Advances in Theoretical Economics

Volume 3, *Issue* 1 2003 *Article* 3

Rationalization and Incomplete Information

Pierpaolo Battigalli[*] Marciano Siniscalchi[†]

[*] Università Bocconi, IEP, pierpaolo.battigalli@uni-bocconi.it

[†] Northwestern University and Princeton University, marciano@northwestern.edu

Rationalization and Incomplete Information

Pierpaolo Battigalli and Marciano Siniscalchi

Abstract

We analyze a family of extensive-form solution procedures for games with incomplete information that do not require the specification of an epistemic type space a la Harsanyi, but can accommodate a (commonly known) collection of explicit restrictions D on first-order beliefs. For any fixed D we obtain a solution called D-rationalizability.

In static games, D-rationalizability characterizes the set of outcomes (combinations of payoff types and strategies) that may occur in any Bayesian equilibrium model consistent with D; these are precisely the outcomes consistent with common certainty of rationality and of the restrictions D. Hence, our approach to the analysis of incomplete-information games is consistent with Harsanyi's, and it may be viewed as capturing the robust implications of Bayesian equilibrium analysis.

In dynamic games, D-rationalizability yields a forward-induction refinement of this set of Bayesian equilibrium outcomes. Focusing on the restriction that first-order beliefs be consistent with a given distribution on terminal nodes, we obtain a refinement of self-confirming equilibrium. In signalling games, this refinement coincides with the Iterated Intuitive Criterion.

KEYWORDS: Incomplete Information, Rationalizability, Bayesian Equilibrium, Self-Confirming Equilibrium, Iterated Intuitive Criterion

1 Introduction

In his seminal contribution, Harsanyi (1967-68) noted that a direct approach to the analysis of games with incomplete information requires modeling each player's entire hierarchy of beliefs: that is, her beliefs about players' payoffs, her beliefs about her opponents' beliefs concerning payoffs, and so on.

Harsanyi proposed to avoid the complexity inherent in this direct approach by introducing the notion of *type space*. The number and variety of path-breaking developments in information economics over the past thirty years demonstrate how effective Harsanyi's suggestion is.

This paper proposes an alternative approach to the analysis of incomplete-information games that (i) is fully consistent with Harsanyi's, (ii) reflects robustness to alternative specifications of the type space consistent with a given incomplete-information environment, and (iii) applies to both static and dynamic games. The basic analytical tool is Δ-*rationalizability*, an extensive form iterative procedure that extends Pearce's (1984) rationalizability concept to games with incomplete information. The procedure embodies a form of forward-induction reasoning, and can accommodate explicit restrictions on first-order beliefs (informally) assumed to be common knowledge. The symbol Δ denotes the set of first-order beliefs satisfying the assumed restrictions; each specification of Δ yields a corresponding solution set.

Our main results may be summarized as follows:

- In *static* games, that is, one-stage games with simultaneous moves, an outcome (combination of payoff types and actions) is Δ-rationalizable if and only if it can be realized in a Bayesian equilibrium consistent with the restrictions Δ. In light of this result, we suggest that Δ-rationalizability implements *robust Bayesian equilibrium analysis.*

- Fix a distribution ζ over outcomes (terminal nodes) in a dynamic game, and assume that Δ reflects the assumption that players' initial beliefs are consistent with ζ. Then, if the set of Δ-rationalizable strategies is nonempty, ζ is a *self-confirming equilibrium distribution.*

- In particular, for signalling games, if ζ and Δ are as above, then the set of Δ-rationalizable strategies is nonempty if and only if ζ is a self-confirming equilibrium distribution that passes the *Iterated Intuitive Criterion* of Cho and Kreps (1987).

In Harsanyi's approach, every element of the type space (*Harsanyi type* henceforth) comprises both a specification of a player's private payoff-relevant information (e.g. a signal, valuation, cost, ability, etc.; *payoff type* henceforth) and an additional parameter (*epistemic type*) that determines that player's hierarchical beliefs as follows. For every player $i = 1, ...n$, one specifies a map p_i associating with each Harsanyi type for Player i a probability distribution over opponents' types. Then, assuming that the maps $p_1, ..., p_n$ are common knowledge among all players, one can retrieve the entire hierarchy of beliefs associated with any Harsanyi type.

This formal construct provides an implicit, but elegant and compact representation of hierarchical beliefs; moreover, it makes it possible to apply standard game-theoretic techniques to incomplete-information games, leading to the notion of Bayesian Nash equilibrium. Furthermore, the Harsanyi approach does not *per se* entail any loss of generality relative to the direct, explicit modeling of hierarchical beliefs. Specifically, the results of Mertens and Zamir (1985) imply that any (coherent) hierarchy of beliefs about payoff types may be represented as an element of an appropriate Harsanyi type space.

But, whereas the formalism of Harsanyi type spaces is not inherently restrictive, specific instances of type spaces may (and typically do) entail restrictions on the players' reasoning processes. In fact, it may be argued that this is often the case in economic applications. In particular, in the vast majority of applications we are aware of, *there is a one-to-one correspondence between payoff types and Harsanyi types.* This places significant restrictions on the players' mutual beliefs.

For instance, it implies that the entire hierarchy of beliefs of an uniformed player (a player with no private information) is common knowledge. Similarly, in the context of Bayesian Nash equilibrium, it implies that *the choice that a player would make conditional on his private information is commonly known.* Loosely speaking, under this "textbook" assumption, there is little genuine strategic uncertainty. Recent experimental evidence that highlights deviations from "standard" predictions of information economics (e.g. Kagel, 1995, in the context of auctions) may be interpreted as suggesting that, on the contrary, strategic uncertainty often influences the choices of rational agents.

More generally, whenever one employs a "small" type space in the analysis of a game with incomplete information, one necessarily introduces some *implicit* restrictions on the players' hierarchical beliefs. Such restrictions may be subtle and hard to fully characterize (see e.g. Bergemann and Morris, 2002, and references therein), especially in the context of extensive games (Battigalli and Siniscalchi, 2002).

As noted above, the Harsanyi approach provides a rich and expressive language to formalize assumptions about players' mutual beliefs in static (i.e. simultaneous-moves) games; however, it offers little guidance as to how to model *belief revision* in dynamic games.

Perhaps partly as a consequence, while there exists a commonly accepted "canonical" notion of Bayesian Nash equilibrium, there seems to be no agreement in the literature (and among textbook authors) on a single notion of "perfect Bayesian equilibrium" for dynamic games with incomplete information. Several alternative definitions of the latter concept have been proposed, each encoding different assumptions about players' beliefs following surprise events.[1]

Moreover, assumptions about players' inferences concerning their opponents' payoff types have long been recognized to be crucial in applications (beginning with Spence's seminal analysis of job-market signalling). In particular, *forward-induction* reasoning often plays a key role. The Harsanyi approach does not *per se* provide the tools required to model this (or any alternative) form of reasoning. Again, the proliferation of "refinements" of the Bayesian equilibrium concept (e.g. for signalling games) might be seen as partly originating from this.

The approach proposed in this paper addresses these concerns. As noted above, the basic analytical tool is Δ-*rationalizability.* The "Δ" in "Δ-rationalizable" indicates a given set of explicit

[1] In this respect, the notion of sequential equilibrium (Kreps and Wilson, 1982) can be viewed as encoding one specific assumption of this sort.

restrictions that rationalizing beliefs are required to satisfy at each step of the procedure. Interesting or appropriate restrictions arise naturally in many applications; we indicate other restrictions of a more general nature below.

We shall now briefly comment on key features of our approach, and discuss our main results.

Consistency with the Harsanyi approach. Extending a result due to Brandenburger and Dekel (1987), we show that, in static games, Δ-rationalizability exactly characterizes the set of outcomes (combinations of payoff types and actions) that may occur in any Bayesian Nash equilibrium consistent with the restrictions Δ.

We suggest the following interpretation. Suppose a modeler is interested in analyzing a static incomplete-information game under the restrictions Δ and the assumption that beliefs about opponents' choice functions mapping types to strategies are correct, as in a Bayesian Nash equilibrium, but without constraining the set of possible epistemic types. The modeler can attempt to characterize all Bayesian Nash equilibrium outcomes of the game, for all possible specifications of the type space. But our results indicate that simply applying the Δ-rationalizability procedure yields the same set of outcomes. Thus, our approach may be viewed as *a tractable way to implement robust Bayesian Nash analysis.*

For dynamic games, Δ-rationalizability may instead be viewed as providing a forward-induction refinement of the set of Bayesian Nash equilibrium outcomes.

No unintended, implicit restrictions on hierarchical beliefs. In static games, Δ-rationalizability is an extension of Bernheim and Pearce's notion that deals with incomplete information and accommodates explicit restrictions on beliefs. Tan and Werlang's (1988) epistemic characterization of rationalizability can be easily adapted to the solution concept we employ.

In dynamic games, Δ-rationalizability can be provided an epistemic characterization, via minor modifications to the arguments in Battigalli and Siniscalchi (2002); some details are provided in Section 3.1. In particular, Δ-rationalizability incorporates a forward-induction criterion, the *best-rationalization principle* (Battigalli, 1996).

Thus, for both static and dynamic games, the precise behavioral and epistemic underpinnings of the procedure can be made explicit. On the other hand, Δ-rationalizability is an algorithm that operates on payoff type–strategy pairs; since one does not need to specify a type space in order to apply it (as is instead the case for Bayesian Nash equilibrium), one never runs the risk of introducing unintended restrictions on hierarchical beliefs.

Universal Type Space. As an alternative to the approach advocated in this paper, a modeler interested in carrying out robust Bayesian Nash equilibrium analysis of specific economic models might consider embedding the set Θ of payoff types in the *universal* type space, formed by taking the 'union' of all Harsanyi type spaces based on Θ (see Footnote 33 for details). Clearly, this approach would also avoid unintended restrictions on beliefs due to the adoption of "small" type spaces. But, while this is a theoretical possibility, we are unaware of successful direct applications of this approach. Indeed, the very richness of universal type spaces is likely to pose an obstacle.

On the other hand, our results imply that, in static games, our approach is equivalent to the one just described: in such games, an outcome is Δ-rationalizable if and only if it is realizable in a Bayesian equilibrium model featuring the "Δ-universal" type space, i.e. the union of all Harsanyi type spaces based on Θ and consistent with the restrictions Δ. Furthermore, for dynamic games, the characterization of Δ-rationalizability mentioned above can be carried out in the context of a Δ-universal epistemic type space.

Thus, our approach avoids both the issues arising from the adoption of "small" type spaces, and the complexity inherent in dealing with the universal type space directly.

Correct Predictions. As was noted above, explicit restrictions on first-order beliefs may arise naturally in specific applications. We also consider a general (i.e., not application-specific) form of "correctness" restriction on beliefs, in the spirit of equilibrium analysis. To motivate it, consider the following learning environment. For each player-role, there is a large population of individuals who are drawn and matched at random to play with individuals from other populations. Populations are heterogeneous with respect to payoff types and beliefs.

Let ζ denote the statistical distribution of combinations of payoff types and terminal histories resulting from the many games played, and suppose the statistic ζ becomes public. Players' beliefs are *stable* if they are consistent with the distribution ζ; that is, if every player's conditional beliefs concerning opponents' types and actions at histories that occur with positive ζ-probability coincide with the corresponding conditional frequencies derived from ζ.

We analyze the implications of the assumption that players' first-order beliefs are stable in this sense, and its interaction with the forward-induction logic of Δ-rationalizability; here, Δ represents the assumption that players' beliefs are consistent with a given distribution ζ. We refer to this variant of the solution concept as ζ-rationalizability.

Self-Confirming Equilibrium and ζ-Rationalizability.[2] We show that a given "feasible" distribution ζ is a self-confirming equilibrium (SCE) distribution if and only if there is a Bayesian equilibrium model consistent with ζ. Since ζ-rationalizability refines the set of Bayesian equilibrium outcomes (see the discussion above), it follows that if the set of ζ-rationalizable outcomes is nonempty, then ζ is a SCE distribution.

The latter result indicates that, in general dynamic games with incomplete information, ζ-rationalizability yields a (forward induction) *refinement* of the SCE concept.

We also observe that ζ-rationalizability can be used as a refinement criterion applicable to any equilibrium concept stronger than SCE: a given equilibrium profile satisfies the criterion if and only if the induced distribution ζ yields a non-empty ζ-rationalizable set.

Iterated Intuitive Criterion. Our last result relates the above mentioned refinement to *the Iterated Intuitive Criterion* of Cho and Kreps (1987) for signalling games. More precisely, we show that, for any feasible distribution ζ on the terminal nodes of a signalling game, ζ is a SCE satisfying the Iterated Intuitive Criterion if and only if the set of ζ-rationalizable outcomes is non-empty. (Of course, the "only if" part of the proposition holds for any stronger equilibrium concept, such as, e.g., the sequential equilibrium.)

[2] On self-confirming equilibria see Fudenberg and Levine (1993) and references therein. We consider self-confirming equilibria with unitary beliefs of the game where each payoff type corresponds to a distinct player. Self-confirming equilibrium is also called "conjectural equilibrium" (Battigalli, 1987, Battigalli and Guaitoli, 1997) or "subjective equilibrium" (Kalai and Lehrer, 1993, 1995).

The rest of the paper is organized as follows. Section 2 contains the game-theoretic set up. In Section 3 we define Δ-rationalizability and provide an epistemic interpretation, relying on our previous work. We also illustrate the solution procedure with some examples. Section 4 relates Δ-rationalizability to Bayesian and self-confirming equilibrium. Section 5 focuses on signaling games and relates Δ-rationalizability to the Iterated Intuitive Criterion. Section 6 discusses some extensions, applications and related papers. The Appendix contains the less instructive proofs and some ancillary results.

2 Game-Theoretic Framework

2.1 Games of Incomplete Information with Observable Actions

To simplify the exposition we limit our analysis to two-person, finite, multistage games with observable actions. This also allows us to use a notation that clearly separates between private information about payoff functions and information about past moves acquired as the play unfolds.[3]

A two-person *game of incomplete information with observable actions* is a structure

$$\Gamma = \left\langle \Theta_1, \Theta_2, A_1, A_2, \overline{\mathcal{H}}, u_1, u_2 \right\rangle$$

given by the following elements:[4]

- For each $i \in \{1, 2\}$, Θ_i is a finite set of possible *payoff-types* (or simply *types*) for player i, and A_i is a finite set of possible *actions* for player i. The opponent of player i is denoted $-i$.
- $\overline{\mathcal{H}} \subseteq \{\phi\} \cup \left(\bigcup_{k \geq 1} (A_1 \times A_2)^k \right)$ is finite a set of *feasible histories* (finite sequences of action pairs) including the *empty history* ϕ. Let $A(h) = \left\{ a \in A : (h, a) \in \overline{\mathcal{H}} \right\}$ denote the set of feasible action pairs given history h; $\overline{\mathcal{H}}$ is such that, for every $h \in \overline{\mathcal{H}}$, every prefix (initial subsequence) of h belongs to $\overline{\mathcal{H}}$ and $A(h)$ is a Cartesian product: that is, $A(h) = A_1(h) \times A_2(h)$, where $A_i(h)$ is the projection of $A(h)$ on A_i. $Z = \{h \in \overline{\mathcal{H}} : A(h) = \emptyset\}$ denotes the set of *terminal* histories and $\mathcal{H} = \overline{\mathcal{H}} \backslash Z$ denotes the set of *non-terminal* histories.
- For each $i \in \{1, 2\}$, $u_i : \Theta_1 \times \Theta_2 \times Z \to \mathbf{R}$ is the *payoff function* for player i ($\mathbf{R}$ denotes the set of real numbers).

Parameter θ_i represents player i's private information about the rules of the game. Note that u_i may depend on the payoff type of the opponent of Player i (this is the case, for example, if i's opponent has private information about the way actions affect i's payoff). Therefore we do not assume private values. On the other hand, we assume for simplicity that constraints on choices are type-independent, which is why we refer to θ_i as the *"payoff-type"* of player i. The *"state of Nature"* $\theta = (\theta_1, \theta_2)$ completely specifies the unknown parameters of the game and the players'

[3] The analysis can be extended to general information structures with perfect recall. See the Discussion section.

[4] See Fudenberg and Tirole (1991, pp 331-332) and Osborne and Rubinstein (1994, pp 231-232).

interactive knowledge about them. Non-terminal histories become common knowledge as soon as they occur. The structure Γ is common knowledge.[5]

The notion of multistage game with observable actions is general enough to cover static games, repeated games, and games with sequential moves as special cases. A game is *static* if it has only one stage, which by definition has simultaneous moves (formally, $\mathcal{H} = \{\phi\}$, $A(\phi) = A_1 \times A_2$). A game Γ has *sequential* (or non-simultaneous) *moves* if for every non-terminal history h there is only one player i, called the *active player*, with more than one feasible action. Games with sequential moves are easily represented by means of game trees and information sets: the set of nodes is $\Theta \times \overline{\mathcal{H}}$, the "prefix of" binary relation on $\overline{\mathcal{H}}$ yields a corresponding partial order on $\Theta \times \overline{\mathcal{H}}$, making it an arborescence. Information sets for player i have the following form:

$$I(\theta_i, h) = \{(\theta_i, \theta'_{-i}, h) : \theta'_{-i} \in \Theta_{-i}\},$$

where i is active at h.

Note that the structure Γ *does not specify players' beliefs about the state of Nature* θ. Therefore, the above definition is different from the standard notion of a Bayesian game. As mentioned in the Introduction, in order to provide a general (albeit implicit) representation of players' beliefs about the state of Nature and of their hierarchies of beliefs, we embed each set Θ_i in a possibly richer set T_i of "Harsanyi-types" and specify belief functions $p_i : T_i \to \Delta(T_{-i})$. For more on this see Section 4.1.

2.2 Strategic Representation

Although this paper analyzes an extensive-form solution concept, it is sometimes analytically convenient to employ a strategic representation of the payoff functions and of the information that transpires as the play unfolds.

A *strategy* for player i is a function $s_i : \mathcal{H} \to A_i$ such that $s_i(h) \in A_i(h)$ for all $h \in \mathcal{H}$. The set of strategies for player i is denoted S_i and $S = S_1 \times S_2$ is the set of strategy pairs. Each pair of strategies $s = (s_1, s_2)$ induces a terminal history $z = O(s)$. Given the outcome function $O : S \to Z$, we can define strategic-form payoff functions $U_i(\theta_i, s_i, \theta_{-i}, s_{-i}) = u_i(\theta_1, \theta_2, O(s_1, s_2))$, $i = 1, 2$. Furthermore, for each history $h \in \mathcal{H}$ we can define the set of strategies consistent with h:

$$S(h) = \{s \in S : h \text{ is a prefix of } O(s)\}.$$

Clearly, $S(\phi) = S$ and $S(h) = S_1(h) \times S_2(h)$ for each h, where $S_i(h)$ is the set of s_i which do not prevent h from being reached. $\Theta_{-i} \times S_{-i}(h)$ is the strategic representation of the information of player i about his opponent if h occurs. Note that if h' is a prefix of h'' then $S(h') \supseteq S(h'')$.

[5] For a generalization to n-person games with possibly infinite horizon, infinite action spaces and type-dependent feasibility constraints, see Battigalli (1999). Chance moves and residual uncertainty about the environment can be modeled by having a pseudo-player c with a constant payoff function. The "type" θ_c of this pseudo-player represents the residual uncertainty about the state of Nature which would remain after pooling the private information of the real players.

2.3 Conditional Beliefs and Explicit Restrictions

Players' beliefs in multistage games can be represented as systems of conditional probabilities; for each history $h \in \mathcal{H}$, player i has a conditional belief $\mu^i(\cdot|h)$ about the types and strategies of his opponent, the conditional beliefs at distinct histories are related to each other *via* Bayes' rule:

Definition 2.1 *(cf. Rényi (1956)) A* conditional probability system *(or CPS) for player i is a collection of conditional beliefs $\mu^i = (\mu^i(\cdot|h))_{h\in\mathcal{H}} \in \prod_{h\in\mathcal{H}} \Delta(\Theta_{-i} \times S_{-i}(h))$ such that for all $\overline{\theta}_{-i} \in \Theta_{-i}$, $\overline{s}_{-i} \in S_{-i}$, $h', h'' \in \mathcal{H}$, if h' is a prefix of h'' (i.e. $S_{-i}(h'') \subseteq S_{-i}(h')$) then*

$$\mu^i(\overline{\theta}_{-i}, \overline{s}_{-i}|h') = \mu^i(\overline{\theta}_{-i}, \overline{s}_{-i}|h'') \left(\sum_{\theta_{-i}\in\Theta_{-i}, s_{-i}\in S_{-i}(h'')} \mu^i(\theta_{-i}, s_{-i}|h') \right). \tag{1}$$

The set of CPSs for player i is denoted by $\Delta^{\mathcal{H}}(\Theta_{-i} \times S_{-i})$.

Note that an element of $\Delta^{\mathcal{H}}(\Theta_{-i} \times S_{-i})$ only describes the *first-order* conditional beliefs of player i. To keep the presentation relatively simple, only such beliefs are explicit in the formal analysis of this paper. However, we shall often refer informally to higher-order beliefs.[6]

A player's beliefs may be assumed to satisfy some restrictions that are not implied by assumptions concerning belief in rationality, or beliefs about such beliefs. Such restrictions may be related to some structural properties of the game model at hand. Our approach accommodates both (i) restrictions on beliefs about payoff types, and (ii) restrictions on beliefs about (payoff types and) behavior. To represent such restrictions we assume that the conditional probability system of payoff-type θ_i of player i belongs to a given, nonempty subset $\Delta^{\theta_i} \subseteq \Delta^{\mathcal{H}}(\Theta_{-i} \times S_{-i})$. Furthermore, our solution concept relies (implicitly) on cross-restrictions on higher-order beliefs about rationality and the restrictions Δ: see Subsection 3.1 for details. We let $\Delta^i = \left(\Delta^{\theta_i}\right)_{\theta_i\in\Theta_i}$ denote the (type-dependent) restrictions for Player i.

The following are examples of restrictions of the first kind:

- Player i believes that each type of the opponent has strictly positive probability.
- It is common knowledge that the set of possible states of nature is $\widehat{\Theta} \subset \Theta_1 \times \Theta_2$; Δ^{θ_i} is the set of CPSs that assign probability zero to the opponent's types inconsistent with θ_i (formally, let $\widehat{\Theta}_{-i}(\theta_i) = \{\theta_{-i} : (\theta_i, \theta_{-i}) \in \widehat{\Theta}\}$, then $\Delta^{\theta_i} = \{\mu^i : \forall h \in \mathcal{H}, \text{supp}\mu^i(\cdot|h) \subseteq \widehat{\Theta}_{-i}(\theta_i) \times S_{-i}\}$).
- The initial beliefs of each type of i about $-i$'s type are derived from a given prior $\rho_i \in \Delta(\Theta)$ (this is a type-dependent restriction, unless ρ_i is a product measure). The assumption of a common prior on Θ is a special case.

The following are examples of restrictions of the second kind:

- The likelihood of a high-ability worker conditional on education is weakly increasing with education.

[6] Infinite hierachies of CPSs are formally analyzed in Battigalli and Siniscalchi (1999).

- In a symmetric first-price auction, there is common certainty that any bid above the lowest valuation wins the object with positive probability.
- Suppose that the individuals playing in the roles 1 and 2 are repeatedly drawn at random from large heterogenous populations and matched to play according to the rules of Γ, a large number of Γ-games has been played and public statistics show that the distribution of types and outcomes is $\zeta \in \Delta(\Theta \times Z)$. Then it may be reasonable to assume that the initial beliefs of each player about his opponent are consistent with ζ. We elaborate on this restriction in Section 4.3.

While we allow for explicit restrictions on first-order beliefs concerning payoff types and/or behavior, we do not consider direct restrictions on behavior, or about higher-order beliefs. At the expense of notational complexity, our approach could be modified to allow for such restrictions.

2.4 Sequential Rationality

A strategy $\hat{s}_i$ is sequentially rational for a player of type $\hat{\theta}_i$ with conditional beliefs μ^i if it maximizes the conditional expected utility of $\hat{\theta}_i$ at every history h consistent with $\hat{s}_i$. Note that this a notion of rationality for plans of actions[7] rather than strategies (see, for example, Reny (1992)). Given a CPS μ^i, a non-terminal history h, a type θ_i and a strategy s_i consistent with h ($s_i \in S_i(h)$) let

$$U_i\left(\theta_i, s_i, \mu^i(\cdot|h)\right) = \sum_{\theta_{-i}\in\Theta_{-i}, s_{-i}\in S_{-i}(h)} U(\theta_i, s_i, \theta_{-i}, s_{-i})\mu^i\left(\theta_{-i}, s_{-i}|h\right)$$

denote the expected payoff for type θ_i from playing s_i given h.

Definition 2.2 *A strategy* $\hat{s}_i$ $(i = 1, 2, ...)$ *is* sequentially rational for type $\hat{\theta}_i$ *with respect to beliefs* $\mu^i \in \Delta^{\mathcal{H}}(\Theta_{-i} \times S_{-i})$, *written* $\hat{s}_i \in r_i(\hat{\theta}_i, \mu^i)$, *if for all* $h \in \mathcal{H}$ *such that* $\hat{s}_i \in S_i(h)$ *and all* $s_i \in S_i(h)$

$$U_i\left(\hat{\theta}_i, \hat{s}_i, \mu^i(\cdot|h)\right) \geq U_i\left(\widehat{\theta}_i, s_i, \mu^i(\cdot|h)\right).$$

It can be shown by a standard dynamic programming argument that the set of maximizers $r_i(\hat{\theta}_i, \mu^i)$ is non-empty for every pair $(\hat{\theta}_i, \mu^i)$.

3 Δ-Rationalizability

This section introduces our main analytical tool, Δ-rationalizability. We discuss its relationship with other notions of rationalizability. We then establish a simple existence result, as well as certain useful properties of the procedure; finally, we present three examples.

[7] Formally, a *plan of action* is a maximal set of strategies consistent with the same histories and prescribing the same actions at such histories.

3.1 Solution Procedure and Interpretation

Fix the belief restrictions $\Delta = \left(\Delta^1, \Delta^2\right)$ $(\Delta^i \subseteq [\Delta^{\mathcal{H}}(\Theta_{-i} \times S_{-i}(h))]^{\Theta_i})$. The solution procedure we define below iteratively eliminates pairs (θ_i, s_i) for each player i.

Definition 3.1 *Consider the following procedure.*

(Step 0) *For every* $i = 1, 2$, *let* $\Sigma^0_{i,\Delta} = \Theta_i \times S_i$.

(Step $n > 0$**)** *For every* $i = 1, 2$, *and for every* $(\theta_i, s_i) \in \Theta_i \times S_i$, *let* $(\theta_i, s_i) \in \Sigma^n_{i,\Delta}$ *if and only if* $(\theta_i, s_i) \in \Sigma^{n-1}_{i,\Delta}$ *and there exists a CPS* $\mu^i \in \Delta^{\theta_i}$ *such that*

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

Finally, let $\Sigma^\infty_{i,\Delta} = \bigcap_{n \geq 0} \Sigma^n_{i,\Delta}$. *We say that strategy* s_i *is* Δ-rationalizable *[*(Δ, n)*-rationalizable] for type* θ_i *if* $(\theta_i, s_i) \in \Sigma^\infty_{i,\Delta}$ *[*$(\theta_i, s_i) \in \Sigma^n_{i,\Delta}$*].*

The procedure just defined modifies extensive-form rationalizability (Pearce, 1984) in two ways. First, the original definition (as applied to the extensive-form representation of Γ) assumes that players' beliefs about their opponents' *types* are consistent with a common prior on the set of states of nature; we only assume that players' beliefs about the opponent's types belong to a given subset of CPSs. Second, we allow for explicit restrictions on players' beliefs about their opponents' *behavior*.

Thus, if Γ has complete information and there are no restrictions on beliefs [i.e. for each player i, $\Theta_i = \{\theta^0_i\}$, and $\Delta^i = \Delta^{\mathcal{H}}(\{\theta^0_{-i}\} \times S_{-i})$], then $\Sigma^\infty_{i,\Delta} = \{\theta^0_i\} \times S^\infty_i$, where S^∞_i is the set of extensive-form rationalizable strategies of player i. Similarly, if Γ has incomplete information and Δ represents the assumption that initial beliefs about the opponent's type are derived from a given prior distribution $\rho \in \Delta(\Theta)$, Δ-rationalizability is equivalent to the application of Pearce's procedure to the extensive-form representation of Γ with a common prior ρ.[8]

Shimoji and Watson (1998) show that Pearce's solution procedure can characterized as the iterated removal of strategies that are conditionally dominated at some information set. By a straightforward extension of their result, one can show that, if no explicit restrictions on beliefs are imposed,[9] the procedure of Definition 3.1 corresponds to the iterated removal of pairs (θ_i, s_i) such that s_i is conditionally dominated for type θ_i at some history h.[10]

As a special case, consider a *static* game and suppose that there are no explicit restrictions on beliefs. Then, by a straightforward extension of known results, one can show that the our

[8] In Section 6 we briefly discuss how to extend the notion of Δ-rationalizability to games with more than two players. Here we only note that the extension would differ from Pearce's definition even in the above mentioned cases, due to different assumptions about belief revision with multiple opponents. For more on this see Battigalli (1996).

[9] That is, $\Delta^{\theta_i} = \Delta(\Theta_{-i} \times S_{-i})$ for every i and θ_i.

[10] Formally, (θ_i, s_i) is deleted at step n if and only if there exist a history $h \in \mathcal{H}$ and a mixed strategy $\mu_i \in \Delta(S_i)$ with Supp$\mu_i \subseteq \{s'_i \in S_i(h) : (\theta_i, s'_i) \in \Sigma^{n-1}_i\}$ such that $U_i(\theta_i, s_i, \theta_{-i}, s_{-i}) < U_i(\theta_i, \mu_i, \theta_{-i}, s_{-i})$ for all $(\theta_{-i}, s_{-i}) \in [\Theta_{-i} \times S_{-i}(h)] \cap \Sigma^{n-1}_{-i}$, where Σ^{n-1}_i is the set of pairs that survived through step $n - 1$.

notion of rationalizability coincides with the iterated removal of pairs (θ_i, a_i) such that a_i is strictly dominated for type θ_i by some mixed action $\alpha_i \in \Delta(A_i)$; in particular, if the game has complete information, we obtain the standard notion of rationalizability. Finally, suppose that there is only one rationalizable action $b_i(\theta_i)$ for each type θ_i. Then, for each state of nature θ, $b(\theta)$ must be the unique rationalizable profile – hence the unique Nash equilibrium – of the normal-form game $G(\theta) = \langle A_1, A_2, u_1(\theta, \cdot), u_2(\theta, \cdot)\rangle$. Thus, in this particular case, the behavioral profile b must be an ex-post equilibrium of the incomplete information game.[11]

We emphasize that the procedure in Definition 3.1 incorporates a forward-induction criterion, and therefore it typically refines other versions of the rationalizability solution concept proposed for extensive-form games (for more on this, see the discussion in Section 6).

An exact epistemic characterization of Δ-rationalizability can be provided within the framework of universal[12] type spaces for dynamic games. Here we only present an informal description.[13]

Let us say that player i *strongly believes* event $\mathcal{E}$ if he is initially certain of $\mathcal{E}$ and would also be certain of $\mathcal{E}$ conditional on every history h whose occurrence does not contradict $\mathcal{E}$. There is *mutual* strong belief in $\mathcal{E}$ if each payer strongly believes $\mathcal{E}$. For any given $\Delta = (\Delta^1, \Delta^2)$, consider the following set of assumptions:

$\mathcal{A}^0$: every player i is rational and her beliefs satisfy the set of restrictions Δ^i,
$\mathcal{A}^1$: there is mutual strong belief in $\mathcal{A}^0$,
$\mathcal{A}^2$: there is mutual strong belief in $\mathcal{A}^0 \cap \mathcal{A}^1$,
...
$\mathcal{A}^{n+1}$: there is mutual strong belief in $\mathcal{A}^0 \cap \mathcal{A}^1 \cap ... \cap \mathcal{A}^n$,
...

It can be shown that, in a universal type space, an arbitrarily given profile of payoff types and strategies $((\theta_1, s_1), (\theta_2, s_2))$ is consistent with the set of assumptions $\bigcap_{n\geq 0} \mathcal{A}^n$ [or $\mathcal{A}^0 \cap \mathcal{A}^1 \cap ... \cap \mathcal{A}^n$] if and only if s_i is Δ-rationalizable [or $(\Delta, n+1)$-rationalizable] for type θ_i, $i = 1, 2$. Furthermore, it can be shown that a profile of payoff types and strategies is jointly consistent with the assumption that the restrictions Δ are "common knowledge"[14] and the set of assumptions $\bigcap_{n\geq 0} \mathcal{A}^n$ if and only if it each strategy in the profile is Δ-rationalizable for the corresponding payoff type.

[11] Possibly for this reason, some authors call this solution concept *ex-post rationalizability* and refer to related notions of dominance as *ex-post dominance* (e.g., Bergemann and Morris, 2002). We do not adopt this terminology, because it is suggestive of a different concept, namely the rationalizability correspondence defined on the class of complete-information games $\{G(\theta); \theta \in \Theta\}$. The graph of this correspondence is contained in the rationalizable set Σ^∞ and the inclusion may be strict.

[12] Or, more generally, "belief-complete": see Battigalli and Siniscalchi (2002).

[13] We label assumptions $\mathcal{A}^0$, $\mathcal{A}^1$, $\mathcal{A}^2$, ..., where $\mathcal{A}^n$ refers to beliefs of order n. For a formal characterization result see Battigalli and Siniscalchi (2002), where these assumptions are formally represented as subsets of a space of states of the world. $\mathcal{A}^j \cap \mathcal{A}^k$ informally denotes the conjunction of assumptions labeled $\mathcal{A}^j$ and $\mathcal{A}^k$.

[14] More precisely, say that the restrictions Δ are "common knowledge" if, for each $i = 1, 2$, (1) i would believe $[\mu^{-i} \in \Delta^{-i}]$ at each $h \in \mathcal{H}$, (2) $-i$ would believe (1) at each $h \in \mathcal{H}$, ... , (2k+1) i would believe (2k) at each $h \in \mathcal{H}$, (2k+2) $-i$ would believe (2k+1) at each $h \in \mathcal{H}$, and so on.

This characterization indicates that Δ-rationalizability relies on specific assumptions about belief revision: a player initially believes that her opponent's behavior is consistent with the set of assumptions $\bigcap_{n\geq 0}\mathcal{A}^n$, and continues to believe that this is the case as long as the opponent's observed actions are indeed consistent with $\bigcap_{n\geq 0}\mathcal{A}^n$. In the event that an action inconsistent with $\bigcap_{n\geq 0}\mathcal{A}^n$ is observed, she "falls back" on the most restrictive set of assumptions $\mathcal{A}^0\cap\mathcal{A}^1\cap\ldots\cap\mathcal{A}^k$ consistent with the observed actions of the opponent. For example, suppose that Player i observes a history h consistent with Player $-i$ being rational and holding beliefs in Δ^{-i}; also suppose that h could *not* occur if, in addition to being rational and holding beliefs in Δ^{-i}, Player $-i$ strongly believed that Player i is herself rational and holds beliefs in Δ^i. Then the assumptions above imply that Player i must believe at h that Player $-i$ is rational and holds beliefs in Δ^{-i}. In a sense, it is assumed that a player always tries to rationalize the observed actions of her opponent ascribing to her the highest "degree of strategic sophistication" consistent with such actions. Thus, these assumptions reflect a specific form of forward-induction reasoning.

In general, the restrictions on beliefs represented by Δ may be inconsistent with some of the assumptions $\mathcal{A}^1$, $\mathcal{A}^2$, ...$\mathcal{A}^n$, ... In this case the set of Δ-rationalizable strategies is empty. On the other hand, the finiteness assumption yields a simple existence result:[15]

Remark 3.2 *Suppose that, for each $i=1,2$, Δ^i only represents restrictions on i's beliefs about the payoff type of his opponent (i.e., for each θ_i there is a non-empty subset $\mathbf{P}^{\theta_i}\subseteq\Delta(\Theta_{-i})$ such that $\Delta^{\theta_i}=\{\mu^i : \mathrm{marg}_{\Theta_{-i}}\mu^i(\cdot|\phi)\in\mathbf{P}^{\theta_i}\}$). Then, for each player i and every θ_i, the set of Δ-rationalizable strategies for type θ_i is non-empty ($\forall i$, $\mathrm{proj}_{\Theta_i}\Sigma^\infty_{i,\Delta}=\Theta_i$).*

We also report here two remarks that will be useful later.

Remark 3.3 *The sequence $\{\Sigma^n_{1,\Delta}\times\Sigma^n_{2,\Delta}\}_{n>0}$ is weakly decreasing and such that $\Sigma^k_{1,\Delta}\times\Sigma^k_{2,\Delta}=\Sigma^{k+1}_{1,\Delta}\times\Sigma^{k+1}_{2,\Delta}$ implies $\Sigma^k_{1,\Delta}\times\Sigma^k_{2,\Delta}=\Sigma^n_{1,\Delta}\times\Sigma^n_{2,\Delta}$ for all $n\geq k$. Since $\Theta\times S$ is finite, there is some N such that $\Sigma^N_{1,\Delta}\times\Sigma^N_{2,\Delta}=\Sigma^n_{1,\Delta}\times\Sigma^n_{2,\Delta}$ for all $n\geq N$. Therefore, for each i and $(\theta_i,s_i)\in\Sigma^\infty_{i,\Delta}$ there is some $\mu^i\in\Delta^{\theta_i}$, such that $\mu^i(\Sigma^\infty_{-i,\Delta}|\phi)=1$ and $s_i\in r_i(\theta_i,\mu^i)$.*

Remark 3.4 *Suppose that restriction Δ^i implies that player i initially assigns positive probability to each payoff-type of the opponent ($\forall\theta_i,\forall\mu^i\in\Delta^{\theta_i}$, $\forall\theta_{-i}\in\Theta_{-i}$, $\mu^i(\{\theta_{-i}\}\times S_{-i}|\phi)>0$). Then, if there is a Δ-rationalizable* pair *for player i, there must be a Δ-rationalizable strategy for each payoff-type θ_{-i} of player $-i$ ($\Sigma^\infty_{i,\Delta}\neq\emptyset\Rightarrow\mathrm{proj}_{\Theta_{-i}}\Sigma^\infty_{-i,\Delta}=\Theta_{-i}$).*

Proof. Any $(\theta_i,s_i)\in\Sigma^\infty_{i,\Delta}$ is such that s_i is a sequential best reply to some belief $\mu^i\in\Delta^{\theta_i}$ such that $\mu^i(\Sigma^\infty_{-i,\Delta}|\phi)=1$ (see Remark 3.3). Fix an arbitrary $\theta_{-i}\in\Theta_{-i}$; by assumption $\mu^i(\{\theta_{-i}\}\times S_{-i}|\phi)>0$. Therefore $\mu^i(\Sigma^\infty_{-i,\Delta}\cap(\{\theta_{-i}\}\times S_{-i})|\phi)>0$. This implies that there must be some s_{-i} such that $(\theta_{-i},s_{-i})\in\Sigma^\infty_{-i,\Delta}$. ■

A note on terminology. Whenever we assume *no* explicit restrictions on beliefs, we omit the symbol Δ.

[15] In Section 5 we obtain necessary and sufficient conditions for existence of Δ-rationalizable strategies in signaling games when beliefs are assumed to be consistent with a given distribution $\zeta\in\Delta(\Theta\times Z)$.

3.2 Examples

We present three examples illustrating the solution procedure. The first one exhibits a sort of no-rationalizable-trade result.[16] Here, explicit restrictions on beliefs do not play any role. In the second example, the efficient allocation of an object is implemented in Δ-rationalizable strategies with a two-stage mechanism. The (type-dependent) restrictions reflect the assumption that it is common knowledge between the agents who values the object the most.[17] In the third example, we apply Δ-rationalizability to the well-known Beer-Quiche game, assuming that beliefs agree with the outcome distribution ζ induced by the (Quiche,Quiche) pooling equilibrium (which does not satisfy the Intuitive Criterion). It turns out that the set of ζ-rationalizable strategies in this case is empty.

3.2.1 Co-authorship

Player 1 is an author who must decide whether to carry out a project alone or propose joint work to another author, Player 2. Player 2 can either accept or reject and then carry out her own project alone. We let θ_i denote the ability of author i and we assume $\Theta_i = \{\theta^1, ..., \theta^K\}$, with $0 < \theta^1 < \theta^2 < ... < \theta^K$. In order to prepare a proposal some paperwork is needed and this has a small cost of ε for Player 1 ($0 < \varepsilon < 1$). The proposal *per se* does not reveal any information because it takes the same form independently of the ability of Player 1.[18] Joint work yields a fixed surplus δ. Let **YES** (**NO**) denote the strategy of accepting (rejecting) if Author 1 proposes. Payoffs are determined by the following table.

at (θ_1, θ_2)	**NO**	**YES**
ALONE	θ_1, θ_2	θ_1, θ_2
PROPOSE	$(\theta_1 - \varepsilon), \theta_2$	$\left(\frac{\theta_1+\theta_2+\delta}{2} - \varepsilon\right), \frac{\theta_1+\theta_2+\delta}{2}$

We do not assume any restriction on beliefs; however, we make the following assumptions about the parameters of the game:

(1) *Author 1 has no incentive to propose a project if he is certain that Author 2 has the same quality as himself or lower:*

$$\frac{\theta + \theta + \delta}{2} - \varepsilon < \theta \text{ (i.e. } \delta < 2\varepsilon)$$

(2) *Author 2 has no incentive to accept a proposal if she is certain that Author 1 has lower quality than herself:*

$$\frac{\theta^k + \theta^{k-1} + \delta}{2} < \theta^k \text{ (i.e. } \delta < \theta^k - \theta^{k-1}) \text{ for all } k = 2, ..., K$$

[16] On rationalizable trade see Morris and Skiadas (2000). Our example also bears some resemblance to models of disclosure. See Okuno-Fujiwara *et al.* (1990), Battigalli (1999, Section 5.2) and references therein.

[17] Cf. Perry and Reny (1999).

[18] Hence the game fits our simplifying assumption that the set of feasible actions is type-independent.

We say that *joint work is rationalizable* if $(\theta_1,\text{PROPOSE};\theta_2,\mathbf{YES})$ is rationalizable for some $\theta_1,\theta_2 \in \{\theta^1,...,\theta^K\}$. We then have the following

Result: *Under the above assumptions, joint work is **not** rationalizable.*

Proof. Define $\Theta_1^n = \{\theta : (\theta, \text{PROPOSE}) \in \Sigma_1^n\}$, $\Theta_2^n = \{\theta : (\theta, \mathbf{YES}) \in \Sigma_2^n\}$. We first prove the following claim: *for every* $n \geq 1$, *either* $\Theta_1^{n+1} \cap \Theta_2^{n+1} = \emptyset$ *or* $\max\Theta_1^{n+1} < \max\Theta_2^n \leq \max\Theta_1^{n-1}$ (since Θ_i is finite, $\max\Theta_i^\ell$ is well defined if and only if $\Theta_i^\ell \neq \emptyset$).

Suppose that $\Theta_1^{n+1} \cap \Theta_2^{n+1} \neq \emptyset$. Then $\Theta_i^{n-1} \supseteq \Theta_i^n \supseteq \Theta_i^{n+1} \neq \emptyset$. Since $\Theta_1^{n-1} \neq \emptyset$, **YES** is n-rationalizable for θ_2 only if YES is a best response for θ_2 to PROPOSE given a belief concentrated on Θ_1^{n-1}. By Assumption (2), the latter holds only if $\theta_2 \leq \max\Theta_1^{n-1}$. Thus $\max\Theta_2^n \leq \max\Theta_1^{n-1}$.

PROPOSE is $(n+1)$-rationalizable for θ_1 only if it is a best response for θ_1 to a belief concentrated on $\Sigma_2^n = \{(\theta_2,s_2) : s_2 = \mathbf{YES} \Rightarrow \theta_2 \in \Theta_2^n\}$. By assumption (1), the latter holds only if $\theta_1 < \max\Theta_2^n$. Thus $\max\Theta_1^{n+1} < \max\Theta_2^n \leq \max\Theta_1^{n-1}$, and the claim is proved.

To complete the analysis of the game, note that, since the game is finite, there exists $N \geq 0$ such that $\Theta_i^n = \Theta_i^N$ for all $n \geq N$. The claim (with $n = N+1$) implies that $\Theta_1^N \cap \Theta_2^N = \emptyset$ (otherwise $\max\Theta_1^{N+2} < \max\Theta_2^{N+1} \leq \max\Theta_1^N$, which contradicts the choice of N). Suppose that both sets Θ_i^N are non-empty, and let $\bar{\theta}_i = \max\Theta_i^N$ for $i = 1,2$; then either $\bar{\theta}_1 < \bar{\theta}_2$ or $\bar{\theta}_1 > \bar{\theta}_2$. If $\bar{\theta}_1 < \bar{\theta}_2$, then Assumption (2) implies that $(\bar{\theta}_2, \mathbf{YES})$ is not rationalizable; if instead $\bar{\theta}_1 > \bar{\theta}_2$, then Assumption (1) implies that $(\bar{\theta}_1, \text{PROPOSE})$ is not rationalizable. Thus, in either case, we obtain a contradiction; hence, one of the sets Θ_i^N is empty, i.e., joint work is not rationalizable. ■

3.2.2 King Solomon's Dilemma with Incomplete Information

Consider the following situation. A planner wishes to allocate an object to one of two agents, at *zero* cost to them. Agent i's value for the object is denoted by θ_i, and is private information to her. One of the agents has strictly higher valuation, and furthermore each agent knows whether she is the high-valuation agent, but the planner does not. All this is commonly known among the agents and the planner. In order to allocate the object to the right agent, the latter must design a mechanism that does not rely on the identity of the agents.

We model the situation as seen by the agents. Let $\Theta_1 = \{2,3\}$ and $\Theta_2 = \{1,2\}$, and consider the following restrictions on beliefs:

1. $\Delta^{1_2} = \{\mu^1 : \mu^1(\{1\} \times S_2|\phi) = 1\}$ and $\Delta^{1_3} = \{\mu^1 : \forall\theta_2,\ \mu^1(\{\theta_2\} \times S_2|\phi) > 0\}$;
2. $\Delta^{2_1} = \{\mu^2 : \forall\theta_1,\ \mu^2(\{\theta_1\} \times S_1|\phi) > 0\}$ and $\Delta^{2_2} = \{\mu^2 : \mu^2(\{3\} \times S_1|\phi) = 1\}$.

Thus, Agent 1 is w.l.o.g. the high-valuation agent. Type 2 of Agent 1 and type 1 of Agent 2 are uncertain about the other agent's type; the only restriction on their beliefs is that they assign positive probability to either type of their opponent. On the other hand, the above restrictions reflect the assumption that type 2 of Agent 2 "knows" that she is the low-valuation agent, and is therefore certain that Agent 1's type equals 3. Similarly, type 2 of Agent 1 "knows" that she is the high-valuation agent, and is therefore certain that Agent 2's type equals 1.

	Out	In1	In2	In3
Out	$0 \quad 0$	$0 \quad 1$	$0 \quad 1$	$0 \quad 1$
	$0 \quad 0$	$0 \quad 2$	$0 \quad 2$	$0 \quad 2$
In1	$2 \quad 0$	$\frac{1}{4} - e \quad -e$	$-e \quad -1 - e$	$-e \quad -2 - e$
	$3 \quad 0$	$\frac{1}{2} - e \quad \frac{1}{4} - e$	$-e \quad -e$	$-e \quad -1 - e$
In2	$2 \quad 0$	$-e \quad -e$	$-e \quad -\frac{1}{4} - e$	$-e \quad -2 - e$
	$3 \quad 0$	$1 - e \quad -e$	$\frac{1}{4} - e \quad -e$	$-e \quad -1 - e$
In3	$2 \quad 0$	$-1 - e \quad -e$	$-1 - e \quad -e$	$-\frac{1}{4} - e \quad -\frac{1}{2} - e$
	$3 \quad 0$	$-e \quad -e$	$-e \quad -e$	$-e \quad -\frac{1}{4} - e$

Table 1: King Solomon's Dilemma

The planner utilizes a first-price auction preceded by an opt-in stage. Specifically, the mechanism works as follows. At time 0, agents simultaneously decide whether or not to participate in the auction; if only one does, the object is awarded to that agent, at no cost to her, and if none does, the object remains in the hands of the planner. If *both* agents choose to participate, an entry fee $e \in (0, \frac{1}{4})$ is levied, and agents participate in a first-price, sealed-bid auction. To simplify the analysis, assume that the set of possible bids is $\{1, 2, 3\}$, reflecting the valuations of the agents; also, in the case of a tie, each agent receives the object with probability $\frac{1}{4}$ (so the object is not awarded with probability $\frac{1}{2}$).

The set of actions is thus $A = \{\text{In,Out,1,2,3}\}$, and the set of non-terminal histories is $\mathcal{H} = \{\phi, (\text{In,In})\}$; the strategy set for each agent $i = 1, 2$ is $S_i = \{\text{Out}n : n = 1...3\} \cup \{\text{In}n : n = 1...3\}$.

Table 1 indicates the reduced-form payoffs to both agents, where "Out" denotes the equivalence class of strategies $\{\text{Out}n : n = 1...3\}$. Corresponding to each strategy profile, we provide a 2x2 matrix that indicates the payoff to, in counterclockwise order starting from the top left cell: Type 2 of Agent 1, Type 3 of Agent 1, Type 2 of Agent 2, and Type 1 of Agent 2.

We now analyze the mechanism using Δ-rationalizability; at each step, we indicate which strategies can be eliminated.

Step 1. We can eliminate the strategy In3 for both types of Agent 2, because it is not sequentially rational: the strategy In2 does strictly better conditional upon participating in the auction (henceforth: "at the auction stage"), regardless of Agent 1's bid. Moreover, In3 is also conditionally dominated at the auction stage for type 2 of Agent 1. It can be verified that all other strategies are $(\Delta, 1)$-rationalizable for each type.

Step 2. Since In1 and In2 are $(\Delta, 1)$-rationalizable for Agent 2, whereas In3 is not, $(\Delta, 2)$-rationalizability requires that Agent 1's justifying beliefs μ^1 satisfy in particular $\mu^1(\Theta_2 \times \{\text{In3}\}|(\text{In})) = 0$. Given this restriction, In2 is the unique conditional best response for type 3 at the auction stage, so we can eliminate In1 and In3 for this type. Furthermore, In2 strictly dominates Out for the same type. As a result, the only $(\Delta, 2)$-rationalizable strategy for Agent 1's type 3 is In2. There are no further eliminations; in particular, note that Out is rationalized for type 2 of Agent 1 by the belief that both types of Agent 2 play In2.

Step 3. Consider Agent 2's type 2 first. This type is certain (at ϕ) that Agent 1's type is 3, and we have just argued that the latter has a unique $(\Delta, 2)$-rationalizable strategy, namely In2.

Arguing as above, Agent 2's beliefs must satisfy $\mu^2(\{(3,\text{In2})\}|\phi) = 1$, and the unique best response to this belief is Out. Now consider Agent 2's type 1. Out is rationalizable for this type as well, as is In1;[19] however, since we have eliminated In3 for both types of Agent 1, In2 is no longer justifiable. Thus, the only $(\Delta, 3)$-rationalizable strategies for Agent 2's type 1 are Out and In1.

Steps 4 and 5. We have already established that In2 is the unique Δ-rationalizable strategy for Agent 1's type 3, so turn to type 2. This type of Agent 1 is certain that Agent 2's type is 1, and we have just seen that this type can only play Out or In1. Relative to these two strategies of Agent 2, Out is strictly dominated by In1 for Agent 1's type 2. Furthermore, since this agent's beliefs must satisfy $\mu^1(\{(1,\text{In1})\}|(\text{In})) = 1$, In1 is strictly better than In2 in the auction. Therefore, the unique $(\Delta, 4)$-rationalizable strategy for Agent 1's type 2 is In1. It is then clear that Agent 2's only $(\Delta, 5)$-rationalizable strategy is Out.

Thus, both types of the low-valuation agent choose not to participate in the auction, so that, by participating, the high-valuation agent secures the object at zero cost.

3.2.3 Beer-Quiche Revisited

The game depicted in Figure 1 corresponds to the well-known Beer-Quiche example used by Cho and Kreps (1987) to motivate the "Intuitive Criterion", perhaps the best-known equilibrium refinement for signalling games. Cho and Kreps analyze a standard extensive form game with a common prior; in their example, the probabilities of the **s**urly and **w**impish types θ^s and θ^w of the sender are 0.9 and 0.1 respectively. They show that only the equilibrium in which each type chooses B satisfies the Intuitive Criterion.

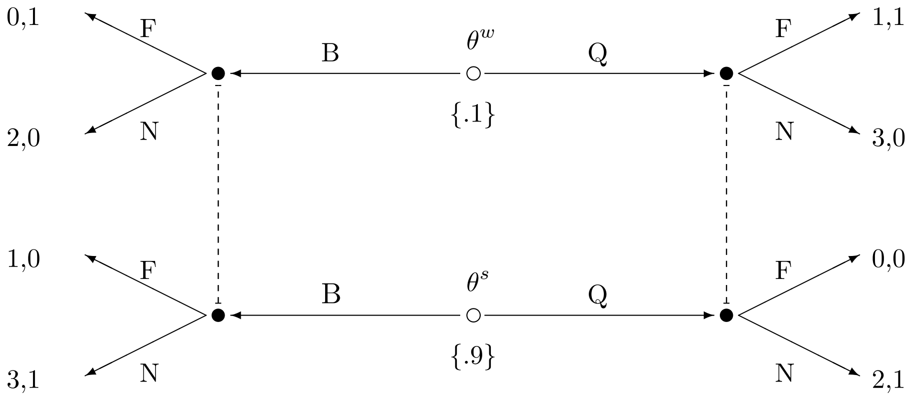

Figure 1: Beer-Quiche

Here we obtain an analogous result with Δ-rationalizability; specifically, we consider the restriction that that players' initial beliefs are consistent with the distribution over terminal nodes

[19] In particular, for In1, a belief consistent with $(\Delta, 2)$-rationalizability is given by $\mu^2(\{(2,\text{Out})\}|\phi) = .99 = 1 - \mu^2(\{(3,\text{In2})\}|\phi)$ and $\mu^2(\{(3,\text{In2})\}|(\text{In})) = 1$

induced by the "bad" equilibrium in which both types choose Q. We show that, for this specification of Δ, the set of Δ-rationalizable profiles is empty. This suggests that the "bad" equilibrium is inconsistent with the forward-induction logic embodied in Δ-rationalizability, and indicates that the latter may be viewed as a refinement of equilibria in extensive games. Section 5 develops this observation.

To clarify our notation, $\Theta_1 = \{\theta^s, \theta^w\}$, whereas Θ_2 is a singleton and will be omitted for simplicity; $S_1 = \{B, Q\}$; and $S_2 = \{FF, FN, NF, NN\}$, where FF is the strategy "Fight if Beer, Fight if Quiche", FN is "Fight if Beer, do Not fight if Quiche", and similarly for NF and NN. The set of non-terminal histories is $\mathcal{H} = \{\phi, (B), (Q)\}$.

Formally, let

$$\Delta^{\theta_1} = \{\mu^1 : \mu^1(\{s_2 : s_2(Q) = N\}|\phi) = 1\},\ \theta_1 = \theta^s, \theta^w$$

and

$$\Delta^2 = \{\mu^2 : \mu^2(\{(\theta^s, Q)\}|\phi) = 0.9 = 1 - \mu^2(\{(\theta^w, Q)\}|\phi)\}.$$

It is easy to check that $(\Delta, 1)$-rationalizability imposes no restriction on type θ^s's behavior; however, type θ^w strictly prefers Q to B, given that Q is his favorite breakfast and ensures that no Fight will ensue. As for Player 2, note that Bayes' rule implies that $\mu^2(\{(\theta^s, Q)\}|(Q)) = 0.9$, so it is sequentially rational for 2 to choose N after Q; however, no restrictions are placed on 2's choice after B. To summarize, $\Sigma^1_{1,\Delta} = \{(\theta^s, B), (\theta^s, Q), (\theta^w, Q)\}$ and $\Sigma^1_{2,\Delta} = \{FN, NN\}$.

In Step 2, nothing further is eliminated for Player 1, so $\Sigma^2_{1,\Delta} = \Sigma^1_{1,\Delta}$. However, Player 2 will now choose N after B as well, because, if his beliefs are consistent with $(\Delta, 1)$-rationalizability, upon observing B, he must conclude that Player 1's type is θ^s; formally, since $\Sigma^1_{1,\Delta} \cap [\Theta_1 \times \{B\}] = \{(\theta^s, B)\}$, Player 2's beliefs must satisfy $\mu^2(\{(\theta^s, B)\}|\phi) = 1$. This may be viewed as a forward-induction restriction on (off-equilibrium) beliefs. To summarize, $\Sigma^2_{2,\Delta} = \{NN\}$.

Now $(\Delta, 3)$-rationalizability clearly implies that each type of Player 1 will have his favorite breakfast, as no Fight will ensue in any case: thus, $\Sigma^3_{1,\Delta} = \{(\theta^s, B), (\theta^w, Q)\}$. Nothing changes as far as 2 is concerned: $\Sigma^3_{2,\Delta} = \Sigma^2_{2,\Delta}$.

Finally, we reach a contradiction in Step 4. According to the definition of $(\Delta, 4)$-rationalizability, $\mu^2(\Sigma^3_{1,\Delta}|\phi) = 1$; this implies that, in particular, $\mu^2(\{(\theta^s, Q)\}|\phi) = 0$. But this belief is not an element of Δ_2, because the restrictions for Player 2 require that $\mu^2(\{(\theta^s, Q)\}|\phi) = 0.9$. It follows that $\Sigma^4_{2,\Delta} = \emptyset$, and thus also $\Sigma^k_{i,\Delta} = \emptyset$ for all $k \geq 5$ and $i = 1, 2$.

Thus, the assumption that players expect the "bad" equilibrium outcome to obtain is inconsistent with the logic of Δ-rationalizability.

4 Rationalization and Equilibrium

This section relates our approach to the analysis of incomplete-information games to equilibrium analysis. As noted in the Introduction, we show that our approach is fully consistent with Harsanyi's, and may in fact be interpreted as a way to identify the robust implications of standard Bayesian equilibrium analysis.

To elaborate, recall that a type in the sense of Harsanyi encodes a player's private information about the external state of Nature (the unknown parameters of the game) and also his epistemic type, i.e., his infinite hierarchy of beliefs about the state of Nature and the beliefs of others. In the standard model of a Bayesian game, these hierarchies of beliefs are derived from a common prior,[20] but this need not be the case in general. We now show that, if the set of epistemic types is not restricted, in every *static* (simultaneous-moves) game of incomplete information, Δ-rationalizability exactly characterizes the set of outcomes realized in some Bayesian equilibrium consistent with restrictions Δ. This equivalence result can be extended to dynamic games by considering weak versions of the rationalizability solution concept and the perfect Bayesian equilibrium concept (for more on this, see the Discussion section).

On the other hand, in *dynamic* games, Δ-rationalizability refines the set of Bayesian equilibrium outcomes. The main reason is that Δ-rationalizability includes a forward-induction principle which is absent from the Bayesian equilibrium concept.

We will further analyze the relationship between Δ-rationalizability and equilibrium refinements in Section 5. As a preliminary step, we present here a result that has some independent interest. We consider a situation where each player's initial beliefs about his opponent are consistent with some statistical distribution $\zeta \in \Delta(\Theta \times Z)$. We show that there is a Bayesian equilibrium where players' beliefs are consistent with ζ if and only if ζ is a self-confirming equilibrium distribution of the extensive form game with prior $\text{marg}_\Theta \zeta$ and player set $\Theta_1 \cup \Theta_2$.[21]

A consequence of our results is that Δ-rationalizability can be regarded as a forward-induction refinement of self-confirming equilibrium.

4.1 Bayesian Models and Equilibria

Let us fix an incomplete-information game

$$\Gamma = \left\langle \Theta_1, \Theta_2, A_1, A_2, \overline{\mathcal{H}}, u_1, u_2 \right\rangle .$$

As noted in Section 2.1, unlike the standard notion of a Bayesian game, the incomplete-information game Γ does not specify players' beliefs about the state of Nature θ, about each other's beliefs, etc. In order to provide a general representation of players' beliefs, we explicitly introduce epistemic types; a "Harsanyi type" t_i for Player i is then modeled as a pair consisting as a payoff type θ_i and an epistemic type e_i. The following definition provides the details, and also indicates the appropriate notion of Bayesian Nash equilibrium.

Definition 4.1 *A (finite)* Bayesian model of Γ *is a tuple*

$$\mathcal{M} = \left\langle \Gamma, (E_i, T_i, p_i, b_i)_{i \in \{1,2\}} \right\rangle$$

such that, for each player $i = 1, 2$: (1) E_i is a finite set, (2) $T_i \subseteq \Theta_i \times E_i$ is such that $\text{proj}_{\Theta_i} T_i = \Theta_i$, (3) $p_i : T_i \to \Delta(T_{-i})$, and (4) $b_i : T_i \to S_i$. We let $p_i^{t_i} \in \Delta(T_{-i})$ denote the belief of t_i.

[20] We mean a common prior on the set of states of the world, where a *state of the world* comprises a state of Nature and an epistemic state (hence an implicit hierarchy of beliefs) for each player.

[21] What makes the result non-obvious is that we consider self-confirming equilibria with "unitary beliefs," whereby each strategy played with positive probability by θ_i is justified by the same belief $\mu^{\theta_i} \in \Delta(\Theta_{-i} \times S_{-i})$.

$\mathcal{M}$ *is an* equilibrium *model if, for each player i,*

$$\forall(\theta_i, e_i) \in T_i, b_i(\theta_i, e_i) \in \arg\max_{s_i \in S_i} \sum_{(\theta_{-i}, e_{-i}) \in T_{-i}} p_i^{(\theta_i, e_i)}(\theta_{-i}, e_{-i}) U_i(\theta_i, s_i, \theta_{-i}, b_{-i}(\theta_{-i}, e_{-i})).$$

$\mathcal{M}$ *is* consistent *with the restrictions $\Delta = (\Delta^1, \Delta^2)$ if for each player $i \in \{1, 2\}$ and type $t_i = (\theta_i, e_i) \in T_i$ there is a CPS $\mu^{t_i} \in \Delta^{\theta_i}$ such that, for all (θ_{-i}, s_{-i}),*

$$\sum_{e_{-i}: b_{-i}(\theta_{-i}, e_{-i}) = s_{-i}} p_i^{t_i}(\theta_{-i}, e_{-i}) = \mu^{t_i}(\theta_{-i}, s_{-i} | \phi)$$

We say that a pair (θ_i, s_i) is realizable *in the Bayesian model $\mathcal{M}$ if there is an epistemic type e_i in $\mathcal{M}$ such that $s_i = b_i(\theta_i, e_i)$.*

In other words, we obtain a Bayesian model of Γ by embedding Θ in a type space *à la* Harsanyi (parts (1)-(3)) and then appending to it a behavioral profile (part 4)). The mappings $p_i : T_i \to \Delta(T_{-i})$ implicitly determine the hierarchy of beliefs corresponding to each Harsanyi type. The behavioral profile b specifies a strategy for each Harsanyi type. The belief and behavioral mappings $(p_i, b_i)_{i \in \{1,2\}}$ implicitly determine the hierarchy of beliefs about payoff-types *and* strategies corresponding to each Harsanyi type. The sub-structure $BG_\Gamma = \langle \Gamma, (E_i, T_i, p_i)_{i \in \{1,2\}} \rangle$ is a Bayesian game based on Γ.[22] An equilibrium model of Γ is given by a Bayesian game BG_Γ and a Bayesian equilibrium of BG_Γ.

4.2 Rationalizability and Bayesian Equilibrium

The relationship between Δ-rationalizability and Bayesian equilibrium is characterized in the following two propositions.

Proposition 4.2 *Fix a* static *(simultaneous-moves) game Γ and restrictions Δ. If a pair (θ_i, a_i) is realizable in a Bayesian equilibrium model of Γ consistent with Δ, then a_i is Δ-rationalizable for type θ_i.*

Proof. Let $\mathcal{M} = \langle \Gamma, (E_i, T_i, p_i, b_i)_{i \in \{1,2\}} \rangle$ be a Bayesian equilibrium model of Γ consistent with Δ and let Σ_i^* be the set of realizable pairs for player i; that is,

$$\Sigma_i^* = \{(\theta_i, a_i) : \exists e_i, a_i = b_i(\theta_i, e_i)\}.$$

We will prove that $\Sigma_i^* \subseteq \Sigma_{i,\Delta}^n$ for all $i =$ 1,2, $n = 1, 2, ...$. For every Harsanyi type $t_i \in T_i$ define the corresponding belief $\mu^{t_i} \in \Delta(\Theta_{-i} \times A_{-i})$ as follows:

$$\mu^{t_i}(\theta_{-i}, a_{-i}) = \sum_{e_{-i}: a_{-i} = b_{-i}(\theta_{-i}, e_{-i})} p_i^{t_i}(\theta_{-i}, e_{-i})$$

[22] In the standard extensive-form representation of BG_Γ, for each i there is a "prior" $p_i^0 \in \Delta(T_1 \times T_2)$ such that $p_i^{t_i}(t_{-i}) = p_i^0(t_{-i} | t_i)$. We refrain from using priors because they are formally unnecessary and may obscure the incomplete-information interpretation of the mathematical structure.

Since $\mathcal{M}$ is a Bayesian equilibrium model consistent with Δ, for every $t_i = (\theta_i, e_i) \in T_i$, $\mu^{t_i} \in \Delta^{\theta_i}$, and for every $(\theta_i, a_i) \in \Sigma_i^*$ there is some e_i such that $a_i = b_i(\theta_i, e_i) \in r_i(\theta_i, \mu^{(\theta_i, e_i)})$. Therefore,

$$\Sigma_i^* \subseteq \Sigma_{i,\Delta}^1,\ i = 1, 2.$$

Suppose by way of induction that

$$\Sigma_j^* \subseteq \Sigma_{j,\Delta}^n,\ j = 1, 2.$$

Let $(\theta_i, a_i) \in \Sigma_i^*$ and fix e_i such that $a_i = b_i(\theta_i, e_i)$. Then $a_i \in r_i(\theta_i, \mu^{(\theta_i, e_i)})$, where $\mu^{(\theta_i, e_i)} \in \Delta_i(\theta_i)$. By definition $\mu^{(\theta_i, e_i)}(\Sigma_{-i}^*) = 1$. By the inductive assumption, $\Sigma_{-i}^* \subseteq \Sigma_{-i,\Delta}^n$; thus, $\mu^{(\theta_i, e_i)}(\Sigma_{-i,\Delta}^n) = 1$. We conclude that $(\theta_i, a_i) \in \Sigma_{i,\Delta}^{n+1}$. ■

Proposition 4.3 *Fix a game Γ and restrictions Δ. Suppose that, for every player i and every payoff type θ_i, there is a strategy s_i that is Δ-rationalizable for θ_i ($\forall i$, $\mathrm{proj}_{\Theta_i} \Sigma_{i,\Delta}^\infty = \Theta_i$). Then there is a Bayesian equilibrium model $\mathcal{M}$ of Γ consistent with Δ such that, for any arbitrary pair (θ_i, s_i), s_i is Δ-rationalizable for θ_i if and only if (θ_i, s_i) is realizable in $\mathcal{M}$.*

Proof of Proposition 4.3. By assumption, $\mathrm{proj}_{\Theta_i} \Sigma_{i,\Delta}^\infty = \Theta_i$. By finiteness of Γ, there is some $N \geq 0$ such that $\Sigma_{i,\Delta}^\infty = \Sigma_{i,\Delta}^n$ for all $n \geq N$, $i = 1, 2$ (see Remark 3.3). Therefore, for every $\theta_i \in \Theta_i$ there is a strategy s_i and a CPS $\mu^{(\theta_i, s_i)} \in \Delta^{\theta_i}$ such that $\mu^{(\theta_i, s_i)}(\Sigma_{-i,\Delta}^\infty | \phi) = 1$, $(\theta_i, s_i) \in \Sigma_{i,\Delta}^\infty$ and $s_i \in r_i(\theta_i, \mu^{(\theta_i, s_i)})$. Let

- $E_i = S_i$, $T_i = \Sigma_{i,\Delta}^\infty \subseteq \Theta_i \times E_i$,
- $p_i^{t_i}(\theta_{-i}, s_{-i}) = \mu^{t_i}(\theta_{-i}, s_{-i} | \phi)$ for all $t_i \in T_i$ and $(\theta_{-i}, s_{-i}) \in T_{-i}$.
- $b_i(\theta_i, s_i) = s_i$ for all $(\theta_i, s_i) \in T_i$.

This defines a Bayesian model $\mathcal{M} = \left\langle \Gamma, (E_i, T_i, p_i, b_i)_{i \in \{1,2\}} \right\rangle$ such that

$$T_i = \Sigma_{i,\Delta}^\infty = \{(\theta_i, s_i) : \exists e_i, s_i = b_i(\theta_i, e_i)\};$$

therefore, s_i is Δ-rationalizable for θ_i if and only if (θ_i, s_i) is realizable in $\mathcal{M}$. By construction, $\mathcal{M}$ is an equilibrium model consistent with the restrictions Δ. [To see this more explicitly, note that, for each i and $t_i = (\theta_i, s_i) \in T_i = \Sigma_{i,\Delta}^\infty$, we have $\mu^{t_i} \in \Delta^{\theta_i}$, $\mu^{t_i}(T_{-i} | \phi) = 1$, $p_i^{t_i}(\theta_{-i}, s_{-i}) = \mu^{t_i}(\theta_{-i}, s_{-i} | \phi)$ for all $(\theta_{-i}, s_{-i}) \in T_{-i}$. Therefore $\mathcal{M}$ is consistent with Δ. Furthermore, $s_i \in r_i(\theta_i, \mu^{t_i})$ and $s_{-i} = b_{-i}(\theta_{-i}, s_{-i})$ for all $(\theta_{-i}, s_{-i}) \in T_{-i}$. Since a sequential best reply must also be an ex-ante best reply, we have

$$b_i(t_i) = s_i \in \arg\max_{s_i'} \sum_{(\theta_{-i}, s_{-i}) \in T_{-i}} p_i^{t_i}(\theta_{-i}, s_{-i}) U_i(\theta_i, s_i', \theta_{-i}, b_{-i}(\theta_{-i}, s_{-i})),$$

showing that $\mathcal{M}$ is an equilibrium model.] ■

Thus, as anticipated in the Introduction, Propositions 4.2 and 4.3 jointly imply that, *in static games, Δ-rationalizability exactly characterizes the set of Bayesian equilibrium outcomes consistent with restrictions Δ.*

We emphasize that the Bayesian equilibrium model constructed in the proof of Proposition 4.3 (i) does *not* necessarily admit a common prior on the set T of profiles of Harsanyi types, and (ii) allows for the possibility that two distinct Harsanyi types of a player have the same payoff component and hold the same hierarchy of beliefs about Θ, but play different strategies. Both aspects deserve further comment; for simplicity, assume there are no explicit restrictions Δ.

To fix ideas, consider the special case of games with *complete* information (for which Θ is a singleton). For such games, by definition, a Bayesian equilibrium model featuring a common prior on the set T of (payoff-irrelevant) types is a *correlated equilibrium* (see Aumann, 1987). Since there exist complete-information games in which certain actions are rationalizable, but never played in any correlated equilibrium, it follows that the result of Proposition 4.3 does not hold if one restricts attention to common-prior type spaces.

Also, for complete-information games, there exists only one (degenerate) hierarchy of beliefs on Θ. Thus, a Bayesian equilibrium model such that players who have the same type and hold the same hierarchical beliefs about Θ also play the same strategy is simply a pure-strategy Nash equilibrium. Thus, for complete-information games, the result of Proposition 4.3 clearly does not hold under the additional assumption just mentioned.

For games with incomplete information, we offer the following observations. Regarding common priors, the restrictions Δ may be chosen so as to reflect the assumption that a common prior on Θ, the set of payoff-relevant type profiles, is exogenously given—as is often stipulated in applications. Thus, our results do accommodate this possibility. We remark that the existence of a common prior on T is neither necessary nor sufficient for the existence of a common prior on Θ. Further comments and references on this issue may be found in the Discussion section.

Next, say that a Bayesian model exhibits *indirect payoff-relevance* if distinct Harsanyi types correspond to distinct hierarchies of beliefs about payoff types. In view of our focus on the robust implications of Bayesian equilibrium analysis, we are unconvinced that one should restrict attention to Bayesian models with this property.[23] Indeed, we are not aware of any application in which indirect payoff-relevance is explicitly invoked. However, it may be interesting to ascertain whether our equivalence result can be strengthened so as to incorporate this restriction.

We conjecture that, if Θ is not a singleton, then Proposition 4.3 remains true under the indirect payoff-relevance restriction. We are able to prove a somewhat simpler result, that implies that the conjecture is correct for 'almost all' static games with two-sided incomplete information.

To clarify our terminology, fix $(\Theta_i, A_i)_{i\in\{1,2\}}$; then a static game Γ is parametrized by the payoff functions and can be regarded as a point in $\mathbf{R}^{\Theta\times A} \times \mathbf{R}^{\Theta\times A}$. We say that a statement holds for almost all static games if, for each $\Theta \times A$ the set of static games with domain $\Theta \times A$ for which it does not hold is nowhere dense (i.e., its closure has empty interior).

Proposition 4.4 *For almost all static games with at least two payoff types for each player, there is a Bayesian equilibrium model $\mathcal{M}$ such that (1) distinct Harsanyi types have distinct first-order beliefs about the opponent's payoff type $[t_i' \neq t_i'' \Rightarrow \mathrm{marg}_{\Theta_{-i}} p_i^{t_i'} \neq \mathrm{marg}_{\Theta_{-i}} p_i^{t_i''}]$ and (2) an arbitrary pair (θ_i, a_i) is rationalizable if and only if it is realizable in $\mathcal{M}$.*

[23] Our comments on universal type spaces are also pertinent to this issue: interested readers are referred to the Discussion section.

Propositions 4.2 and 4.4 imply that, in almost all static games with two-sided incomplete information, the set of rationalizable outcomes coincides with the set of outcomes realizable in Bayesian equilibrium models that exhibit indirect payoff-relevance.

We remark that Proposition 4.4 can be extended to dynamic games, adopting the appropriate notion of genericity.

4.3 Self-Confirming and Bayesian Equilibrium

We now relate Δ-rationalizability to an extensive-form equilibrium concept, namely self-confirming equilibrium (cf. Fudenberg and Levine (1993)). We do so indirectly, by establishing a form of outcome equivalence between self-confirming and Bayesian equilibrium that may be of independent interest.

4.3.1 Agreement of Beliefs with an Outcome Distribution

Loosely speaking, in a self-confirming equilibrium each player best-responds to her beliefs, and such beliefs are confirmed by whatever evidence she can get about the private information and behavior of her opponents. Specifically, we consider a situation where this evidence is given by reliable statistics about the frequencies of occurrence of combinations of payoff types and terminal histories.

Thus, in order to define self-confirming equilibrium, we need to formalize the assumption that a player's beliefs about her opponent *agree with an outcome distribution* $\zeta \in \Delta(\Theta \times Z)$.

The intuition behind this notion of agreement is as follows. The probability distribution ζ encodes information about payoff types and choices made by *both* players at certain histories. On the other hand, a belief $\mu^i \in \Delta(\Theta_{-i} \times S_{-i})$ held by Player i (at the beginning of the game) encodes information about the payoff type and choices made by Player $-i$ only. However, if ζ reflects the assumptions that (i) payoff types are independent, and (ii) action choices at each non-terminal history are stochastically independent, then information about Player $-i$ at a given non-terminal history *can* be retrieved from ζ by conditioning. If this is the case, it is meaningful to require agreement between the conditional probabilities derived from ζ and μ^i at every non-terminal history reached with positive ζ-probability. This is precisely the notion of agreement we adopt.

It is convenient to introduce additional terminology and notation. First, in the spirit of equilibrium analysis, we shall often refer to a probability distribution $\mu_i \in \Delta(\Theta_i \times S_i)$ either as a *distributional strategy* for Player i (see Milgrom and Weber, 1985), or as a belief held by Player $-i$ about Player i.

Second, an outcome distribution ζ will be deemed *feasible* if $\mathrm{marg}_\Theta \zeta$ is strictly positive and ζ is generated by a product of distributional strategies $(\mu_1 \times \mu_2) \in \Delta((\theta_1 \times S_1) \times (\Theta_2 \times S_2))$. This captures assumptions (i) and (ii) in the preceding paragraph. In the Beer-Quiche example, the distribution $\zeta = \frac{9}{10}[(\theta^s, Q, N)] + \frac{1}{10}[(\theta^w, Q, N)]$ is feasible because it is generated by any product of distributional strategies $\mu_1 \times \mu_2$ with $\mu_1 = \frac{9}{10}[(\theta^s, Q)] + \frac{1}{10}[(\theta^w, Q)]$ and $\mu_2 \in \Delta(\{FN, NN\})$.

Third, we define the probability of non-terminal histories, as well as other related probabilities.

Denote by $\preceq$ ($\prec$) the (asymmetric) "prefix of" relation on $\overline{\mathcal{H}}$ and let

$$\zeta(h) = \sum_{\theta,z:h\prec z} \zeta(\theta,z), \quad \zeta(\theta_i,h) = \sum_{\theta',z:\theta'_i=\theta_i,h\prec z} \zeta(\theta',z), \quad \zeta(\theta_i,h,a_{-i}) = \sum_{z:\exists a_i\in A_i(h),(h,a_i,a_{-i})\preceq z} \zeta(\theta_i,z).$$

Finally, we define the collection of strategies of Player i that are consistent with history h and choose a specific action $a_i \in A_i(h)$ at h:

$$\forall h \in \mathcal{H},\ a_i \in A_i(h): \quad S_i(h,a_i) = \{s_i \in S_i(h) : s_i(h) = a_i\}.$$

We can now formalize the notion of agreement with an outcome distribution ζ.

Definition 4.5 *A distributional strategy (or belief) $\mu_{-i} \in \Delta(\Theta_{-i}\times S_{-i})$* agrees *with ζ (is consistent with ζ) if for all $h \in \mathcal{H}$, $\theta_{-i} \in \Theta_{-i}$, $a_{-i} \in A_{-i}(h)$,*

$$\zeta(h) > 0 \Rightarrow \frac{\mu_{-i}(\{\theta_{-i}\} \times S_{-i}(h,a_{-i}))}{\mu_{-i}(\Theta_{-i} \times S_{-i}(h))} = \frac{\zeta(\theta_{-i},h,a_{-i})}{\zeta(h)}. \tag{2}$$

A Bayesian model $\mathcal{M}$ of Γ is consistent with an outcome distribution $\zeta \in \Delta(\Theta\times Z)$ if it is consistent with the following restrictions:

$$\Delta^{\theta_i} = \{\mu^i \in \Delta^{\mathcal{H}}(\Theta_{-i} \times S_{-i}) : \mu^i(\cdot|\phi) \textit{ agrees with } \zeta\},\ i = 1,2, \theta_i \in \Theta_i\ .$$

Thus, a Bayesian model $\mathcal{M}$ is consistent with ζ if, for $i = 1,2$, the distributional strategy for Player $-i$ derived from each belief p^{t_i} in $\mathcal{M}$ agrees with ζ.

The following remark clarifies the implications of the notion of agreement.

Remark 4.6 *Let $\mu_{-i} \in \Delta(\Theta_{-i} \times S_{-i})$ be a distributional strategy that agrees with a given $\zeta \in \Delta(\Theta \times Z)$. Then:*
(1) $\mathrm{marg}_{\Theta_{-i}} \mu_{-i} = \mathrm{marg}_{\Theta_{-i}} \zeta$;
(2) for all $h \in \mathcal{H}$, $\theta_{-i} \in \Theta_{-i}$, $a_{-i} \in A_{-i}(h)$: $\zeta(\theta_{-i},h) > 0 \Rightarrow \frac{\mu_{-i}(\{\theta_{-i}\}\times S_{-i}(h,a_{-i}))}{\mu_{-i}(\{\theta_{-i}\}\times S_{-i}(h))} = \frac{\zeta(\theta_{-i},h,a_{-i})}{\zeta(\theta_{-i},h)}$.

4.3.2 Characterization of Self-Confirming Equilibrium Outcomes

Self-confirming equilibrium can now be defined.

Definition 4.7 *A feasible distribution $\zeta \in \Delta(\Theta\times Z)$ is a* self-confirming equilibrium *(SCE) if there are distributional strategies $\mu_i \in \Delta(\Theta_i \times S_i)$ and $(\mu^{\theta_i})_{\theta_i\in\Theta_i} \in [\Delta(\Theta_{-i} \times S_{-i})]^{\Theta_i}$, $i = 1,2$, such that, for each player i,*
(1) $\forall\theta_i, s_i$, if $\mu_i(\theta_i,s_i) > 0$ then

$$s_i \in \arg\max_{s'_i} \sum_{\theta_{-i},s_{-i}} U_i(\theta_i,\theta_{-i},s'_i,s_{-i})\mu^{\theta_i}(\theta_{-i},s_{-i});$$

(2) $\forall\theta_i$, μ^{θ_i} agrees with ζ;
(3) μ_i is agrees with ζ.

Condition (1) of this definition deserves further comment. Note that every strategy played with positive μ_i-probability by type θ_i is rationalized using the *same* belief $\mu_{-i}^{\theta_i}$, as if we were checking whether type θ_i plays a randomized best response.[24] A more congenial interpretation of the randomness of behavior is to assume that the game Γ is played by individuals drawn at random from large heterogeneous populations (one for each player role), and that different individuals with the same payoff type (preferences and/or abilities) may play different strategies. According to this interpretation, the probability $\mu_i(\theta_i, s_i)/\zeta(\theta_i)$ represents the relative frequency of strategy s_i in the sub-population of individuals whose payoff type is θ_i.

This interpretation would call for a weakening of Condition (1) whereby different strategies may be justified by possibly different beliefs, as in a Bayesian equilibrium.[25] However, the following proposition (the main result of this subsection) shows that, even with the stronger definition of SCE given above, if a Bayesian equilibrium is consistent with ζ, then ζ is an SCE distribution. It turns out that the converse is also true.[26]

Proposition 4.8 *$\zeta \in \Delta(\Theta \times Z)$ is an SCE distribution if and only if there is a Bayesian equilibrium model consistent with ζ.*

Sketch of proof. *(If)* Fix a Bayesian equilibrium model $\mathcal{M}$. From the beliefs of each Harsanyi type t_i we obtain a corresponding distributional strategy for Player $-i$. If $\mathcal{M}$ is consistent with ζ then any product of such distributional strategies for i and $-i$ induces distribution ζ on $\Theta \times Z$. The key observation is that any pair (θ_i, s_i) with positive probability according to some opponent's type t_{-i} is such that s_i is a (ex ante) best response for θ_i to beliefs that agree with ζ. Fix two pairs (θ_i, s_i) and (θ_i, s_i'). Since the beliefs justifying s_i and s_i' both agree with ζ, it can be shown that the belief justifying s_i also justifies s_i'. Therefore we can recover from $\mathcal{M}$ an array of distributional strategies supporting ζ as an SCE distribution.

(Only if). Given an array of distributional strategies $(\mu_i, \mu_i^{\theta_i})_{i=1,2,\theta_i\in\Theta_i}$ supporting ζ as an SCE, construct a Bayesian model $\mathcal{M}$ by setting $E_i = S_i$, $T_i = \bigcup_{\theta_{-i}\in\Theta_{-i}} \operatorname{supp} \mu^{\theta_{-i}} \subset \Theta_i \times S_i$, $p_i^{(\theta_i,s_i)} = \mu^{\theta_i}$. One can then show that conditions (1)-(3) of Definition 4.7 imply that $\mathcal{M}$ is an equilibrium model consistent with ζ. ■

4.3.3 Comments

The "if " result, together with Proposition 4.3, clarifies the relationship between self-confirming equilibrium and Δ-rationalizability. Specifically, it implies that *Δ-rationalizability yields a forward-induction refinement of self-confirming equilibrium.*[27] We show in Section 5 that, in signaling games, this refinement corresponds to the Iterated Intuitive Criterion.

The "only if" result crucially relies on the assumption that beliefs agree with the "true" distribution on $\Theta \times Z$. With coarser feedback about the outcome of the interaction (corresponding

[24] This corresponds to the definition of self-confirming equilibrium with "unitary beliefs" of Fudenberg and Levine (1993), as applied to the extensive form representation where the player set is $\Theta_1 \cup \Theta_2$.

[25] This corresponds to the definition of a *"type heterogeneous SCE"* in Dekel *et al.* (2003).

[26] The converse extends to incomplete information games similar results due to Battigalli (1987), Fudenberg and Levine (1993) and Kalai and Lehrer (1993).

[27] Reny (1992) puts forward a similar refinement of Nash equilibrium.

to weaker definitions of SCE), the result may fail. For example, in some contexts it is plausible to assume that only the distribution of actions (terminal histories) can be observed ex post. It can be shown that in this case the "only if" part of Proposition 4.8 fails (see Example 3 in Dekel *et al.,* 2003). Here we consider a different case. Suppose that individuals can observe ex post the distribution of payoffs (e.g. because there are public statistics on income distribution), but not the actual distribution $\zeta \in (\Theta \times Z)$ that generates it. Then the definition of SCE must be modified as follows: ζ is a *SCE* distribution* if it is generated by a pair of distributional strategies such that each type chooses a best response to some belief that "agrees" with the distribution of payoffs generated from ζ. The following example shows that there are SCE* distributions ζ for which there is no Bayesian equilibrium model consitent with ζ.

Consider the incomplete information game given by the following matrices, where Player 1 chooses the row and Player 2 has private information:

θ_2'	ℓ	r
u	2,2	2,1
d	0,1	4,0

θ_2''	ℓ	r
u	2,1	2,2
d	0,0	4,1

Let $\zeta(\theta_2', u, \ell) = \frac{1}{3}$ and $\zeta(\theta_2'', u, r) = \frac{2}{3}$. Then ζ is a SCE* distribution that generates a distribution on payoffs concentrated on $(2, 2)$. In particular, distribution ζ obtains if each type of Player 2 chooses its dominating action and Player 1 best responds to a belief μ^1 such that $\mu^1(\theta_2', \ell) > \frac{1}{2}$, $\mu^1(\theta_2'', r) = 1 - \mu^1(\theta_2', \ell)$. Such belief does not agree with ζ but it agrees with the observed distribution of payoffs. Note that, in every Bayesian equilibrium model where Player 1 assigns probability $\frac{2}{3}$ to θ_2'', Player 1 chooses d. Therefore there is no Bayesian equilibrium model consistent with ζ.[28]

5 Rationalization and the Iterated Intuitive Criterion

We now focus on the implications of Δ-rationalizability for signalling games. We first provide an alternative characterization of the procedure for this class of games. The main result of this section then states that Δ-rationalizability characterizes self-confirming equilibrium outcomes consistent with the Iterated Intuitive Criterion of Cho and Kreps (1987). The Beer-Quiche example of Section 3 illustrates this result.

Recall that a *signalling game* is a two-person, two stage game with sequential moves 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 incomplete information already implies that the set of feasible actions of the Sender is the same for each payoff-

[28] Note that there must be *some* Bayesian equilibrium model where Player 1 plays u, θ_2' plays ℓ and θ_2'' plays r, because this profile is rationalizable (with no restrictions on beliefs; see Proposition 4.3). We only claim that any such model cannot be consistent with the given distribution ζ. There are examples of self-confirming distributions that violate common belief in rationality and hence are inconsistent with any Bayesian equilibrium model. Example 3 in Dekel *et al.* (2003) is a case in point. The working paper version of the present article contained another example of this sort.

Object	**Notation**	**Remarks**
Payoff-Types for Player 1	$\theta \in \Theta = \Theta_1$	
Actions	$m \in M = A_1$, $a \in A = A_2$	$S_1 = M$ $S_2 = A^M$
Partial histories	$\mathcal{H} = \{\phi\} \cup M$	
Outcome distribution	$\zeta \in \Delta(\Theta \times M \times A)$	

Table 2: Notation for Signalling Games

type. We also assume that the set of feasible actions for the Receiver is independent of the signal.[29] Table 2 summarizes our notation for signalling games.

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

For any given outcome distribution $\zeta \in \Delta(\Theta \times M \times A)$ 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 feasible, $\zeta(m|\theta)$ and $\zeta(m, a|\theta)$ are always well-defined, because $\zeta(\theta) > 0$ for all θ.

Δ-rationalizability admits a simplified characterization in signalling games. Fix an outcome distribution ζ. Denote by $\Sigma^k_{1,\zeta}$ and $S^k_{2,\zeta}$, $k = 0, 1, \ldots$, the sets of type-message pairs and strategies obtained from step k of the procedure specified in Definition 3.1, when first order beliefs are assumed to agree with ζ, i.e. assuming the restrictions[30]

$\Delta^1 = \left\{\mu^1 \in \Delta(S_2) : \forall m, \forall a, \zeta(m) > 0 \Rightarrow \mu^1(\{s_2 : s_2(m) = a\}) = \zeta(a|m)\right\}$,

$\Delta^2 = \left\{\mu^2 \in \Delta^{\mathcal{H}}(\Sigma_1) : \forall(\theta, m), \mu^2(\theta, m|\phi) = \zeta(\theta, m)\right\}$.

For notational simplicity, we refer to this procedure as *ζ-rationalizability*. Now consider the following iterative definition, which, loosely speaking, identifies types that may send a given message, and actions that may be played in response to a given message. For every $m \in M$, let $\Theta^0(m; \zeta) = \Theta$ and $A^0(m; \zeta) = A$; then, for $k > 0$, let

$$\Theta^k(m;\zeta) = \left\{ \theta \in \Theta^{k-1}(m;\zeta) \ : \ \exists \pi_2 \in [\Delta(A)]^M \text{ s.t. } \begin{array}{l} \forall m', \forall a, \zeta(m') > 0 \Rightarrow \pi_2(a|m') = \zeta(a|m') \\ \forall m', \pi_2(A^{k-1}(m';\zeta)|m') = 1 \\ m \in \arg\max_{m'} \sum_a u_1(\theta, m', a)\pi_2(a|m') \end{array} \right\}$$

and

$$A^k(m;\zeta) = \left\{ a \in A^{k-1}(m;\zeta) : \exists \nu^m \in \Delta(\Theta) \text{ s.t. } \begin{array}{l} \Theta^{k-1}(m;\zeta) \neq \emptyset \Rightarrow \nu^m(\Theta^{k-1}(m;\zeta)) = 1; \\ \zeta(m) > 0 \Rightarrow \forall \theta, \nu^m(\theta) = \zeta(\theta|m); \\ a \in \arg\max_{a'} \sum_\theta u_2(\theta, m, a')\nu^m(\theta) \end{array} \right\}.$$

We then have:

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

[30] Since the restrictions for the Sender are type-indepedent, we omit θ from the notation. Note also that $\Delta^{\mathcal{H}}(S_2)$ is isomorphic to $\Delta(S_2)$ because in a signaling game $S_2(h) = S_2$ for all $h \in \mathcal{H}$.

Lemma 5.1 *For all $k \geq 0$, θ, m and s_2,*
(i) $(\theta, m) \in \Sigma^k_{1,\zeta}$ if and only if $\theta \in \Theta^k(m;\zeta)$
(ii) $s_2 \in S^k_{2,\zeta}$ if and only if, for all m', $s_2(m') \in A^k(m';\zeta)$.

Cho and Kreps (1987) put forward the Iterated Intuitive Criterion (IIC) as a test for sequential equilibria, but the same criterion can be naturally and more generally be applied to self-confirming equilibria (cf. Kohlberg, 1990, p. 23, footnote 17). We now provide the details.

For any outcome distribution 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,

$$BR_2(\overline{\Theta}, m) = \bigcup_{\nu \in \Delta(\overline{\Theta})} \left\{ \arg\max_{a\in A} \sum_\theta \nu(\theta) u_2(\theta,m,a) \right\}$$

denotes the set of best responses to beliefs concentrated on $\overline{\Theta}$ given message m.

Let $IA^0(m;\zeta) = A$ and $I\Theta^0(m;\zeta) = \Theta$ and, for all $n = 0, 1, 2, ...$ define

$$\begin{aligned} IA^{n+1}(m;\zeta) &= \begin{cases} BR_2(I\Theta^n(m;\zeta), m), & \text{if } I\Theta^n(m;\zeta) \neq \emptyset \\ IA^n(m;\zeta), & \text{if } I\Theta^n(m;\zeta) = \emptyset. \end{cases}, \\ I\Theta^{n+1}(m;\zeta) &= \left\{ \theta \in I\Theta^n(m;\zeta) : u_1^\zeta(\theta) \leq \max_{a \in IA^n(m;\zeta)} u_1(\theta,m,a) \right\}. \end{aligned}$$

We then have:

Definition 5.2 *A SCE outcome distribution ζ satisfies the IIC 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 $a \in \bigcap_{n>0} IA^n(m;\zeta)$ such that $u_1(\theta,m,a) \leq u_1^\zeta(\theta)$.*[31]

Proposition 5.3 *ζ is an SCE outcome distribution satisfying the IIC if and only if there is a ζ-rationalizable strategy for the Receiver and a ζ-rationalizable message for each type θ of the Sender ($S^\infty_{2,\zeta} \neq \emptyset$ and $\mathrm{proj}_\Theta \Sigma^\infty_{1,\zeta} = \Theta$).*[32]

Proof. We provide only a sketch here. A complete proof can be found in the Appendix.

[31] In the Appendix we show that the definition in the text is equivalent to the original one due to Cho and Kreps (1987).

[32] Furthermore, one can show that ζ is a SCE outcome distribution satisfying the (non iterated) Intuitive Criterion if and only if $\Sigma^4_{1,\zeta} \times S^4_{2,\zeta} \neq \emptyset$.

(If) If $S^{\infty}_{2,\zeta} \neq \emptyset$ and $\mathrm{proj}_{\Theta}\Sigma^{\infty}_{1,\zeta} = \Theta$, the hypothesis of Proposition 4.3 is satisfied; hence there is a Bayesian equilibrium consistent with ζ. By Proposition 4.8, ζ must be an SCE distribution. Furthermore, if $\zeta(\theta, m^*) > 0$, there is a ζ-rationalizable belief of the Receiver assigning positive prior probability to (θ, m^*), which means that m^* must be a ζ-rationalizable message for type θ, or (by Lemma 5.1) $\theta \in \bigcap_k \Theta^k(m^*; \zeta)$.

One of the main steps of the proof consists in showing that if $\theta \in \bigcap_k \Theta^k(m^*; \zeta)$ for each pair (θ, m^*) with $\zeta(\theta, m^*) > 0$, then for every *off-the-ζ-path* message m and every step k, $\Theta^k(m; \zeta) = I\Theta^k(m; \zeta)$ and $A^k(m; \zeta) = IA^k(m; \zeta)$. This in turn implies that for every off-the-ζ-path message m and type θ there is an action $a \in \bigcap_k IA^k(m; \zeta)$ such that $u_1^{\zeta}(\theta) \geq u_1(\theta, m, a)$, i.e. ζ satisfies the IIC.

(Only if) If ζ is an SCE satisfying the IIC, then it can be shown that, as above, for every off-the-ζ-path message m and every step k, $\Theta^k(m; \zeta) = I\Theta^k(m; \zeta)$ and $A^k(m; \zeta) = IA^k(m; \zeta)$; furthermore, $\bigcap_{k>0} IA^k(m; \zeta) \neq \emptyset$ (see Definition 5.2). Therefore there is some ζ-rationalizable response to every message m off the ζ-path. On the other hand, the ζ-rationalizable responses to any message m^* on the ζ-path are simply the best responses to belief $\zeta(\cdot|m) \in \Delta(\Theta)$. Since there is a ζ-rationalizable response for every message, $S^{\infty}_{2,\zeta}$ is not empty. By Remark 3.4, this implies $\mathrm{proj}_{\Theta}\Sigma^{\infty}_{1,\zeta} = \Theta$. ■

6 Discussion

In this section we jointly discuss the related literature and some extensions and applications of the basic framework.

Common Priors. As we observed in Section 4.1, the notion of Δ-rationalizability is consistent with, but does not necessarily reflect the assumption that the players' hierarchical beliefs about each other's payoff types and strategies are generated by a common prior on some set T of Harsanyi types. We have also noted that, on the other hand, our approach may incorporate the assumption of a common prior on the set Θ of *payoff-relevant* types.

Recall that, in Harsanyi's approach, a type represents a player's hierarchical beliefs about Θ (and, for a given equilibrium profile b, her hierachical beliefs about choices as well). Thus, in a genuine incomplete-information setting, the common-prior assumption should be evaluated on the basis of its implications for the players' hierarchical beliefs. Recent work by Bonanno and Nehring (1999), Samet (1999a,b), and Feinberg (2000) clarifies that assuming the existence of a common prior on T is equivalent to imposing a rather strong restriction on mutual beliefs: in particular, players cannot "agree to disagree" on the probability of events related to payoff types.

Unless this sort of assumption arises naturally in a specific application, if one is interested in the robust implications of Bayesian equilibrium analysis, it seems appropriate not to insist on requiring that the common-prior assumption be satisfied.

Further discussion of these issues may be found in Aumann (1998), Gul (1998), Morris (1995), and Bergemann and Morris (2002).

Universal Type Space and Spaces of Hierarchical Beliefs. In order to simplify the discussion, assume that the explicit restrictions Δ concern the players' beliefs about payoff types. Our approach is then equivalent to carrying out standard Bayesian equilibrium analysis in the context of the Δ-universal type space, formed by taking the 'union' of all Harsanyi type spaces based on Θ and consistent with Δ.[33] More precisely, an immediate consequence of Propositions 4.2 and 4.3 is that, in static games, every Δ-rationalizable outcome is realizable in an equilibrium of the Bayesian game with the Δ-universal type space.

Also, Mertens and Zamir (1985) and Brandenburger and Dekel (1993) provide constructive characterizations of the type space consisting of all coherent hierarchies of beliefs on a given set X. We note that, if $X = \Theta$, the resulting type space is *smaller* than the universal type space described above. Furthermore, carrying out Bayesian equilibrium analysis of a specific situation by embedding Θ in the Mertens-Zamir / Brandenburger-Dekel type space incorporates the indirect payoff-relevance restriction that player types who share the same payoff component and hierarchy of beliefs on Θ play the same strategy; see the discussion following Proposition 4.3. In other words, this alternative approach reflects a weaker notion of robustness than the one we adopt. See also the discussion in Bergemann and Morris (2002, subsection 2.1.2).

n players

The main issue in this more general case is whether we should assume that a player regards the types and (strategic) choices of distinct opponents as stochastically independent. If not, the extension of the solution concept and results is straightforward. If instead independence is assumed, subtle issues arise pertaining to players' beliefs after unexpected events. In order to fully reflect the implications of stochastic independence, marginal beliefs about the types and strategies of distinct opponents should be updated independently of one another; joint (conditional) beliefs should then be derived as product measures. This can be added as an explicit restriction of beliefs. But then it also makes sense to assume that the observed behavior of opponent j is rationalized independently of the behavior of opponent k. Thus, if k's behavior is clearly irrational, but j's behavior is not, i should keep believing that j is rational. We refer the reader to Battigalli (1996, 1999) and references therein for details.

Infinite games and more general information structures

For the sake of simplicity, we have restricted our attention to finite games with observable actions. Battigalli (1999) provides an analysis of Δ-rationalizability in games with possibly infinite type and action spaces, and shows how to extend the analysis to games with perfect recall.

These extensions affect some of our results. Propositions 4.2 and 4.3 continue to hold as stated, whereas the genericity qualification in Proposition 4.4 relies on finiteness. We conjecture that Proposition 4.8 holds for all games with perfect recall and can also be extended to sufficiently regular infinite games. Similarly, we conjecture that Proposition 5.3 can be extended to sufficiently regular infinite signaling games.

[33] Fix a collection of type spaces based on Θ and consistent with Δ, $\{\mathcal{T}^k = (T_i^k, p_i^k)_{i\in\{1,2\}}, k \in K\}$, where K is some index set. It may be assumed without loss of generality that the sets T_i^k $(k \in K)$ are disjoint. The union of the $\mathcal{T}^k$'s is the type space $\mathcal{T} = (\bigcup_{k\in K} T_i, p_i)_{i\in\{1,2\}}$ where p_i is obtained from the belief functions p_i^k in the obvious way [for all $k \in K$, $t_i \in T_i^k$, $F_{-i} \subset T_{-i}$, $p_i^{t_i}(F_{-i}) = p_i^{k,t_i}(F_{-i} \cap T_{-i}^k)$]. For more general restrictions Δ concerning beliefs about both strategies and types, type spaces based on $\Theta \times S$ must be considered.

Weak rationalizability and perfect Bayesian equilibrium

We have focused on a notion of rationalizability for dynamic games whereby players try to rationalize the observed behavior of the opponent. Alternatively, one may wish to analyze a weaker notion of extensive form rationalizability that does *not* satisfy this forward-induction requirement, but instead only reflects the assumption that there is common certainty of sequential rationality at the beginning of the game. We refer the reader to Ben Porath (1997), Battigalli and Siniscalchi (1999) and Battigalli (1999) for the epistemic and procedural characterizations of this "*weak (extensive form) rationalizability*" solution concept. Weak rationalizability can be related to a version of the Bayesian equilibrium concept whereby players are assumed to carry out *sequential* best responses to their beliefs. We call such equilibria "weakly perfect Bayesian". Battigalli (1999, Proposition 3.10) provides a generalization of Propositions 4.2 and 4.3 stating that an outcome is realizable in a weakly perfect Bayesian equilibrium consistent with Δ if and only if it is weakly Δ-rationalizable. On the other hand, the analog of Proposition 4.8 ("only if") does not hold: a self-confirming equilibrium distribution ζ need not be consistent with any weakly perfect Bayesian equilibrium, because it may be supported by beliefs whereby players ascribe to their opponents irrational behavior at unreached histories. In other words, an SCE distribution ζ may be inconsistent with the assumption of initial common certainty of *sequential* rationality.[34]

Other definitions of self-confirming equilibrium

Recall that it is possible to interpret the incomplete-information model as a grand game with random matching of players drawn from heterogeneous populations, assuming that the distribution of outcomes (types and terminal histories) is observed ex post. Our definition of self-confirming equilibrium is motivated by this interpretation; in particular, an SCE is supposed to represent a stable distribution in such an environment. Different interpretations call for different definitions. As we mentioned, relaxing the assumptions about information feedback one obtains weaker definitions of SCE; as a result, the "only if" part of Proposition 4.8 is lost. Definitions of SCE related to different environments can be found in Battigalli (1987), Fudenberg and Levine (1993), Kalai and Lehrer (1993, 1995), Battigalli and Guaitoli (1997) and Dekel *et al.* (2003). The latter paper considers a class of environments where the state of nature is drawn at random at each repetition of the game according to an objective distribution $\rho \in \Delta(\Theta)$; they show that it is very difficult to find information structures that allow the players to learn to play a given Bayesian equilibrium with incorrect beliefs about the state of nature. The reason is quite simple: if the information feedback is too poor they need not come to have correct conjectures about the opponents, if it is too rich they need not hold on to their (incorrect) beliefs about the state of nature.[35]

Iterated Intuitive Criterion

[34] By the above mentioned equivalence result, ζ is an SCE distribution consistent with common certainty of (agreement with ζ and) sequential rationality if and only if there is some weakly perfect Bayesian equilibrium model consistent with ζ. For related definitions of *rationalizable SCE* see Rubinstein and Wolinsky (1994) (static games) and in particular Dekel *et al.* (1999).

[35] For an environment where the state of nature θ is determined once and for all and the same players interact repeatedly, Battigalli and Guaitoli (1997) put forward a notion of "conjectural equilibrium" at a given state θ. Dekel *et al.* (2003) also briefly consider this case.

Our analysis of the IIC is inspired by the work of Sobel *et al.* (1990). In particular, Proposition 5.3 is somewhat similar to their Proposition 2, which relates the IIC to extensive-form rationalizability in an auxiliary game where the messages on the equilibrium path are coalesced into a fictitious 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 3.1; here the 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. But the similarities with Sobel *et al.* (1990) allow us to adapt some of their arguments in order to prove Lemma 5.1 and Proposition 5.3. Christian Ewerhart (private communication) provided a related characterization of the Intuitive Criterion.

Applications of Δ*-rationalizability*

Battigalli (1999) provides applications of Δ-rationalizability to models of disclosure, job market signaling and reputation. The latter application uses the "weak" version of the concept (see comment above) and relies on previous work by Watson (1993) and Battigalli and Watson (1997). Battigalli and Siniscalchi (2003) and Dekel and Wolinsky (2003) apply Δ-rationalizability to the analysis of first price auctions. Δ-rationalizability can be used to replicate Perry and Reny's "General Solution to King Solomon's Dilemma" (1999). Our second example in Section 3.2 is inspired by their work.

Morris and Skiadas (2000) provide necessary and sufficient conditions for rationalizable trade among two players. Once a small trading cost is introduced, their analysis may be interpreted as an application of Δ-rationalizability.[36]

Finally, Bergemann and Morris (2002) analyze robust mechanism design under incomplete information. In order to formalize the notion of robustness, they consider various specifications of the agents' type spaces, including the universal space formed by taking the union of all Harsanyi type spaces. Thus, while their paper and ours focus on different issues and adopt different approaches and techniques, they share a common motivation. Furthermore, while most of their analysis involves equilibrium notions, some of their results can be interpreted as pertaining to Δ-rationalizable (truthful) implementation.[37]

7 Colophon

Part of this paper is a revision of Battigalli (1999), an unpublished manuscript. All the additional material has been jointly developed by Pierpaolo Battigalli and Marciano Siniscalchi. Part of the material concerning the Iterated Intuitive Criterion was originally contained in the working paper version of Battigalli and Siniscalchi (2002). We thank Adam Brandenburger, Drew Fudenberg,

[36] Formally, fix nonempty sets $\widehat{\Theta}_{-i}(\theta_i) \subset \Theta_{-i}$ for $i = 1, 2$ and $\theta_i \in \Theta_i$; then $\Delta^{\theta_i} = \{\mu^i : \mathrm{supp}\mu^i = \widehat{\Theta}_{-i}(\theta_i) \times S_{-i}\}$. Morris and Skiadas actually, specify a prior p_i on Θ for each player i, but they observe that their proof only relies on whether or not certain pairs of types are in the support of the prior.

[37] For instance, one such result states that a social choice correspondence is (truthfully) implementable in an equilibrium of the Bayesian model with the universal type space if and only if it is (truthfully) implementable in rationalizable strategies. This can be seen as a consequence of our Propositions 4.2 and 4.3 (and also of Proposition 3.10 in Battigalli (1999)). We observe that Bergemann and Morris employ a different terminology than we do: see Footnote 11.

Joel Watson, two anonymous referees and the Editor for helpful comments. This material is based upon work supported by the National Science Foundation under Grant No. 9911490. Pierpaolo Battigalli also thanks MIUR and Bocconi University for financial support.

Pierpaolo Battigalli, Bocconi University and IGIER
pierpaolo.battigalli@uni-bocconi.it

Marciano Siniscalchi, Northwestern University and Princeton University
marciano@northwestern.edu

8 Appendix

8.1 Proof of Proposition 4.4

We show that, for almost all games, the justifying beliefs $\mu^{(\theta_i,a_i)}$ in the proof of Proposition 4.3 can be chosen so as to ensure that $\mathrm{marg}_{\Theta_{-i}}\,\mu^{(\theta_i,a_i)} \neq \mathrm{marg}_{\Theta_{-i}}\,\mu^{(\theta'_i,a'_i)}$ whenever $(\theta_i,a_i) \neq (\theta'_i,a'_i)$. The Bayesian model constructed in that proof will then satisfy the appropriate restriction on first-order beliefs about Θ.

Throughout this proof, we indicate the best-response correspondence and the set of rationalizable choices for player i in the game $u \in \mathbf{R}^{\Theta\times A}\times\mathbf{R}^{\Theta\times A}$ by $r_i(\cdot,\cdot,u)$ and $\Sigma_i^\infty(u)$ respectively. By standard arguments $r_i(\theta_i,\mu,u)$ is upper-hemicontinuous in (μ,u) (of course, the finite set A_i is endowed with the discrete topology).

Say that a game $u \in \mathbf{R}^{\Theta\times A}\times\mathbf{R}^{\Theta\times A}$ satisfies the *strict best-response property* (SBRP) iff, for all i and $(\theta_i,a_i) \in \Sigma_i^\infty(u)$, there exists $\mu \in \Delta(\Theta_{-i}\times A_{-i})$ such that $\mu(\Sigma_{-i}^\infty(u)) = 1$ and $r_i(\theta_i,\mu,u) = \{a_i\}$.

We claim that the set $\mathit{NSBR} \subseteq \mathbf{R}^{\Theta\times A}\times\mathbf{R}^{\Theta\times A}$ of games that do not satisfy SBRP is nowhere dense**.** To see this, first note that NSBR can be decomposed as follows: let $\mathcal{C}$ denote the collection of all subsets $\Sigma = \Sigma_1\times\Sigma_2$ with $\Sigma_i \subseteq \Theta_i\times A_i$, $\mathrm{proj}_{\Theta_i}\Sigma_i = \Theta_i$ for $i = 1,2$ (see Remark 3.2); then

$$\mathit{NSBR} = \bigcup_{\Sigma\in\mathcal{C}}\bigcup_{i=1,2}\bigcup_{(\theta_i,a_i)\in\Sigma_i}\{u : \Sigma^\infty(u) = \Sigma, \forall\mu\in\Delta(\Sigma_{-i}),\ r_i(\theta_i,\mu,u)\neq\{a_i\}\}\,.$$

Let

$$\mathit{NSBR}(\Sigma,\theta_i,a_i) \equiv \{u : \Sigma^\infty(u) = \Sigma, \forall\mu\in\Delta(\Sigma_{-i}),\ r_i(\theta_i,\mu,u)\neq\{a_i\}\}\,;$$

then

$$\mathit{NSBR}(\Sigma,\theta_i,a_i) \subseteq \mathit{NS}(\Sigma,\theta_i,a_i)\cap \mathit{BR}(\Sigma,\theta_i,a_i),$$

where

$$\begin{aligned}\mathit{NS}(\Sigma,\theta_i,a_i) &= \{u : \forall\mu\in\Delta(\Sigma_{-i}),\ \exists a_i^\mu\in r_i(\theta_i,\mu,u)\backslash\{a_i\}\}\,,\\ \mathit{BR}(\Sigma,\theta_i,a_i) &= \{u : \exists\nu\in\Delta(\Sigma_{-i}),\ a_i\in r_i(\theta_i,\nu,u)\}\,.\end{aligned}$$

We claim that $\mathit{NS}(\Sigma,\theta_i,a_i)\cap \mathit{BR}(\Sigma,\theta_i,a_i)$ is a closed set with empty interior.

To see that this set is closed consider a sequence $\{u^n\}_{n\geq1} \subseteq \mathit{NS}(\Sigma,\theta_i,a_i)\cap \mathit{BR}(\Sigma,\theta_i,a_i)$ such that $u^n \to u$. Then for all $\mu\in\Delta(\Sigma_{-i})$ and all n there are $a_i^{\mu,n}\neq a_i$ and ν^n such that $a_i^{\mu,n}\in r_i(\theta_i,\mu,u^n)$

and $a_i \in r_i(\theta_i, \nu^n, u^n)$. By compactness of $A_i \times \Delta(\Sigma_{-i})$ we may as well assume that, for all μ there is $(a_i^{\mu}, \nu) \in A_i \times \Delta(\Sigma_{-i})$ such that $(a_i^{\mu,n}, \nu^n) \to (a_i^{\mu}, \nu)$ for some $(a_i^{\mu}, \nu) \in A_i \times \Delta(\Sigma_{-i})$, which in turn implies that $a_i^{\mu,n} = a_i^{\mu}$ for n large enough; hence $a_i^{\mu} \neq a_i$. By upper-hemicontinuity of $r_i(\theta_i, \cdot, \cdot)$, $a_i^{\mu} \in r_i(\theta_i, \mu, u)$ and $a_i \in r_i(\theta_i, \nu, u)$. Therefore $u \in NS(\Sigma, \theta_i, a_i) \cap BR(\Sigma, \theta_i, a_i)$.

$NS(\Sigma, \theta_i, a_i) \cap BR(\Sigma, \theta_i, a_i)$ has empty interior if for every $u \in NS(\Sigma, \theta_i, a_i) \cap BR(\Sigma, \theta_i, a_i)$ and every $\varepsilon > 0$, there is at least one game $u' \notin NS(\Sigma, \theta_i, a_i) \cap BR(\Sigma, \theta_i, a_i)$ such that $\|u - u'\| < \varepsilon$, where $\|\cdot\|$ is the sup-norm. Fix $u \in NS(\Sigma, \theta_i, a_i) \cap BR(\Sigma, \theta_i, a_i)$, $\varepsilon > 0$ and define v as follows:

$$\begin{array}{rcl} u'_{-i} & = & u_{-i} \\ u'_i(\theta'_i, a'_i, \theta_{-i}, a_{-i}) & = & \left\{ \begin{array}{ll} u_i(\theta'_i, a'_i, \theta_{-i}, a_{-i}), & \text{if } (\theta'_i, a'_i) \neq (\theta_i, a_i) \\ u_i(\theta'_i, a'_i, \theta_{-i}, a_{-i}) + \frac{\varepsilon}{2}, & \text{if } (\theta'_i, a'_i) = (\theta_i, a_i) \end{array} \right. \end{array}$$

By definition there is some $\nu \in \Delta(\Sigma_{-i})$ such that $a_i \in r_i(\theta_i, \nu, u)$. By construction $\|u - u'\| < \varepsilon$ and $\{a_i\} = r_i(\theta_i, \nu, u')$. Therefore u' has the required property.

Thus, each set $NSBR(\Sigma, \theta_i, a_i)$ is contained in a closed set with empty interior, hence it is nowhere dense. This implies that the finite union $NSBR = \bigcup_{\Sigma \in \mathcal{C}} \bigcup_{i=1,2} \bigcup_{(\theta_i, a_i) \in \Sigma_i} NSBR(\Sigma, \theta_i, a_i)$ is nowhere dense as required.

The proof will be completed by showing that, for any game u that satisfies the strict best-response property, justifying beliefs can be chosen as indicated above. Enumerate the rationalizable action-type pairs of Player i: thus, $\Sigma_i^{\infty} = \{(\theta_i^k, a_i^k) : k = 1, ..., K_i\}$, for some $K_i \geq 1$. Now argue by induction.

For $k = 1$, choose a belief $\mu^1 \in \Delta(\Theta_{-i} \times A_{-i})$ such that $a_i^1 \in r_i(\theta_i^1, \mu^1)$ and $\mu^1(\Sigma_{-i}^{\infty}) = 1$. For $k > 1$, pick an arbitrary belief μ such that $r_i(\theta_i^k, \mu) = \{a_i^k\}$ and $\mu(\Sigma_{-i}^{\infty}) = 1$. If $\text{marg}_{\Theta_{-i}} \mu^j \neq \text{marg}_{\Theta_{-i}} \mu$ for all $j = 1, ..., k-1$, then let $\mu^{k+1} = \mu$; otherwise, μ can be slightly perturbed so as to obtain a belief μ^{k+1} with all required properties. To see this, pick some (θ_{-i}, a_{-i}) such that $\mu(\theta_{-i}, a_{-i}) > 0$. By assumption there is another payoff type $\theta'_{-i} \neq \theta_{-i}$; furthermore, there is at least one a'_{-i} such that $(\theta'_{-i}, a'_{-i}) \in \Sigma_{-i}^{\infty}$ (see Remark 3.2). For any $\varepsilon \in (0, \mu(\theta_{-i}, a_{-i}))$, let μ_{ε} be the belief which coincide with μ except that $\mu_{\varepsilon}(\theta_{-i}, a_{-i}) = \mu(\theta_{-i}, a_{-i}) - \varepsilon$, $\mu_{\varepsilon}(\theta'_{-i}, a'_{-i}) = \mu(\theta'_{-i}, a'_{-i}) + \varepsilon$. Note that $\mu_{\varepsilon} \in \Delta(\Sigma_{-i}^{\infty})$ and $\text{marg}_{\Theta_{-i}} \mu \neq \text{marg}_{\Theta_{-i}} \mu_{\varepsilon}$. Since a_i is a strict best response to μ for type θ_i, there is some $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon})$, $a_i \in r_i(\theta_i, \mu_{\varepsilon})$. Clearly ε can be chosen so that $\text{marg}_{\Theta_{-i}} \mu^j \neq \text{marg}_{\Theta_{-i}} \mu_{\varepsilon}$ for all $j = 1, ..., k-1$.■

8.2 Proof of Proposition 4.8 (Self-Confirming and Bayesian Equilibrium)

(If) Fix a Bayesian equilibrium model

$$\mathcal{M} = \left\langle \Gamma, (E_i, T_i, p_i, b_i)_{i \in \{1,2\}} \right\rangle$$

consistent with the *feasible* distribution ζ, and an arbitrary Harsanyi type $\bar{t}_{-i} \in T_{-i}$. Let μ_i be the distributional strategy obtained from $p^{\bar{t}_{-i}}$:

$$\forall \theta_i, \forall s_i, \quad \mu_i(\theta_i, s_i) = \sum_{e_i : b_i(\theta_i, e_i) = s_i} p^{\bar{t}_{-i}}(\theta_i, e_i).$$

For any θ_i, pick $\overline{s}_i$ and $\overline{e}_i$ such that $\mu_i(\theta_i, \overline{s}_i) > 0$ and $b_i(\theta_i, \overline{e}_i) = \overline{s}_i$ (such $\overline{s}_i$ and $\overline{e}_i$ must exist because ζ assigns positive marginal probability to every payoff type and $\mathcal{M}$ is consistent with ζ). Let μ^{θ_i} be the distributional strategy of $-i$ obtained from $p^{(\theta_i, \overline{e}_i)}$:

$$\forall \theta_{-i}, \forall s_{-i}, \quad \mu^{\theta_i}(\theta_{-i}, s_{-i}) = \sum_{e_{-i}: b_{-i}(\theta_{-i}, e_{-i}) = s_{-i}} p^{(\theta_i, \overline{e}_i)}(\theta_{-i}, e_{-i})$$

We claim that the collection of probability measures thus constructed satisfies conditions (1)-(3) of Definition 4.7 and hence ζ is an SCE distribution.

First note that, by construction and consistency of $\mathcal{M}$ with ζ, distribution $\zeta \in \Delta(\Theta_1 \times \Theta_2 \times Z)$ is induced by any product measure of the form $(\mu_i \times \mu^{\theta_i}) \in \Delta((\Theta_i \times S_i) \times (\Theta_{-i} \times S_{-i}))$, $i = 1, 2$, $\theta_i \in \Theta_i$. This yields conditions (2) and (3). We must now show that (1) also holds. Notice that Definition 4.7 requires that, for each type θ_i, *every* strategy s_i such that $\mu_i(\theta_i, s_i) > 0$ be a best response to the belief $\mu^{\theta_i} \in \Delta(\Theta_{-i} \times S_{-i})$: the same belief about the opponent must rationalize every strategy that receives positive weight. But μ^{θ_i} is constructed starting with a *fixed* strategy $\bar{s}_i$ such that $\mu_i(\theta_i, \bar{s}_i) > 0$, so we must show that μ^{θ_i} also rationalizes any *arbitrary* s_i such that $\mu_i(\theta_i, s_i) > 0$.

To this end, recall that, by construction, $b_i(\theta_i, \overline{e}_i) = \overline{s}_i$. Also, consider an arbitrary $s_i \in S_i$ and recall that, by construction, $\mu_i(\theta_i, s_i) > 0$ implies that there is some e_i with $b_i(\theta_i, e_i) = s_i$. The Bayesian equilibrium conditions for Harsanyi types (θ_i, e_i) and $(\theta_i, \overline{e}_i)$ ensure that

$$\begin{aligned} & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, s_i, b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, e_i)}(\theta_{-i}, e_{-i}) \\ \geq & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, \overline{s}_i, b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, e_i)}(\theta_{-i}, e_{-i}), \end{aligned} \tag{3}$$

$$\begin{aligned} \forall s_i', & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, s_i', b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, \overline{e}_i)}(\theta_{-i}, e_{-i}) \\ \leq & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, \overline{s}_i, b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, \overline{e}_i)}(\theta_{-i}, e_{-i}). \end{aligned} \tag{4}$$

Since μ^{θ_i} is derived from $p^{(\theta_i, \overline{e}_i)}$ we have

$$\begin{aligned} \forall s_i', & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, s_i', b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, \overline{e}_i)}(\theta_{-i}, e_{-i}) \\ = & \sum_{\theta_{-i}, s_{-i}} U_i(\theta_i, \theta_{-i}, s_i', \theta_{-i}, s_{-i}) \mu_i^{\theta_i}(\theta_{-i}, s_{-i}) \end{aligned} \tag{5}$$

Claim: *Consistency of $\mathcal{M}$ with ζ implies that the expected utility from playing $\overline{s}_i$ (resp. s_i) is the same for Harsanyi types (θ_i, e_i) and $(\theta_i, \overline{e}_i)$:*

$$\begin{aligned} & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, \overline{s}_i, b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, e_i)}(\theta_{-i}, e_{-i}) \\ = & \sum_{\theta_{-i}, e_{-i}} U_i(\theta_i, \theta_{-i}, \overline{s}_i, b_{-i}(\theta_{-i}, e_{-i})) p^{(\theta_i, \overline{e}_i)}(\theta_{-i}, e_{-i}) \end{aligned} \tag{6}$$

$$\begin{aligned} & \sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,e_i)}(\theta_{-i},e_{-i}) \\ = & \sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,\overline{e}_i)}(\theta_{-i},e_{-i}) \end{aligned} \tag{7}$$

Proof of the claim. For every h such that $p_i^{(\theta_i,e_i)}(\{t_{-i} : b_{-i}(t_{-i}) \in S_{-i}(h)\}) > 0$ and action pair $a = (a_1, a_2) \in A(h)$ define

$$\Pr[a|h,\theta_{-i};\theta_i,e_i,s_i] = \begin{cases} 0, & \text{if } s_i(h) \neq a_i, \\ \frac{p_i^{(\theta_i,e_i)}(\{(\theta_{-i},e_{-i}):b_{-i}(\theta_{-i},e_{-i})\in S_{-i}(h,a_{-i})\})}{p_i^{(\theta_i,e_i)}(\{(\theta_{-i},e_{-i}):b_{-i}(\theta_{-i},e_{-i})\in S_{-i}(h)\})}, & \text{if } s_i(h) = a_i. \end{cases}$$

Then the probability of $(\theta_{-i}, a^1, ..., a^K)$ induced by strategy s_i given belief $p_i^{(\theta_i,e_i)}$ is

$$\begin{aligned} & \Pr[\theta_{-i}, a^1, ..., a^K|\theta_i, e_i, s_i] \\ = & \left(\sum_{e_{-i}} p^{(\theta_i,e_i)}(\theta_{-i},e_{-i})\right) \Pr[a^1|\phi,\theta_{-i};\theta_i,e_i,s_i] \prod_{k=2}^{K} \Pr[a^k|(a^1,...,a^{k-1})\theta_{-i};\theta_i,e_i,s_i]. \end{aligned}$$

By consistency of $\mathcal{M}$ with the feasible distribution ζ, we have

$$\left(\sum_{e_{-i}} p^{(\theta_i,e_i)}(\theta_{-i},e_{-i})\right) = \zeta(\theta_{-i})$$

$$\Pr[a|h,\theta_{-i};\theta_i,e_i,s_i] = \zeta(\theta_{-i},h,a_{-i})/\zeta(\theta_{-i},h), \text{ if } \zeta(\theta_{-i},h) > 0,$$

$$[s_i \in S_i(h) \text{ and } p_i^{(\theta_i,e_i)}(\{(\theta_{-i},e_{-i}) : b_{-i}(\theta_{-i},e_{-i}) \in S_{-i}(h)\}) > 0] \Rightarrow \zeta(\theta_{-i},h) > 0$$

[to obtain the latter implication, assume that the antecedent in brackets holds and note that (a) $\mu_i(\theta_i, s_i) > 0$, hence $\mu_i(\Theta_i \times S_i(h)) > 0$, (b) $\mu^{\theta_i}(\{\theta_{-i}\} \times S_{-i}(h)) = p_i^{(\theta_i,e_i)}(\{(\theta_{-i},e_{-i}) : b_{-i}(\theta_{-i},e_{-i}) \in S_{-i}(h)\}) > 0$, (c) $\zeta(\theta_{-i},h) = \mu_i(\Theta_i \times S_i(h)) \times \mu^{\theta_i}(\{\theta_{-i}\} \times S_{-i}(h))$]. Analogous equations hold for Harsanyi type $(\theta_i, \overline{e}_i)$. Therefore

$$\forall z \in Z, \Pr[\theta_{-i}, z|\theta_i, e_i, s_i] = \Pr[\theta_{-i}, z|\theta_i, \overline{e}_i, s_i].$$

Since

$$\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,e_i)}(\theta_{-i},e_{-i}) = \sum_{\theta_{-i},z} u_i(\theta_i,\theta_{-i},z) \Pr[\theta_{-i},z|\theta_i,e_i,s_i],$$

$$\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,\overline{e}_i)}(\theta_{-i},e_{-i}) = \sum_{\theta_{-i},z} u_i(\theta_i,\theta_{-i},z) \Pr[\theta_{-i},z|\theta_i,\overline{e}_i,s_i],$$

we obtain eq. (7). Eq. (6) follows from a similar argument. ■

The inequalities and equalities (3), (4), (5), (6) and (7) imply that s_i is a best reply for θ_i to μ^{θ_i} :

$$\begin{aligned}
&\sum_{\theta_{-i},s_{-i}} U_i(\theta_i,\theta_{-i},s_i,s_{-i})\mu^{\theta_i}(\theta_{-i},s_{-i}) \overset{(5)}{=}\\
&\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,\bar{e}_i)}(\theta_{-i},e_{-i}) \overset{(7)}{=}\\
&\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,e_i)}(\theta_{-i},e_{-i}) \overset{(3)}{\geq}\\
&\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},\bar{s}_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,e_i)}(\theta_{-i},e_{-i}) \overset{(6)}{=}\\
&\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},\bar{s}_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,\bar{e}_i)}(\theta_{-i},e_{-i}) \overset{(4)}{\geq}\\
&\sum_{\theta_{-i},e_{-i}} U_i(\theta_i,\theta_{-i},s'_i,b_{-i}(\theta_{-i},e_{-i}))p^{(\theta_i,\bar{e}_i)}(\theta_{-i},e_{-i}) \overset{(5)}{=}\\
&\sum_{\theta_{-i},s_{-i}} U_i(\theta_i,\theta_{-i},s'_i,s_{-i})\mu^{\theta_i}(\theta_{-i},s_{-i}).
\end{aligned}$$

(Only if) Let ζ be an SCE distribution and let $(\mu_i,(\mu^{\theta_i})_{\theta_i\in\Theta_i})_{i=1,2}$ satisfy conditions (1)-(3) of Definition 4.7. Define $\mathcal{M}$ as follows:

- $E_i = S_i$,
- $T_i = \bigcup_{\theta_{-i}\in\Theta_{-i}} \operatorname{supp}\mu^{\theta_{-i}}$,
- $\forall(\theta_i,s_i,\theta_{-i},s_{-i})\in T_i\times T_{-i}$, $b_i(\theta_i,s_i)=s_i$, $p^{(\theta_i,s_i)}(\theta_{-i},s_{-i})=\mu^{\theta_i}(\theta_{-i},s_{-i})$.

First note that the definition of T_i implies $\operatorname{proj}_{\Theta_i} T_i = \Theta_i$, because the distributional strategies $\mu^{\theta_{-i}}$ $(\theta_{-i}\in\Theta_{-i})$ agree with ζ and $\zeta(\theta_i)>0$ for every θ_i. Therefore $\mathcal{M}$ is indeed a Bayesian model of Γ. We claim that $\mathcal{M}$ is an equilibrium model consistent with ζ. Consistency of $\mathcal{M}$ with ζ is implied by conditions (2) and (3) of Definition 4.7. Now we prove that consistency of $\mathcal{M}$ with ζ and condition (1) imply that for every $(\theta_i,s_i)\in T_i$, s_i is a (possibly non-sequential) best response for θ_i given belief $p^{(\theta_i,s_i)}$.

Fix $(\theta_i,s_i)\in T_i$. Then $(\theta_i,s_i)\in\operatorname{supp}\mu^{\theta_{-i}}$ for some θ_{-i}. Let $\pi_i(\cdot|\theta_i,\cdot)\in[\Delta(A_i(h))]^{\mathcal{H}}$ be the behavioral strategy for θ_i derived from μ_i and μ^{θ_i} as follows:

$$\pi_i(a_i|\theta_i,h)=\begin{cases}\mu_i(\{\theta_i\}\times S_i(h,a_i))/\mu_i(\{\theta_i\}\times S_i(h)), & \text{if } \mu_i(\{\theta_i\}\times S_i(h))\mu_{-i}(S_{-i}(h))>0\\ 1/|A_i(h)|, & \text{otherwise.}\end{cases}$$

By construction, behavioral strategy $\pi_i(\cdot|\theta_i,\cdot)$ and the mixed strategy $\mu_i(\cdot|\theta_i)$ derived from μ_i yield the same expected utility for θ_i against belief μ^{θ_i}, as they both induce the distribution

$\zeta(\cdot|\theta_i) \in \Delta(\Theta_{-i} \times Z)$. By condition (1) $\mu_i(\cdot|\theta_i)$ is a mixed best response for θ_i to belief μ^{θ_i}. Therefore $\pi_i(\cdot|\theta_i, \cdot)$ is a locally randomized best response. This means that every pure strategy in the support of $\pi_i(\cdot|\theta_i, \cdot)$ is also a best response, where the support is defined as follows:[38]

$$\operatorname{supp} \pi_i(\cdot|\theta_i, \cdot) = \left\{ s_i' : \prod_{h \in \mathcal{H}} \pi_i(s_i'(h)|\theta_i, h) > 0 \right\}.$$

Since μ_i and μ^{θ_i} agree with ζ, $\mu_i(\{\theta_i\} \times S_i(h))\mu_{-i}(S_{-i}(h)) = \zeta(\theta_i, h)$. Since also $\mu^{\theta_{-i}}$ agrees with ζ, $\zeta(\theta_i, h) > 0$ implies $\zeta(\theta_i, h, s_i(h)) > 0$ (recall that $\mu^{\theta_{-i}}(\theta_i, s_i) > 0$). Therefore

$$\begin{aligned} \zeta(\theta_i, h) &> 0 \Rightarrow \pi_i(s_i(h)|\theta_i, h) = \frac{\zeta(\theta_i, h, s_i(h))}{\zeta(\theta_i, h)} > 0, \\ \zeta(\theta_i, h) &= 0 \Rightarrow \pi_i(s_i(h)|\theta_i, h) = 1/|A_i(h)| > 0, \end{aligned}$$

which means that $s_i \in \operatorname{supp} \pi_i(\cdot|\theta_i, \cdot)$ and s_i must be a best response to μ^{θ_i} for type θ_i. By construction, $\mu^{\theta_i} = p^{(\theta_i, s_i)}$. ■

8.3 ζ-Rationalizability and Iterated Intuitive Criterion

We collect here the proofs of results contained in Section 5 5 and of some ancillary statements. We start with the characterization of ζ-rationalizability in signaling games. Next we show that our definition of the IIC is equivalent to the original definition of Cho and Kreps (1987). Finally we provide a complete proof of the main proposition relating ζ-rationalizability to the IIC.

To simplify the notation we omit from proofs the reference to outcome distribution ζ, whenever this causes no misunderstanding. However, we keep reference to ζ in statements like definitions, remarks and claims.

8.3.1 Proof of Lemma 5.1 (ζ-rationalizability in Signaling Games)

The claim of the Lemma is trivially true for $k = 0$. Suppose it is true for some $k \geq 0$.

(i) Pick $(\theta, m) \in \Sigma_1^{k+1}$; then $(\theta, m) \in \Sigma_1^k$, so that by the induction hypothesis $\theta \in \Theta^k(m)$; moreover, there exists $\mu \in \Delta^1$ such that $\mu(S_2^k) = 1$ and $m \in r_1(\theta, \mu)$. Since Δ^1 is the set of beliefs about the receiver that agree with ζ, the behavioral representation π_2^μ of μ satisfies $\pi_2^\mu(a|m') = \zeta(a|m')$ for all a and m' such that $\zeta(m') > 0$: moreover, $\pi_2^\mu(A^k(m')|m') = \mu(\{s_2 : s_2(m') \in A^k(m')\}) = \mu(S_2^k) = 1$, where the second equality follows from the induction hypothesis. Finally, $m \in \arg\max_{m'} \sum_a u_1(\theta, m', a)\pi_2^\mu(a|m')$; thus, $\theta \in \Theta^{k+1}(m)$.

Conversely, pick $m \in M$ and $\theta \in \Theta^{k+1}(m)$; then $\theta \in \Theta^k(m)$ and the induction hypothesis implies that $(\theta, m) \in \Sigma_1^k$. Moreover, there exists a behavioral strategy $\pi_2 \in [\Delta(A)]^M$ whose mixed representation μ^{π_2} satisfies $\mu^{\pi_2} \in \Delta^1$ (by the first condition in the definition of $\Theta^{k+1}(m)$) and

$$\begin{aligned} \mu^{\pi_2}(S_2^k) &= \mu^{\pi_2}(\{s_2 : \forall m', s_2(m') \in A^k(m')\}) \\ &= \sum_{(a(m'))_{m' \in M} : \forall m', a(m') \in A^k(m')} \prod_{m'} \pi_2(a(m')|m') = 1, \end{aligned}$$

[38] This is the support of the mixed strategy derived from $\pi_i(\cdot, |\theta_i, \cdot)$, which may well be different from $\mu_i(\cdot|\theta_i)$.

where the first equality follows from the induction hypothesis and the second follows from the second condition in the definition of $\Theta^{k+1}(m)$. Finally, $m \in r_1(\theta, \mu^{\pi_2})$ because m is a best response to π_2 for θ (third condition in the definition of $\Theta^{k+1}(m)$). Thus, $(\theta, m) \in \Sigma_1^{k+1}$ as required.

(ii) Similarly, pick $s_2 \in S_2^{k+1}$; then $s_2 \in S_2^k$, so by the induction hypothesis $s_2(m) \in A^k(m)$ for all $m \in M$; moreover, for every $m \in M$, there exists $\mu \in \Delta^2 \subset \Delta^{\mathcal{H}}(\Theta \times M)$ such that $s_2 \in r_2(\mu)$ and $\Sigma_1^k \cap (\Theta \times \{m\}) \neq \emptyset \Rightarrow \mu(\Sigma_1^k|m) = 1$. But, again by the induction hypothesis, $\Sigma_1^k \cap (\Theta \times \{m\}) \neq \emptyset$ iff $\Theta^k(m) \neq \emptyset$. Hence, $\Theta^k(m) \neq \emptyset$ implies $\mu(\Theta^k(m) \times \{m\}|m) = 1$, and the belief $\nu^{\mu,m} \in \Delta(\Theta)$ defined by $\nu^{\mu,m}(\theta) = \mu((\theta, m)|m)$ satisfies the requirements in the definition of $A^{k+1}(m)$: thus, $s_2(m) \in A^{k+1}(m)$.

Conversely, suppose that $s_2(m) \in A^{k+1}(m)$ for all $m \in M$ (so in particular $A^{k+1}(m) \neq \emptyset$). By definition $A^{k+1}(m) \subseteq A^k(m)$; therefore $s_2(m) \in A^k(m)$ for all m. Thus, by the induction hypothesis, $s_2 \in S_2^k$. For each m, let ν^m be the belief satisfying the requirements in the definition of $s_2(m) \in A^{k+1}(m)$. Now define a CPS $\mu \in \Delta^{\mathcal{H}}(\Theta \times M)$ by $\mu((\theta, m)|\phi) = \zeta(\theta, m)$, $\mu((\theta, m)|m) = \zeta(\theta|m)$ for every $m \in M$ with $\zeta(m) > 0$, and $\mu((\theta, m)|m) = \nu^m(\theta)$ for every $m \in M$ with $\zeta(m) = 0$. Then $\mu \in \Delta^2$ and $s_2 \in r_2(\mu)$. Moreover, for every $m \in M$, $\Theta^k(m) \neq \emptyset \Leftrightarrow [\Sigma_1^k \cap (\Theta \times \{m\}) \Rightarrow \nu^m(\Theta^k(m)) = 1]$. By the induction hypothesis, $\nu^m(\Theta^k(m)) = 1$ implies $\mu(\Sigma_1^k|m) = 1$. Therefore $\mu(\Sigma_1^k|\phi) = 1$ and $\Sigma_1^k \cap (\Theta \times \{m\}) \Rightarrow \mu(\Sigma_1^k|m) = 1$ We conclude that $s_2 \in S_2^{k+1}$. ■

8.3.2 Definitions of the Iterated Intuitive Criterion

For the reader's convenience we include the definition of the Iterated Intuititve Criterion[39] due to Cho and Kreps (1987) and we also repeat here the alternative procedure used in the definition of the main text.

Definition 8.1 (Cho-Kreps) *Fix a (self-confirming) equilibrium distribution $\zeta \in \Delta(\Theta \times Z)$ and a message $m \in M$ such that $\zeta(m) = 0$. Let $\overline{A}^{-1}(m, \zeta) = A$, $\overline{\Theta}^0(m; \zeta) = \Theta$. For all $k = 0, 1, 2, ...$ define*

$$\overline{A}^{2k+1}(m;\zeta) = \begin{cases} BR_2(\overline{\Theta}^{2k}(m;\zeta), m), & \textit{if } \overline{\Theta}^{2k}(m;\zeta) \neq \emptyset \\ \overline{A}^{2k-1}(m;\zeta), & \textit{if } \overline{\Theta}^{2k}(m;\zeta) = \emptyset. \end{cases},$$

$$\overline{\Theta}^{2(k+1)}(m;\zeta) = \left\{ \theta \in \overline{\Theta}^{2k}(m;\zeta) : u_1^\zeta(\theta) \leq \max\nolimits_{a \in \overline{A}^{2k+1}(m;\zeta)} u_1(\theta, m, a) \right\}.$$

Distribution ζ satisfies the Iterated Intuitive Criterion *(IIC) 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 $a \in \bigcap_{k>0} \overline{A}^{2k+1}(m; \zeta)$ such that $u_1(\theta, m, a) \leq u_1^\zeta(\theta)$.*

Definition 8.2 (Alternative procedure) *Let $IA^0(m; \zeta) = A$ and $I\Theta^0(m; \zeta) = \Theta$ and, for all $n = 0, 1, 2, ...$ define*

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

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

[39] Cf. Fudenberg and Tirole (1991, Definition 11.5). We number superscripts in a different way: in odd steps ($n = 2k + 1$) we eliminate actions of the Receivers, in even steps ($n = 2k$) we eliminate payoff types of the Sender.

The following is easily shown by induction:

Remark 8.3 *Fix a SCE outcome distribution ζ and message m with $\zeta(m) = 0$.*
(a) $\overline{\Theta}^{2(k+1)}(m;\zeta) \subseteq \overline{\Theta}^{2k}(m;\zeta)$, $\overline{A}^{2k+1}(m;\zeta) \subseteq \overline{A}^{2k-1}(m;\zeta)$, $I\Theta^{k+1}(m;\zeta) \subseteq I\Theta^{k}(m;\zeta)$, $IA^{k+1}(m;\zeta) \subseteq IA^{k}(m;\zeta)$.
(b) If the Receiver has no conditionally dominated action, that is, $BR_2(\Theta, m) = A$, then $\overline{\Theta}^{2k}(m;\zeta) = I\Theta^{2k}(m;\zeta)$ and $\overline{A}^{2k+1}(m;\zeta) = IA^{2k+1}(m,\zeta)$ for all $k = 0, 1, 2, ...$.

The next remark shows that the two procedures defined above are closely related also in signalling games with conditionally dominated actions.

Remark 8.4 *Fix a SCE outcome distribution ζ and message m with $\zeta(m) = 0$. For all $k = 0, 1, 2, ...$, $\overline{\Theta}^{2k}(m;\zeta) = I\Theta^{2k}(m;\zeta)$, and if $\overline{\Theta}^{2k}(m;\zeta) \neq \emptyset$ then $\overline{A}^{2k+1}(m,\zeta) = IA^{2k+1}(m;\zeta)$.*

Proof of Remark 8.4. The statement is trivially true for $k = 0$. Suppose it is true for a given k. We have to consider two cases.

(i) If $\overline{\Theta}^{2k}(m) \neq \emptyset$, then by definition and the inductive hypothesis

$$\overline{A}^{2k+1}(m) = BR_2(I\Theta^{2k}(m), m) = IA^{2k+1}(m);$$

hence

$$\overline{\Theta}^{2k+2}(m) = \left\{\theta \in I\Theta^{2k}(m) : u_1^{\zeta}(\theta) \leq \max_{a \in IA^{2k+1}(m)} u_1(\theta, m, a)\right\}.$$

Since $IA^{2k+1}(m) \subseteq IA^{2k}(m)$, we have

$$\max_{a \in IA^{2k+1}(m;\zeta)} u_1(\theta, m, a) \leq \max_{a \in IA^{2k}(m)} u_1(\theta, m, a).$$

The equality and inequality above yield $\overline{\Theta}^{2k+2}(m) \subseteq I\Theta^{2k+1}(m)$. By inspection of the definition of $I\Theta^{2k+2}(m)$, we obtain $\overline{\Theta}^{2k+2}(m) = I\Theta^{2k+2}(m)$.

(ii) If $\overline{\Theta}^{2k}(m) = \emptyset$, then by definition and the inductive hypothesis

$$I\Theta^{2k+2}(m) = I\Theta^{2k+1}(m) = \emptyset = \overline{\Theta}^{2k+2}(m).$$

This concludes the proof of the remark. ■

Proposition 8.5 *An (SCE) outcome distribution ζ satisfies the IIC according to Definition 8.1 if and only if it satisfies the IIC according to Definition 5.2*

Proof. First note that, since the defined sequences of subsets are weakly decreasing, by finiteness there exists some K such that $\overline{A}^{2K+1}(m) = \bigcap_{k>0} \overline{A}^{2k+1}(m)$, $IA^{2K+1}(m) = \bigcap_{n>0} IA^{n}(m)$.

Suppose that ζ does *not* satisfy Definition 8.1 (Cho-Kreps). Then there is some (θ, m) such that $\zeta(m) = 0$ and

$$u_1^\zeta(\theta) < \min_{a \in \overline{A}^{2K+1}(m)} u_1(\theta, m, a).$$

Taking into account that $\overline{A}^{2K+1}(m) \subseteq \overline{A}^{2k+1}(m)$ for all k, the inequality above implies

$$\forall k \geq 0,\ u_1^\zeta(\theta) \leq \max_{a \in \overline{A}^{2k+1}(m)} u_1(\theta, m, a).$$

which in turn implies $\theta \in \overline{\Theta}^{2K}(m)$. [This is proved by induction: $\theta \in \overline{\Theta}^0(m)$; suppose that $\theta \in \overline{\Theta}^{2k}(m)$, then $u_1^\zeta(\theta) \leq \max_{a \in \overline{A}^{2k+1}(m)} u_1(\theta, m, a)$ implies that $\theta \in \overline{\Theta}^{2k+2}(m)$.] Thus, by Remark 8.4, $\overline{A}^{2K+1}(m) = IA^{2K+1}(m)$ and $u_1^\zeta(\theta) < \min_{a \in IA^{2K+1}(m)} u_1(\theta, m, a)$, that is, there is no action $a \in \bigcap_{n>0} IA^n(m)$ such that $u_1(\theta, m, a) \leq u_1^\zeta(\theta)$. Hnece ζ does not satisfy our Definition 5.2.

The converse is proved by a similar argument. If ζ does not satisfy our modified criterion, there is some (θ, m) such that $\zeta(m) = 0$ and

$$u_1^\zeta(\theta) < \min_{a \in IA^{2K+1}(m)} u_1(\theta, m, a).$$

The above inequality implies $\theta \in I\Theta^{2k}(m)$. By Remark 8.4 we obtain $IA^{2K+1}(m) = \overline{A}^{2K+1}(m)$ and $u_1(\theta) < \min_{a \in \overline{A}^{2K+1}(m)} u_1(\theta, m, a)$. Therefore ζ does not satisfy the IIC. ■

8.3.3 Proof of Proposition 5.3

We prove the proposition through a sequence of claims. The first one is a corollary of Propositions 4.3 and 4.8:

Claim 1. *If* $S^\infty_{2,\zeta} \neq \emptyset$ *and* $\mathrm{proj}_\Theta \Sigma^\infty_{1,\zeta} = \Theta$, *then* ζ *is an SCE distribution.*

Claim 2. *If* $S^\infty_{2,\zeta} \neq \emptyset$*, then* $\forall \theta, \forall m^*$, $\zeta(\theta, m^*) > 0 \Rightarrow \theta \in \bigcap_k \Theta^k(m^*; \zeta)$.

Proof of Claim 2. Suppose that $\zeta(\theta, m^*) > 0$ and pick any $s_2 \in S_2^\infty$. This strategy is a sequential best reply to some CPS μ^2 such that $\mu^2(\Sigma^\infty_{1,\zeta}|\phi) = 1$ and $\mu^2(\theta, m^*|\phi) = \zeta(\theta, m^*) > 0$. Therefore $(\theta, m^*) \in \Sigma_1^\infty$. By Lemma 5.1, this implies $\theta \in \bigcap_k \Theta^k(m^*)$. ■

Claim 3. *Suppose that* $\forall m^*$, $\zeta(\theta, m^*) > 0 \Rightarrow \theta \in \bigcap_k \Theta^k(m^*; \zeta)$. *Then, for every message* m *with* $\zeta(m) = 0$,

$$\forall k \geq 0,\ \Theta^k(m; \zeta) = I\Theta^k(m; \zeta) \text{ and } A^k(m; \zeta) = IA^k(m; \zeta). \tag{8}$$

Proof of Claim 3. Eq. (8) holds trivially for $k = 0$. Suppose it holds for some k.

Pick $\theta \in \Theta^{k+1}(m)$. From the definition of $\Theta^{k+1}(m)$, we must have $\sum_a u_1(\theta, m, a)\pi_2(a|m) \geq u_1^\zeta(\theta)$ and $1 = \pi_2(A^k(m)|m)$, for some $\pi_2 \in [\Delta(A)]^M$; hence, there must be an action $a \in A^k(m) = IA^k(m)$ such that $u_1(\theta, m, a) \geq u_1^\zeta(\theta)$, and therefore $\theta \in I\Theta^{k+1}(m)$.

Conversely, pick $\theta \in I\Theta^{k+1}(m)$. Then there exists $a^m \in IA^k(m) = A^k(m)$ such that $u_1(\theta, m, a^m) \geq u_1^\zeta(\theta)$. Fix any message m^* such that $\zeta(\theta, m^*) > 0$ (such a message exists because $\zeta(\theta) > 0$). By assumption, $\theta \in \Theta^{k+1}(m^*)$. Therefore there is some π_2^* that agrees with ζ and for every message m' satisfies

$$\pi_2^*(A^k(m')|m') = 1 \text{ and } u_1^\zeta(\theta) = \sum_a u_1(\theta, m^*, a)\pi_2^*(a|m^*) \geq \sum_a u_1(\theta, m', a)\pi_2^*(a|m').$$

Define π_2 as follows: (i) $\pi_2(a^m|m) = 1$, (ii) $\forall m' \neq m$, $\pi_2(\cdot|m') = \pi_2^*(\cdot|m')$. By construction, π_2 satisfies the three requirements for $\theta \in \Theta^{k+1}(m)$. In particular, m is a best response to π_2 for type θ because

$$\forall m' \neq m,\ u_1(\theta, m, a^m) \geq u_1^\zeta(\theta) \geq \sum_a u_1(\theta, m', a)\pi_2(a|m'). \tag{9}$$

To show that $A^{k+1}(m) = IA^{k+1}(m)$ we consider two cases. (1) If $\Theta^k(m) \neq \emptyset$, then

$$A^{k+1}(m) = BR(\Theta^k(m), m) = BR(I\Theta^k(m), m).$$

(2) If $\Theta^k(m) = \emptyset$, then $A^{k+1}(m) = A^k(m)$, and $I\Theta^k(m) = \emptyset$ (inductive hypothesis) which implies $IA^{k+1}(m) = IA^k(m)$. Since $A^k(m) = IA^k(m)$ (inductive hypothesis), we obtain $A^{k+1}(m) = IA^{k+1}(m)$. ■

Claim 4. *If ζ is an SCE distribution satisfying the IIC then, for every message m with $\zeta(m) = 0$, eq. (8) holds, and furthermore*

$$\forall k, \forall m',\ \zeta(m') > 0 \Rightarrow A^k(m', \zeta) = \arg\max_a \sum_\theta u_2(\theta, m', a)\zeta(\theta|m) \neq \emptyset. \tag{10}$$

Proof of Claim 4. Let ζ be an SCE distribution satisfying the IIC. By definition,

$$\forall m', \forall a',\ \zeta(m', a') > 0 \Rightarrow a' \in \arg\max_a \sum_\theta u_2(\theta, m', a)\zeta(\theta|m); \tag{11}$$

furthermore, for every message m' off the equilibrium path and every type θ there is an action $a(\theta, m') \in \bigcap_k IA^k(m')$ such that $u_1^\zeta(\theta) \geq u_1(\theta, m', a(\theta, m'))$. Fix m with $\zeta(m) = 0$ and assume, by way of induction, that eq. (8) holds for some k.

As in the proof of Claim 3, it can be shown that $\Theta^{k+1}(m) \subseteq I\Theta^{k+1}(m)$. Conversely, let $\theta \in I\Theta^{k+1}(m)$. Then there exists $a^m \in IA^k(m) = A^k(m)$ such that $u_1(\theta, m, a^m) \geq u_1^\zeta(\theta)$. Define π_2 as follows: (i) $\pi_2(a^m|m) = 1$, $\forall m' \neq m$, $\zeta(m') = 0 \Rightarrow \pi_2(a(\theta, m')|m') = 1$, (ii) $\zeta(m') > 0 \Rightarrow \forall a, \pi_2(a|m') = \zeta(a|m')$. Conjecture π_2 agrees with ζ and message m is a best response to π_2 for type θ because eq. (9) in the proof of Claim 3 holds. For every m' with $\zeta(m') = 0$, $\pi_2(A^k(m')|m') = \pi_2(IA^k(m')|m') = 1$. For every m' with $\zeta(m') > 0$, eq. (11) and the inductive hypothesis imply $\pi_2(A^k(m')|m') = 1$. Therefore, π_2 satisfies all the conditions for $\theta \in \Theta^{k+1}(m)$.

The proof of $A^{k+1}(m) = IA^{k+1}(m)$ is the same as for Claim 3. We only have to show that eq. (10) holds for $k+1$; in particular, it is sufficient to show that $\theta \in \Theta^k(m')$ for every for every pair (θ, m') with $\zeta(\theta, m') > 0$. Pick such a pair (θ, m') (hence, $m' \neq m$) and define a corresponding conjecture π_2' as follows: (i) $\pi_2'(a(\theta, m)|m) = 1$, (ii) $\pi_2'(\cdot|m'') = \pi_2(\cdot|m'')$ for $m'' \neq m$. By construction, π_2' agrees with ζ and m' is a best reply to π_2' for type θ. As for conjecture

π_2, eq. (11) and the inductive hypothesis imply that $\pi_2'(A^k(m'')|m'') = 1$ for all m''. Thus $\theta \in \Theta^{k+1}(m') \subseteq \Theta^k(m')$. ■

The proof of the proposition follows quite easily from these claims. Suppose that $S_2^\infty \neq \emptyset$ and $\mathrm{proj}_\Theta \Sigma_1^\infty = \Theta$. Then ζ is an SCE distribution (Claim 1), $\forall\theta, \forall m^*$, $\zeta(\theta, m^*) > 0 \Rightarrow \theta \in \bigcap_k \Theta^k(m^*)$ (Claim 2) and eq. (8) holds for every m with $\zeta(m) = 0$ (Claim 3). By finiteness, there is some K such that $\Theta^K(m') = \bigcap_k \Theta^k(m')$ and $A^K(m') = \bigcap_k A^k(m')$ for all m'. For any θ, there is some m^* with $\zeta(\theta, m^*) > 0$. Then $\theta \in \Theta^K(m^*)$ and m^* is best reply for θ to a conjecture π_2^* that agrees with ζ and satisfies $\pi_2^*(\bigcap_k IA^k(m')|m') = \pi_2^*(\bigcap_k A^k(m')|m') = 1$ for all m'. This implies that for every m with $\zeta(m) = 0$ there is an action $a \in \bigcap_k IA^k(m)$ such that $u_1(\theta, m, a) \leq u_1^\zeta(\theta)$. Hence ζ satisfies the IIC.

Now suppose that ζ is an SCE distribution satisfying the IIC. Then eq. (11) holds (Claim 4) and $\bigcap_k A^k(m') \neq \emptyset$ for all m' with $\zeta(m') > 0$. Furthermore, $\bigcap_k A^k(m) \neq \emptyset$ for all m with $\zeta(m) = 0$. Thus, Lemma 5.1 implies $S_2^\infty \neq \emptyset$, which in turn implies $\mathrm{proj}_\Theta \Sigma_1^\infty = \Theta$ (Remark 3.4). ■

References

[1] AUMANN, R.J. (1987): “Correlated Equilibrium as an Expression of Bayesian Rationality,” *Econometrica,* **55**, 1-18.

[2] AUMANN, R.J. (1998): “Reply to Gul,” *Econometrica,* **66**, 929–938.

[3] BATTIGALLI, P. (1987): *Comportamento razionale ed equilibrio nei giochi e nelle situazioni sociali*, undergraduate dissertation, Università Bocconi, Milano.

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

[5] BATTIGALLI, P. (1999): “Rationalizability in Incomplete Information Games,” EUI working paper 99/15.

[6] BATTIGALLI, P. and D. GUAITOLI (1997): “Conjectural Equilibria and Rationalizability in a Game with Incomplete Information,” in (P.Battigalli, A. Montesano and F.Panunzi, Eds.) *Decisions, Games and Markets.* Dordrecht: Kluwer Academic Publishers.

[7] BATTIGALLI, P. and M. SINISCALCHI (1999): “Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games,” *Journal of Economic Theory*, **88**, 188-230.

[8] BATTIGALLI, P. and M. SINISCALCHI (2002): “Strong Belief and Forward Induction Reasoning,”*Journal of Economic Theory*, **106**, 356-391.

[9] BATTIGALLI, P. and M. SINISCALCHI (2003): “Rationalizable Bidding in First Price Auctions,” *Games and Economic Behavior*, in print.

[10] BATTIGALLI, P. and J. WATSON (1997): “On ‘Reputation’ Refinements with Heterogeneous Beliefs,” *Econometrica*, **65**, 369-374.

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

[12] BERGEMANN, D. and S. MORRIS (2002): "Robust Mechanism Design,"

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

[14] BONANNO, G. and K. NEHRING (1999): "How to Make Sense of the Common Prior Assumption Under Incomplete Information," *International Journal of Game Theory,* **28**, 409-434.

[15] BRANDENBURGER, A. and E. DEKEL (1987): "Rationalizability and Correlated Equilibria," *Econometrica,* **55,** 1391-1402.

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

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

[18] DEKEL, E., D. FUDENBERG and D. LEVINE (1999): "Payoff Information and Self-Confirming Equilibrium," *Journal of Economic Theory,* **89**, 165-185.

[19] DEKEL, E., D. FUDENBERG and D. LEVINE (2003): "Learning to Play Bayesian Games,"*Games and Economic Behavior*, in print.

[20] 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.

[21] DEKEL, E. and A. WOLINSKY (2003): "Rationalizable Outcomes of Large Independent Private-Value First-Price Discrete Auctions," *Games and Economic Behavior*, in print.

[22] FEINBERG, Y. (2000): "Characterizing Common Priors in the Form of Posteriors," *Journal of Economic Theory*, **91**, 127-179.

[23] FORGES, F. (1993): "Five Legitimate Definitions of Correlated Equilibrium in Games with Incomplete Information," *Theory and Decision*, **35**, 277-310.

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

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

[26] GUL, F. (1998): "A Comment on Aumann's Bayesian View," *Econometrica,* **66**, 923–928.

[27] 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.

[28] KAGEL, J. (1995): "Auctions: A Survey of Experimental Research," in (J. Kagel and A. Roth Eds) *The Handbook of Experimental Economics.* Princeton: University Press.

[29] KALAI, E. and E. LEHRER (1993): "Subjective Equilibrium in Repeated Games," *Econometrica*, **61**, 11231-1240.

[30] KALAI, E. and E. LEHRER (1995): "Subjective Games and Equilibria," *Games and Economic Behavior*, **8**, 123-163.

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

[32] 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.

[33] MILGROM, P. and R. WEBER (1985): "Distributional Strategies for Games with Incomplete Information," *Mathematics of Operation Research,* **10**, 619-632.

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

[35] MORRIS, S. (1995): "The Common Prior Assumption in Economic Theory," *Economics and Philosophy*, **11**, 227-253.

[36] MORRIS, S. and C. SKIADAS (2000) "Rationalizable Trade," *Games and Economic Behavior*, **31**, 311-323.

[37] OKUNO-FUJIWARA, M., A. POSTELWAITE, and K. SUZUMURA (1990): "Strategic Information Revelation," *Review of Economic Studies,* **57**, 25-47.

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

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

[40] PERRY, M. and P. RENY (1999): "A General Solution to King Solomon's Dilemma," *Games and Economic Behavior,* **26**, 279-285.

[41] RABIN M. (1994): "Incorporating Behavioral Assumptions into Game Theory," in J. Friedman (ed.) *Problems of Coordination in Economic Activity.* Dortrecht: Kluwer Academic Publishers.

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

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

[44] RUBINSTEIN, A. and A.WOLINSKY (1994): "Rationalizable Conjectural Equilibrium: Between Nash and Rationalizability," *Games and Economic Behavior,* **6**, 299-311.

[45] SAMET, D. (1998a): "Iterated Expectations and Common Priors," *Games and Economic Behavior*, **24**, 131-141.

[46] SAMET, D. (1998b): "Common Priors and Separation of Convex Sets," *Games and Economic Behavior*, **24**, 172-174.

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

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

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

[50] WATSON, J. (1993): "A 'Reputation' Refinement without Equilibrium," *Econometrica,* **61**, 199-205.