, Condition (2) in Definition 9 and the choice of $\nu^{\theta,m}$ for $\zeta(m) = 0$ ensure that $s_2 \in r_2(\nu^\theta)$. Finally, Condition (2) also implies that ζ satisfies Equation (8), and Conditions (1) and (3) imply

that it satisfies Equations (5), (6) and (7) as well. Therefore, by Claim 3b, for every $s_2 \in \text{supp } \mu^\theta$ there exists $\tau_2^\theta(s_2) \in T_2$ such that $(s_2, \tau_2^\theta(s_2)) \in \text{proj}_{\Omega_2} IC_2^\zeta$.

By Lemma 12, there exists t_1 such that

$$\forall s_2 \in S_2, \quad g_{1,\phi}(t_1)(\{(s_2, \tau_2^\theta(s_2))\}) = \mu^\theta(s_2);$$

therefore, $\zeta(\theta, m) > 0$ implies that $(\theta, m, t_1) \in \text{proj}_{\Omega_1} IC^\zeta$. ■

Claim 5. $\{(\theta, m) : \zeta(\theta, m) > 0\} \subset \text{proj}_{\Sigma_1} IC^\zeta$ if and only if $[\zeta]_2 \cap \text{B}_{2,\phi}(IC^\zeta) \neq \emptyset$.

The proof is straightforward, and is therefore omitted.

To prove Part (1), note first that, if $[\zeta]_2 \cap \text{B}_{2,\phi}(R_1 \cap [\zeta]_1) \neq \emptyset$, the proof of Claim 3a shows that Equations (5) and (6) hold, whereas Equation (7) is replaced by the weaker condition $\zeta(m) = 0 \Rightarrow \forall \theta, \exists a(m, \theta) \in A : \; u_1(\theta, m, a) \leq u_1^\zeta(\theta)$. This implies that it is possible to define a tuple of behavioral strategies $\{\pi_2^\theta\}$ as in the "only if" part of the proof of Claim 4, so that Conditions (1) and (3) of Definition 9 hold.

Also, the proof of Claim 3b shows that, if $(s_2, t_2) \in \text{proj}_{\Omega_2} R_2 \cap [\zeta]_2$, then there exists a CPS ν that agrees with ζ and such that $s_2 \in r_2(\nu)$. Then, as in the proof of the "only if" part of Claim 4, $t_1 \in \text{proj}_{T_1}[\zeta]_1 \cap \text{B}_{1,\phi}(R_2 \cap [\zeta]_2)$ implies that, whenever $\zeta(m, a) > 0$, there exists $s_2 \in \text{supp } f_{1,\phi}(t_1)$ such that $s_2(m) = a \in \arg\max_{a'} \sum_\theta \zeta(\theta|m) u_2(\theta, m, a')$. Hence, Condition (2) also hold, and the proof of Proposition 11 is complete. ■

References

[1] G. Asheim and M. Dufwenberg, Admissibility and Common Belief, Memorandum N. 07/2000, Department of Economics, University of Oslo, 2000.

[2] R. J. Aumann, Backward Induction and Common Knowledge of Rationality, *Games Econ. Behav.* **8** (1995), 6-19.

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

[4] P. Battigalli, Strategic Rationality Orderings and the Best Rationalization Principle, *Games Econ. Behav.* **13** (1996), 178-200.

[5] P. Battigalli, On Rationalizability in Extensive Games, *J. Econ. Theory* **74** (1997), 40-61.

[6] P. Battigalli and M. Siniscalchi, An Epistemic Characterization of Extensive-Form Rationalizability, Social Science Working Paper 1009, California Institute of Technology, 1997.

[7] P. Battigalli and M. Siniscalchi, Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games, *J. Econ. Theory* **88** (1999), 188-230.

[8] P. Battigalli and M. Siniscalchi, "Rationalization in Incomplete-Information Games," in preparation, 2001.

[9] E. Ben-Porath, Rationality, Nash Equilibrium and Backwards Induction in Perfect Information Games, *Rev. Econ. Stud.* **64** (1997), 23-46.

[10] O. Board, Algorithmic Characterization of Rationalizability in Extensive-Form Games, mimeo, Brasenose College, Oxford University, 1998.

[11] A. Brandenburger, On the Existence of a 'Complete' Belief Model, working paper 99-056, Harvard Business School, 1998.

[12] A. Brandenburger and E. Dekel, Hierarchies of Beliefs and Common Knowledge, *J. Econ. Theory* **59** (1993), 189-198.

[13] A. Brandenburger and H.J. Keisler,An Impossibility Theorem on Beliefs in Games, working paper 00-10, Harvard Business School, 1999.

[14] A. Brandenburger and H.J. Keisler,Epistemic Conditions for Iterated Admissibility, mimeo, Harvard Business School, 2000.

[15] I. K. Cho and D. Kreps, Signalling Games and Stable Equilibria, *Quart. J. Econ.* **102** (1987), 179-221.

[16] R. Fagin, J. Halpern, Y. Moses and M. Vardi, "Reasoning About Knowledge," MIT Press, Cambridge MA, 1995.

[17] D. Fudenberg and D. K. LevineSelf-Confirming Equilibrium," *Econometrica* **61** (1993), 523-545.

[18] D. Fudenberg and J. Tirole, "Game Theory," MIT Press, Cambridge MA, 1991.

[19] P. Gärdenfors, "Knowledge in Flux," MIT Press, Cambridge MA, 1988.

[20] J. Harsanyi, Games of Incomplete Information Played by Bayesian Players. Parts I, II, III, *Management Science* **14** (1967-68), 159-182, 320-334, 486-502.

[21] E. Kohlberg, Refinement of Nash Equilibrium: The Main Ideas, in "Game Theory and Applications" (T. Ichiishi, A. Neyman and Y. Tauman. Eds.), Academic Press, San Diego, 1990.

[22] E. Kohlberg and J.F. Mertens, On the Strategic Stability of Equilibria, *Econometrica* **54** (1986), 1003-1037.

[23] J.F. Mertens and S. Zamir, Formulation of Bayesian Analysis for Games with Incomplete Information, *Int. J. Game Theory* **14** (1985), 1-29.

[24] M. Osborne and A. Rubinstein, "A Course in Game Theory", MIT Press, Cambridge MA, 1994.

[25] D. Pearce, Rationalizable Strategic Behavior and the Problem of Perfection, *Econometrica* **52** (1984), 1029-1050.

[26] P. Reny, Rationality, common knowledge and the theory of games, mimeo, Department of Economics, Princeton University, 1985.

[27] P. Reny, Backward Induction, Normal-Form Perfection and Explicable Equilibria, *Econometrica* **60** (1992), 626-649.

[28] A. Rênyi, On a New Axiomatic Theory of Probability, *Acta Mathematica Academiae Scientiarum Hungaricae* **6** (1955), 285-335.

[29] A. Rubinstein, Comments on the Interpretation of Game Theory, *Econometrica* **59** (1991), 909-904.

[30] J. Sobel, L. Stole, and I. Zapater, Fixed-Equilibrium Rationalizability in Signaling Games, *J. Econ. Theory* **52** (1990), 304-331.

[31] R. Stalnaker, Knowledge, Belief and Counterfactual Reasoning in Games, *Econ. Philos.* **12** (1996), 133-163.

[32] R. Stalnaker, Belief Revision in Games: Forward and Backward Induction, *Math. Soc. Sci.* **36** (1998), 31-56.

[33] E. van Damme, "Stability and Perfection of Nash Equilibria" (2nd Edition), Springer Verlag, Berlin, 1991.