EUROPEAN UNIVERSITY INSTITUTE
DEPARTMENT OF ECONOMICS

EUI Working Paper ECO No. 99/15

Interactive Beliefs and Forward Induction

Pierpaolo Battigalli and Marciano Siniscalchi

BADIA FIESOLANA, SAN DOMENICO (FI)

Printed in Italy in May 1999
European University Institute
Badia Fiesolana
I-50016 San Domenico (FI)
Italy

Interactive Beliefs and Forward Induction

Pierpaolo Battigalli
Princeton University
and
European University Institute

Marciano Siniscalchi
Princeton University

March 1999[*]

Abstract

We provide an epistemic analysis of forward induction in games with complete and incomplete information. We suggest that forward induction 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. We also investigate the robustness of rationalizability to slight payoff uncertainity.

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

[*] This is a significantly expanded and revised version of Battigalli and Siniscalchi [10]. We thank the Editor and three anonymous referees 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. Email: `battigal@iue.it`, `marciano@princeton.edu`.

1 Introduction

Forward induction[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 (see Stalnaker [46], [47]) and hence their behavior.

In particular, we adopt a model of interactive conditional beliefs based on Battigalli and Siniscalchi [12] 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."

A player's belief revision policy may be governed by several working hypotheses concerning her opponents' rationality and strategic reasoning. For instance, she may believe that her opponents are rational until she observes a dominated move. Additionally, she may strongly believe that each opponent is rational *and* strongly believes that his own opponents are rational. Thus, if she receives information falsifying the latter hypothesis but not former, she continues to believe that her opponents are rational. Loosely speaking, we can use strong belief to formalize the assumption that a player interprets unexpected moves of her opponents in

[1] A partial list of references includes Banks and Sobel [6], Ben-Porath and Dekel [15], Cho and Kreps [20], Kohlberg and Mertens [32], McLennan [34], Van Damme [49]. In a non-equilibium setting, see Asheim and Dufwenberg [1], Battigalli [7, 8], Pearce [38], Reny [40].

[2] In a different formal setting, Stalnaker [47] independently introduced the notion of "robust belief," which captures a similar intuition.

a manner consistent with the highest possible "degree of strategic sophistication."

Our main results build on this intuition to provide epistemic characterizations of *extensive-form rationalizability* (Battigalli [7, 8], Pearce [38], Shimoji and Watson [44])[3] and the iterated *intuitive criterion* of Cho and Kreps [20]—perhaps the best-known "forward induction" equilibrium refinement for signalling games.

Since extensive-form rationalizability induces the backward induction outcome in generic perfect information games (Battigalli [8]; for a related result, see Reny [40]), our analysis additionally provides sufficient epistemic conditions for *backward induction.*

Finally, Dekel and Fudenberg [22] have emphasized the lack of robustness of solution procedures capturing forward induction reasoning to "slight" payoff uncertainty. We can derive and interpret related results in our setting, again building on the notion of strong belief.

Belief Revision and Type Spaces

Following Battigalli and Siniscalchi [12], we model beliefs as (infinite hierarchies of) conditional probability systems (Rênyi [41]) corresponding to *epistemic types* (cf. Mertens and Zamir [35]). A *state* in our model comprises a profile of strategies and epistemic types. The former encode each player's dispositions to *act*; the latter, each player's dispositions to *hold conditional beliefs* about her opponents' own strategies and types. Thus, assumptions about the players' rationality and reasoning processes correspond to events in our state space, and the players' epistemic attitudes towards them are formalized by means of systems of conditional probabilities—a representation that is both consistent with Bayesian decision making and familiar to economists.

[3] Bernheim's [16] "subgame rationalizability" is a weakening of subgame perfection and does not satisfy any forward induction criterion.

In the same spirit, we note that the notion of strong belief allows us to formulate assumptions about a player's belief revision process in a familiar setting.[4]

Unlike the usual probability-one belief operator, strong belief does not satisfy the monotonicity and conjunction properties. As we will discuss in Section 3, this implies that, in order to carry out our analysis of forward induction in a *neutral* setting, it is easiest to employ *belief-complete* epistemic models. In any such model, for any player i, every conditional probability system on the sets of strategies and epistemic types of Player i's opponents correspond to (at least) one of Player i's epistemic types (cf. Brandenburger [18]). The existence of belief-complete models follows from Battigalli and Siniscalchi [12].

Our characterization results involve higher-order strong beliefs. However, simply iterating the strong belief operator (as may be suggested by an analogy with the notion of "common certainty of rationality") leads to contradictions. This point is illustrated in Section 4 by means of an example.

Finally, we note that we analyze a general version of extensive-form rationalizability which allows for *payoff uncertainty* and *exogenous restrictions on first-order beliefs* (see Battigalli [9]).

Related Literature on the Epistemic Analysis of Dynamic Games and Forward Induction

Finite (hence incomplete) extensive form type spaces are introduced in Ben Porath [14] to characterize common certainty of rationality at the beginning of a perfect information game. Battigalli and Siniscalchi [12] provide a general analysis of (finite and infinite) type spaces for extensive form games and show the existence of a belief-complete type space, the building block of our analysis.

[4] Belief revision (mostly in a single-person setting) has been studied extensively by philosophers. See e.g. Gärdenfors [28], Grove [29], Stalnaker [46], [47], and references therein.

Stalnaker [46] puts forward a related normal form, finite epistemic model, which can also be used to analyze extensive form reasoning. This model is used by Stalnaker [47] to provide a brief discussion of forward induction and by Board [17] to characterize some extensive form solution concepts, including extensive form rationalizability. 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.[5] Stalnaker and Board are thus forced to qualify their characterization results with the proviso that the incomplete model at hand contains "enough" epistemic types to allow for forward induction reasoning in the game under consideration.[6]

In the context of a partitional, normal form model, Asheim and Dufwenberg [1] investigate the consequences of common knowledge of cautious rationality, where the latter is not defined as a property of strategy-type pairs, but rather as a property of types. They characterize an iterated deletion procedure which captures certain aspects of forward induction.

Aumann [2] and related papers, such as Aumann [3], Samet [43] and Balkenborg and Winter [5] use a partitional epistemic model to provide sufficient conditions for the backward induction outcome in generic perfect information games. In the epistemic models of these papers, a state of the world describes the players' strategies (dispositions to act) and their initial epistemic state, but it does *not* describe how a player would revise her beliefs, should she *learn* that a particular node has been reached. On the other hand, a theory of belief revision is implicit in Aumann's [2] notion of "rationality." For more on this we refer to the discussions in Section 5 of Stalnaker [47] and Section 6 of Battigalli and Siniscalchi [12].

[5] We doubt that belief-complete models à la Stalnaker exist at all.

[6] This is made precise by Board [17], who also builds on Battigalli [7] as well as on Battigalli and Siniscalchi [10], the previous version of this paper.

The remainder of the paper is organized as follows. Notation is introduced in Section 2. Section 3 motivates the notion of strong belief by means of an example, provides the formal definition and discusses its properties. Section 4 draws the connections between strong belief and forward induction; a characterization of extensive-form rationalizability for complete-information games is also included here. Section 5 deals with games with payoff uncertainty. It contains our general characterization result, a characterization of the intuitive criterion and the analysis of robustness with respect to slight payoff uncertainty. All proofs are contained in the Appendix.

2 The Model

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 [12].

2.1 Notation for Extensive–Form Games with Complete Information

In the first part of this paper, we focus on finite games with complete information. As was mentioned in the Introduction, we shall subsequently enrich the formal setup to accommodate payoff uncertainty.

In order to keep notation at a minimum, our analysis shall deal mainly with multistage games with *observable actions* (Fudenberg and Tirole [27], §3.3; Osborne and Rubinstein [37], Chap. 6), although most of our results can be extended to general games. We also note that the majority of dynamic games of interest in economics fits within this framework (allowing for payoff uncertainty.)

We shall be interested in the following primitive objects: a set $I = \{1, \ldots, |I|\}$ of players, a collection $\mathcal{H}$ of *partial histories*,[7] including

[7] Histories are sequences of consecutive action profiles.

the *empty history* ϕ, a collection of *terminal histories* $\mathcal{Z}$, and a payoff function $u_i : \mathcal{Z} \to \mathbb{R}$ for every player $i \in I$. As the game progresses, each player is informed of the partial history that has just occurred. At some stages there can be simultaneous moves. If there is only one active player at each stage, 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. 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 the partial or terminal 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.

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

It is convenient to introduce two additional pieces of notation. For every strategy s_i, $\mathcal{H}(s_i) = \{h \in \mathcal{H} : s_i \in S_i(h)\}$ denotes the collection of partial histories consistent with s_i. For every partial history h and strategy s_i, s_i^h denotes the strategy consistent with h which coincides with s_i on the set of partial histories not preceding h (thus, $h \in \mathcal{H}_i(s_i)$ implies $s_i^h = s_i$).[8]

2.2 Conditional Probability Systems

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* (Myerson [36], Rênyi [41]).

[8] Assume without loss of generality that each player chooses an action immediately after every partial history. Then s_i^h is defined as follows. For all $h' \in \mathcal{H}$, if either $(h', (a_j)_{j \in I})$ comes before h or $(h', (a_j)_{j \in I}) = h$, then $s_i^h(h') = a_i$. Otherwise, $s_i^h(h') = s_i(h')$.

Fix a player $i \in I$. For a given measure space $(X_i, \mathcal{A}_i)$, consider a non-empty, finite or countable collection $\mathcal{B}_i \subset \mathcal{A}_i$ 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 a "discrete" component of x.

Definition 1 *A* conditional probability system *(or CPS) on $(X_i, \mathcal{A}_i, \mathcal{B}_i)$ is*[9] *a mapping*

$$\mu(\cdot|\cdot) : \mathcal{A}_i \times \mathcal{B}_i \to [0,1]$$

satisfying the following axioms:

Axiom 1 *For all $B \in \mathcal{B}_i$, $\mu(B|B) = 1$.*

Axiom 2 *For all $B \in \mathcal{B}_i$, $\mu(\cdot|B)$ is a probability measure on $(X_i, \mathcal{A}_i)$.*

Axiom 3 *For all $A \in \mathcal{A}_i$, $B, C \in \mathcal{B}_i$, if $A \subset B \subset C$ then $\mu(A|B)\mu(B|C) = \mu(A|C)$.*

The set of probability measures on $(X_i, \mathcal{A}_i)$ will be denoted by $\Delta(X_i)$; we shall endow it with the topology of weak convergence of measures. The set of conditional probability systems on $(X_i, \mathcal{A}_i, \mathcal{B}_i)$ can be regarded as a subset of $[\Delta(X_i)]^{\mathcal{B}_i}$, 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 beliefs* about her opponents' behavior may be represented by taking $X_i = S_{-i}$ and $\mathcal{B}_i = \{B \subset S_{-i} : B = S_{-i}(h) \text{ for some } h \in \mathcal{H}\}$. We denote the collection of CPSs on $(S_{-i}, \mathcal{B}_i)$ thus defined by $\Delta^{\mathcal{H}}(S_{-i})$. Since S_{-i} and $\mathcal{H}$ are finite, $\Delta^{\mathcal{H}}(S_{-i})$ is easily seen to be a closed subset of the Euclidean $|\mathcal{H}| \cdot |S_{-i}|$-space.

To represent Player i's *higher-order beliefs*, we will consider a (finite or infinite) set of "possible worlds" $\Omega = \prod_{i \in I} \Omega_i$, where $\Omega_i \subset S_i \times Y_i$ and

[9] The tuple $(X, \mathcal{A}, \mathcal{B}, \mu)$ is called *conditional probability space* by Rênyi [41]. When X is finite, $\mathcal{A} = 2^X$, $\mathcal{B} = 2^X \backslash \{\emptyset\}$, we obtain Myerson's [36] conditional probability systems.

$\text{proj}_{S_i}\Omega_i = S_i$. Elements of the sets Y_i will be interpreted as epistemic types. As will be clear momentarily, it is convenient to assume that each Y_i is a Polish (i.e. separable and completely metrizable) space.

To represent Player i's hierarchical beliefs about her opponents, we use the following structure: let $X_i = \Omega_{-i}$, let $\mathcal{A}_i$ be the Borel sigma algebra on Ω_{-i} and

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

The set of CPSs on $(\Omega_{-i}, \mathcal{B}_i)$ will be denoted by $\Delta^{\mathcal{H}}(\Omega_{-i})$. Similarly, to represent Player i's hierarchical beliefs about the prevailing state of the world (including her own strategy and beliefs, as well as her opponents'), let $X_i = \Omega$, let $\mathcal{A}$ be the Borel sigma algebra on Ω and

$$\mathcal{B} = \{B \in \mathcal{A}\text{: } B = \{(s, y) \in \Omega : s \in S(h)\} \text{ for some } h \in \mathcal{H}\}.$$

The set of CPSs on $(\Omega, \mathcal{B})$ is denoted $\Delta^{\mathcal{H}}(\Omega)$.

Note that Ω_{-i} and Ω are Polish spaces in the respective product topologies; also, the finite collections $\mathcal{B}_i$ and $\mathcal{B}$ consist of sets that are both *open and closed* in the respective topologies. Battigalli and Siniscalchi [12] show that, under these conditions, $\Delta^{\mathcal{H}}(\Omega_{-i})$ and $\Delta^{\mathcal{H}}(\Omega)$ are closed subsets of the Polish spaces $[\Delta(\Omega_{-i})]^{\mathcal{B}_i}$ and, respectively, $[\Delta(\Omega)]^{\mathcal{B}}$. Hence, they are Polish spaces in the relative topology.

2.3 Epistemic Models

We next introduce our basic representation of hierarchical conditional beliefs.

Definition 2 *(cf. Ben Porath [14]) A* type space *on* $(\mathcal{H}, S(\cdot), I)$ *is a tuple* $\mathcal{T} = (\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i\in I})$ *such that, for every* $i \in I$, T_i *is a Polish space and*

i. Ω_i *is a closed subset of* $S_i \times T_i$ *such that* $\text{proj}_{S_i}\Omega_i = S_i$*;*

ii. $g_i : T_i \to \Delta^{\mathcal{H}}(\Omega_{-i})$ *is a continuous mapping.*

For any $i \in I$, the elements of the set T_i are referred to as Player i's epistemic types. *A type space is* compact *if all the sets T_i, $i \in I$, are compact topological spaces.*

Thus, at any "possible world" $\omega = (s_i, t_i)_{i \in I} \in \Omega$, we specify each player i's *dispositions to act* (her strategy s_i) and *dispositions to believe* (her system of conditional probabilities $g_i(t_i) = (g_{i,h}(t_i))_{h \in \mathcal{H}}$). These dispositions also include what a player *would* do and think at histories that are inconsistent with ω.[10]

As is traditional in the epistemic analysis of games, we complete a player's system of conditional beliefs by assuming that she is certain of her strategy and epistemic type. More specifically, we assume that for every state of the world $((s_i, t_i), \omega_{-i})$ and every history h, Player i would be certain of t_i given h and would also be certain of s_i given h provided that s_i is consistent with h, i.e. $s_i \in S_i(h)$. We also assume that if $s_i \notin S_i(h)$ Player i would still be certain that her continuation strategy agrees with s_i. (The latter assumption is immaterial for our analysis, but we include it for completeness).

Formally, Player i's conditional beliefs on $(\Omega, \mathcal{B})$ are given by a continuous mapping

$$g_i^* = (g_{i,h}^*)_{h \in \mathcal{H}} : \Omega_i \to \Delta^{\mathcal{H}}(\Omega)$$

derived from g_i by the following formula: for all $(s_i, t_i) \in \Omega_i$, $h \in \mathcal{H}$, $E \in \mathcal{A}$,

$$g_{i,h}^*(s_i, t_i)(E) = g_{i,h}\left(\left\{\omega_{-i} \in \Omega_{-i} : ((s_i^h, t_i), \omega_{-i}) \in E\right\}\right) \tag{1}$$

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 [35] and Brandenburger and Dekel [19] answered this question in the affirmative when beliefs are represented by probability measures on a compact

[10] History h is inconsistent with (or counterfactual at) $\omega = (s, t)$ if $h \notin S(h)$.

Hausdorff or Polish space; Battigalli and Siniscalchi [12] 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 $(\mathcal{H}, S(\cdot), I)$ is a type space $\mathcal{T} = (\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i \in I})$ such that, for every $i \in I$, $\Omega_i = S_i \times T_i$ and $g_i : T_i \to \Delta^{\mathcal{H}}(\prod_{j \neq i} S_j \times T_j)$ is* onto.[11]

It is shown in [12] that a belief-complete type space may always be constructed (at least for finite 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.[12] Finally, since we assume that the set of external states S is finite and hence compact, the belief-complete type space thus constructed is also compact.

3 Forward Induction and Strong Belief

With the basic framework notation in place, we now turn to the main focus of this paper, *forward induction* reasoning. We begin by specifying the notion of rationality we adopt.

3.1 Sequential Rationality

We take the view that a strategy $s_i \in S_i$ for Player i should be optimal, given Player i's beliefs, conditional upon any history consistent with s_i;

[11] We use "complete" in the same sense as Brandenburger [18], who shows (in a different framework) that a (belief-) complete, filter-theoretic type space does not exists. Of course, this notion of completeness is not to be confused with the topological one.

[12] [12] uses a slightly different definition of type space. But all the arguments in [12] can be easily adapted to the present framework.

we do not impose restrictions on the action specified at histories that cannot obtain if Player i follows the strategy s_i. This is a sequential best response property which applies to plans of actions[13] as well as strategies (see, for example, [42] and [40]).

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

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

For any CPS $\mu_i \in \Delta^{\mathcal{H}}(S_{-i})$, let $r_i(\mu_i)$ denote the set of sequential best replies to μ_i.

It can be shown by standard arguments that r_i is a nonempty-valued and upper-hemicontinuous correspondence. It is convenient to introduce the following additional notation. Fix a type space $(\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i\in I})$. For every player $i \in I$, let $f_i = (f_{i,h})_{h\in\mathcal{H}} : T_i \to [\Delta(S_{-i})]^{\mathcal{H}}$ denote her first-order belief mapping, that is, for all $t_i \in T_i$ and $h \in \mathcal{H}$,

$$f_{i,h}(t_i) = \mathrm{marg}_{S_{-i}} g_{i,h}(t_i)$$

(recall that $\mathrm{proj}_{S_{-i}}\Omega_{-i} = S_{-i}$). It is easy to see that $f_i(t_i) \in \Delta^{\mathcal{H}}(S_{-i})$ for every $t_i \in T_i$; also, f_i is continuous.

Finally, we can introduce our key behavioral axiom. We say that Player i is *rational* at a state $\omega = (s, t)$ in $\mathcal{T}$ if and only if $s_i \in r_i(f_i(t_i))$. Then the event

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

corresponds to the statement, “Player i is rational.” (Note that R_i is closed because the correspondence $r_i \circ f_i$ is upper hemicontinuous.) We

[13] 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.

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").

A word of caution. Events are defined with reference to a specific type space. In the following, we shall ensure that the type space we refer to is clear from the context and notation.

3.2 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 non–terminal 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."

Given a type space $\mathcal{T} = (\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i\in I})$, for every $i \in I$ and $h \in \mathcal{H}$, define the event[14]

$$\mathrm{B}_{i,h}(E) = \{(s,t) \in \Omega \ : \ g^*_{i,h}(s_i, t_i)(E) = 1\}$$

which corresponds to the statement "Player i would be certain of E, were she to observe history h." Observe that this definition incorporates the natural requirement that a player only be certain of events which are consistent with her own (continuation) strategy and epistemic type (recall how g^*_i was derived from g_i).

For each player i and history $h \in \mathcal{H}$, the definition identify a set–to–set operator $\mathrm{B}_{i,h} \ : \mathcal{A} \to \mathcal{A}$ which satisfies the usual properties of falsifiable beliefs (see, for example, Chapter 3 of Fagin *et al* [24]); in particular, it satisfies

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

[14] For any measurable subset $E \subset \Omega$, $\mathrm{B}_{i,h}(E)$ is closed, hence measurable; this follows from the continuity of $g^*_{i,h}$, via an application of the *portmanteau* theorem.

3.3 Strong Belief and Forward Induction

The conditional belief operator $\mathrm{B}_{i,h}$ is the natural extension to the present dynamic setting of the belief operator used in the analysis of normal–form games. It features prominently in the analysis of several problems in the theory of extensive games (see Battigalli and Siniscalchi [12] and references therein). However, we shall presently argue that the logic of forward induction suggests an alternative notion of belief. The game depicted in Figure 1 illustrates this point.

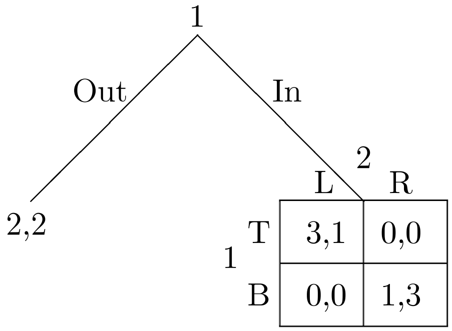

Figure 1: The Battle of the Sexes with an Outside Option

The usual "forward induction analysis" of this game runs as follows. Observe first that the profile (OutB, R) (where "OutB" stands for "Play Out at the empty history, and B if the simultaneous–moves subgame is ever reached") 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. The threat is credible in the (weak) sense that (B,R) constitutes a Nash equilibrium of the subgame. However, Player 2's threat is not entirely convincing, according to forward induction reasoning: after all, 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. The conclusion is that (OutB, R) is *not* stable with respect to forward induction reasoning. 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. In order to make sense of it, we introduce a "sufficiently rich" type space. Note that here $I = \{1, 2\}$, $S_1 = \{\text{OutT}, \text{OutB}, \text{InT}, \text{InB}, \text{B}\}$ and $S_2 = \{\text{L}, \text{R}\}$; "(In)" denotes the partial history in which Player 1 chooses In at the beginning of the game — in other words, (In) is the root of the simultaneous–moves subgame, so $\mathcal{H} = \{\phi, (\text{In})\}$; also, $S(\text{In}) = \{\text{InT}, \text{InB}\} \times \{\text{L}, \text{R}\}$.

Table 1 describes a type space for the game under consideration; we shall denote it by $\mathcal{T}$.

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 1: The Type Space $\mathcal{T}$

The table specifies the sets $T_1 = \{t_1^1, t_1^2\}$ and $T_2 = \{t_2^1, t_2^2, t_3^3\}$ 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, $\omega_1(5) = (\text{ InT}, t_1^2)$. Similarly, we will use the notation $\omega(n_1, n_2) = (\omega_1(n_1), \omega_2(n_2))$.

3.3.1 Step 1: Initial Common Certainty of Rationality

Consider state $\omega(3, 2)$, where the unstable subgame–perfect equilibrium profile (OutB,R) is played.[15] Note first that both players are certain at ϕ

[15] More precisely, at $\omega(3, 2)$ Player 1 chooses Out, but the Nash equilibrium (B,R) *would* be played *if* the subgame was reached.

that the prevailing state is indeed $\omega(3,2)$: that is, $\omega(3,2) \in \mathrm{B}_{i,\phi}(\{\omega(3,2)\})$ for $i = 1,2$. Also, $\omega(3,2) \in R$; hence, by monotonicity,

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

In words, at $\omega(3,2)$ there is *initial common certainty of the opponent's rationality.*

We have thus exhibited a type space, $\mathcal{T}$, and a state, $\omega(3,2)$, where the unstable profile (OutB,R) is played, and yet players are rational, they *initially* recognize this, and indeed they are *initially* quite "sophisticated" (they recognize that they recognize each other's rationality, and so on).

3.3.2 Step 2: Forward Induction and Belief Revision

A closer look at Table 1 shows why initial common certainty of the opponent's rationality may fail to yield the forward induction outcome. In state $\omega(3,2)$ Player 2 would be certain at (In) that $\omega_1 = \omega_1(1)$. However, at $\omega_1(1)$[16] Player 1 is not rational, because InB is strictly dominated. Thus, $\omega(3,2) \notin \mathrm{B}_{2,(\mathrm{In})}(R_1)$.

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. Note that this is a restriction on how Player 2 should *revise his beliefs* upon observing that his initial conjecture was incorrect. As we have just shown, this restriction is violated at $\omega(3,2)$.

But note that type t_2^3 of Player 2 holds beliefs consistent with initial common certainty of the opponent's rationality; moreover, at (In), this type assigns probability one to $\omega_1(5)$, which is consistent with both Player 1's rationality and the observed history of play. Thus, there are states in $\mathcal{T}$ where the above restriction is satisfied.

[16] By this we mean "at any state $\omega = (\omega_1, \omega_2)$ such that $\omega_1 = \omega_1(1)$."

3.3.3 Step 3: Strong Belief

The preceding discussion suggests that forward induction is related to the idea that players may formulate *working hypotheses* at the beginning of the game (e.g. "My opponent is rational,") and subsequently *maintain them insofar as they are not explicitly contradicted*—even as they revise their beliefs. We presently propose a notion of "belief" which formalizes this idea.[17]

We shall 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. Formally, for any type space $\mathcal{T} = (\mathcal{H}, S(\cdot), I, (\Omega_i, T_i, g_i)_{i \in I})$, define the operator $\mathrm{SB}_i : \mathcal{A} \to \mathcal{A}$ by $\mathrm{SB}_i(\emptyset) = \emptyset$ and

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

for all events $E \in \mathcal{A} \backslash \{\emptyset\}$. Note that $\mathrm{SB}_i(E) \subset \mathrm{B}_{i,\phi}(E)$ for all $E \in \mathcal{A}$; that is, strong belief implies initial certainty.

Reverting to our example, we have $\mathrm{SB}_2(R_1) = \{\omega(n_1, 3) \ : \ n_1 = 1 \ldots 5\}$. Note that $R_2 = \Omega$, so

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

If we now add the further assumption that Player 1 is certain, at the beginning of the game, that Player 2 is rational and strongly believes that Player 1 is rational, we obtain

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

i.e. we identify the forward induction solution.

[17] An analogous notion (called "robust belief") was independently put forth by Stalnaker [47].

3.4 Monotonicity, Conjunction, and the Pitfalls of Incomplete Type Spaces

Consider any two events E, F defined in an arbitrary type space $\mathcal{T}$. Fix a player $i \in I$, suppose that $E \subset F$, and consider a state $\omega = (s,t) \in \mathrm{SB}_i(E)$.

By definition, $g^*_{i,h}(s_i,t_i)(E) = 1$ for all histories h consistent with E. This clearly implies that, *at these histories*, $g^*_{i,h}(s_i,t_i)(F) = 1$; however, since $F \supset E$, there may be a history h' consistent with F but *not* consistent with E. Thus, Player i may or may not assign probability one to F conditional upon reaching h' in state ω, without prejudice to the assumption that $\omega \in \mathrm{SB}_i(E)$.

In general, *strong belief is not a monotonic operator.* An entirely similar reasoning shows that *it need not satisfy conjunction.* As we shall demonstrate in the next Section, this is relevant to our analysis.

Here we wish to point out another important consequence of the failure of these properties: *analyzing an extensive-form game in the framework of an incomplete type space introduces extraneous 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 2.

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 2: 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.[18] 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 $R_1' = \{\omega(n_1, n_2) \ : \ n_1 = 3, 4, \ n_2 = 1, 2\}$. Next, note that $\mathrm{SB}_2'(R_1) = \{\omega(n_1, 2) \ : \ n_1 = 1 \ldots 4\}$: 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, there is 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

$$R_1' \cap R_2' \cap \mathrm{SB}_2'(R_1') \cap \mathrm{B}_{1,\phi}' \left(R_2' \cap \mathrm{SB}_2'(R_1')\right) = \{\omega(3,2), \omega(4,2)\},$$

where both $\omega(3,2)$ and $\omega(4,2)$ yield the "unstable" equilibrium 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

$$\begin{array}{rcl} \mathrm{SB}_2(R_1') & = & \mathrm{SB}_2(R_1 \cap \Omega') = \{(n_1, 2) \ : \ n_1 = 1, \ldots, 5\} \neq \\ & \neq & \mathrm{SB}_2(R_1) \cap \mathrm{SB}_2(\Omega') = \emptyset. \end{array}$$

and

$$R_1' \cap \mathrm{SB}_2'(R_1') = (R_1 \cap \Omega') \cap \mathrm{SB}_2(R_1 \cap \Omega') \subsetneq R_1 \cap \mathrm{SB}_2(R_1).$$

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.

[18] 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.

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.*[19]

4 Belief Revision, Strong Belief and Rationalizability

We argued in the preceding section that the notion of strong belief plays a central rôle in forward induction reasoning. In accordance with standard practice in the literature on the epistemic foundations of solution concepts, we now investigate the implications of *iterated* (strong) *beliefs* about the players' rationality.

4.1 Preliminaries

In light of the remarks at the end of Section 3, we state our assumptions and results in the "epistemologically neutral" setting provided by belief-complete type spaces.

Our objective is to identify the behavioral implications of assumptions pertaining to the players' rationality and conditional beliefs. Our characterization results are thus statements concerning the projection of

[19] Alternatively, one may carry out the analysis in the context of a belief-incomplete, but "sufficiently rich" type space—i.e. one that contains "enough" epistemic types to formalize the variant of forward induction reasoning one is interested in: see e.g. Board [17]. 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.

the corresponding events onto the set S of strategy profiles. More explicitly, let E be the set of states of the world (in a belief-complete type space) satisfying a given collection of assumptions A^0, A^1, ..., and let S^* be the set of strategy profiles selected by a given solution concept. Then

$$S^* = \mathrm{proj}_S E$$

means that $(s_i)_{i\in I} \in S^*$ if and only if there is a profile of conceivable epistemic types $(t_i)_{i\in I}$ such that the assumptions $A^0, A^1, ...$ are satisfied at the state of the world $\omega = (s_i, t_i)_{i\in I}$.

The epistemic assumptions we consider only restrict a player's beliefs about her opponents' behavior and their beliefs. That is, for instance, we do not *explicitly* require that a player be certain, or strongly believe, that she is rational.[20] This approach emphasizes those aspects of strategic reasoning that are most familiar to economists and game theorists; it is also the most natural approach given the structure of our epistemic model.

We introduce the following auxiliary operators to simplify notation: for any event $E \in \mathcal{A}$ and for any history $h \in \mathcal{H}$ let

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

For example, if $I = \{1, 2\}$ and $E = R$, then $\mathrm{SB}(R) = \mathrm{SB}_1(R_2) \cap \mathrm{SB}_2(R_1)$.

We also introduce a uniform notation for the n-fold composition of operators. Formally, 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 $\mathcal{O}^n(E) = \mathcal{O}(\mathcal{O}^{n-1}(E))$.

[20] However, in our epistemic model, a player is certain of her own actual (continuation) strategy and beliefs at each state. This guarantees that, if she is rational, then she is certain of this at each history consistent with her strategy; indeed, the converse is also true. This feature is shared by most models in the literature on epistemic foundations of normal and extensive-form solution concepts.

4.2 The Benchmark: Common Certainty of Rationality

The notion of *common certainty of (the opponents') rationality* is central in the analysis of normal-form games. A straightforward extension to dynamic games is also possible (see Ben-Porath [14] and Battigalli and Siniscalchi [12]).

Definition 5 *Fix a history $h \in \mathcal{H}$.*

(Step 0) *For every $i \in I$, let $W^0_{i,h} = S_i(h)$. Also, let $W^0_{-i,h} = \prod_{j \neq i} W^0_{i,h}$ and $W^0_h = \prod_{i \in I} W^0_{i,h}$.*

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

i. $s_i \in r_i(\mu)$;

ii. $\mu(W^{n-1}_{-i,h} | S_{-i}(h)) = 1$.

Also let $W^n_{-i,h} = \prod_{j \neq i} W^n_{i,h}$ and $W^n_h = \prod_{i \in I} W^n_{i,h}$.

Finally, let $W^\infty_h = \bigcap_{k \geq 0} W^n_h$. For $h = \phi$, the strategy profiles in W^∞_ϕ are said to be weakly rationalizable.[21]

Building on Ben Porath [14], Battigalli and Siniscalchi [12] show that the W^∞_h solution is characterized by common certainty of rationality conditional on h; we state their result below to facilitate comparisons with the assumptions and solution concepts we consider in this paper.

For any history $h \in \mathcal{H}$, let $[h] = \{(s, t) \in \Omega \ : \ s \in S(h)\}$. Also recall that $\mathrm{B}^0_h(R) = R$.

[21] Weak rationalizability is a well-known solution procedure for extensive games (see e.g. Ben Porath (1997), Dekel and Gul (1997) and references therein). In generic perfect information games it first eliminates the weakly dominated strategies and then iteratively deletes strictly dominated strategies, a procedure first analyzed by Dekel and Fudenberg (1990).

Proposition 1 *Fix a history $h \in \mathcal{H}$. Then, for any belief-complete type space,*

(i) for all $n \geq 0$, $W_h^{n+1} = \mathrm{proj}_S\left(\bigcap_{m=0}^{n} \mathrm{B}_h^n(R) \cap [h]\right)$.

(ii) If the type space is also compact, then $W_h^{\infty} = \mathrm{proj}_S\left(\bigcap_{n\geq 1} \mathrm{B}_h^n(R) \cap [h]\right)$

In particular, W_{ϕ}^{∞} is the set of strategy profiles consistent with common certainty of rationality. As was noted in Section 3, initial common certainty of rationality is consistent with the profile (OutB, R) in the Battle of the Sexes with an outside option. Also, in that game, $W_{(\mathrm{In})}^{\infty} = \{(\mathrm{InT,L})\} \neq \emptyset$; by Proposition 1, this implies that, in any complete model, there are states in which Player 1 chooses In at the beginning of the game, and there is common certainty of rationality in the subgame. Thus, common certainty of rationality is possible conditional on every history.

This is not the case for the game in Figure 2, which we shall refer to throughout this section.

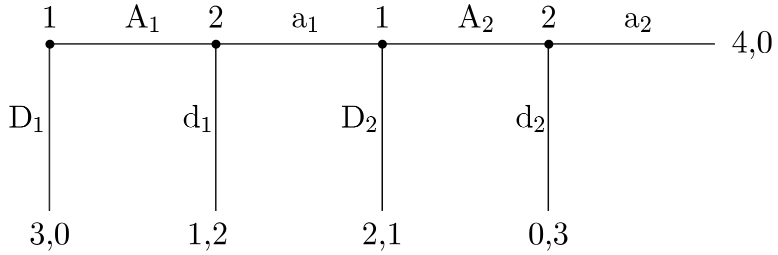

Figure 2: A perfect information game (Reny [40])

The backward induction solution is of course $(\mathrm{D_1D_2}, \mathrm{d_1d_2})$, so the only history consistent with backward induction is ϕ. It is easy to check that $W_{(\mathrm{A}_1)}^{\infty} = \emptyset$: indeed, in this example there cannot be common certainty of rationality conditional on any history off the backward induction path (cf. Reny [39] and Ben Porath [14]).

4.3 A Caveat: Iterated Strong Beliefs and Failures of Conjunction

Motivated by the observations in Section 3, we now turn to the central notion of iterated strong beliefs.

A formal analogy with Proposition 1 might suggest considering assumptions of the form $\bigcap_{m=0}^{n} \mathrm{SB}^m(R)$. However, consider the event

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

The key observation is that, although $\mathrm{SB}(R) \neq \emptyset$ and $\mathrm{SB}(\mathrm{SB}(R)) \neq \emptyset$ in any complete model, it may still be the case that $\mathrm{SB}(R)\cap\mathrm{SB}(\mathrm{SB}(R)) = \emptyset$, even if $R \cap \mathrm{SB}(R) \neq \emptyset$.[22] Thus, one may have $\bigcap_{m=0}^{2} \mathrm{SB}^m(R) = \emptyset$.

The game in Figure 2 offers an example. It can be checked[23] that $\mathrm{proj}_S R = (S_1 \setminus \{\mathrm{A_1D_2}\}) \times (S_2 \setminus \{\mathrm{a_1a_2}\})$ and $\mathrm{proj}_S R \cap \mathrm{SB}(R) = \{\mathrm{D_1D_2}, \mathrm{D_1A_2}\} \times \{\mathrm{a_1d_2}\}$. Although history $(\mathrm{A_1})$ is consistent with R_1 and with the assumption $\mathrm{SB}_1(R_2)$ (which implies no behavioral restriction), 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 $\mathrm{A_1}$, which implies that $\mathrm{SB}(R) \cap \mathrm{SB}(\mathrm{SB}(R)) = \emptyset$.

4.4 Strong Belief and the Best Rationalization Principle

The notion of strong belief allows us to provide a rigorous formulation of the *best rationalization principle* (Battigalli [7]) and emphasize its impli-

[22] This is consistent with the general observation that the strong belief operator need not satisfy the conjunction property (see Section 3).

[23] Formally, both equalities follow from Proposition 3. The intuition is that $\mathrm{A_1D_2}$ is strictly dominated for Player 1, and $\mathrm{a_1a_2}$ is not sequentially rational; the further assumption that players strongly believe that these strategies will not be chosen eliminates $\mathrm{A_1A_2}$, $\mathrm{d_1d_2}$ and $\mathrm{d_1a_2}$.

cations as a theory of belief revision in extensive games.

The best rationalization principle requires that players' beliefs conditional upon observing a history $h \in H$ be consistent with the highest degree of "strategic sophistication" of their opponents.

Formally, define the auxiliary "correct strong belief" operator CSB : $\mathcal{A} \to \mathcal{A}$ by

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

for any $E \in \mathcal{A}$. Also let $\mathrm{CSB}^\infty(E) = \bigcap_{n \geq 0} \mathrm{CSB}^n(E)$.

For every $n \geq 0$, we associate the event $\mathrm{CSB}^n(R)$ with n-th order strategic sophistication. Unraveling the above definition allows us to be precise as to the formal content of the best rationalization principle.

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

A first-order strategically sophisticated player is rational, and also maintains 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 $\mathrm{CSB}^2(R) = R \cap \mathrm{SB}(R) \cap \mathrm{SB}(\mathrm{CSB}^1(R))$. Note that, since $\mathrm{CSB}^1(R) \subset \mathrm{CSB}^0(R) = R$, the difficulties described in Subsection 4.3 do not arise.

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 $n \geq 0$,

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

which may now be seen to capture the intuition behind the best rationalization principle.[24]

We are now ready to state our main characterization result. The following procedure is a straightforward adaptation of *extensive form rationalizability* (see Pearce [38] and Battigalli [8]).

Definition 6 *Consider the following procedure.*

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

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

i. $s_i \in r_i(\mu)$;

ii. for all $h \in \mathcal{H}$, if $S_{-i}^{n-1} \cap S_{-i}(h) \neq \emptyset$, then $\mu(S_{-i}^{n-1}|S_{-i}(h)) = 1$.

Also let $S_{-i}^n = \prod_{j \neq i} S_i^n$ and $S^n = \prod_{i \in I} S_i^n$.[25]

Finally, let $S^\infty = \bigcap_{k \geq 0} S^n$. The strategy profiles in S^∞ are said to be extensive-form rationalizable.

As a preliminary result, we investigate the connection between extensive-form rationalizability, the procedure of Definition 5, and common certainty of rationality.

Proposition 2 *For all $h \in \mathcal{H}$ and $n \geq 1$, $S^n \cap S(h) \subset W_h^n$. Therefore, if history h is consistent with extensive form rationalizability, it is also*

[24] The nested collection of events $\{\mathrm{CSB}^n(R)\}_{n=0}^\infty$ is analogous to a (sub)system of spheres in the sense of Grove [29]. Systems of spheres can be used to formalize the notion of epistemic entrenchment (see Gardenfors [28], Chapter 4).

[25] It can be shown that the computational complexity of the procedure can be reduced by checking the restrictions on believes only at histories consistent with the given strategy (cf. Battigalli [8]). Computationally less demanding characterizations of extensive form rationalizability are analyzed in Shimoji and Watson [44].

consistent with common certainty of rationality; that is, $S^\infty \cap S(h) \neq \emptyset$ *implies* $\bigcap_{n\geq 0} \mathrm{B}_h^n(R) \neq \emptyset$.[26]

The main result of this section states that rationality and the best rationalization principle completely characterize extensive-form rationalizability.

Proposition 3 *For any belief-complete type space,*

(i) for any $n \geq 0$, $S^{n+1} = \mathrm{proj}_S \mathrm{CSB}^n(R)$*;*

(ii) if the type space is also compact, then $S^\infty = \mathrm{proj}_S \mathrm{CSB}^\infty(R)$.

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

Proposition 3 is a corollary of a more general result proved in the Appendix. Here we outline the main argument.

Sketch of Proof: observe first that, for any player $i \in I$, given any CPS $\delta_i \in \Delta^{\mathcal{H}}(S_{-i})$, we can construct a CPS $\mu_i \in \Delta^{\mathcal{H}}(S_{-i} \times T_{-i})$ by associating a type $t_j(s_j) \in T_j$ to each $s_j \in S_j$ for every $j \neq i$ and letting $\mu_i(\{(s_j, t_j(s_j))_{j\neq i}\}|S_{-i}(h) \times T_{-i}) = \delta_i(\{(s_j)_{j\neq i}\}|S_{-i}(h))$; Lemma 10 provides the details. Then, since g_i is onto, we can find a type $t_i \in T_i$ such that $g_i(t_i) = \mu_i$ and hence $f_i(t_i) = \delta_i$.

Step 0 of Part (i) follows. To establish the inductive step, use the representation $\mathrm{CSB}^n(R) = R \cap \bigcap_{m=0}^{n-1} \mathrm{SB}(\mathrm{CSB}^m(R))$. Lemma 11 yields a related representation of S^n. Fix a player $i \in I$ and a strategy $s_i \in S^{n+1}$, and let δ_i be the first-order CPS justifying s_i. If $\mathrm{proj}_{S_j} \mathrm{CSB}^m(R) = S_j^{m+1}$ for $m = 0 \ldots n-1$ and $j \neq i$, then we can associate each $s_j \in S_j$, $j \neq i$, with a type $t_j(s_j) \in T_j$ so as to ensure that $(s_j, t_j(s_j))_{j\neq i} \in \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^m(R)$ whenever $(s_j)_{j\neq i} \in S_{-i}^{m+1}$,

[26] Note that the Proposition only provides a *sufficient* condition. Reny [39] provides an example where a non-extensive-form-rationalizable history is consistent with common certainity of rationality (his discussion does not employ a formal epistemic model and the example illustrates a slightly different point).

for $m = 0 \ldots n-1$. Using this construction, δ_i can then be extended to a CPS μ_i on Ω_{-i} such that $\mu_i(\mathrm{proj}_{\Omega_{-i}}\mathrm{CSB}^m(R)|S_{-i}(h) \times T_{-i}) = 1$ whenever $[S_{-i}(h) \times T_{-i}] \cap \mathrm{proj}_{\Omega_{-i}}\mathrm{CSB}^m(R) \neq \emptyset$. It follows that any type t_i such that $g_i(t_i) = \mu_i$ satisfies $(s_i, t_i) \in \mathrm{proj}_{\Omega_i}\mathrm{CSB}^n(R)$. The other direction is a matter of checking the definitions.

Finally, Part (ii) follows from compactness.

4.5 Strong Beliefs and Backward Induction

Battigalli [8] shows that, in generic perfect information games, extensive-form rationalizability is outcome-equivalent to backward induction (for a related result, see Reny [40]). Note that, since S is finite and $S^{n+1} \subset S^n$, there is some $N \geq 0$ such that $S^\infty = S^N$. Hence, Proposition 3 also provides a set of *sufficient* epistemic conditions for the backward induction outcome:

Proposition 4 *Suppose the game under consideration has 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 $s \in \mathrm{proj}_S\mathrm{CSB}^N(R)$ induces the unique backward induction outcome.*

We emphasize that our results provide an explicit set of conditions on the players' beliefs revision processes leading to backward induction play.

It should also be noted that our 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 [40]). 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 $\mathrm{A}_1\mathrm{D}_2$ is strictly dominated by $\mathrm{D}_1\mathrm{D}_2$ for Player 1, whereas $\mathrm{A}_1\mathrm{A}_2$ may at least be justified by the "unsophisticated" belief that Player 2 will irrationally play $\mathrm{a}_1\mathrm{a}_2$.

5 Interactive Epistemology in Dynamic Games with Payoff Uncertainty

We now extend our analysis to multistage games with observable actions in which at least one player does not know some payoff-relevant aspect of the game, such as an opponent's preferences over (lotteries on) $\mathcal{Z}$, or the mapping between $\mathcal{Z}$ and the relevant space of consequences.

We address three issues related to payoff uncertainty and forward induction. First, we characterize a variant of extensive-form rationalizability which accounts for payoff uncertainty and, possibly, exogenous restrictions on first-order beliefs. Second, we provide an epistemic characterization of the *intuitive criterion* of Cho and Kreps [20], perhaps the best-known equilibrium refinement for signalling games. Finally, we provide a rigorous epistemic analysis of the robustness of forward induction reasoning to "slight" payoff uncertainty.

5.1 Preliminaries

In order to model payoff uncertainty, we associate with each player $i \in I$ a nonempty, finite set Θ_i of conceivable *payoff-types*. Each element $\theta_i \in \Theta_i$ represents Player i's private information about the unknown payoff-relevant aspects of the game. Correspondingly, we assume that payoffs associated with terminal nodes depend on players' payoff-types: formally, for each player $i \in I$, the payoff function is a map $u_i : \prod_{j \in I} \Theta_j \times \mathcal{Z} \to \mathbb{R}$.

Minor modifications in our notation are sufficient to allow for payoff uncertainty: Table 3 briefly summarizes the required changes.[27]

[27] 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 [30], because it contains no description of players' interactive beliefs about payoff-types. However, once we have specified a type space $(\mathcal{H}, \Sigma(\cdot), I, (\Omega_i, T_i, g_i)_{i \in I})$, we can define the set of Harsanyi-types for each player $i \in I$ to be $\Theta_i \times T_i$; the belief mapping $\overline{g}_i : \Theta_i \times T_i \to \Delta(\prod_{j \in I} \Theta_j \times T_j)$ may be obtained from $g^*_{i,\phi}$ by marginalization. This construction yields a game with incomplete information, as defined by Harsanyi. Of course, Harsanyi-consistency (i.e. the possibility to derive beliefs at each state from a common prior) is satisfied only in

We point out that, as our definition of the map g_i^* suggests, we assume that each player is certain of her payoff-type as well as her (continuation) strategy and epistemic type at each state.

In the present setting, a type space on $(\mathcal{H}, \Sigma(\cdot), I)$ is said to be *complete* if and only if $\Omega_i = \Sigma_i \times T_i$ and g_i is *onto* for every $i \in I$. The belief operators $\mathrm{B}_{i,h}$ and SB_i, B_h, SB and CSB are defined as in Sections 3 and 4.

Object	**Notation**	**Definition**
Payoff Type-Strategy Pairs	Σ_i, Σ_{-i}, Σ	$\Sigma_i = \Theta_i \times S_i$ $\Sigma = \prod_{i \in I} \Sigma_i$, $\Sigma_{-i} = \prod_{j \neq i} \Sigma_j$
Pairs consistent with $h \in \mathcal{H}$	$\Sigma_i(h)$, $\Sigma_{-i}(h)$, $\Sigma(h)$	$\Sigma_i(h) = \Theta_i \times S_i(h)$, etc.
Strategic-form Payoffs	$U_i : \Sigma_i \times \Sigma_{-i} \to \mathbb{R}$	$U_i(\theta_i, s_i, \theta_{-i}, s_{-i}) = u_i(\theta_i, \theta_{-i}, z)$ where $(s_i, s_{-i}) \in S(z)$
Type Space on $(\mathcal{H}, \Sigma(\cdot), I)$	$(\mathcal{H}, \Sigma(\cdot), I, (\Omega_i, T_i, g_i)_{i \in I})$	$\Omega_i \subset \Sigma_i \times T_i$, etc. $g_i : T_i \to \Delta^{\mathcal{H}}(\Omega_{-i})$ continuous
Induced Beliefs on Ω	$g_i^* : \Omega_i \to \Delta^{\mathcal{H}}(\Omega)$	$g_{i,h}^*(\theta_i, s_i, t_i)(E) =$ $= g_{i,h}(t_i)\left(\{\omega_{-i} : ((\theta_i, s_i^h, t_i), \omega_{-i}) \in E\}\right).$
First-Order Beliefs	$f_i : T_i \to \Delta^{\mathcal{H}}(\Sigma_{-i})$	$f_{i,h}(t_i) = \mathrm{marg}_{\Sigma_{-i}} g_{i,h}(t_i)$

Table 3: Notation for Games with Payoff Uncertainty

We also need to modify the notion of *rationality* to reflect payoff uncertainty. 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* θ_i if and only if, for every $h \in \mathcal{H}(s_i)$ and every $s_i' \in S_i(h)$,

$$\sum_{\sigma_{-i} \in \Sigma_{-i}} [U_i(\theta_i, s_i, \sigma_{-i}) - U_i(\theta, s_i', \sigma_{-i})] \mu_i(\{\sigma_{-i}\} | \Sigma_{-i}(h)) \geq 0$$

For any CPS $\mu_i \in \Delta^{\mathcal{H}}(S_{-i})$, let $r_i(\mu_i)$ denote the set of type-strategy pairs $(\theta_i, s_i) \in \Sigma_i$ such that s_i is a sequential best reply to μ_i for θ_i.[28] The event "Player i is *rational*" is

$$R_i = \{(\sigma_i, t_i, \omega_{-i}) \in \Omega : \sigma_i \in r_i(f_i(t_i))\}.$$

special cases. The existence of a common prior on Θ is an even more special case.

[28] Note that $r_i(\cdot)$ is a nonempty-valued, upper hemicontinuous correspondence.

5.2 Forward Induction in Games with Payoff Uncertainty

In Section 4 we provided an epistemic characterization of extensive form rationalizability (EFR) based on the notions of strong belief and best rationalization. We now define and characterize a solution procedure which extends EFR in two ways: first, we introduce payoff uncertainty; second, we allow for restrictions on players' first-order beliefs.

Such restrictions are commonplace in applications featuring payoff uncertainty; for instance, one often assumes the existence of a common prior on the set Θ of payoff-types. More generally, certain features of players' beliefs may appear to be salient in a given applied context (e.g. Cho [21], Watson [50], Battigalli and Watson [13], Battigalli [9]). The procedure we characterize combines these restrictions with forward induction reasoning.

We begin by specifying the type of restrictions we consider. For every player $i \in I$ and history $h \in \mathcal{H}$ fix a nonempty closed subset $\Delta_{i,h} \subset \Delta(\Sigma_{-i})$ and let $\Delta_i = \Delta^{\mathcal{H}}(\Sigma_{-i}) \cap \prod_{h \in \mathcal{H}} \Delta_{i,h}$. We call a subset of CPSs of this form *regular*.[29]

For any given collection of regular subsets $\Delta = (\Delta_i)_{i \in I}$, we define a solution procedure that iteratively eliminates (payoff-type, strategy) pairs for each player i:

Definition 7 *Consider the following procedure.*

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

(Step $n > 0$) *For every $i \in I$, and for every $\sigma_i \in \Sigma_i$, let $\sigma_i \in \Sigma^n_{i,\Delta}$ if and only if $\Sigma_i \in \Sigma^{n-1}_{i,\Delta}$ and there exists a CPS $\mu \in \Delta_i$ such that*

i. $\sigma_i \in r_i(\mu)$;

[29] A version of our result applies to general restrictions, whereby each Δ_i is an arbitrary subset of $\Delta^{\mathcal{H}}(\Sigma_{-i})$. We restrict attention to regular subsets of CPSs for expository reasons.

ii. for all $h \in \mathcal{H}$, *if* $\Sigma^{n-1}_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset$, *then* $\mu(\Sigma^{n-1}_{-i,\Delta} | \Sigma_{-i}(h)) = 1$.

Also let $\Sigma^n_{-i,\Delta} = \prod_{j \neq i} \Sigma^n_{i,\Delta}$ *and* $\Sigma^n_\Delta = \prod_{i \in I} S^n_i$.

Finally, let $\Sigma^\infty_\Delta = \bigcap_{n \geq 0} \Sigma^n_\Delta$.

Denote by $E(\Delta)$ the event that the players' first order beliefs satisfy the restrictions given by $\Delta = (\Delta_i)_{i \in I}$; that is,

$$E_i(\Delta_i) = \{(\sigma_i, t_i, \omega_{-i}) \in \Omega : f_i(t_i) \in \Delta_i\}, \; E(\Delta) = \bigcap_{i \in I} E_i(\Delta_i).$$

Proposition 5 *Fix a collection* $\Delta = (\Delta_i)_{i \in I}$ *of regular subsets of CPSs. Then, for any belief-complete type space,*

(i) for any $n \geq 0$, $\Sigma^{n+1}_\Delta = \mathrm{proj}_\Sigma \mathrm{CSB}^n(R \cap E(\Delta))$*;*

(ii) if the type space is also compact, then $\Sigma^\infty_\Delta = \mathrm{proj}_\Sigma \mathrm{CSB}^\infty(R \cap E(\Delta))$

Intuitively, the procedure described here is characterized by the assumption that players apply the best rationalization principle but do so in a manner consistent with the assumed restrictions Δ. Thus, Player i's own first-order beliefs are an element of Δ_i; she adopts the working assumption that her opponents are rational *and their beliefs are elements of* $\prod_{j \neq i} \Delta_j$; and so on.

The following subsections rely on this general characterization result.

5.3 Strong Belief and the Intuitive Criterion

We now consider a (finite) signaling game, that is, a two-person game with observable actions and uncertainty about the payoff-type of Player 1, where Player 1 (the Sender) is active at the first stage and Player 2 (the Receiver) is active at the second stage. Our definition of game with payoff uncertainty already implies that the set of feasible messages is the same for each payoff-type. We also assume that the set of feasible actions for the Receiver is independent of the signal.[30] Table 4 summarizes our notation for signalling games.

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

Table 4: 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. Behavioral strategies are defined as in Kreps and Wilson [33].

In this framework, an external state is given by a tuple $\sigma = (\theta, m, s_2) \in \Theta \times M \times A^M$ 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: $\zeta(\theta)$, $\zeta(m)$, $\zeta(m, a)$, $\zeta(m|\theta)$, $\zeta(m, a|\theta)$, $\zeta(\theta|m)$, $\zeta(a|m)$. 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.

Definition 8 *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 $\pi_2^\theta(a|m) = \zeta(a|m)$.

[30] Removing these assumptions is straightforward but implies a more complex notation.

Our definition of self-confirming equilibrium outcome agrees with the definition of self-confirming equilibrium with unitary beliefs put forward by Fudenberg and Levine [26], if each incarnation θ of the Sender is regarded as an individual player selected by chance with probability $\pi_0(\theta)$. Clearly, every sequential equilibrium outcome (Kreps and Wilson [33]) 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 [20] put forward the (Iterated) 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 [31], 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 $\emptyset \neq \overline{\Theta} \subseteq \Theta$ and message m, $BR_2(\overline{\Theta}, m)$ is the set of best responses to beliefs concentrated on $\overline{\Theta}$ given message m. Consider the following procedure.

Definition 9 *(Modified Iterated Intuitive Criterion) Fix a self-confirming equilibrium outcome ζ and a message $m \in M$ such that $\zeta(m) = 0$. Let $I\Theta^0(m;\zeta) = \Theta$ and $IA^0(m;\zeta) = A$. For all $k = 0, 1, 2, ...$ define*

$$I\Theta^{k+1}(m;\zeta) = \left\{\theta \in I\Theta^k(m;\zeta) : u_1^\zeta(\theta) \leq \max\nolimits_{a\in IA^k(m;\zeta)} u_1(\theta, m, a)\right\},$$

$$IA^{k+1}(m;\zeta) = \begin{cases} BR_2(I\Theta^k(m;\zeta), m), & \text{if } I\Theta^k(m;\zeta) \neq \emptyset \\ IA^k(m;\zeta), & \text{if } I\Theta^k(m;\zeta) = \emptyset. \end{cases}$$

Outcome ζ satisfies the Iterated Intuitive Criterion 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 $a \in \bigcap_k IA^k(m;\zeta)$ such that $u_1(\theta, m, a) \leq u_1^\zeta(\theta)$.

As in the original treatment by Cho and Kreps [20], a candidate outcome *fails* the modified IIC if a Sender's type may deviate to an off-equilibrium message and "reasonably" expect to obtain a higher payoff than she receives according to ζ. The "textbook" definition (e.g. Fudenberg and Tirole [27], p. 449) iteratively strikes out dominated responses by the Receiver *first* and type-message pairs of the Sender *next*; our modification requires that these steps be carried out simultaneously. We wish to treat assumptions about each player's strategic sophistication symmetrically, as we have done so far.[31] In any case, if the Receiver has no dominated actions (i.e. $BR_2(\Theta, m) = A$ for all m,) the two procedures coincide.

Cho and Kreps [20] 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.

Sobel *et al.* ([45], Proposition 2) relate the Iterated Intuitive Criterion to extensive form rationalizability in an auxiliary game where the messages on-the-equilibrium-path are coalesced into a message m^ζ that yields the equilibrium payoff $u_1^\zeta(\theta)$ to each incarnation θ of the Sender. Our result relies instead on the procedure in Definition 7; the exogenous restrictions on first-order beliefs Δ_i are chosen to reflect the assumption that Player i's prior beliefs "agree" with the outcome distribution ζ. Proposition 5 can then be invoked to provide an epistemic characterization of the Iterated Intuitive Criterion that helps clarify Cho and Kreps' [20] informal statements.

We say that Player i's beliefs about her opponent $-i$ *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 about the Receiver agree with ζ" is

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

[31] However, one can easily formulate a variant of Proposition 5 to accommodate the usual definition of the procedure. We also note that Cho and Kreps do not appear to favor any specific order of elimination (cf. [20], p. 196.)

Similarly,

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

Part (1) of the following proposition is a preliminary step of some independent interest, similar in spirit to Theorem A in Aumann and Brandenburger [4]. Part (2) is our characterization result.

Proposition 6 *Fix a π_0-feasible outcome ζ.*
(1) If $\bigcap_{i=1,2} R_i \cap [\zeta]_i \cap \mathrm{B}_{i,\phi}\left(R_{-i} \cap [\zeta]_{-i}\right) \neq \emptyset$ in some type space, then ζ is a self-confirming equilibrium outcome.[32]
(2) For any belief-complete and compact type space, $\mathrm{CSB}^{\infty}\left(R \cap [\zeta]_1 \cap [\zeta]_2\right) \neq \emptyset$ if and only if ζ is a self-confirming equilibrium outcome satisfying the Iterated Intuitive Criterion.

5.4 Robustness of Rationalizability with Respect to Payoff Uncertainty

We conclude this section with a collection of results pertaining to the robustness of forward induction reasoning to "slight" payoff uncertainty. Our analysis is similar in spirit to that of Fudenberg, Kreps and Levine [25] and, especially, Dekel and Fudenberg [22]; however, our arguments do not involve payoff perturbations and limiting arguments. Rather, we relate robustness (or lack thereof) to specific assumptions about belief revision policies.

As in the first reference cited above, we embed a complete information game within a richer one featuring payoff uncertainty. Specifically, fix a game IG with payoff uncertainty, a profile of payoff-types $\theta^0 \in \Theta$, and denote by G_{θ^0} the complete information game corresponding to θ^0.

[32] The hypothesis in (1) can be replaced by

$$\bigcap_{i=1,2} \mathrm{B}_{i,\phi}\left(R \cap [\zeta]_1 \cap [\zeta]_2\right) \neq \emptyset.$$

Also the converse of (1) is true if the Receiver has no conditionally dominated action.

We can then apply the procedures defined in Section 4 to the latter game. In particular, we shall focus on the sequences of sets $\{W_\phi^n\}_{n\geq 0}$ and $\{S^n\}_{n\geq 0}$; in order to emphasize the dependence on θ^0, we shall use the notation $\{W_{\theta^0}^n\}_{n\geq 0}$ and $\{S_{\theta^0}^n\}_{n\geq 0}$ respectively.

Our objective is to relate weak and extensive-form rationalizability in G_{θ^0} with assumptions about rationality and belief revision in IG. As a preliminary observation, intuition suggests that analyzing the complete information game G_{θ^0} should be equivalent to analyzing the game IG focusing on states where (0) the profile of payoff-types is θ^0, (1) every player $i \in I$ would be certain of (0) conditional on every history $h \in \mathcal{H}$, ... $(k+1)$ every player i would be certain of (k) conditional on every history $h \in \mathcal{H}$ ... .

The following result validates this intuition and derives its implications for weak and extensive-form rationalizability. To capture assumptions (0), (1), ... we consider the iterations of operator $\mathrm{B}_\mathcal{H}$ defined by

$$\mathrm{B}_\mathcal{H}(E) = \bigcap_{h\in\mathcal{H}} \mathrm{B}_h(E)$$

and we denote by $[\theta^0]$ the set of states in which the profile of payoff–types is θ^0: that is, $[\theta^0] = \{(\theta, s, t) \in \Omega \ : \ \theta = \theta^0\}$.

Proposition 7 *Let IG be a game with payoff uncertainty and fix $\theta^0 \in \Theta$. Then, in any belief-complete type space, for all $n \geq 0$,*

(i) $\mathrm{proj}_S \left(\bigcap_{k=0}^n \mathrm{B}_\phi^k(R) \cap \bigcap_{k\geq 0} \mathrm{B}_\mathcal{H}^k([\theta^0])\right) = W_{\theta^0}^{n+1}$*;*

(ii) $\mathrm{proj}_S \left(\mathrm{CSB}^n(R) \cap \bigcap_{k\geq 0} \mathrm{B}_\mathcal{H}^k([\theta^0])\right) = S_{\theta^0}^{n+1}$.

In the setting of Proposition 7, the assumption that the profile of payoff-types is θ^0 is accorded the highest "epistemic priority:" it is maintained throughout the game, even at histories where the event $R \cap [\theta^0]$ is falsified (furthermore, this is common certainty).

We now focus on games with *private values*: that is, we assume that, for all $i \in I$, u_i is independent of θ_{-i}. Our next result may be interpreted as stating that, in such games, assigning the same epistemic priority to the events R and $[\theta^0]$ (as well as to assumptions concerning the players' beliefs about them) is actually sufficient to obtain a characterization of weak and extensive-form rationalizability.

Proposition 8 *Let IG be a game with private values and fix $\theta^0 \in \Theta$. Then, in any belief-complete type space, for all $n \geq 0$,*

(i) $\mathrm{proj}_S \left(\bigcap_{k=0}^{n} \mathrm{B}_{\phi}^{k}(R \cap [\theta^0]) \right) = W_{\theta^0}^{n+1}$.

(ii) $\mathrm{proj}_S \left(\mathrm{CSB}^n (R \cap [\theta^0]) \right) = S_{\theta^0}^{n+1}$.

However, the assumption that the profile of payoff-types is θ^0 (and that this is common certainty) may conceivably be accorded a *low* epistemic priority. We interpret this as a form of "slight" payoff uncertainty. It is then natural to ask whether forward induction reasoning generally retains its bite in this setting. The main result of this subsection shows that, if payoff uncertainty is "diffuse," albeit slight, then the answer is negative.

More specifically, we analyze the implications of iterated correct strong belief in rationality at states in which there is common certainty *conditional on the initial history alone* that the payoff-type profile is θ^0.

In order to model "diffuse" payoff uncertainty, we assume that the game IG is *rich*: for all $j \in I$, $s_j \in S_j$ and $\mu \in \Delta^{\mathcal{H}}(\Sigma_{-j})$ there is some $\theta_j \in \Theta_j$ such that s_j is a sequential best response to μ for θ_j.[33] Embedding a complete information game within a rich game with payoff uncertainty is similar in spirit to considering "elaborations" of a given extensive game, as in Fudenberg, Kreps and Levine [25].[34]

[33] Note that a sufficient condition for richness is that each Θ_j contains an indifferent payoff-type, i.e. some θ_j^* such that $u_j(\theta_j^*, \cdot)$ is constant. Alternatively, it is sufficient to assume that, for every player i and $s_i \in S_i$, there is a type $\theta_i(s_i)$ such that s_i is weakly dominant for $\theta_i(s_i)$.

[34] On the other hand, we emphasize that our assumptions require that players assign probability zero to payoff-type profiles other than θ^0, conditional on the initial history.

Proposition 9 *Let IG be a game with private values, and fix $\theta^0 \in \Theta$. If IG is rich, then, in any belief-complete type space, for all $n \geq 0$,*

$$\mathrm{proj}_S \left(\mathrm{CSB}^n(R) \cap \left(\bigcap_{k=0}^{n} \mathrm{B}^k_\phi([\theta^0]) \right) \right) = W^{n+1}_{\theta^0}.$$

As was noted above, this result is related to Dekel and Fudenberg's [22] analysis of the robustness of iterated weak dominance with respect to "slightly incomplete information." Indeed, the procedure they characterize coincides with $\left\{ W^n_{\theta^0} \right\}_{n \geq 1}$ if G_{θ^0} is a perfect information game without ties between payoffs at terminal nodes (cf. Ben Porath [14]).

6 Conclusions

We provide some indications pertaining to extensions of our results and directions for further research.

While we have restricted our attention to games with observable actions, our characterization of extensive-form rationalizability immediately extends to general extensive games; we refer the interested reader to the previous version of this paper [10]. Similar remarks apply to the procedure of Definition 7.

According to the notion of rationalizability discussed here, a player may have correlated beliefs about his opponents. Moreover, the best rationalization principle, as axiomatized here, also reflects a notion of "correlated" belief revision: for instance, if a player observes an irrational move by one of her opponents, she is *not* required to maintain her belief in the rationality of her other opponents.

This is perfectly consistent with a noncooperative approach (e.g. Stalnaker [46] and [47]), especially outside the realm of equilibrium analysis. Also, our results imply that neither aspect is actually crucial in order to obtain the backward induction outcome in generic perfect information games.

However, at the heart the best rationalization principle is the assumption that players tend to attribute the *highest possible degree of strategic sophistication* to their opponents. Hence, a notion of stochastic independence and independent revision of beliefs about distinct opponents seems to be called for, perhaps even as a matter of consistency.

In order to focus on the somewhat more basic notions of strong belief and best rationalization, we have relegated these issues to a companion paper [11]. There we characterize a solution concept proposed by Battigalli [7] and (modulo some technical differences) Reny [40], which incorporates both forward induction ideas and independent beliefs.

In light of our analysis of the iterated Intuitive Criterion, it seems natural to investigate other refinements for signalling games, such as divinity (Banks and Sobel [6]), D1 (Cho and Kreps [20]) and related notions. More broadly, one may construct a test of equilibrium outcomes in general extensive games based on the best rationalization principle.

References

[1] ASHEIM, G. and M. DUFWENBERG (1988): “Admissibility and Common Knowledge,” mimeo, University of Oslo and Uppsala University.

[2] AUMANN, R.J. (1995): “Backward Induction and Common Knowledge of Rationality,” *Games and Economic Behavior*, **8** , 6-19.

[3] AUMANN, R.J. (1996): “Reply to Binmore,” *Games and Economic Behavior*, **17**, 138-146.

[4] AUMANN, R. and A. BRANDENBURGER (1995): “Epistemic conditions for Nash equilibrium,” *Econometrica,* **63**, 1161-1180.

[5] BALKENBORG, D. and E. WINTER (1997): “A Necessary and Sufficient Epistemic Condition for Playing Backward Induction,” *Journal of Mathematical Economics*, **27**, 325-345.

[6] BANKS, J. and J. SOBEL (1987): "Equilibrium Selection in Signalling Games," *Econometrica,* **55**, 647-662.

[7] BATTIGALLI, P. (1996): "Strategic Rationality Orderings and the Best Rationalization Principle," *Games and Economic Behavior,* **13**, 178-200.

[8] BATTIGALLI, P. (1997): "On Rationalizability in Extensive Games," *Journal of Economic Theory*, **74**, 40-61.

[9] BATTIGALLI, P. (1998): "Rationalizability in Incomplete Information Games," mimeo, Princeton University.

[10] BATTIGALLI, P. and M. SINISCALCHI (1997): "An Epistemic Characterization of Extensive Form Rationalizability," Social Science Working Paper 1009, California Institute of Technology.

[11] BATTIGALLI, P. and M. SINISCALCHI (1997): "Interactive Beliefs, Epistemic Independence and Explicability," mimeo, Princeton University and Stanford University.

[12] BATTIGALLI, P. and M. SINISCALCHI (1998): "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games," working paper ECO 98/29, European University Institute.

[13] BATTIGALLI, P. and J. WATSON (1997): "On 'Reputation' Refinements with Heterogeneous Beliefs," *Econometrica*, **65**, 369-374.

[14] BEN-PORATH, E. (1997): "Rationality, Nash Equilibrium and Backwards Induction in Perfect Information Games," *Review of Economic Studies,* **64**, 23-46.

[15] BEN-PORATH, E. and E. DEKEL (1992): "Signalling Future Actions and the Potential for Sacrifice," *Journal of Economic Theory,* **57**, 36-51.

[16] BERNHEIM, D. (1984): "Rationalizable Strategic Behavior," *Econometrica,* **52**, 1002-1028.

[17] BOARD, O. (1998): "Algorithmic Characterization of Rationalizability in Extensive Form Games," mimeo, Brasenose College, Oxford University.

[18] BRANDENBURGER, A. (1998): "On the Existence of a 'Complete' Belief Model," working paper 99-056, Harvard Business School.

[19] BRANDENBURGER, A. and E. DEKEL (1993): "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory*, **59**, 189-198.

[20] CHO, I.K. and D. KREPS (1987): "Signalling Games and Stable Equilibria," *Quarterly Journal of Economics*, **102**, 179-221.

[21] CHO, I.K. (1994): "Stationarity, Rationalizability and Bargaining," *Review of Economic Studies,* **61**, 357-374.

[22] DEKEL, E. and D. FUDENBERG (1990): "Rational Play under Payoff Uncertainty," *Journal of Economic Theory*, **52**, 243-267.

[23] DEKEL, E. and F. GUL (1997): "Rationality and Knowledge in Game Theory," in *Advances in Economics and Econometrics* (D. Kreps and K. Wallis, Eds.). Cambridge UK: Cambridge University Press.

[24] FAGIN, R., J. HALPERN, Y. MOSES and M. VARDI (1995): *Reasoning About Knowledge.* Cambridge MA: MIT Press.

[25] FUDENBERG, D., KREPS, D. and D.K. LEVINE (1988): "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory*, **44**, 354-380.

[26] FUDENBERG, D. and D. K. LEVINE (1993): "Self-Confirming Equilibrium," *Econometrica*, **61**, 523-545.

[27] FUDENBERG D. and J. TIROLE (1991): *Game Theory.* Cambridge MA: MIT Press.

[28] GÄRDENFORS, P. (1988): *Knowledge in Flux.* Cambridge MA: MIT Press.

[29] GROVE, A. (1986): “Two Modellings For Theory Change,” Auckland Philosophy Papers 13.

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

[31] KOHLBERG, E. (1990): “Refinement of Nash Equilibrium: The Main Ideas,” in *Game Theory and Applications,* ed. by T. Ichiishi, A. Neyman and Y. Tauman T. San Diego: Academic Press.

[32] KOHLBERG, E. and J.F. MERTENS (1986): “On the Strategic Stability of Equilibria,” *Econometrica,* **54**, 1003-1037.

[33] KREPS, D. and R. WILSON (1982): “Sequential Equilibria,” *Econometrica*, **50**, 863-894.

[34] McLENNAN, A. (1985): “Justifiable Beliefs in Sequential Equilibrium,” *Econometrica,* **53**, 889-904.

[35] MERTENS J.F. and S. ZAMIR (1985): “Formulation of Bayesian Analysis for Games with Incomplete Information,” *International Journal of Game Theory*, **14**, 1-29.

[36] MYERSON, R. (1986): “Multistage Games with Communication,” *Econometrica,* **54**, 323-358.

[37] OSBORNE, M. and A.RUBINSTEIN (1994): *A Course in Game Theory.* Cambridge MA: MIT

[38] PEARCE, D. (1984): “Rationalizable Strategic Behavior and the Problem of Perfection,” *Econometrica,* **52**, 1029-1050.

[39] RENY, P. (1985). “Rationality, common knowledge and the theory of games,” mimeo, Department of Economics, Princeton University.

[40] RENY, P. (1992): “Backward Induction, Normal Form Perfection and Explicable Equilibria,” *Econometrica,* **60**, 626-649.

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

[42] RUBINSTEIN, A. (1991): "Comments on the Interpretation of Game Theory," *Econometrica*, **59**, 909-904.

[43] SAMET, D. (1996): "Hypothetical Knowledge and Games with Perfect Information," *Games and Economic Behavior*, **17**, 230-251.

[44] SHIMOJI, M. and J. WATSON (1998). Conditional Dominance, Rationalizability, and Game Forms. *Journal of Economic Theory,* **83**, 161-195.

[45] SOBEL, J., L. STOLE and I. ZAPATER (1990): "Fixed-Equilibrium Rationalizability in Signaling Games," *Journal of Economic Theory,* **52**, 304-331.

[46] STALNAKER, R. (1996): "Knowledge, Belief and Counterfactual Reasoning in Games," *Economics and Philosophy*, **12**, 133-163.

[47] STALNAKER, R. (1998): "Belief Revision in Games: Forward and Backward Induction," *Mathematical Social Sciences*, **36**, 31-56.

[48] TAN, T. and S. WERLANG (1988): "The Bayesian Foundation of Solution Concepts of Games," *Journal of Economic Theory,* **45**, 370-391.

[49] van DAMME, E. (1989): "Stable Equilibria and Forward Induction," *Journal of Economic Theory,* **48**, 476-496.

[50] WATSON, J. (1996): "Reputation in Repeated Games with no Discounting," *Games and Economic Behavior*, **15**, 82-109.

7 Appendix: Proofs

7.1 Main Characterization Results

Observe that Proposition 3 follows from Proposition 5 by assuming that, for each player $i \in I$, Θ_i consists of a single element (so there is a one-to-one correspondence between Σ_i and S_i, and we need not distinguish between the two sets) and $\Delta_i = \Delta^{\mathcal{H}}(S_{-i})$.

Proposition 8 does *not* follow from Proposition 5, but the proofs are very similar. We shall emphasize the proof of Proposition 5, and note the modifications required to establish Proposition 8.

We begin with two preliminary results.

Lemma 10 *Fix a map $\tau_{-i} : \Sigma_{-i} \to T_{-i}$. Also, fix a first-order CPS $\delta_i \in \Delta^{\mathcal{H}}(\Sigma_{-i})$. 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*

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

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

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

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

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

$$g_{i,h}(t_i)\left((\sigma_{-i}, \tau_{-i}(\sigma_{-i}))\right) = \mu_i((\sigma_{-i}, \tau_{-i}(\sigma_{-i}))|\Sigma_{-i}(h)\times T_{-i}) = \delta_i(\sigma_{-i}|\Sigma_{-i}(h))$$

for all $\sigma_{-i} \in \Sigma$ and $h \in \mathcal{H}$. ■

The next lemma provides an alternative characterization of $\{\Sigma^n_\Delta\}^\infty_{n=0}$, where $\Delta = (\Delta_i)_{i \in I}$ is any regular collection of subsets of CPSs.

Lemma 11 *Suppose Δ is regular. For every $i \in I$ and $n \geq 1$, $\sigma_i \in \Sigma^n_{i,\Delta}$ if and only if there exists a CPS $\mu \in \Delta_i$ such that $\sigma_i \in r_i(\mu)$ and*

$$\forall m = 0, \ldots, n-1,\ \forall h \in \mathcal{H}: \quad \Sigma^m_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset \Rightarrow \mu(\Sigma^m_{-i,\Delta} | \Sigma_{-i}(h)) = 1 \tag{2}$$

Proof: The statement is obvious for $n = 1$. Now pick $n \geq 2$ and assume it is true for $m = 0, \ldots, n-1$. If $\sigma_i \in r_i(\mu)$ for some $\mu \in \Delta_i$ satisfying (2), then $\sigma_i \in \Sigma^{n-1}_{i,\Delta}$ by the induction hypothesis, because $\Sigma^m_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset \Rightarrow \mu(\Sigma^m_{-i,\Delta} | \Sigma_{-i}(h)) = 1$ for $m = 0 \ldots n-2$; moreover, since also $\Sigma^{n-1}_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset \Rightarrow \mu(\Sigma^{n-1}_{-i,\Delta} | \Sigma_{-i}(h)) = 1$, and $\sigma_i \in r_i(\mu)$, we conclude $\sigma_i \in \Sigma^n_{i,\Delta}$.

In the other direction, suppose $\sigma_i \in \Sigma^n_{i,\Delta}$. Then also $\sigma_i \in \Sigma^m_{i,\Delta}$ for $m = 0, \ldots, n-1$, so we can find CPSs $\mu^m \in \Delta_i$, $m = 0, \ldots, n-1$, such that, for each such m, $\sigma_i \in r_i(\mu^m)$ and, for any $h \in \mathcal{H}$, $\Sigma^m_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset$ implies $\mu^m(\Sigma^m_{-i,\Delta} | \Sigma_{-i}(h)) = 1$. Now construct a new CPS μ as follows: for any $h \in \mathcal{H}$, let $m(h) = \max\{m = 0, \ldots, n-1 \ : \ \Sigma^m_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset\}$, and define $\mu(\cdot | \Sigma_{-i}(h)) = \mu^{m(h)}(\cdot | \Sigma_{-i}(h))$. It is easy to verify that this is a well-defined CPS, i.e. $\mu \in \Delta^{\mathcal{H}}(\Sigma_{-i}))$ (for a similar construction, see e.g. Battigalli [8]).

By construction, $\mu(\cdot | \Sigma_{-i}(h)) \in \Delta_{i,h}$ for all h. By definition of regularity, $\Delta_i = \Delta^{\mathcal{H}}(\Sigma_{-i}) \cap \prod_{h \in \mathcal{H}} \Delta_{i,h}$. Therefore $\mu \in \Delta_i$. Moreover, clearly $\sigma_i \in r_i(\mu)$. Finally, μ satisfies (2), which concludes the proof. ■

Note that Lemma 11 also applies to games with complete information (take Θ_i to be a singleton for each player); hence, in the setting of Section 5.4, it applies to the game G_{θ^0} and the sets $S^n_{\theta^0}$, $n = 0, 1, \ldots$.

We can finally prove our main result.

Proof of Proposition 5: To prove (i), we proceed by induction, assuming first that the sets appearing in the statement are nonempty.

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

Conversely, for each $i \in I$ and $\sigma_i \in \Sigma_i$, pick $\tau^0_i(\sigma_i) \in T_i$ arbitrarily. Now fix $\sigma \in \Sigma^1_\Delta$, and for each player $i \in I$, let $\mu_i \in \Delta_i$ be such that $\sigma_i \in r_i(\mu_i)$. Now Lemma 10 yields a type $\tau^1_i(\sigma_i) \in T_i$ such that $g_{i,h}(\tau^1_i(\sigma_i))(\{(\sigma'_j, \tau^0_j(\sigma_j))_{j \neq i}\}) = \mu_i(\sigma'_{-i}|\Sigma_{-i}(h))$ for every $\sigma'_{-i} \in \Sigma_{-i}$, and hence $f_i(\tau^1_i(\sigma_i)) = \mu_i$. Thus, $(\sigma_i, \tau^1_i(\sigma_i))_{i \in I} \in R \cap E(\Delta)$.

Finally, for each $i \in I$, we complete the definition of the function $\tau^1_i(\cdot)$ by letting $\tau^1_i(\sigma_i) = \tau^0_i(\sigma_i)$ for $\sigma_i \in \Sigma_i \setminus \Sigma^1_{i,\Delta}$.

(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 $\tau^{m+1}_i : \Sigma_i \to T_i$ such that $(\sigma_i, \tau^{m+1}_i(\sigma_i))_{i \in I} \in \mathrm{CSB}^m(R \cap E(\Delta))$ whenever $\sigma \in \Sigma^{m+1}_\Delta$. Finally, let the functions $\tau^0_i(\cdot)$ be defined as above.

Note that, for any event $E \in \mathcal{A}$ and $n \geq 1$,

$$\mathrm{CSB}^n(E) = E \cap \bigcap_{i \in I} \left\{ \bigcap_{m=0}^{n-1} \mathrm{SB}_i(\Omega_i \times [\mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^m(E)]) \right\} \tag{3}$$

Also note that, for any $i \in I$, $h \in \mathcal{H}$ and event E such that $\mathrm{proj}_{\Omega_i} E = \Omega_i$,

$$E \cap (\Sigma(h) \times T) \neq \emptyset \quad \Leftrightarrow \quad [\mathrm{proj}_{\Sigma_{-i}} E] \cap \Sigma_{-i}(h) \neq \emptyset \tag{4}$$

Now consider $(\sigma, t) \in \mathrm{CSB}^n(R \cap E(\Delta))$ and fix $i \in I$. By Equation 3 (taking $E = R \cap E(\Delta)$) we conclude that $\sigma_i \in r_i(f_i(t_i))$ and $f_i(t_i) \in \Delta_i$; also, for any $m = 0, \ldots, n-1$, the induction hypothesis and Equation 4 imply that, for any $h \in \mathcal{H}$, $\Sigma^{m+1}_{-i,\Delta} \cap \Sigma_{-i}(h) = [\mathrm{proj}_{\Sigma_{-i}} \mathrm{CSB}^m(R \cap E(\Delta)] \cap \Sigma_{-i}(h) \neq \emptyset$ if and only if $[\Omega_i \times \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^m(R \cap E(\Delta))] \cap (\Sigma(h) \times T) \neq \emptyset$. Now Equation 3 and the definition of strong belief implies that, for any $h \in \mathcal{H}$ satisfying the latter condition for some

$m = 0, \dots, n-1$, $g_{i,h}(t_i)(\mathrm{proj}_{\Omega_{-i}}\mathrm{CSB}^m(R \cap E(\Delta))) = 1$. This implies $f_i(t_i)(\mathrm{proj}_{\Sigma_{-i}}\mathrm{CSB}^m(R \cap E(\Delta))|\Sigma_{-i}(h)) = 1$; in turn, the induction hypothesis implies $f_i(t_i)(\Sigma^{m+1}_{-i,\Delta}|\Sigma_{-i}(h)) = 1$. Hence, Lemma 11 implies that $\sigma_i \in \Sigma^{n+1}_{i,\Delta}$.

For the converse implication, begin by defining

$$m_i(\sigma_i) = \max\{m = 0, \dots, n \ : \ \sigma_i \in \Sigma^m_{i,\Delta}\}$$

for every $i \in I$ and $\sigma_i \in \Sigma_i$; recall that $\Sigma^0_{i,\Delta} = \Sigma_i$, so $m_i(\cdot)$ is well-defined for every $\sigma_i \in \Sigma_i$. Now consider $\sigma \in \Sigma^{n+1}_{\Delta}$ and fix a player $i \in I$. By Lemma 11, we can find a CPS $\mu_i \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ satisfying Equation 2. By Equation 4 and the induction hypothesis, for $h \in \mathcal{H}$ and $m = 0, \dots, n-1$, $[\Omega_i \times \mathrm{proj}_{\Omega_{-i}}\mathrm{CSB}^m(R \cap E(\Delta))] \cap (\Sigma(h) \times T) \neq \emptyset$ if and only if $\Sigma^{m+1}_{-i,\Delta} \cap \Sigma_{-i}(h) \neq \emptyset$. But if the latter inequality holds, $\mu_i(\Sigma^{m+1}_{-i,\Delta}|\Sigma_{-i}(h)) = 1$ by Equation 2.

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

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

Lemma 10 now yields a type $\tau_i^{n+1}(\sigma_i) \in T_i$ such that

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

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

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

because, for all $j \neq i$: (a) $m_j(\sigma'_j) \geq m+1$ if $\sigma'_{-i} \in \Sigma^{m+1}_{-i,\Delta}$; (b) if $m_j(\sigma'_j) \geq 1$ then, by the induction hypothesis,

$$(\sigma'_j, \tau_j^{m_j(\sigma'_j)}(\sigma'_j)) \in \mathrm{proj}_{\Omega_j}\mathrm{CSB}^{m_j(\sigma'_j)-1}(R \cap E(\Delta));$$

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

$$g^*_{i,h}(\sigma_i, \tau_i^{n+1}(\sigma_i))(\Omega_i \times \mathrm{proj}_{\Omega_{-i}}\mathrm{CSB}^m(R \cap E(\Delta))) = 1$$

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

Moreover, since by construction $f_i(\tau_i^{n+1}(\sigma_i)) = \mu_i$, we also have $\sigma_i \in r_i(f_i(\tau_i^{n+1}(\sigma_i)))$ and $f_i(\tau_i^{n+1}(\sigma_i)) \in \Delta_i$.

Repeating the argument for every $i \in I$ yields a profile of types $(\tau_i^{n+1}(\sigma_i))_{i \in I}$ which, by Equation 3, satisfies $(\sigma_i, \tau_i^{n+1}(\sigma_i))_{i \in I} \in \text{CSB}^n(R \cap E(\Delta))$. To complete the induction step, for each $i \in I$ we now define the function $\tau_i^{n+1}(\cdot)$ for $\sigma_i \in \Sigma_i \setminus \Sigma^{n+1}_{i,\Delta}$ by letting $\tau_i^{n+1}(\sigma_i) = \tau_i^n(\sigma_i)$ for any such strategy σ_i.

The argument just given shows that if one of the sets appearing in the statement of (i) is nonempty, so is the other one. Hence, the proof of (i) is complete.

For Part (ii), assume first that $\Sigma^\infty_\Delta \neq \emptyset$. Then $\Sigma^n_\Delta \neq \emptyset$ for all $n \geq 0$; hence $\text{CSB}^n(R \cap E(\Delta)) \neq \emptyset$ for $n \geq 0$ by Part (i). Then $\text{CSB}^\infty(R \cap E(\Delta))$ is nonempty, because T is compact by assumption and the nested, nonempty closed sets $\{\text{CSB}^n(R \cap E(\Delta))\}_{n \geq 0}$ form a family with the finite intersection property.

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

Next, let N be the smallest integer such that $\Sigma^N_\Delta = \Sigma^\infty_\Delta$ (which must exist because Σ is finite). Pick any $\sigma \in \Sigma^N_\Delta = \Sigma^\infty_\Delta$ and consider the sequence of sets $M(m, \sigma) = \text{CSB}^{(N-1)+m}(R \cap E(\Delta)) \cap (\{\sigma\} \times T)$, $m \geq 0$ (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 $\emptyset \neq \bigcap_{m \geq 0} M(m, \sigma) \subset \text{CSB}^\infty(R \cap E(\Delta))$, so $\Sigma^\infty_\Delta \subset \text{proj}_\Sigma \text{CSB}^\infty(R \cap E(\Delta))$.

If $\Sigma^\infty_\Delta = \emptyset$, let N be the smallest integer such that $\Sigma^N_\Delta = \emptyset$. Since $\Sigma^N_\Delta = \text{proj}_\Sigma \text{CSB}^{N-1}(R \cap E(\Delta))$, we conclude that $\text{CSB}^{N-1}(R \cap E(\Delta)) = \emptyset$, so $\text{CSB}^\infty(R \cap E(\Delta)) = \emptyset$, and again $\Sigma^\infty_\Delta = \text{proj}_\Sigma \text{CSB}^\infty(R \cap E(\Delta))$. ■

The following observations allow one to modify the preceding argument to prove Part (ii) of Proposition 8. Fix a player $i \in I$ and a profile of payoff-types $\theta^0 \in \Theta$.

First, note that, for any CPS $\mu_i \in \Delta^{\mathcal{H}}(\Sigma_{-i})$, one can define a "marginal" CPS $\mu_i^S \in \Delta^{\mathcal{H}}(S_{-i})$ by letting $\mu_i^S(\{s_{-i}\}|\Sigma_{-i}(h)) = \mu_i(\Theta_{-i} \times \{s_{-i}\}|\Sigma_{-i}(h))$ for each $h \in \mathcal{H}$; conversely, for any CPS $\nu_i^S \in \Delta^{\mathcal{H}}(S_{-i})$, one can define a CPS $\nu_i \in \Delta^{\mathcal{H}}(\Sigma_{-i})$ by letting $\nu_i(\{(\theta_{-i}^0, s_{-i})\}|\Sigma_{-i}(h)) = \nu_i^S(\{s_{-i}\}|S_{-i}(h))$ for each $h \in \mathcal{H}$.

With these definitions, for any strategy $s_i \in S_i$, by the private values assumption $(\theta_i^0, s_i) \in r_i(\mu_i)$ implies $s_i \in r_{i,\theta^0}(\mu_i^S)$, and conversely $s_i \in r_{i,\theta^0}(\nu_i^S)$ implies $(s_i, \theta_i^0) \in r_i(\nu_i)$, where $r_{i,\theta^0}(\cdot)$ denotes Player i's best response correspondence in the game G_{θ^0}. This allows one to adapt Step 0 in the above proof of Part (i) and show that $S_{\theta^0}^1 = W_{\theta^0}^1 = \text{proj}_S R \cap [\theta^0]$. We leave the rest of the proof of Part (i) to the reader.

As for the proof of Part (ii), note that, since $\Sigma(h) = \Theta \times S(h)$ for all $h \in \mathcal{H}$, $E \cap (\Sigma(h) \times T) \neq \emptyset$ if and only if $[\text{proj}_{S_{-i}} E] \cap S_{-i}(h) \neq \emptyset$ for any event E such that $\text{proj}_{S_i} E = S_i$. The inductive step in the proof of Part (i) of Proposition 5 may then be easily adapted to the present context. Again we leave the details to the reader.

7.2 Other Proofs

Proposition 2

Proof. The equality $S^1 \cap S(h) = W_h^1$ holds by definition. Suppose by way of induction that $S^n \cap S(h) \subset W_h^n$ and let $s \in S^{n+1} \cap S(h)$. Since $S^{n+1} \subset S^n$ it follows that $S_{-i}^n \cap S_{-i}(h) \neq \emptyset$. By definition of S^{n+1}, for each $i \in I$ there is some $\mu_i \in \Delta^{\mathcal{H}}(S_{-i})$ such that $s_i \in r_i(\mu_i)$ and $\mu_i(S_{-i}^n|S_{-i}(h)) = 1$. By the inductive hypothesis $S_{-i}^n \cap S_{-i}(h) \subset W_{-i,h}^n$. Therefore $\mu_i(W_{-i,h}^n|S_{-i}(h)) \geq \mu_i(S_{-i}^n|S_{-i}(h)) = 1$. Since $s_i \in r_i(\mu_i) \cap S_i(h)$ and $\mu_i(W_{-i,h}^n|S_{-i}(h)) = 1$ for all $i \in I$, then $s \in W_h^{n+1}$. This conclude the proof of the first statement.

It follows that if $S^\infty \cap S(h) \neq \emptyset$ also $W_h^\infty \neq \emptyset$. But $W_h^\infty = \text{proj}_S \bigcap_{n \geq 0} \mathrm{B}_h^n(R)$. Therefore $S^\infty \cap S(h) \neq \emptyset$ implies $\bigcap_{n \geq 0} \mathrm{B}_h^n(R) \neq \emptyset$. ■

Proposition 6

Proof. (1) Fix $\overline{\omega} = (\overline{\theta}, \overline{m}, \overline{t}_1, \overline{s}_2, \overline{t}_2) \in \bigcap_{i=1,2} R_i \cap [\zeta]_i \cap \mathrm{B}_{i,\phi}\left(R_{-i} \cap [\zeta]_{-i}\right)$.

Since $\overline{\omega} \in R_2 \cap [\zeta]_2 \cap \mathrm{B}_{2,\phi}\left(R_1 \cap [\zeta]_1\right)$, for each $\theta \in \Theta$ and each m with $\zeta(m|\theta) > 0$, there is some epistemic type $t_1^{\theta,m}$ such that $(\theta, m, t_1^{\theta,m})$ is in the support of $g_{2,\phi}(\overline{t}_2)$, the Sender is rational and her beliefs agree with ζ at $(\theta, m, t_1^{\theta,m}, \overline{s}_2, \overline{t}_2)$. For any such θ and m, let $\pi_2^{\theta,m}$ be the behavioral representation of $f_{1,\phi}(t_1^{\theta,m}) = \mathrm{marg}_{S_2} g_{1,\phi}(t_1^{\theta,m})$, that is,

$$\forall m' \in M, \forall a \in A,\ \pi_2^{\theta,m}(a|m') = f_{1,\phi}(t_1^{\theta,m})\left(\{s_2 : s_2(m') = a\}\right).$$

By agreement with ζ, $\pi_2^{\theta,m}(\cdot|m') = \mathrm{marg}_A \zeta(\cdot|m')$ whenever $\zeta(m') > 0$. Therefore, we can fix an arbitrary $\bar{m} \in M$ with $\zeta(\bar{m}) > 0$, define $\pi_2^{\theta} \equiv \pi_2^{\theta,\bar{m}}$ and conclude that, for all m, m' such that $\zeta(m) > 0$ and $\zeta(m') > 0$, $\pi_2^{\theta}(.|m') = \pi_2^{\theta,m}(.|m') = \mathrm{marg}_A \zeta(.|m')$. This implies that, for any equilibrium message m, the expected payoff calculated using either π_2^{θ} or $\pi_2^{\theta,m}$ is $u_1^{\zeta}(\theta)$.

We claim that each message m such that $\zeta(m|\theta) > 0$ is a best response for θ to π_2^{θ}, that is, condition (1) of Definition 8 is satisfied. Clearly, each such m is a best response to $\pi_2^{\theta,m}$ for θ because the Sender is rational at state $(\theta, m, t_1^{\theta,m}, \overline{s}_2, \overline{t}_2)$. The Sender's expected payoff following m is $u_1^{\zeta}(\theta)$; but note that

$$\forall m' \in M,\ u_1^{\zeta}(\theta) = \sum_a u_1(\theta, \bar{m}, a)\pi_2^{\theta}(a|\bar{m}) \geq \sum_a u_1(\theta, m', a)\pi_2^{\theta}(a|m')$$

because $\bar{m}$ is itself an equilibrium message. This proves the claim.

Repeating the process for each θ we obtain the required *tuple* of behavior strategies $\left(\pi_2^{\theta}\right)_{\theta\in\Theta}$.

Let π_2 be the behavioral representation of $f_{1,\phi}(\overline{t}_1)$. Since the Sender's beliefs agree with ζ and the Sender believes that the Receiver is rational and also has beliefs agreeing with ζ, it must be the case that for each m with $\zeta(m) > 0$, $\pi_2(\cdot|m) = \mathrm{marg}_A \zeta(\cdot|m)$ and for each a with $\pi_2(a|m) > 0$, a is a best reply to belief $\mathrm{marg}_{\Theta} \zeta(\cdot|m)$ given m. Therefore also condition (2) of Definition 8 is satisfied.

(2) First notice that

$$\mathrm{CSB}\left(R\cap[\zeta]_1\cap[\zeta]_2\right)\subset\bigcap_{i=1,2}R_i\cap[\zeta]_i\cap\mathrm{B}_{i,\phi}\left(R_{-i}\cap[\zeta]_{-i}\right).$$

Therefore, by Part (1), $\mathrm{CSB}^\infty\left(R\cap[\zeta]_1\cap[\zeta]_2\right)\neq\emptyset$ implies that ζ is a self-confirming equilibrium outcome. To complete the proof of part (2) we rely on a specialization of our main characterization result, Proposition 5, to the present setting. We introduce the following notation.

Let $\Sigma_1^0=\Sigma_1=\Theta\times M$, $S_2^0=S_2$,

$\Delta_1(\zeta)=\{\mu\in\Delta(S_2):\forall m,\forall a,\zeta(m)>0\Rightarrow\mu(\{s_2:s_2(m)=a\})=\zeta(a|m)\}$,

$\Delta_2^{\mathcal{H}}(\zeta)=\left\{\mu\in\Delta^{\mathcal{H}}(\Sigma_1):\forall(\theta,m),\mu(\theta,m|\Sigma_1)=\zeta(\theta,m)\right\}$.

$\Delta_1(\zeta)$ is the set of initial first order beliefs of the Sender about the Receiver that agree with ζ, $\Delta_2^{\mathcal{H}}(\zeta)$ has a similar meaning. In particular, observe that these restrictions on beliefs are regular. It is convenient to have a special notation for the system of beliefs derived from some CPS μ on $(\Sigma_1,\mathcal{H})$: $\nu_\mu\in[\Delta(\Theta)]^M$ satisfies $\nu_\mu(\theta|m)=\mu(\theta,m|\Sigma_1(m))$ for all (θ,m). For every $\mu\in\Delta(S_2)$, we let $BR_1(\mu,\theta)$ denote the set messages that maximize the expected payoff of type θ of the Sender against μ. Similarly, for every message m and belief $\nu^m\in\Delta(\Theta)$, $BR_2(\nu^m,m)$ is the set of Receiver's best responses. Then the iterative deletion procedure corresponding to the sequence of events $\mathrm{CSB}^k\left(R\cap[\zeta]_1\cap[\zeta]_2\right)$ is, for all $k=0,1,...$

$\Sigma_1^{k+1}=\left\{(\theta,m)\in\Sigma_1^k:\exists\mu\in\Delta_1(\zeta),\ m\in BR_1(\mu,\theta),\ \mu(S_2^k)=1\right\}$,

$S_2^{k+1}=\{s_2\in S_2^k:\exists\mu\in\Delta_2^{\mathcal{H}}(\zeta),\forall h\in\mathcal{H},\ \Sigma_1^k\cap\Sigma_1(h)\neq\emptyset\Rightarrow\mu(\Sigma_1^k|\Sigma_1(h))=1,\forall m\in M,s_2(m)\in BR_2\left(\nu_\mu(\cdot|m),m\right)\}$.

The characterization result yields $\mathrm{proj}_\Sigma\mathrm{CSB}^k\left(R\cap[\zeta]_1\cap[\zeta]_2\right)=\Sigma_1^{k+1}\times S_2^{k+1}$.

Now, for every step k of the procedure and every message m, let $\Theta^k(m)$ and $A^k(m)$ respectively denote the types and responses consistent with step k given message m, that is,

$\Theta^k(m)=\left\{\theta\in\Theta:(\theta,m)\in\Sigma_1^k\right\}$,

$A^k(m)=\left\{a\in A:\exists s_2\in S_2^k,s_2(m)=a\right\}$.

(2.a) We first prove the following "*decomposition property*": $s_2 \in S_2^k$ if and only if $s_2(m) \in A^k(m)$ for all m. The "only if" part is true by definition. Now suppose that $s_2(m) \in A^k(m)$ for all m. Then we can find strategies $s_2^m \in S_2^k$ and CPSs $\mu^m \in \Delta^{\mathcal{H}}(\Sigma_1)$ $(m \in M)$ such that, for all m, $s_2(m) = s_2^m(m) \in BR_2(\nu_{\mu^m}(\cdot|m), m)$, $\mu^m(\cdot|\Sigma_1) =$marg$_{\Theta\times M}\zeta$ (recall that $\Sigma_1 := \Theta \times M$), $\mu^m(\Sigma_1^{k-1}|\Sigma_1) = 1$, $\mu^m(\Sigma_1^{k-1}|\Sigma_1(m')) = 1$ whenever $\Sigma_1^{k-1} \cap \Sigma_1(m') \neq \emptyset$. Construct $\mu(\cdot|\cdot)$ as follows: $\mu(\cdot|\Sigma_1) =$ marg$_{\Theta\times M}\zeta$, $\mu(\theta, m|\Sigma_1(m)) = \zeta(\theta|m)$ for all θ and m with $\zeta(m) > 0$, and $\mu(\cdot|\Sigma_1(m)) = \mu^m(\cdot|\Sigma_1(m))$ for all m with $\zeta(m) = 0$. It can be checked that $\mu \in \Delta_2^{\mathcal{H}}(\zeta)$, $s_2(m) \in BR_2(\nu_\mu(\cdot|m), m)$ for all m, $\mu(\Sigma_1^{k-1}|\Sigma_1) = 1$ and $\mu(\Sigma_1^{k-1}|\Sigma_1(m)) = 1$ whenever $\Sigma_1^{k-1} \cap \Sigma_1(m) \neq \emptyset$. Therefore $s_2 \in S_2^k$.

(2.b) Next we prove a property of the $\Theta^k(m)$, $A^k(m)$ sequences: $\Theta^k(m) = \emptyset$ implies $A^{k+1}(m) = A^k(m)$. $\Theta^k(m) = \emptyset$ is equivalent to $\Sigma_1^k \cap \Sigma_1(m) = \emptyset$. Suppose this condition holds. We only have to prove that in this case $A^k(m) \subset A^{k+1}(m)$. Let $a \in A^k(m)$. Then there are $s_2 \in S_2^{k-1}$ and μ satisfying the conditions for $s_2 \in S_2^k$ such that $s_2(m) = a$. In particular, $a \in BR_2(\nu_\mu(\cdot|m), m)$. Now pick a strategy s_2' and a CPS μ' satisfying the conditions for $s_2' \in S_2^{k+1}$. Construct a new CPS μ^* which coincides with μ' for all $h \in \mathcal{H}\backslash\{m\}$ and coincides with μ for $h = m$. Let s_2^* be the strategy choosing a after m and $s_2'(m')$ for $m' \neq m$. Then s_2^* and μ^* satisfy the conditions for $s_2^* \in S_2^{k+1}$. (In particular, μ^* is a CPS satisfying the required conditions because $\mu^*(\Sigma_1^k|\Sigma_1) = 1$ and $\Sigma_1^k \cap \Sigma_1(m) = \emptyset$.) Therefore $a \in A^{k+1}(m)$.

(2.c) In order to prove part (2) of the proposition it is sufficient to show that if (A) either ζ is a self-confirming equilibrium outcome passing the Iterated Intuitive Criterion (IIC) or $\bigcap_k \Sigma_1^k \times S_2^k \neq \emptyset$, then (B) for all $m \in M$ and $k = 0, 1, ...,$

$$\zeta(m) = 0 \Rightarrow [I\Theta^k(m;\zeta) = \Theta^k(m) \text{ and } IA^k(m;\zeta) = A^k(m)]. \qquad (5)$$

Suppose assumption (A) holds. By definition, 5 holds for $k = 0$.

Assume that 5 holds for all $k = 0, 1, ..., n$ and fix a message m with $\zeta(m) = 0$ (a message "off-the-path").

($\Theta^{n+1}(m) \subset I\Theta^{n+1}(m;\zeta)$) If $\theta \in \Theta^{n+1}(m)$, there is a conjecture $\mu \in \Delta_1(\zeta)$ such that $\mu(S_2^n) = 1$ and $m \in BR_1(\mu, \theta)$. The behavioral representation of μ is a π_2^μ such that for all $m' \in M$, if $\zeta(m') > 0$, $\pi_2^\mu(\cdot|m') = \zeta(\cdot|m')$, and if $\zeta(m') = 0$, supp $\pi_2^\mu(\cdot|m') \subset A^n(m')$. By the inductive hypothesis $A^n(m) = IA^n(m;\zeta)$. Therefore there is some $a \in \mathrm{supp}\pi_2^\mu(\cdot|m) \subset IA^n(m;\zeta)$ such that $u_1(\theta, m, a) \geq u_1^\zeta(\theta)$, which implies $\theta \in I\Theta^{n+1}(m;\zeta)$.

($I\Theta^{n+1}(m;\zeta) \subset \Theta^{n+1}(m)$) **Claim:** By assumption (A) and the inductive hypothesis, for every payoff-type θ there is mapping $a'(\cdot)$ such that for all messages m' off-the-path $a'(m') \in A^n(m') = IA^n(m';\zeta)$ and $u_1^\zeta(\theta) \geq u_1(\theta, m', a'(m'))$.

The claim is obvious if ζ satisfies the IIC. Suppose that $\bigcap_k \Sigma_1^k \times S_2^k \neq \emptyset$. Then, in particular, $S_2^{n+1} \neq \emptyset$ and it must be possible to find Receiver's beliefs $\mu \in \Delta^{\mathcal{H}}(\Sigma_1)$ such that $\mu(\theta, m'|\Sigma_1) = \zeta(\theta, m')$ for all m' and $\mu(\Sigma_1^n|\Sigma_1) = 1$. If we had $u_1(\theta, m', a) > u_1^\zeta(\theta)$ for all m' off-the-path and actions $a \in A^n(m')$, then $\mu(\Sigma_1^n|\Sigma_1) = 1$ would imply that $\mu(\theta, m^*|\Sigma_1) = 0 < \zeta(\theta, m^*)$ for all on-the-path messages m^* and no belief rationalizing strategies in S_2^{n+1} would exist. This establishes the claim.

Now let $\theta \in I\Theta^{n+1}(m;\zeta)$. Then there is an action $a^* \in A^n(m)$ such that $u_1(\theta, m, a^*) \geq u_1^\zeta(\theta) \geq u_1(\theta, m', a'(m'))$ for all m' with $\zeta(m') = 0$. Define $\mu^* \in \Delta(S_2)$ as follows: for all $s_2 \in S_2$, $\mu^*(s_2) = \prod_{m^*:\zeta(m^*)>0} \zeta(s_2(m^*)|m^*)$ if $s_2(m) = a^*$ and $s_2(m') = a'(m')$ for all $m' \neq m$ with $\zeta(m') = 0$; $\mu^*(s_2) = 0$ otherwise. By construction, $m \in BR_1(\mu^*, \theta)$ and $\mu^* \in \Delta_1(\zeta)$. Furthermore, $\mu^*(s_2) > 0$ implies that $s_2(m') \in A^n(m')$ for all m'. By the "decomposition property" proved above, $\mu^*(S_2^n) = 1$. Therefore $\theta \in \Theta^{n+1}(m)$.

($A^{n+1}(m) = IA^{n+1}(m;\zeta)$) Suppose that $\Theta^n(m) = I\Theta^n(m;\zeta) = \emptyset$. We proved in part (2.b) above that in this case $A^{n+1}(m) = A^n(m)$. By the inductive hypothesis and the definition of $IA^{n+1}(m;\zeta)$, $A^{n+1}(m) = IA^n(m;\zeta) = IA^{n+1}(m;\zeta)$.

Now suppose that $\Theta^n(m) = I\Theta^n(m;\zeta) \neq \emptyset$. By definition, every $a \in A^{n+1}(m)$ is a best response to some belief $\nu(\cdot|m)$ concentrated on $\Theta^n(m)$. Thus $A^{n+1}(m) \subset BR_2(I\Theta^n(m;\zeta), m) = IA^{n+1}(m;\zeta)$. Let

$a \in IA^{n+1}(m;\zeta) = BR_2(I\Theta^n(m;\zeta), m)$. By the inductive hypothesis, $a \in BR_2(\nu(\cdot|m), m)$ for some belief $\nu(\cdot|m)$ concentrated on $\Theta^n(m) = I\Theta^n(m;\zeta)$. Using the same procedure as in part (2.b) of this proof, we can find a strategy s_2^* and a CPS μ^* such that $s_2^*(m) = a$, $\nu_{\mu^*}(\cdot|m) = \nu(\cdot|m)$ and satisfying the conditions for $s_2^* \in S_2^{n+1}$. (In particular, μ^* is a CPS because by agreement with ζ it must assign zero probability to off-equilibrium-path message m. Thus Bayes' rule does not restrict the value of $\mu^*(\cdot|\Sigma_1(m))$.) Therefore $a \in A^{n+1}(m)$. ■

Proposition 7

Proof. Given a belief-complete type space for game IG we derive a belief-complete type space for game G_{θ^0} as follows:

For all $k = 0, 1, ...$ and $i \in I$, let $T^0_{\theta^0,i} = T_i$,

$$T^{k+1}_{\theta^0,i} = \left\{ t_i \in T^k_{\theta^0,i} : \forall h \in \mathcal{H}, g_{i,h}(t_i) \left(\prod_{j \neq i} \{\theta^0_j\} \times S_j \times T^k_{\theta^0,j} \right) = 1 \right\}$$

and

$$T_{\theta^0,i} = \bigcap_{k \geq 0} T^k_{\theta^0,i}.$$

We take $T_{\theta^0,i}$ to be Player i's space of epistemic types in game G_{θ^0} and define the belief mapping $g_{\theta^0,i} : T_{\theta^0,i} \to \Delta^{\mathcal{H}} \left(\prod_{j \neq i} S_j \times T_{\theta^0,j} \right)$ so that, for all $t_i \in T_{\theta^0,i}$, $g_{\theta^0,i}(t_i)$ is the CPS satisfying

$$\begin{aligned} \forall h &\in \mathcal{H}, \forall s_{-i} \in S_{-i}, \forall K_{-i} \subset T_{\theta^0,-i} \text{ (measurable)}, \\ g_{\theta^0,i,h}(t_i)\left(\{s_{-i}\} \times K_{-i}\right) &= g_{i,h}(t_i)\left(\{\theta^0_{-i}, s_{-i}\} \times K_{-i}\right). \end{aligned}$$

(we abuse notation in writing ordered tuples and Cartesian products: the meaning is obvious). By construction $(T_{\theta^0,i}, g_{\theta^0,i})_{i \in I}$ defines a belief-complete type space for game G_{θ^0} and for all $i \in I$, $(s,t) \in S \times T_{\theta^0}$, $h \in \mathcal{H}$, $E \subset S \times T_{\theta^0}$ (measurable),

$$(s,t) \in \mathrm{B}_{\theta^0,i,h}(E) \Leftrightarrow (\theta^0, s, t) \in \mathrm{B}_{i,h}(\{\theta^0\} \times E)$$

and

$$(s,t) \in R_{\theta^0} \Leftrightarrow (\theta^0, s, t) \in R \cap [\theta^0],$$

where $\mathrm{B}_{\theta^0,i,h}$ and R_{θ^0} denote the (i,h)-belief operator and the rationality event in the type space for game G_{θ^0}. By Propositions 1 and 3 these equivalences imply the thesis. ■

Proposition 9

Proof. The statement is obviously true for $n = 0$. Suppose it is true for index $n-1$. It can be easily shown by induction that, for every event E, $\mathrm{CSB}^n(E) \subset \bigcap_{k=0}^{n} \mathrm{B}_\phi^k(E)$, which implies

$$\mathrm{CSB}^n(R) \cap \left(\bigcap_{k=0}^{n} \mathrm{B}_\phi^k([\theta^0]) \right) \subset \bigcap_{k=0}^{n} \left(\mathrm{B}_\phi^k(R) \cap \mathrm{B}_\phi^k([\theta^0]) \right) = \bigcap_{k=0}^{n} \left(\mathrm{B}_\phi^k(R \cap [\theta^0]) \right).$$

Therefore, by Proposition 8, $\mathrm{proj}_S \left(\mathrm{CSB}^n(R) \cap \left(\bigcap_{k=0}^{n} \mathrm{B}_\phi^k([\theta^0]) \right) \right) \subset W_{\theta^0}^{n+1}$.

Assume that $s \in W_{\theta^0}^{n+1}$. Fix $i \in I$. By assumption there exists a CPS $\nu \in \Delta^{\mathcal{H}}(S_{-i})$ such that $s_i \in r_{\theta^0,i}(\nu)$ ($r_{\theta^0,i}$ is Player i's best response correspondence in G_{θ^0}) and $\nu \left(W_{\theta^0,-i}^n | S_{-i} \right) = 1$. We now construct a CPS $\mu \in \Delta^{\mathcal{H}}(\Omega_{-i})$ having ν as marginal CPS on S_{-i}.

By the induction hypothesis, $\mathrm{proj}_{S_{-i}} \left(\mathrm{CSB}^{n-1}(R) \cap \left(\bigcap_{k=0}^{n-1} \mathrm{B}_\phi^k([\theta^0]) \right) \right) = W_{\theta^0,-i}^n$. Hence, for any $s_{-i} \in W_{\theta^0,-i}^n$ we can find $\theta_{-i}(s_{-i}) \in \Theta_{-i}$ and $t_{-i}(s_{-i}) \in T_{-i}$ such that $(\theta_{-i}(s_{-i}), s_{-i}, t_{-i}(s_{-i})) \in \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^{n-1}(R) \cap \left(\bigcap_{k=0}^{n-1} \mathrm{B}_\phi^k([\theta^0]) \right)$.

Since the game IG is rich, $\mathrm{proj}_{S_{-i}} \Sigma^n = S_{-i}$, where Σ^n is the result of the procedure in Definition 7 when there are no restrictions on first order beliefs. By Proposition 5 $\mathrm{proj}_\Sigma \mathrm{CSB}^{n-1}(R) = \Sigma$. Therefore $\mathrm{proj}_{S_{-i}} \mathrm{CSB}^{n-1}(R) = S_{-i}$, and for every $s_{-i} \in S_{-i} \setminus W_{\theta^0,-i}^n$ we can find $\theta_{-i}(s_{-i}) \in \Theta_{-i}$ and $t_{-i}(s_{-i}) \in T_{-i}$ such that $(\theta_{-i}(s_{-i}), s_{-i}, t_{-i}(s_{-i})) \in \mathrm{proj}_{\Omega_{-i}} \mathrm{CSB}^{n-1}(R)$.

We have thus defined a map $s_{-i} \longmapsto (\theta_{-i}(s_{-i}), s_{-i}, t_{-i}(s_{-i}))$ which provides an embedding of S_{-i} into Ω_{-i}. As in the proof of Lemma 10, we can then construct a CPS $\mu \in \Delta^{\mathcal{H}}(\Omega_{-i})$ such that, for all $s_{-i} \in$

S_{-i} and $h \in \mathcal{H}$, $\mu(\{\theta_{-i}(s_{-i}), s_{-i}, t_{-i}(s_{-i})|\Sigma_{-i}(h) \times T_{-i}) = \nu(s_{-i}|S_{-i}(h))$. Therefore,

$$\begin{aligned} \mu\left(\text{proj}_{\Omega_{-i}}\text{CSB}^{n-1}(R) \cap \left(\bigcap_{k=0}^{n-1} \text{B}_\phi^k([\theta^0])\right)|\Sigma_{-i}(\phi) \times T_{-i}\right) &= \\ \nu(W^n_{\theta^0,-i}|S_{-i}(\phi)) &= 1 \end{aligned}$$

and $\mu(\text{proj}_{\Omega_{-i}}\text{CSB}^{n-1}(R)|\Sigma_{-i}(h) \times T_{-i}) = 1$ for all $h \in \mathcal{H}$.

Since we are considering a belief-complete space there is an epistemic type $t_i \in T_i$ such that $g_i(t_i) = \mu$. By the private values assumption $(\theta_i^0, s_i) \in r_i(f_i(t_i))$.

Repeat the same construction for each player and let (s, t) be the tuple of strategies and epistemic types thus obtained. As in the proof of Proposition 3, it now follows that

$$(\theta^0, s, t) \in \text{CSB}^n(R) \cap \left(\bigcap_{k=0}^{n} \text{B}_\phi^k([\theta^0])\right).$$

This concludes the proof. ■