

Algorithmic Characterization of Rationalizability in Extensive Form Games*

Oliver Board†

ojboard@pitt.edu

July 2005

Abstract

We construct a dynamic epistemic model for extensive form games, which generates a hierarchy of beliefs for each player over her opponents' strategies and beliefs, and tells us how those beliefs will be revised as the game proceeds. We use the model to analyze the implications of the assumption that the players possess common (true) belief in rationality, thus extending the concept of rationalizability to extensive form games.

1 Introduction

This paper seeks to examine the implications of common belief in rationality in extensive form games. It was once thought that the notorious backward induction argument provided a precise characterization of these implications, at least in games of perfect information. The implausibility of the backward induction outcome in games such as the repeated prisoner's dilemma was attributed to the strength of the assumptions made. And once the assumption of common belief (or knowledge) is relaxed even a little, Kreps *et al.* [18] showed that cooperation until the final rounds can become a rational response. But later work questioned the very validity of the backward induction argument. Binmore [8], Pettit and Sugden [22] and Reny [23] were among the first to take this line, and argued that, even if there is common belief in rationality at the beginning of an extensive form game, there may not be at each of the subsequent information sets. Indeed, backward induction typically implies

*This paper is a much revised version of Board [11]. Helpful comments from Michael Bacharach, Paolo Battigalli, Adam Brandenburger, Amanda Friedenberg, Matthias Hild and Bob Stalnaker are gratefully acknowledged.

†Department of Economics, University of Pittsburgh, Pittsburgh, PA 15232.

that certain information sets will not be reached. If the backward induction argument is correct, these information sets are therefore not consistent with common belief in rationality. But the argument assumes that there is common belief in rationality at *every* information set in the game. The lesson to be learned from this resolution of the backward induction paradox is that analysis of rational play in extensive form games requires careful consideration not just of the players' beliefs at the beginning of the game, but also of how these beliefs change as the game progresses.

Unlike the backward induction argument, however, most solution concepts in game theory make no explicit reference to players' rationality or beliefs. Nash equilibrium, for instance, is defined purely in terms of conditions on the player's strategy sets. A notable exception is the notion of rationalizability, developed by Bernheim [7] and Pearce [21]. A strategy is said to be rationalizable if it is consistent with common belief in rationality. The idea that game theoretic solution concepts could be characterized epistemically (that is, by a set of restrictions on the players' beliefs and behavior) was developed further by Aumann [2], who showed that rational players with a common prior over the space of uncertainty will play according to a correlated equilibrium distribution; and that every correlated equilibrium distribution is consistent with rationality of the players and the common prior assumption. Aumann used his *information partition* model (Aumann [1]) to provide a precise description of each player's beliefs about the game (i.e. about which strategies would be played), and about each other. Although Bernheim and Pearce did not employ any such formal model of interactive epistemology, the results of their analysis of strategic form games were later proved by Tan and Werlang [28] and Stalnaker [24] in the context of such a model.

But the information partition model of Aumann and the alternative, hierarchical, model of interactive epistemology used by Tan and Werlang (see e.g. Mertens and Zamir [19] and Brandenburger and Dekel [9]) are static: they tell us what each player believes about her opponents' beliefs, but they cannot tell us what she will believe at future information sets, or what she believes her opponents' will believe. Hence they are not rich enough to analyze rational play in extensive form games. Dynamic models have been developed to serve precisely this purpose, most notably by Battigalli and Siniscalchi [5] and Stalnaker [25] (see also Board [12]). In this paper we use such a model to give a precise characterization of rationalizability in extensive form games.

Section 2 gives a brief discussion of related literature. In section 3 we develop the formal model of beliefs in extensive form games, and in section 4 we use that model to give a precise characterization of the implications of common belief in rationality, in terms of an iterated deletion algorithm. Section 5 concludes.

2 Related literature

Most closely related to the current project is the work of Stalnaker [25] and [26], who uses a model of belief revision which is a special case of that presented here. It is shown in Board [12] that Stalnaker’s model makes rather strong introspection assumptions which we do not require¹. Furthermore, Stalnaker analyzes only strategic form games. On the other hand, he uses the belief revision component to examine several alternative notions of rationality in addition to the basic notion we consider. These stronger notions pick up elements of extensive form reasoning even in the strategic form of the game.

Battigalli and Siniscalchi [6] use a very different kind of model, built up from infinite hierarchies of conditional probability systems. A *conditional probability system* describes a player’s beliefs at each stage of an extensive form game, and each level of the hierarchy describes the player’s beliefs about every level beneath it. Thus their hierarchical structures provide an explicit model of beliefs and beliefs about beliefs throughout the game tree. They use the structures to investigate the implications of common belief, and of a stronger concept, common *strong* belief, in rationality. Unlike our paper, they consider incomplete information games, where the players may be uncertain of each other’s payoffs. But they restrict their attention to games with *observable actions*, where at each stage everyone observes the actions of the previous stage. This assumption is for the sake of tractability, and not imposed by any limitations of their model.

Brandenburger and Keisler [10] use a similar model to Battigalli and Siniscalchi, with *lexicographic probability systems* playing the role of conditional probability systems. A lexicographic probability system is a (finite) sequence of probability measures. Like Stalnaker, they focus on the strategic form of the game, and derive epistemic conditions for iterated deletion of weakly dominated strategies. But their results also shed light on the extensive form procedures of backward and forward induction.

Feinberg [15] and [16] develops a rich language which can be employed to describe what he calls ‘subjective’ reasoning in extensive form games, and also to describe the structure of the game itself, including payoffs. An system of axioms is used to prove theorems in the language, and semantic structures provide truth conditions. A player is represented by a different hypothetical identity at every information set at which she is on move. Belief is a property of these identities, and only implicitly of players. And beliefs of a player’s future identities are not derived from those of her

¹We conjecture that they are not required for Stalnaker’s results either.

past identities: there is no belief revision component to the logic. Feinberg uses his framework to analyze backward and forward induction, as well as provide epistemic characterizations of Nash equilibrium and sequential equilibrium and to introduce a new concept, the *reasonable solution* of a game.

For a more detailed discussion of some of these papers and comprehensive surveys of many earlier results, see Dekel and Gul [14] and Battigalli and Bonanno [4].

3 Beliefs in extensive form games

The analysis of this paper is restricted to finite extensive form games of complete information and perfect recall. The description of such a game specifies the following five elements (see also Osborne and Rubinstein [20]):

- a finite set N of players.
- a finite set H of sequences, which satisfies: (i) $\emptyset \in H$; and (ii) if $(a^k)_{k=1,\dots,K} \in H$ and $L < K$, then $(a^k)_{k=1,\dots,L} \in H$. Each $h \in H$ is a *history*, and each component of h is an *action* taken by a player. The set of histories defines the game tree, with each element h representing a node of the tree, the node that is reached if that history is played. A history $(a^k)_{k=1,\dots,K} \in H$ is *terminal* if there is no a^{K+1} such that $(a^k)_{k=1,\dots,K+1} \in H$. The set of actions available after the nonterminal history h is denoted $A(h) = \{a : (h, a) \in H\}$, and the set of terminal histories is denoted Z .
- a function $\iota : H \setminus Z \rightarrow N$ that assigns to each nonterminal history the player whose turn to move it is.
- a partition \mathcal{I} of $H \setminus Z$ that divides all the nonterminal histories into *information sets*. The cell $\mathcal{I}(h)$ of \mathcal{I} that contains h identifies the nonterminal histories that the player on move cannot distinguish from h based on the information available to her at h . It is required that for every history in a given cell of the partition, the same player is on move and the same actions are available, i.e. if $h' \in \mathcal{I}(h)$, then $\iota(h) = \iota(h')$ and $A(h) = A(h')$. This is implied by the fact that each player knows when it is her turn to move, and what actions are available to her. Thus for any information set $I \in \mathcal{I}$ we can write $\iota(I)$ for the player on move, and A_I for the actions available to her, and we can partition \mathcal{I} into sets $\mathcal{I}_i = \iota^{-1}(i)$.

To characterize perfect recall, let $X_i(h)$ denote player i 's *experience* at a given history h .

$X_i(h)$ is the sequence of information sets that player i encounters in the history h and the actions she takes at them, in the order that these events occur. For each player i , if $h, h' \in I$ for some $I \in \mathcal{I}_i$, then $X_i(h) = X_i(h')$.

- a utility function $U_i : Z \rightarrow \mathbb{R}$ for each player i , which assigns an expected utility value to each terminal history.

The collection $\langle N, H, \iota, \mathcal{I}, (U_i)_{i \in N} \rangle$ defines an extensive form game, Γ .

It will be convenient to use the following additional notation. Let $A_{\mathcal{I}} = \times_{I \in \mathcal{I}} A_I$ be the set of *action profiles*, which specify an action $a_I \in A_I$ for every information set $I \in \mathcal{I}$, and let A_{-I} be the set of action profiles at every information set other than I (so that $A_I \times A_{-I} = A_{\mathcal{I}}$). For a given action profile $a_{\mathcal{I}} \in A_{\mathcal{I}}$, let $h(a_{\mathcal{I}})$ be the history *induced by* $a_{\mathcal{I}}$, i.e. $h(a_{\mathcal{I}})$ is a sequence of actions of the form $(a_{\mathcal{I}}(\mathcal{I}(\emptyset)), a_{\mathcal{I}}(\mathcal{I}(a_{\mathcal{I}}(\emptyset))), \dots)$ such that $h(a_{\mathcal{I}}) \in Z$. We can write $u_i(a_{\mathcal{I}}) = U_i(h(a_{\mathcal{I}}))$, where u_i is player i 's strategic form utility function. Finally, for a given information set I , let $A_{\mathcal{I}}(I)$ be the set of action profiles *consistent with* I , i.e. $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ if there is some sequence of actions $(a_{\mathcal{I}}(\mathcal{I}(\emptyset)), a_{\mathcal{I}}(\mathcal{I}(a_{\mathcal{I}}(\emptyset))), \dots) \in I$.

As we discussed in the introduction, in order to analyze rational play in extensive form games, it is crucial to have a precise model not only of the players' beliefs but also of the way these beliefs are revised as the game proceeds. Traditional theories of belief revision, such as Bayes' rule, have concentrated on modeling how beliefs change when new information is learned that is compatible with one's existing beliefs. But such a focus is too narrow for our purposes: in order to model counterfactual reasoning in games, we will need to know how beliefs change or would change in the event of surprises, when information is learned that contradicts what is currently believed. In this case, some of these existing beliefs must be given up, and the problem is that there is a multitude of ways to select just how this should be done.

Board [12] develops a multi-agent logic of belief revision; the language of that logic can be used to describe players' beliefs in extensive form games. We start with a set of primitive formulas, $\Phi = \{a_I \mid a_I \in A_I \text{ for some } I \in \mathcal{I}\}$. The primitive formulas describe which actions are taken at each information set, so that a_I denotes the sentence "action a_I is chosen at information set I "². The language \mathcal{L} is the smallest set of formulas such that:

²Or "action a_I *was* / *will be* / *would have been* / *would be* chosen at information set I ". There is no notion of time in our logic, so sentences should be interpreted as past, present or future, indicative or subjunctive depending on the viewpoint.

- (a) if $\phi \in \Phi$, then $\phi \in \mathcal{L}$;
- (b) if $\phi, \psi \in \mathcal{L}$, then $\neg\phi \in \mathcal{L}$ and $\phi \wedge \psi \in \mathcal{L}$;
- (c) if $\phi, \psi \in \mathcal{L}$, then $B_i\phi \in \mathcal{L}$, $C\phi \in \mathcal{L}$ and $B_i^\phi\psi \in \mathcal{L}$, for $i \in N$.

With slight abuse of notation, we shall use $a_{\mathcal{I}}$ to denote the sentence “action profile $a_{\mathcal{I}}$ is chosen”, and I to denote the sentence “information set I is reached”. Formally, $a_{\mathcal{I}}$ and I are abbreviations for longer sentences containing only primitive formulas, negations and conjunctions. B_i represents player i ’s beliefs before the start of the game, and B_i^ϕ her beliefs after she learns that ϕ is the case. Finally, C is the common (prior) belief operator.

Truth conditions are assigned to the formulas of \mathcal{L} by means of a *model*. A model M for an extensive form game Γ is a tuple $\langle W, f, \preceq \rangle$ where

- W is a non-empty set of *possibles worlds*;
- $f : W \rightarrow A_{\mathcal{I}}$ is an *action function*;
- \preceq is a vector of *plausibility orderings*, one for each player at every world.

Models work in the same way as the belief revision structures used in Board [12]. The action function plays the role of the interpretation in a belief revision structure, and specifies, for each world, which action will (or would) be taken at every information set in the game. We shall use $f_I(w)$ to denote the action taken at information set I in world w . The structure of the game implies that $f_I(w) \in A_I$, for all I, w . Note that none of the facts about the structure of the game are included in the model. The implication is that all these facts are true at every world in the model (and hence are common belief among the players). This corresponds to the assumption that the game is one of complete information.

\preceq_i^w denotes the plausibility ordering of player i at world w , and encodes her beliefs and her belief revision policy. $x \preceq_i^w y$ means that from the point of view of player i at world w , world x is at least as plausible as world y . Intuitively, the player considers possible only the worlds which are most plausible according to her ordering: we call these worlds *accessible*; the remainder of the ordering is used to construct her revised beliefs, as we shall see. We impose two constraints on the form of the \preceq_i^w relations. Let $W_i^w = \{x \mid x \preceq_i^w y \text{ for some } y\}$; W_i^w is the set of worlds which are *conceivable* to i at world w , though not necessarily accessible. Then, we assume that:

R1 for all i, w : \preceq_i^w is complete and transitive on W_i^w ;

R2 for all i, w : \preceq_i^w is well-founded.

R1 ensures that each plausibility ordering divides all the worlds into ordered equivalence classes; the inconceivable worlds, i.e. those not in W_i^w , are a class unto themselves and are to be considered least plausible. If \preceq_i^w is well-founded (**R2**), then there are no infinitely descending sequences of the form $\dots w_n \prec_i^w \prec w_{n-1} \prec_i^w \dots \prec_i^w w_0$ (where $x \prec_i^w y$ if and only if $x \preceq_i^w y$ and not $y \preceq_i^w x$). This guarantees that for every nonempty set $X \subseteq W_i^w$, $\min_i^w(X \cap W_i^w) \neq \emptyset$, where \min_i^w is defined in the obvious way (i.e. $\min_i^w(X) = \{x \in X \mid \text{for all } y \in X, x \preceq_i^w y\}$); intuitively, it says that if there are any conceivable worlds in a certain set, then there is a most plausible world in that set. Well-foundedness is satisfied automatically in the case where W is finite. Henceforth we shall assume that all models satisfy **R1** and **R2**.

The model of the game allows us to assign truth conditions to every formula in the language. Let $[\phi]$ denote the set of worlds at which ϕ is true. Truth is assigned to primitive formulas as follows: $[a_I] = \{w \mid f_I(w) = a\}$. Negations and conjunctions are dealt with in the obvious way: $[\neg\phi] = [W \setminus \phi]$ and $[\phi \wedge \psi] = [\phi] \cap [\psi]$. $B_i\phi$ is true precisely if ϕ is true at every world w accessible to i before she learns anything: $[B_i\phi] = \{w \mid \min_i^w(W_i^w) \subseteq [\phi]\}$; and $B_i^\phi\psi$ is true precisely if ψ is true at every world accessible to her after she learns that ϕ : $[B_i^\phi\psi] = \{w \mid \min_i^w([\phi] \cap W_i^w) \subseteq [\psi]\}$. Finally, to define the truth conditions for $C\phi$, let $E\phi$ abbreviate $\bigwedge_{i \in N} B_i\phi$, let $E^0\phi$ abbreviate ϕ , and let $E^k\phi$ abbreviate $EE^{k-1}\phi$ for $k = 1, 2, \dots$. Then $[C\phi] = \bigcap_{k=1,2,\dots} [E^k\phi]$.

There is, however, a problem with this account of belief revision: the method just described calculates each player's beliefs at a given information set by revising her original beliefs (as represented by \preceq_i^w) with the information that the information set has been reached. But a given history may pass through several information sets of the player, and beliefs should be revised at each information set. There is in general no guarantee that the beliefs generated by a sequence of such revisions will be the same as the beliefs generated by revising just once. But in games of perfect recall this may be a reasonable assumption to make. In such games, the information received by a given player as the game progresses has a particular property: each new piece of information implies all of the previous pieces. If a history passes through more than one information set of a given player, these information sets can be strictly ordered in terms of precedence, and the set of histories consistent with a given information set is always a subset of those consistent with every previous information set. And if ψ logically implies ϕ , it may be reasonable to assume that learning ϕ and then ψ will generate the same beliefs as if one learns ψ at first: in both cases the same information is learned. This simplifying assumption saves us the trouble of dealing with iterated belief

revisions. Whether the single-revision process is appropriate for modeling beliefs at information sets in games of imperfect recall is an open question and beyond the scope of this paper.

The results of Board [12] give us a precise understanding of the formal language \mathcal{L} : we can provide an axiomatic characterization of the formulas which are true at every world of every model of a particular game. Theorem 5 of Board [12] states that the axiom system BRS^C is sound and complete with respect to the class of all belief revision structures which satisfy **R1** and **R2**, i.e. a formula is true at every world of every such belief revision structure if and only if it is provable in BRS^C . But the models described above are more restrictive than belief revision structures: unlike the interpretation of a belief revision structure, the action function used here to tell us which actions are played at each information set cannot assign arbitrary truth values to primitive formulas. One and only one action must be chosen at each information set, so that if $w \in [a_I]$ it must be the case that $w \in [-a'_I]$ for all $a'_I \neq a_I$. To provide a syntactic counterpart of this semantic restriction we add the axiom **Game**, which tells us which combinations actions are consistent with the rules of the game. For example, for a game with only four possible action profiles, $a_I^1, a_I^2, a_I^3, a_I^4$, **Game** would be $a_I^1 \vee a_I^2 \vee a_I^3 \vee a_I^4$. $BRS^C + \mathbf{Game}$ is sound and complete with respect to the class of all models satisfying **R1** and **R2**.

4 Rationalizability

To characterize rationality in extensive form games, we must compute the players' beliefs at each information set at which they are on move. The information they learn as the game progresses is given by the information structure of the game, as specified by the information sets \mathcal{I} . Specifically, at information set $I \in \mathcal{I}_i$, player i learns that she must be at one of the histories in I , i.e. that one of the action profiles in $A_{\mathcal{I}}(I)$ has been chosen.

But to make sense of the definition of rationality given below, we must also make sure that each player has true belief at a given information set about what action she is choosing at that information set (see Board [13] for a more detailed discussion of this point). There are two ways of doing this: the first is to add an additional constraint to the models: for all i , if $I \in \mathcal{I}_i$ then $\min_i^w([I] \cap W_i^w) \subseteq [f_I(w)]$. This constraint says that at every world player i considers possible when she learns that information set I has been reached, her action at that information set is the same as it is in the actual world. The syntactic counterpart is the axiom schema $a_I \Rightarrow B_i^I a_I$. A problem with this approach is that it imposes restrictions not only on the beliefs at information set

I , but also at beliefs prior to that. To see why, suppose that a player is moving at two successive information sets, I_1 and I_2 , and that she chooses action a_{I_1} and a_{I_2} respectively, with a_{I_1} leading to I_2 . At I_1 she is assumed to believe (correctly) that she is choosing a_{I_1} . So she learns nothing when I_2 is reached, and hence her beliefs do not change. But at the second information set she assumed to believe (again correctly) that she is choosing a_{I_2} . It follows that she must have already believed this at I_1 ! More generally, the implication is that players must have true beliefs about their actions at every future information set compatible with their current beliefs. To put it another way, they are not allowed to change their minds unless they are surprised. Of course, this may be a reasonable assumption to make in many (or even most) circumstances, but it is not good modeling practice to hide such an assumption in the formalism.

For this reason, we adopt the second approach, and assume that the a player learns what action she will choose at a given information set when that information set is reached. Of course we are not suggesting that the player is told what to do, but rather that she does not necessarily know what she is going to do until required to make the choice. The B_i^ϕ operators represent the player's beliefs after deliberation³, when the player has figured out what she will do, but we do not want encode the outcome of this deliberation process into the prior beliefs. According to this second approach, player i 's beliefs at any information set $I \in \mathcal{I}_i$ at which she is on move are therefore given by $B_i^{I \wedge a_I}$, where a_I is the action she chooses at I . In terms of the model of the game, i learns that the true world must lie in the set $[I] \cap [a_I]$. The set of worlds accessible to her at world w after receiving this information is obtained by taking the \preceq_i^w -minimal worlds in $[I] \cap [f_I(w)] \cap W_i^w$.

To define rationality in the standard way, as expected utility maximization, we must first explain how each agent's probabilistic beliefs are derived. Given a (prior) probability measure p_i on the set of worlds W , define the conditional probability measure $p_{i,I}^w$ as follows: for any $E \subseteq W$,

$$p_{i,I}^w(E) = \frac{p_i(E \cap \min_i^w([I] \cap [f_I(w)] \cap W_i^w))}{p_i(\min_i^w([I] \cap [f_I(w)] \cap W_i^w))}.$$

$p_{i,I}^w$ is obtained from p_i by conditioning player i 's information, since $\min_i^w([I] \cap [f_I(w)] \cap W_i^w)$ is the set of worlds which agent i considers possible in world w at information set I . The probability which player i assigns to any formula $\phi \in \mathcal{L}$ in world w at information set $I \in \mathcal{I}_i$ is given by $p_{i,I}^w([\phi])$. Of course, this expression may not be well defined, since there is nothing to guarantee that the denominator is greater than zero. To avoid technical issues that are not relevant for the

³See Aumann [2] (p. 8) for a detailed discussion of this point.

ongoing discussion, we shall simply assume that in such a case the player is not rational. In effect, we are claiming that rational players should not rule out any information sets *a priori* (though they may certainly do so once the game is in progress).

An action profile is rationalizable if it is consistent with common knowledge of rationality among the players. We build up the definition of rationalizability in several stages. First, an action is defined as rational if it maximizes the expected utility of the player who takes that action.

Definition 1 *Suppose $I \in \mathcal{I}_i$. $a_I \in A_I$ is rational with respect to p_i at world w if $p_{i,I}^w$ is well defined and*

$$\sum_{a_{-I} \in A_{-I}} p_{i,I}^w([a_{-I}]) \cdot u_i(a_I, a_{-I}) \geq \sum_{a_{-I} \in A_{-I}} p_{i,I}^w([a_{-I}]) \cdot u_i(a'_I, a_{-I})$$

for all $a'_I \in A_I$.

There is an important difference between this notion of rationality at an information set and the concept employed elsewhere in the literature. It is usually assumed that strategies⁴ rather than actions are the objects of choice, and hence the objects of rationality. A strategy is said to be rational at a given information set if it yields the highest expected utility of all those strategies which are consistent with that information set's being reached. But it is unclear when if ever players will actually make a choice between the various strategies available to them. Although we could think of a hypothetical pre-play stage when such choices are made, it seems more appropriate and more accurate to think of the players as making their choices as and when they are on move. Indeed, this is the approach that majority of the work in this area takes⁵. And at each information set a player chooses only part of her strategy, the part which specifies what she does at that information set. It is these choices that should be assessed as rational. For assessing the entire strategy at a particular information set carries with it the substantive assumption that the player on move has control over her choices at all future information sets. To see why, suppose we say that a particular strategy choice (rather than just the action choice) is rational at some information set. Presumably we mean that, among the strategies that are consistent with that information set's being reached, the strategy chosen maximizes expected utility⁶. All of these strategies specify what actions will

⁴Or sometimes *plans of action*, which specify actions only at nodes not ruled out by the player's previous actions. See e.g. Reny [23].

⁵A notable exception is the work of Stalnaker: he discusses this issue in [27] (p. 315), and shows that, under certain assumptions, the two approaches are equivalent.

⁶See e.g. Gul [17] "... rational players choose strategies s_i such that s_i is optimal at [an information set] against some conjecture that reaches [that information set] whenever s_i reaches [that information set]" (p. 15). The majority of papers in the Bayesian tradition adopt a similar definition of rationality.

be taken at future information sets as well as at the current information set. If the player cannot control what she does at these information sets while on move at the current information set, then she cannot choose among these strategies. This assumption of self control does not follow from rationality alone; rationality alone does not even imply that a player *knows* what she will do at future information sets!⁷ The example in Figure 2 below may shed further light on this issue.

Next, we define what it is for a player to be rational. It is not immediately clear how to do this. In particular, we can distinguish *reached-node rationality*, where a player is rational if each action she actually plays is rational; *own-node rationality*, where a player is rational if each of her actions at nodes not ruled out by her previous behavior is rational; and *all-node rationality*, where a player is rational only if her actions are rational at every information set at which she is on move. Reached-node rationality does not seem strong enough, especially if we are thinking about what it means for one player to know that another is rational. Suppose for instance that the second player does not get a chance to move because of an action taken by the first. This makes the second player (vacuously) reached-node rational. And yet intuitively we would expect the first player to be able to make inferences about what the second would do if given the chance to move. For a similar reason own-node rationality will not suffice either. If a player believes herself to be rational, this ought to impose restrictions on what she believes she will or would do at future nodes. But own-node rationality says nothing about her behavior at those nodes which are ruled out by her past actions⁸. Thus in what follows we shall adopt the concept of all-node rationality, and say that a player is rational if she chooses actions that are rational at every information set at which she is on move:

Definition 2 *Player i is rational at world w if there is some p_i such that, for all $I \in \mathcal{I}_i$, $f_I(w)$ is rational with respect to p_i at world w .*

Let Rat_i denote the sentence “player i is rational”; Rat_i is true at world w precisely if player i is rational at world w . Note that we are not introducing any new formulas into the language: whether or not a player is rational is determined completely by her choice of actions and her first-order beliefs, i.e. her beliefs about which actions will be chosen. Thus Rat_i is simply an abbreviation for

⁷Nor even at past information sets: in games of imperfect recall, players can forget their previous action choices.

⁸The importance of the distinction between own-node rationality and all-node rationality arises only because we take actions as the fundamental objects of choice. If players are assumed to choose between the strategies (or plans of action) available to them at a given node, it is specified what they will do at all relevant future nodes. In this case the two concepts yield path-equivalent results.

a long formula of the language. Let $Rat = \bigwedge_{i \in N} Rat_i$, and $CTBR = C(Rat) \wedge Rat$ ($CTBR$ stands for “there is common true belief in rationality”).

A action profile is said to be rationalizable if it is consistent with common true belief in rationality. Formally,

Definition 3 *For any game Γ , an action profile $a_{\mathcal{I}} \in A_{\mathcal{I}}$ is rationalizable if there is some model of M with a world $w \in [CTBR]$ such that $f(w) = a_{\mathcal{I}}$.*

The following theorem provides a characterization of the set of rationalizable strategies. First we define, for each information set $I \in \mathcal{I}$ a sequence of action sets D_I^1, D_I^2, \dots , where

$$D_I^1 = \{a_I \in A_I \mid \text{there is no } \alpha'_I \in \Delta A_I \text{ such that } u_{i(I)}(\alpha'_I, a_{-I}) > u_{i(I)}(a_I, a_{-I}) \text{ for all } a_{-I} \in A_{-I}(I)\}$$

$$D_I^{m+1} = \{a_I \in A_I \mid \text{there is no } \alpha'_I \in \Delta A_I \text{ such that } u_{i(I)}(\alpha'_I, a_{-I}) > u_{i(I)}(a_I, a_{-I}) \text{ for all } a_{-I} \in D_{-I}^m\}$$

for $m = 1, 2, \dots$ (where ΔA_I is the set of probability measures on A_I , and the definition of u_i is extended in the usual way). D_I is the limit of this sequence: $D_I = \bigcap_{m=1}^{\infty} D_I^m$, and $D_{\mathcal{I}}$ is the set of corresponding action profiles. $D_{\mathcal{I}}$ is the set of actions profiles which survive a certain iterated elimination procedure, the generalization of iterated deletion of strictly dominated strategies to extensive form games: every action that is strictly dominated at any information set for the player on move at that information set is deleted in the first round, and the standard procedure is applied to what is left of the whole game.

Theorem 1 *For any game Γ , a action profile $a_{\mathcal{I}}$ is rationalizable if and only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$.*

The proof of Theorem 1 is given in the appendix, but the intuition behind the result is straightforward. It follows from the definition of rationality that no rational player will play an action that is strictly dominated at any information set at which she is on move. This accounts for the first round of deletion. But we cannot apply iterated deletion at any of these information sets: unless all the players believed with positive probability at the start of the game that a particular information set *would be* reached, there may no longer be common belief in rationality if that information set *is* reached. And it is common belief that drives iterated deletion. Nevertheless, there is common belief at the start of the game, so we can apply iterated deletion to the game as a whole.

Two examples will illustrate the strength and the weakness of the deletion procedure. Figure 1 shows the familiar entry deterrence game, with an entrant (E) first deciding whether to enter the

market (**i**) or stay out (**o**), and then an incumbent (*I*) deciding whether fight (**f**) or acquiesce (**a**) if entry occurs.

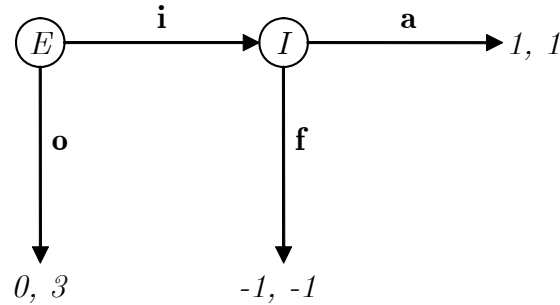


Figure 1: Entry deterrence

Let 1 and 2 denote the two information sets. Neither of the actions available to *E* is strictly dominated at information set 1, so $D_1^1 = \{\mathbf{o}, \mathbf{i}\}$. But **f** is strictly dominated by **a** for *I* at information set 2, so $D_2^1 = \{\mathbf{a}\}$. Now, given that only actions in D_2^1 are chosen, **o** is strictly dominated by **i** for *E* at information set 1. No more actions can be deleted, so $D_{\mathcal{I}} = \{(\mathbf{i}, \mathbf{a})\}$. This simple example shows how the information structure of the game is used to eliminate actions which would survive if the iterated deletion procedure were applied to the strategic form of the game. Furthermore, in this game, rationalizability is stronger than Nash equilibrium.

Figure 2 depicts a single-person decision problem⁹. Self (*S*) wants to coordinate her actions to maximize her payoff.

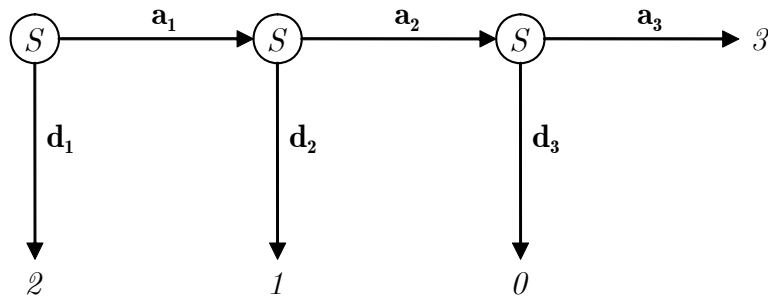


Figure 2: Self coordination

Label the information sets 1, 2 and 3 in order. On the first round, \mathbf{d}_3 is dominated by \mathbf{a}_3 at information set 3, but no other action is dominated at the information set at which it is chosen. But no more actions can be deleted on the second round: \mathbf{d}_2 survives because it does just as well as \mathbf{a}_2 against $(\mathbf{d}_1, \mathbf{a}_3)$. Thus $D_{\mathcal{I}} = \{(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3), (\mathbf{d}_1, \mathbf{a}_2, \mathbf{a}_3), (\mathbf{a}_1, \mathbf{d}_2, \mathbf{a}_3), (\mathbf{d}_1, \mathbf{d}_2, \mathbf{a}_3)\}$. This seemingly

⁹This is a one-player version of Figure 2 in Stalnaker [25].

paradoxical result arises because we do not assume that players can commit to their action choices at future nodes. Consider action profile $(\mathbf{d}_1, \mathbf{d}_2, \mathbf{a}_3)$. How can this be consistent with common true belief in rationality? Suppose that if information set 2 were reached, S would no longer believe herself to be rational, but rather that she would play \mathbf{d}_3 if information set 3 were reached. Then the rational thing to do at information set 2 is to play \mathbf{d}_2 ; and if she believes that she is rational and would have these beliefs at information set 2, the rational thing to do at information set 1 is to play \mathbf{d}_1 . It follows that if information set 2 is reached, S is certainly right to doubt her own rationality: according to the beliefs just described she has just chosen an irrational action.

This example does not rely on lack of introspection: at information set 1, S has no doubt about what her choices will be throughout the game, and these beliefs are correct. If she plays \mathbf{a}_1 she surprises herself and her beliefs must be revised. Rather the issue is one of self control: at a given information set, S can control her action only at that information set, and not at future information sets as well.

5 Conclusions

This paper uses the framework of Board [12] to construct models which describe players' beliefs in extensive form games. Their beliefs about the game and about each other are expressed at the beginning the game and at every information set. These models are used to analyze rational play, and Theorem 1 describes the implications of common (true) belief in rationality.

We believe that our approach has two key strengths. The first is transparency: although the models we use to prove Theorem 1 are based around the rather obscure notion of a possible world, they can be used to provide truth conditions for a formal language of belief revision which has a straightforward interpretation. Furthermore, the properties of this language can be clarified by means of an axiom system: formulas of the language that are true at every world of every model are precisely those that are provable in the axiom system. Thus Theorem 1 can be translated into the formal language. The “only if” part of that theorem tells us that the formula $CTBR \Rightarrow D_{\mathcal{I}}$ is true at every world of every model. It follows that it is provable in the axiom system $BRS^C + \mathbf{Game}$: we have a set of precise conditions which are sufficient to derive our result. The “if” part of the theorem tells us that, for any action profile $a_{\mathcal{I}} \in D_{\mathcal{I}}$, there is some world of some model at which $a_{\mathcal{I}} \wedge CTBR$ is true; thus $a_{\mathcal{I}} \wedge CTBR$ is logically consistent according to $BRS^C + \mathbf{Game}$.

The second strength is flexibility. The axiom system used is minimal in the sense that it

imposes a weak set of conditions on the beliefs and belief revision policies of rational players (at least, in relation to most of the related literature). But extra axioms can be added, along with the corresponding restrictions on the models so that the tight link between truth and provability is retained, and their effects can be examined. For instance we could examine whether the strong introspection assumption (that players are fully aware of all their beliefs, past present and future) implicitly adopted by Stalnaker [25] affects our result. It would also be interesting to analyze the impact of Battigalli’s [3] *best rationalization principle*¹⁰. According to this principle, players should believe each other to be rational as long as those beliefs are consistent with the observed pattern of behavior; subject to that constraint, they should believe that they believe they are all rational as long as it is consistent to do so, and so on. It is not possible to represent this principle in an arbitrary model of a given game: we need to ensure that there *enough* worlds in the model so that if an action is consistent with iterated belief in rationality of a certain depth, then there is a world in the model being used in which that action is played and there is iterated belief in rationality of a certain depth. The *canonical structure* described in Board [12] contains a world for *every* logically consistent set of beliefs, so it provides an ideal tool for investigating the best rationalization principle. We conjecture that adding the principle to the assumption of common belief in rationality would allow us to carry out iterated deletion of actions at each information set, rather than just one round of deletion followed by iterated deletion in the strategic form. In perfect information games, this would generate the backward induction procedure. Finally, we could consider Stalnaker’s [25] notion of *perfect* rationality, according to which every action profile of one’s opponents is taken into account in the expected utility calculation. We conjecture that common belief in perfect rationality plus the best rationalization principle would give an epistemic characterization of iterated deletion of weakly dominated strategies.

References

- [1] AUMANN, R. J. (1976), “Agreeing to disagree”, *Annals of Statistics* **4**, 1236–1239.
- [2] AUMANN, R. J. (1987), “Correlated Equilibrium as an Expression of Bayesian Rationality”, *Econometrica* **55**, 1–18.

¹⁰This would not be a new project: much of the analysis of Battigalli and Siniscalchi [6] is based around this principle.

- [3] BATTIGALLI, P. (1996), “Strategic Rationality Orderings and the Best Rationalization Principle”, *Games and Economic Behavior* **13**, 178–200.
- [4] BATTIGALLI, P. AND G. BONANNO (1998), “Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory,” *Research in Economics* **53**, 149–225.
- [5] BATTIGALLI, P. AND M. SINISCALCHI (1999), “Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games”, *Journal of Economic Theory* **88**, 188–230.
- [6] BATTIGALLI, P. AND M. SINISCALCHI (2001): “Strong Belief and Forward Induction Reasoning”, forthcoming, *Journal of Economic Theory*.
- [7] BERNHEIM, B. D. (1984), “Rationalizable Strategic Behavior”, *Econometrica* **52**, 1007–1028.
- [8] BINMORE, K. (1987), “Modelling Rational Players I”, *Economics and Philosophy* **3**, 179–214.
- [9] BRANDENBURGER, A. AND E. DEKEL (1993), “Hierarchies of Beliefs and Common Knowledge”, *Journal of Economic Theory* **59**, 189–198.
- [10] BRANDENBURGER, A. AND H. J. KEISLER (2002), “Epistemic Conditions for Iterated Admissibility”, mimeo, Harvard Business School.
- [11] BOARD, O. J. (1998), “Belief Revision and Rationalizability”, TARK VII, Conference Proceedings, ed. by I. Gilboa.
- [12] BOARD, O. J. (2002), “Dynamic Interactive Epistemology”, mimeo, Department of Economics, University of Oxford.
- [13] BOARD, O. J. (2002), “The Equivalence of Bayes and Causal Rationality in Games”, mimeo, Department of Economics, University of Oxford.
- [14] DEKEL, E. AND F. GUL (1997), “Rationality and Knowledge in Game Theory”, in *Advances in Economics and Econometrics: Theory and Applications: Seventh World Congress, Vol. 1*, ed. by D. M. Kreps and K. W. Wallis. Cambridge University Press, 87–172.
- [15] FEINBERG, Y. (2001), “Epistemic Characterizations of Equilibria and the Reasonable Solution”, mimeo, Stanford Graduate School of Business.
- [16] FEINBERG, Y. (2002), “Subjective Reasoning in Dynamic Games”, mimeo, Stanford Graduate School of Business, Stanford.

- [17] GUL, F. (1996): “Rationality and Coherent Theories of Strategic Behavior”, *Journal of Economic Theory* **70**, 1–31.
- [18] KREPS, D., P. MILGROM, J. ROBERTS, AND R. WILSON (1982), “Rational Cooperation in the Finitely Repeated Prisoners’ Dilemma”, *Journal of Economic Theory* **27**, 245–252.
- [19] MERTENS, J. F. AND S. ZAMIR, (1985), “Formalization of Harsanyi’s notion of ‘type’ and ‘consistency’ in games with incomplete information”, *International Journal of Game Theory* **14**, 1–29.
- [20] OSBORNE, M. J. AND A. RUBINSTEIN (1994), *A Course in Game Theory*. The MIT Press, Cambridge, MA.
- [21] PEARCE, D. G. (1984): “Rationalizable Strategic Behavior and the Problem of Perfection”, *Econometrica* **52**, 1029–1050.
- [22] PETTIT, P. AND R. SUGDEN, (1989), “The Backward Induction Paradox”, *Journal of Philosophy* **86**, 169–182.
- [23] RENY, P. (1992), “Rationality in Extensive Form Games”, *Journal of Economic Perspectives* **6**, 103–118.
- [24] STALNAKER, R. (1994), “On the Evaluation of Solution Concepts”, *Theory and Decision* **37**, 49–73.
- [25] STALNAKER, R. (1996), “Knowledge, Belief and Counterfactual Reasoning in Games”, *Economics and Philosophy* **12**, 133–163.
- [26] STALNAKER, R. (1998), “Belief Revision in Games: Forward and Backward Induction”, *Mathematical Social Sciences* **36**, 31–56.
- [27] STALNAKER, R. (1999), “Extensive and Strategic Form Games: Games and Models for Games”, *Research in Economics* **53**, 293–319.
- [28] TAN, T, AND S. R. C. WERLANG (1988), “The Bayesian Foundations of Solution Concepts of Games”, *Journal of Economic Theory* **45**, 370–391.

A Proof of Theorem 1

First we recall the following lemma (see e.g. Pearce [21]).

Lemma 1 *An action of a player in a finite strategic form game is a best response if and only if it is not strictly dominated.*

We are now in a position to prove the main theorem.

(if) To prove the “if” statement, we must construct, for arbitrary $a_{\mathcal{I}} \in D_{\mathcal{I}}$, a model of Γ in which there a world $w \in [CTBR]$ such that $f(w) = a_{\mathcal{I}}$. Let $W = A_{\mathcal{I}}^{11}$, and for all $a_{\mathcal{I}}$, let $f(a_{\mathcal{I}}) = a_{\mathcal{I}}$. We show how to construct the plausibility orderings of each player at an arbitrary world $a_{\mathcal{I}}^* \in D$. We do this by constructing a function $k : A_{\mathcal{I}} \rightarrow \mathbb{N}$, which assigns each world a numerical ranking according to plausibility.

First we order each information set $I \in \mathcal{I}_i$: if player i is moving from the n th time at information set I , let $order(I) = n$. Given the assumption of perfect recall, this function is well defined. Now take any I such that $order(I) = 1$. We can think of the actions in D_I as player i 's strategy set in a strategic form game, and the action profiles in D_{-I} as the strategy profiles of her opponent. i 's payoffs are given by $u_i(a_I, a_{-I})$, and her opponent's payoffs are chosen arbitrarily. Since $a_I^* \in D_I$, it is not strictly dominated in this game, and therefore by Lemma 1 there is some probability measure μ_0 over D_{-I} such that a_I^* is a best response to μ_0 . Extend the domain of μ_0 to the whole of $A_{\mathcal{I}}$ in the following way: $\mu_0(a_I, a_{-I}) = \mu_0(a_{-I})$ if $a_I = a_I^*$ and $a_{-I} \in D_{-I}$; $\mu_0(a_I, a_{-I}) = 0$ otherwise. Notice that $\mu_0(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$. Next, construct a probability measure μ_1 over $A_{\mathcal{I}}$ in three steps:

1. if $a_{\mathcal{I}} \notin A_{\mathcal{I}}(I)$ for any $I \in \mathcal{I}_i$ of order 1, let $\mu_1(a_{\mathcal{I}}) = \mu_0(a_{\mathcal{I}})$;
2. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 1, but $\mu_0(A_{\mathcal{I}}(I)) = 0$, let $\mu_1(a_{\mathcal{I}}) = 0$.
3. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 1 and $\mu_0(A_{\mathcal{I}}(I)) > 0$, consider the conditional probability $\mu_0(a_{-I} | A_{\mathcal{I}}(I)) = \frac{\mu_0(a_{-I} \cap A_{\mathcal{I}}(I))}{\mu_0(A_{\mathcal{I}}(I))}$. There is some $a'_I \in A_I$ which is best response to $\mu_0(\cdot | A_{\mathcal{I}}(I))$ (there may be more than one). It must be the case that $a'_I \in D_I$ by Lemma 1, since $\mu_0(\cdot | A_{\mathcal{I}}(I))$ places positive weight only on $a_{-I} \in D_{-I}$. $\mu_1(a_{\mathcal{I}})$ is defined as follows: (i) $\mu_1(a_{\mathcal{I}}) = \mu_0(a_{-I})$ if $a_I = a'_I$ (where a_I is the I th component of $a_{\mathcal{I}}$ and a_{-I}

¹¹Note that we are now using action labels for three purposes: to denote actions themselves, to denote formulas of \mathcal{L} describing which actions are chosen, and to denote worlds. Since we do not use the language \mathcal{L} in this proof, there should be no risk of confusion.

is the $-I$ th component of $a_{\mathcal{I}}$); (ii) $\mu_1(a_{\mathcal{I}}) = 0$ otherwise. Observe that $\mu_1(a_{-I} | A_{\mathcal{I}}(I)) = \mu_0(a_{-I} | A_{\mathcal{I}}(I))$, since $\mu_1(a_{-I}) = \mu_1(a'_I, a_{-I}) = \mu_0(a_{-I})$ for all $a_{-I} \in A_{-I}(I)$, and the a_{-I} 's in $A_{-I}(I)$ partition $A_{\mathcal{I}}(I)$. So a'_I is a best response to $\mu_1(\cdot | A_{\mathcal{I}}(I))$.

This process is well defined since the $A_{\mathcal{I}}(I)$ sets are disjoint. Notice that $\mu_1(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$.

Now construct a probability measure μ_2 , again in three steps:

1. if $a_{\mathcal{I}} \notin A_{\mathcal{I}}(I)$ for any $I \in \mathcal{I}_i$ of order 2, let $\mu_2(a_{\mathcal{I}}) = \mu_1(a_{\mathcal{I}})$;
2. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 2, but $\mu_1(A_{\mathcal{I}}(I)) = 0$, let $\mu_2(a_{\mathcal{I}}) = 0$.
3. if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 2 and $\mu_1(A_{\mathcal{I}}(I)) > 0$, consider the conditional probability $\mu_1(a_{-I} | A_{\mathcal{I}}(I))$. By the same reasoning as before, there is some $a'_I \in D_I$ which is best response to $\mu_1(\cdot | A_{\mathcal{I}}(I))$. Let $\mu_2(a_{\mathcal{I}})$ be defined as follows: (i) $\mu_2(a_{\mathcal{I}}) = \mu_1(a_{-I})$ if $a_I = a'_I$; (ii) $\mu_2(a_{\mathcal{I}}) = 0$ otherwise. Again by the same reasoning as before, we know that a'_I is a best response to $\mu_2(\cdot | A_{\mathcal{I}}(I))$.

Notice that $\mu_2(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$. We have shown that if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 2 and $\mu_2(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu_2(\cdot | A_{\mathcal{I}}(I))$. We want to show also that if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order 1 and $\mu_2(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu_2(\cdot | A_{\mathcal{I}}(I))$. We know that a_I is a best response to $\mu_1(\cdot | A_{\mathcal{I}}(I))$, i.e.

$$\sum_{a_{-I}} \mu_1(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a_I, a_{-I}) \geq \sum_{a_{-I}} \mu_1(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I.$$

Now consider the information sets $I', I'', \dots \in \mathcal{I}_i$ immediately following I , and corresponding subsets of $A_{-I}(I'), A_{-I}(I''), \dots$ of A_{-I} . For every a_{-I} not in one of these subsets, $\mu_1(a_{-I}) = \mu_2(a_{-I})$ (by step 1) and therefore $\mu_1(a_{-I} | A_{\mathcal{I}}(I)) = \mu_2(a_{-I} | A_{\mathcal{I}}(I))$. Next consider every $a_{-I} \in A_{-I}(I')$. If $\mu_1(A_{-I'}(I)) = 0$, then $\mu_1(a_{-I}) = \mu_2(a_{-I})$ (by step 2) and therefore $\mu_1(a_{-I} | A_{\mathcal{I}}(I)) = \mu_2(a_{-I} | A_{\mathcal{I}}(I))$ again. So suppose $\mu_1(A_{-I'}(I)) > 0$. μ_1 and μ_2 generate the same beliefs about actions at every information set except I' , but μ_2 assumes that the action chosen at I' is a best response to those beliefs, while according to μ_1 it can be chosen arbitrarily (step 3). Thus, restricting attention to $a_{-I} \in A_{-I}(I')$, we have:

$$\sum_{a_{-I'} \in A_{-I'}(I')} \mu_2(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a_I, a_{-I}) \geq \sum_{a_{-I'} \in A_{-I'}(I')} \mu_1(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I.$$

On the other hand, if action $a'_I \neq a_I$ is chosen at information set I , information set I' is not reached (given perfect recall) and we have:

$$\sum_{a_{-I'} \in A_{-I'}(I')} \mu_2(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a'_I, a_{-I}) = \sum_{a_{-I'} \in A_{-I'}(I')} \mu_1(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a'_I, a_{-I}) \text{ for all } a'_I \neq a_I.$$

Aggregating across the subsets $A_{-I}(I'), A_{-I}(I''), \dots$ of A_{-I} and every a_{-I} not in one of these subsets, we obtain:

$$\sum_{a_{-I}} \mu_2(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a_I, a_{-I}) \geq \sum_{a_{-I}} \mu_1(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I,$$

as required.

Now construct a probability measure μ_3 by the same procedure, taking each information set $I \in \mathcal{I}_i$ of rank 3. Repeat until every information set in \mathcal{I}_i has been used. We have some μ_k with the property that:

- (i) $\mu_k(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in D_{\mathcal{I}}$;
- (ii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu_k(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu_k(\cdot | A_{\mathcal{I}}(I))$;
- (iii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu_k(a_{\mathcal{I}}) > 0$, then $\mu_k(a_I | A_{\mathcal{I}}(I)) = 1$.

For all $a_{\mathcal{I}}$ such that $\mu_k(a_{\mathcal{I}}) > 0$, let $k(a_{\mathcal{I}}) = 0$, and let $p'_i(a_{\mathcal{I}}) = \mu_k(a_{\mathcal{I}})$.

Now consider every information set $I \in \mathcal{I}_i$ such that $\mu_k(A_{\mathcal{I}}(I)) = 0$. These are the information sets that should not be reached according to the beliefs μ_k . For each such set, I , of lowest order, we can use the same technique as for the construction of μ_k to construct a probability measure μ over $A_{\mathcal{I}}(I)$ with analogous properties to μ_k :

- (i) $\mu(a_{\mathcal{I}}) > 0$ only if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$;
- (ii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu(a_{\mathcal{I}}) > 0$, then a_I is a best response to $\mu(\cdot | A_{\mathcal{I}}(I))$;
- (iii) if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ and $\mu(a_{\mathcal{I}}) > 0$, then $\mu(a_{\mathcal{I}} | A_{\mathcal{I}}(I)) = 1$.

Note that $\mu(a_{\mathcal{I}}) > 0$ only if $\mu_k(a_{\mathcal{I}}) = 0$. For all $a_{\mathcal{I}}$ such that $\mu(a_{\mathcal{I}}) > 0$, let $k(a_{\mathcal{I}}) = \text{order}(I)$ and let $p'_i(a_{\mathcal{I}}) = \mu(a_{\mathcal{I}})$.

Now we take every information set $I \in \mathcal{I}_i$ for which there is no $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ such that $k(a_{\mathcal{I}})$ has been defined, and repeat the process. We continue until there are no information sets left.

The $\preceq_i^{a_{\mathcal{I}}}$ relation is defined as follows: $a'_{\mathcal{I}} \preceq_i^{a_{\mathcal{I}}} a''_{\mathcal{I}}$ if and only if $k(a'_{\mathcal{I}}) \leq k(a''_{\mathcal{I}})$ or $k(a'_{\mathcal{I}})$ is defined and $k(a''_{\mathcal{I}})$ is not. $a'_{\mathcal{I}} \in W_i^{a_{\mathcal{I}}}$ if and only if it has been assigned a rank by $k(\cdot)$, and since \leq is complete and transitive on the natural numbers and $A_{\mathcal{I}}$ is finite, $\preceq_i^{a_{\mathcal{I}}}$ satisfies **R1** and **R2**.

To show that $a_{\mathcal{I}} \in [Rat_i]$ (i.e. that player i is rational at $a_{\mathcal{I}}$), let p_i be the normalization of p'_i so that $p_i(A_{\mathcal{I}}) = 1$, with $p_i(a_{\mathcal{I}}) = 0$ if $p'_i(a_{\mathcal{I}})$ is not defined. For arbitrary $I \in \mathcal{I}_i$, we must compute $p_{i,I}^{a_{\mathcal{I}}}([a_{-I}])$. Consider the set $\min_i^{a_{\mathcal{I}}}([I] \cap [f_I(w)] \cap [W_i^w])$. $[I] = A_{\mathcal{I}}(I)$ and $[f_I(w)] = a_I$, from the definition of W . Furthermore, from the construction of $\preceq_i^{a_{\mathcal{I}}}$, every $\preceq_i^{a_{\mathcal{I}}}$ -minimal element of a given set must have been assigned the same k ranking (and must therefore be in $W_i^{a_{\mathcal{I}}}$). Each of these elements must therefore have been assigned its k rank and its $p'_i(\cdot)$ value (if strictly positive) by the same μ measure (or by μ_k), since by perfect recall, if $a_{\mathcal{I}} \in A_{\mathcal{I}}(I)$ for some $I \in \mathcal{I}_i$ of order n , there is no other $I' \in \mathcal{I}_i$ of order n such that $a_{\mathcal{I}} \in A_{\mathcal{I}}(I')$. So there is some μ or μ_k such that:

$$\begin{aligned} p_{i,I}^w([a_{-I}]) &= \frac{p_i([a_{-I}] \cap \min_i^w([I] \cap [f_I(w)] \cap W_i^w))}{p_i(\min_i^w([I] \cap [f_I(w)] \cap W_i^w))} \\ &= \frac{p'_i(a_{-I} \cap \min_i^w(A_{\mathcal{I}}(I) \cap a_I))}{p'_i(\min_i^w(A_{\mathcal{I}}(I) \cap a_I))} \\ &= \frac{\mu(a_{-I} \cap A_{\mathcal{I}}(I) \cap a_I)}{\mu(A_{\mathcal{I}}(I) \cap a_I)} \\ &= \mu(a_{-I} | A_{\mathcal{I}}(I) \cap a_I) \\ &= \mu(a_{-I} | A_{\mathcal{I}}(I)) \end{aligned}$$

The last inequality follows from (iii) above. From (ii) above, a_I is a best response to $\mu(\cdot | A_{\mathcal{I}}(I))$, i.e.

$$\begin{aligned} \sum_{a_{-I} \in A_{-I}} \mu(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a_I, a_{-I}) &\geq \sum_{a_{-I} \in A_{-I}} \mu(a_{-I} | A_{\mathcal{I}}(I)) \cdot u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I \\ \Rightarrow \sum_{a_{-I} \in A_{-I}} p_{i,I}^w([a_{-I}]) \cdot u_i(a_I, a_{-I}) &\geq \sum_{a_{-I} \in A_{-I}} p_{i,I}^w([a_{-I}]) \cdot u_i(a'_I, a_{-I}) \text{ for all } a'_I \in A_I \end{aligned}$$

The same result holds at every $I \in \mathcal{I}_i$, and so player i is rational at $a_{\mathcal{I}}$ as required.

$\preceq_i^{a_{\mathcal{I}}}$ is defined in the same way for every player i at every world $a_{\mathcal{I}} \in D_{\mathcal{I}}$. If $a_{\mathcal{I}} \notin D_{\mathcal{I}}$, $\preceq_i^{a_{\mathcal{I}}}$ can be defined in any way that satisfies **R1** and **R2**. We have already seen that $D_{\mathcal{I}} \subseteq [Rat]$. For every player i , notice that if $a_{\mathcal{I}} \in D_{\mathcal{I}}$, $a'_{\mathcal{I}} \in \min_i^{a_{\mathcal{I}}}(W_i^{a_{\mathcal{I}}})$ only if $a'_{\mathcal{I}} \in D_{\mathcal{I}}$. It follows from the definition of $[B_i\phi]$ that $D_{\mathcal{I}} \subseteq [B_iRat]$ for all i . So we have $D_{\mathcal{I}} \subseteq [ERat]$, and repeating the argument we obtain $D_{\mathcal{I}} \subseteq [CTBR]$. So we have shown that, for every $a_{\mathcal{I}} \in D_{\mathcal{I}}$, $a_{\mathcal{I}}$ is rationalizable, as required.

(only if) Take any model of Γ . First, we observe that, for all $I \in \mathcal{I}_i$, if $w \in [Rat_i]$, then $f_I(w) \in D_I^1$. This follows immediately from Lemma 1 and the definition of rationality. Thus, for all $w \in [Rat]$, $f(w) \in D_{\mathcal{I}}^1$. Now suppose that for some $I \in \mathcal{I}_i$, $a_I \notin D_I^2$. By Lemma 1, there is no probability measure over D_{-I}^1 to which a_I is a best response. It follows that $D_{-I}^1 \subseteq A_{-I}(I)$, since if there was some $a_{-I} \in D_{-I}^1$ which did not reach I , a_I would not affect the path through the game if a_{-I} were chosen. Hence a_I would be a best response to a_{-I} . So $[Rat] \subseteq [D_{-I}^1] \subseteq [A_{-I}(I)]$, and therefore $[B_i Rat] \subseteq [B_i A_{-I}(I)]$. Now suppose $w \in [B_i Rat]$. We must have $\min_i^w(W_i^w) \subseteq [Rat] \subseteq [A_{-I}(I)]$. But $[A_{-I}(I)] = [I]$, so $\min_i^w(W_i^w) = \min_i^w([I] \cap W_i^w)$. It follows from the definition of $p_{i,I}^w(\cdot)$ that $p_{i,I}^w(a_{-I}) > 0$ only if $a_{-I} \in D_{-I}^1$. So from the definition of rationality, if $w \in [B_i Rat] \cap [Rat_i]$, $f_I(w) \in D_I^2$. Aggregating over players and information sets gives us $[ERat] \cap [Rat] \subseteq [D^2]$, and iteration of the second step yields $[CTBR] \subseteq [D_{\mathcal{I}}]$. Thus if $a_{\mathcal{I}}$ is rationalizable, then $a_{\mathcal{I}} \in D_{\mathcal{I}}$, as required. ■