האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

BACKWARD INDUCTION AND COMMON STRONG BELIEF OF RATIONALITY

By

ITAY ARIELI

Discussion Paper # 535 February 2010

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Backward Induction and Common Strong Belief of Rationality*

Itai Arieli†‡

2010

*A preliminary version of this paper was published under the same name [1].

†Center for the Study of Rationality, Department of Mathematics, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel. E-mail: iarieli@math.huji.ac.il

‡This work is a part of the author's PhD dissertation being written under the direction of Professor R.J. Aumann. The author would like to express a special thanks to Professor Aumann for many hours of conversation and a prominent involvement in the writing process and in conceptual issues related to this paper. Please see also the last paragraph of Section 1.3.

Abstract

In 1995, Aumann showed that in games of perfect information, common knowledge of rationality is consistent and entails the backward induction (BI) outcome. That work has been criticized because it uses "counterfactual" reasoning—what a player "would" do if he reached a node that he *knows* he will not reach, indeed that he himself has excluded by one of his own previous moves.

This paper derives an epistemological characterization of BI that is outwardly reminiscent of Aumann's, but avoids counterfactual reasoning. Specifically, we say that a player *strongly believes* a proposition at a node of the game tree if he believes the proposition unless it is logically inconsistent with that node having been reached. We then show that *common* strong belief of rationality is consistent and entails the BI outcome, where—as with knowledge—the word "common" signifies strong belief, strong belief of strong belief, and so on ad infinitum.

Our result is related to—though not easily derivable from—one obtained by Battigalli and Sinischalchi [7]. Their proof is, however, much deeper; it uses a full-blown semantic model of probabilities, and belief is defined as attribution of probability 1. However, we work with a syntactic model, defining belief directly by a sound and complete set of axioms, and the proof is relatively direct.

1 Introduction

In extensive games of *perfect information* (PI), *backward induction* (BI) is a particularly prominent solution concept. The guiding principle behind BI is repeated application of the principle that when a player must choose between several options, he chooses the option that yields him the highest payoff.

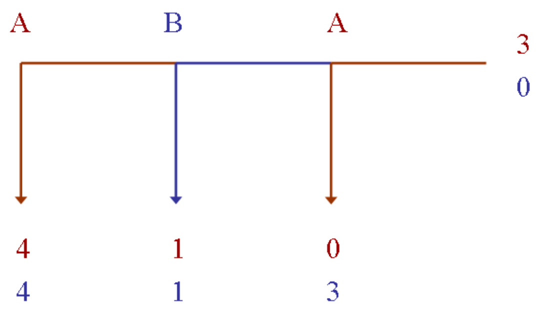

Figure 1:

Call a PI game *generic* if the payoffs of each player are different at different terminal nodes. In a seminal paper, Aumann (1995) gave the following epistemic characterization of BI in PI games:

Theorem 1.1. In a generic PI game, common knowledge[1] of rationality (CKR) is consistent and entails the BI outcome.

Here, "rationality" is defined as follows: Given a node h of the game, a player is *h-rational* if it is *not* the case that he knows that in the subgame starting at h, he can get a higher payoff by changing his strategy.[2] He is *rational* if he is h-rational at each of his nodes h.

Aumann's work has been criticized because his definition of rationality appears too strong; specifically, because it calls for h-rationality even when h has been excluded at a previous node by the very same player who must play at h. To illustrate the difficulty, consider first the game in Figure 1, and suppose that Ann's strategy and Bob's strategy call for them to exit

[1] An assertion is *commonly known* if all players know it, all know that all know it, all know *that,* and so on ad infinitum.

[2] A *strategy* of a player i is a function that assigns an alternative of i at v to each node v of i.

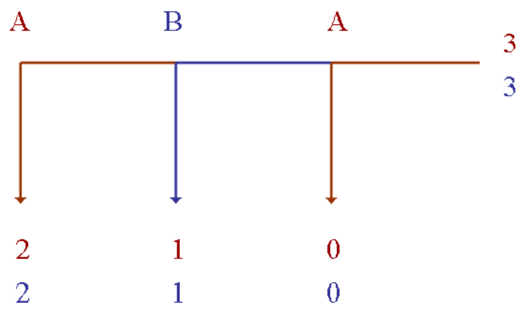

Figure 2:

(play "down") at each of their moves, and both know this. By Aumann's definition, Ann is not rational, because at her second node, she chooses to get 0, whereas she could get 3 by playing "across." But one might argue that it does not matter what she does at her second node, since she herself excluded the possibility of reaching that node by exiting at her first node; and, *that* was rational. So perhaps one should use a weaker (i.e., less demanding) definition of rationality, namely, to call a player "rational" if he is h-rational at all h that he himself has not excluded at a previous node. Henceforth, we call such a node *unexcluded.*

But as a conceptual rebuttal, this example is not fully convincing. Here we still get the BI outcome, even though only the weakened form of rationality obtains. Aumann could say, "OK, perhaps my definition is too strong, perhaps a weaker—and more appropriate—definition of rationality is sufficient to yield my result; but my result, though perhaps not the strongest possible, still points in the right direction—that CK of rationality, however defined, leads to the BI outcome." For a more convincing rebuttal of Aumann, one needs an example in which CK of the weakened version of rationality obtains, and the BI outcome is *not* reached.

Such an example was indeed suggested by Robert Stalnaker [10]; see Figure 2. Again, suppose that both Ann's strategy and Bob's call for them to exit at each of their moves, and both know that. With rationality in Aumann's sense, CKR does not obtain—Ann behaves irrationally at her second node—and his theorem does not require the BI outcome to be reached; and indeed it is not. But CK of the weakened—and apparently more appropriate—version of rationality *does* obtain, and nevertheless the BI outcome is not reached.

To make sense of this, Stalnaker tells the following story: At her first node, Ann must choose between "down" and "across." If she chooses "down," she gets 2. If she chooses "across," then it is Bob's turn; since he exits, she will get 1, which is < 2. So it is indeed rational for her to exit. Moreover, if she plays "across," then Bob will conclude that she is irrational, and so will fear that she would play irrationally also at her second node; and indeed, that is what her strategy prescribes. So it appears that CK of rationality—in a fairly natural sense of the word—*does* obtain here.

But one may ask, if Ann would play "across," would Bob really necessarily conclude that she is irrational? Might he not conclude that she is not playing as he thought she would (to exit)—as indeed is clearly the case—and does not expect him to exit either? Perhaps she expects him to play "across," and then she would play across as well! In that case he would be well-advised indeed to play across.

It is difficult to answer this question with the above kind of verbal, imprecise reasoning. To reason about "counterfactuals"—what Bob "would" do if Ann did something different from what he knows that she actually does—one needs a formal model, which allows events to occur that are "known" not to occur. And indeed, such a model can be constructed if we replace "knowledge" by "belief."

1.1 Counterfactuals and Belief

Counterfactual reasoning underlies game theory; more fundamentally, it underlies decision theory. When a decision maker chooses a over b, he must have some idea of what *would* have happened had he chosen b. In Stalnaker's example, if Ann exits at her first move, she must have some idea of what *would* have happened had she stayed in (played "across").

That's where belief—by which we mean attributing probability 1—comes in. If Bob *knows* that Ann will go out at her first move, then it is difficult to make sense of what he "would" have done had she stayed in. But if he only *believed* that she would exit—did not entirely exclude the possibility that she would stay in—then it makes sense to talk about what he "would" have done had she stayed in; it is simply what he *will* do, if[3] she stays in. Replacing "knowledge" by "belief" enables us to replace the troublesome "would" by "will."

So, let's review Stalnaker's example, substituting "belief" for "knowledge." Suppose that at the beginning of play:

(1) Ann believes that Bob will exit if she stays in;

(2.1) Bob believes that Ann will exit;

(2.2) Bob believes that if both Ann and he stay in, Ann will exit at her second node; and

(3) assertions (1), (2.1), and (2.2) are common belief[4] between Ann and Bob.

With this scenario, at the beginning of play there is indeed common belief

[3]This "if," like all subsequent "if"s in this discussion, signifies the ordinary material implication of mathematics; i.e., "if p, then q" means "q or not p," no more and no less.

[4]That is, both believe them, both believe that both believe them, both believe *that,* and so on ad infinitum.

that both players will play rationally at all unexcluded nodes.[5] So on the face of it, it would appear that Aumann's theorem is indeed an artifact of an excessively strong definition of rationality.

But on closer scrutiny, the explicit setting forth of the above scenario reveals its difficulties. Does (2.2) make sense? By (2.1), the fact that Bob's node was reached means that Ann did something highly unexpected. To be sure, Bob *could* reach the conclusion that Ann is irrational. But why *should* he? If she already did something unexpected—something that requires him to revise his beliefs—why would he not reach the conclusion that if he plays across, then she, too, will play across at her second node? To be sure, he does not *have* to think this, but why shouldn't he?

Well, the reader may ask, why *should* he? Can we formulate some rationality postulate that would enable us to reach that conclusion?

The answer is yes. The key concept is that of "strong belief." Let's say that a player *strongly believes* an assertion, if he believes it *unless it is logically impossible.* Then Stalnaker's "story"—i.e., the above scenario—is inconsistent with *common strong belief of rationality at unexcluded nodes* (CSBRU). Indeed, (2.2) is inconsistent with CSBRU, as we now show.

Assume Bob strongly believes Ann is rational and also believes that Ann will exit at the beginning of the game. Suppose that Ann stayed and it is now Bob's turn to play; can Bob revise his beliefs in a way that still supports Ann's rationality? The answer is yes; if Ann is rational and stayed in, it must be because she believes that Bob will stay in, and then on her last move Ann will stay in, and get 3 instead of 2. So if Bob strongly believes that Ann is rational, he should deduce that if Ann stayed in on her first move, she will also stay in on her last move.

[5]See Stalnaker's "story," set forth in the previous section.

Now assume that Ann believes that Bob strongly believes that she is rational; then by the above reasoning, she must believe that if she were to stay in, Bob would also stay in. So if Bob is rational and strongly believes that Ann is rational, he will stay in, if he gets a chance to play. And if Ann is rational and believes that Bob strongly believes that she is rational, then she will stay in on her first and last node. So in Stalnaker's game, CSBRU entails the BI outcome.

Thus in Stalnaker's game, all three nodes can be reached without contradicting the assumption of rationality. But there are games where some nodes cannot be reached with rational players. Thus in the game of Figure 1, Bob's node cannot be reached if Ann is rational. Nevertheless, CSBRU is possible in this game. Indeed, Bob's node is always unexcluded, since he has no previous nodes; if it is reached, then it is *logically impossible* for Ann to be rational, so under the definition, he still "strongly believes" that she is rational. Thus here, too, as in Stalnaker's game, CSBRU is consistent, and entails the BI outcome.

It is the purpose of this paper to prove that that holds in general; i.e., we have the following:

Main Theorem: *In a generic PI game, CSBRU is consistent and entails the BI outcome.*

1.2 Syntax and Semantics

The theorem just formulated belongs to an area of mathematical game theory called *interactive epistemology.* There are two parallel kinds of formalism for formulating and proving results in this area: the *semantic* and the *syntactic.*

A semantic formalism[6] employs semantic *universes*—sets whose elements are called *states of the world,* or simply *states.* On each universe are defined one or more structures representing the players' knowledge and beliefs (events, partitions, probability distributions, and the like). A particular universe represents a particular realization of epistemic principles, just as a particular group represents a particular realization of the axioms of group theory. To use the semantic formalism to prove a general assertion like the theorem of Aumann cited in Section 1, one establishes the assertion at each state in an arbitrary universe.

Syntactic formalisms are different; they work directly with sentences, rather than with states. There is a formal language, and there are axioms, rules of deduction, tautologies,[7] and formal proofs, using the axioms and rules.

In many contexts, a sentence is a tautology in a semantic formalism—"holds" at each state in an arbitrary universe—if and only if it is provable in the corresponding syntactic formalism—follows logically from the axioms and rules of deduction. Specifically, that is so in the context of knowledge (e.g., Aumann [3]).

Each kind of formalism has advantages. The main advantages of semantic formalisms are practical: they are easier to visualize, and also easier to work with. The main advantage of a syntactic formalism is conceptual: it is more straightforward and transparent—basically it says in plain words what it is that one wants to prove, and then proves it, logically, from explicit assumptions. By contrast, semantic formalisms are devious: to prove something,

[6]We use the indefinite article because in different contexts—like knowledge, probability, and belief—the formalisms are somewhat different. Moreover, the actual realization of the formalisms depends on parameters such as the number of players.

[7]Tautologies are sometimes called theorems

one must first restate it in set-theoretic language, and then establish it in an arbitrary universe. As Professor Dov Samet has put it,[8] if you want to explain it to your mother, say it syntactically; there's no way that she'll understand the semantic formulation.

There is, however, one important respect in which semantic formalisms are formally superior—one kind of task they can perform, that the syntactic formalisms cannot. Namely, they can prove consistency. In syntactic formalisms, one cannot show that a sentence is consistent—that its negation is not a tautology from the axioms. For that, one needs a *model* of the sentence—a state in a semantic universe at which the sentence in question "holds." Indeed, throughout mathematics, all consistency proofs use models, starting with the Bolyai-Lobachevsky proof that Euclid's parallel postulate does not follow from his axioms—i.e., that its negation is consistent with the axioms.

In particular, whereas the second part of our main theorem—that CSBRU entails the BI outcome—may be proved syntactically, its first part—that it is consistent—requires a semantic proof. Note, however, that whereas the *proof* is semantic, the *formulation* is purely syntactic. Indeed, the consistency of an assertion is intrinsically a syntactic notion: it means that the negation of the assertion does not follow from the axioms.

In practice, our proof of the main theorem combines syntactic and semantic methods throughout.

In addition to its transparency, the syntactic formalism has two important advantages in our context, both having to do with the fundamental notion of "strong belief." The first has to do with "belief," the second with that of "strong." What we here want to convey by saying that a player "believes"

[8]Private communication.

something is that he ascribes to it probability 1. The advantage of the syntactic formalism is that to deal with this formally, one does not need the whole gamut of numerical probabilities; rather, one axiomatizes the notion of belief—probability 1—directly, and then works only with those axioms, without reference to other probabilities. By contrast, semantic formalisms for belief that have heretofore been used in game-theoretic contexts allow all numbers between 0 and 1 as probabilities, and so are needlessly complex. In this paper we do develop a semantic formalism for belief and belief revision, which does not use numerical probabilities.

The second–and perhaps more fundamental–advantage of the syntactic formalism has to do with the adjective "strong," which calls for the notion of "provability" to play an important formal role *within* the statement of the result. Of course this paper, like all others in mathematics, is about theorems; all tautologies are provable—what we do in mathematics is prove theorems. But usually, the notion of "tautology" is not part of the statement of the result; the result is stated without involving the notion of provability, and then one simply asserts and proves the statement.

Here the situation is different. Assertions that some specific statements are or are not tautologies become elements in more complex assertions, and these, in turn, become elements in still more complex assertions, and so on. Specifically, CSBRU—common strong belief of rationality at unexcluded nodes—involves the notion of strong belief; and strong belief of a statement means that the statement is believed unless it is logically impossible—i.e., *unless its negation is tautology.* When we talk about *common* strong belief, we are iterating this kind of statement, indeed unboundedly often. Thus, in addition to the usual logical operators and connectives like " not," "or", and "and," we use an additional operator, t, which says that the formula following it is tautology; and whereas this operator is familiar in the metalanguage of logic, it is unusual that it becomes part of the formal language itself, from which new assertions can be formed.

Provability can be treated also within the semantic formalism, but it is considerably more awkward to do so, as we will see presently.

1.3 Battigalli and Sinischalchi

A seminal result that is conceptually closely related to our Main Theorem was established by Battigalli and Sinischalchi [7] (henceforth BS). But, whereas the conceptual content of the BS result is similar to that of ours, its formal statement is devious and round-about. In contrast, our Main Theorem formulates its conceptual content in a transparent and straightforward manner.

The BS result is formulated semantically. To state it and understand its relationship with ours, we start by describing the relationship between semantic and syntactic formalisms more carefully. Each sentence in a syntactic formalism corresponds to a set in each semantic universe—intuitively, the set of states in that universe at which that sentence "holds." Moreover, each logical operator corresponds to a set operation: "and" to intersection, "or" to union, and "not" to complementation (w.r.t. that particular universe); and a sentence that is provable in the syntax corresponds to the entire universe, since it must hold at each state. Conversely, if a sentence in the syntax corresponds in each arbitrary universe to the entire universe, then it is a tautology.

All that is well and good as long as the provability operator, t, is not an element of the syntactic language itself, but only of the metalanguage. As soon as t becomes part of the language itself, the elegant one-one correspondence between syntax and semantics breaks down. Indeed, the operator t

does not correspond to any set operation within a particular universe, since it refers simultaneously to *all* universes. For a sentence to be a tautology means that in *any* universe, the sentence corresponds to the entire universe; and there is no way of saying that within a particular universe.

BS work with semantics, so that is a real obstacle for them. To describe how they circumvent it, we must ourselves make a detour, via the epistemology of knowledge. In that context, there exists a single universe, called the *canonical* universe, such that a given sentence is provable in the syntax if and only if the corresponding set in the canonical universe is the entire canonical universe (see, e.g., Aumann [3]). Thus, once we have a canonical universe, we need no longer to refer to arbitrary universes to formulate the syntax-semantics equivalence; it is enough refer to one specific universe, namely the canonical one. Constructing the canonical universe is not a simple matter; and once constructed, it is not a simple matter to establish the basic property just enunciated. But such an object does exist.

Moreover, the canonical universe enables a valid semantic representation of syntactic sentences involving the provability operator t. Namely, the tautology operator corresponds to a set operator that takes the whole canonical universe to itself, and all its proper subsets to the empty set.

BS started by constructing a probabilistic analogue of the canonical semantic knowledge universe, which they called the *universal type space.* This construction is already very complex and deep, using a full-blown probabilistic semantic formalism, with sigma-fields of events on which numerical probabilities ranging between 0 and 1 are defined; and this in spite of the fact that they, like us, are interested only in belief—i.e., probability 1. It was published by them separately in 1999, three years before the paper with the main result [6]. Then, in 2002, they showed that the subset of the universal type space that corresponds to CSBRU is nonempty and entails the BI outcome (see [7]).

This may be considered a semantic *analogue* of our main theorem. To use it actually to *derive* our main theorem—which is syntactic—one would need to establish a syntax-semantics equivalence similar to that to which we alluded above in the context of knowledge. But BS did not prove, or even formulate, such an equivalence. Thus the conceptual *interpretation* of their result is like that of ours; but both its formulation and proof are far more intricate and difficult.

Nevertheless, there is no question that their contribution is of fundamental importance.

The notion of strong belief presented here, as well as it syntactic representation, were developed by R.J. Aumann and A. Brandenburger in the nineties of the previous century; they also conjectured the result established here, independently of Battigalli and Sinischalchi. However, they did not succeed in proving it fully; specifically, they were unable to establish the consistency of CSBRU, and so did not publish thier research on this topic.

2 Framework

2.1 Language

Start with a generic[9] PI game G. A *node* (a.k.a. *vertex*, or *history,*) of Player i is one at which i is active. A *strategy* of i is a function that assigns an action of i at h to each of i's nodes h. Each strategy s_i of i determines a set $H(s_i)$ of nodes of i, namely, those that s_i *allows* (does not preclude by

[9]This means that for each player, the payoffs at different terminal nodes are different.

an action at a previous node). A *plan* of i is the restriction[10] of a strategy s_i to $H(s_i)$.

In the sequel, we wish to refer to the beliefs of a player at each of his nodes, and also to his prior beliefs, before he observes anything. It is therefore convenient to add to the formalism the belief of each player at the root of the game tree (or at the empty history). Let H_i be the set consisting of the root and the nodes at which Player i is active.

We now construct a formal language. The building blocks are the following:

1. *Atomic sentences.* These have the form "player i uses plan p_i," denoted simply p_i.

2. *Left parentheses* and *right parentheses.*

3. *Connectives and operators of the propositional calculus.* As is known, it is sufficient to take just "or" ($\vee$) and "not" ($\neg$) as primitives, and in terms of them to define "and" ($\wedge$) and "implies" ($\rightarrow$).

4. *Belief modalities.* For each player i and node $h \in H_i$, there is a belief modality, b_i^h. Informally, if g is a formula (see below), then $b_i^h(g)$ means that conditional on the players *other than* i choosing plans that allow h, player i ascribes probability 1 to g at the beginning of play.[11] Verbally, we will describe $b_i^h(g)$ by saying that " i believes g at h;" but that is only a manner of speaking—the more accurate meaning is the above.

[10]Plans are sometimes called "strategies."Here we do not want a strategy to be defined at the nodes that it excludes.

[11]Players are not permitted to condition on their own actions; that would bring us uncomfortably close to counterfactual reasoning.

Definition 2.1. A formula is a finite string obtained by applying the following two rules, in some order, finitely often:

- Every atomic sentence is a formula.
- If f and g are formulas, so are $(f) \vee (g)$, $\neg(f)$, and $b_i^h(f)$, for every non-terminal node h.

The set of all formulas (for the game under consideration) is called the *syntax* of that game, and is denoted χ. Call each formula $f \in \chi$ a *simple* formula

If h and h' are nodes, then $h \succ h'$ means that h follows h' in the game tree (or that h' is a prefix of h). If a is an action at node $h \in H_i$, the formula that expresses "i plays a" (or simply a for short) has the form $\vee p_i$, where the disjunction ranges over all plans of i that call for him to play a at h. Also, "h is reached" (or simply h) is the formula $\wedge d$, where the conjunction ranges over all actions d on the path to h of all players with histories on that path. If L is a set of nodes, then "L is reached" (or simply L) is the formula $\vee h$, where the disjunction ranges over all h in L.

For any node h and player i, an *h-plan* of i is a plan of i that allows h; denote by $P_i(h)$ the set of all i's h-plans. An *opposition h-plan* is a conjunction of plans that allow h, one for each player other than i. An h-plan p_i together with an opposition h-plan p_{-i} determine a terminal node z of the game tree where $z \succ h$. and a payoff $u_i(p_i, p_{-i})$ for i. The set of all opposition h-plans is denoted $P_{-i}(h)$, and the formula that expresses "all players other than i allow h" is:

$$h_i^o = \bigvee_{p_{-i} \in P_{-i}(h)} p_{-i}.$$

2.2 Logic

We now present the axioms and inference rules that govern the internal logic of our language. The axioms are as follows:

(1) The axioms of the propositional calculus.

For every player i:

(2.1) $\bigvee p_i$, where the disjunction is over all plans of player i.

(2.2) $\neg(p_i \wedge q_i)$, where p_i and q_i are different plans of player i.

(3.1) $b_i^h(f \to g) \to (b_i^h f \to b_i^h g)$, where $h \in H_i$.

(3.2) $b_i^h f \to \neg b_i^h \neg f$.

(3.3) $b_i^h f \to b_i^{\hat{h}} b_i^h f$, where $h, \hat{h} \in H_i$.

(3.4) $\neg b_i^h f \to b_i^{\hat{h}} \neg b_i^h f$.

(3.5) $p_i \leftrightarrow b_i^h p_i$ for every $h \in H_i$.

(3.6) $b_i^h h_i^o$ for every $h \in H_i$.

(3.7) $(b_i^h f \wedge \neg b_i^h \neg \hat{h}_i^o) \to b_i^{\hat{h}} f$, where $h, \hat{h} \in H_i$ and $h \prec \hat{h}$.

The system defined by these axioms and rules will be called **AX**.

The inference rules are as follows:

(4.1) *From* $f \to g$ and f *infer* g *(modus ponens).*

(4.2) *From* f *infer* $b_i^h f$ *(generalization).*

Axioms (2.1) and (2.2) express the requirement that every player execute exactly one plan. Axiom schemas (3.1) and (3.2) represent classical modal belief axioms (see, e.g., [8]). Axiom schemas (3.3) through (3.5) combine versions of the "truth" and "introspection" axioms. Briefly, they say that players are sure of their own plans and beliefs. Axiom (3.6) says that at h, player i believes that the other players played to allow h. Axiom (3.7), concerns belief revision; it says that if at h, i believed f and also that the subsequent node $\hat{h}$ "could" occur, then he believes f at $\hat{h}$. This reflects the idea that players update their beliefs in a Bayesian way.

A set of formulas $\mathfrak{L}$ is called a list.

Definition 2.2. As in [3] a list $\mathfrak{L}$ is called *logically closed* if it is closed under modus ponens:

$$f \in \mathfrak{L} \text{ and } f \to g \in \mathfrak{L} \text{ implies } g \in \mathfrak{L}.$$

It is called *epistemically closed* if it is closed under generalization:

$$f \in \mathfrak{L} \text{ implies } b_i^h f \in \mathfrak{L}\ \forall i,\ h \in H_i,$$

and *closed* if it is both logically and epistemically closed. The *closure* of a list $\mathfrak{L}$ is the smallest closed list that includes $\mathfrak{L}$.

A formula f is called *tautology* or of **AX**, denoted by $\vdash_{\mathbf{AX}} f$, if it is in the closure of the list of all formulas having one of the forms (1), (2) or (3).[12] f is *inconsistent* if its negation is tautology; otherwise it is *consistent*. It *entails* g if the formula $f \to g$ is tautology. The formulas $f_1, f_2, \ldots$ are *inconsistent* if the conjunction of some finite subset of them is inconsistent; otherwise they are *consistent*. They *entail* g if the conjunction of some finite subset of them entails g. Denote by $\mathbf{T}$ the set of all tautologies of **AX**. Call each formula $f \in \mathbf{T}$ a *simple* tautology.

2.3 Tautology Calculus

To state the Main Theorem as a formula within our language, we need to incorporate the notion of provability as a formal element of the language.

[12]Alternatively, f is a *tautology* of **AX** iff there exists a finite sequence of formulas whose last formula is f, and each of which is either an axiom or follows from those preceding it through the application of one of the two inference rules.

So, we augment the language by adding a provability modality, denoted t; informally, if g is a formula, then $t(g)$ means that g is a tautology. To the rules that define the formation of formulas (see [8]), we add the following:

- If f is a formula, so is $t(f)$.

We denote the augmented syntax by χ'. The definition of closure (2.2 above) extends verbatim to the augmented syntax

Definition 2.3. A list $\mathfrak{L} \subset \chi'$ is called *tautologically complete* if

$$f \in \mathfrak{L} \text{ implies } t(f) \in L$$

and

$$f \notin \mathfrak{L} \text{ implies } \neg t(f) \in L.$$

Loosely speaking, we would like to extend the concept of "tautology" to the augmented syntax in such a way so that $t(f)$ is a tautology whenever f is a tautology and $\neg t(f)$ is a tautology whenever f is not a tautology. Therefore, for every formula f in χ', either $t(f)$ is a tautology or $\neg t(f)$ is a tautology. Thus, one may expect that the list of tautologies in the augmented language will include all the formulas of the form $t(f)$ where f is a basic tautology and all the formulas $\neg t(f)$ where f is a basic formula that is not a basic tautology. By the following Lemma, there exists a unique such list which is strongly closed and satisfies this requirement.

Lemma 2.1. There exists a unique list $\mathbf{T}' \subseteq \chi'$ with $\mathbf{T}' \cap \chi = \mathbf{T}$ that is closed and tautologically complete.

the proof of the Lemma is relegated to an Appendix.

A formula $f \in \chi'$ is a tautology in the augmented axiomatization if it is in $\mathbf{T}'$. Write $\vdash_{\mathbf{AX}'} f$ for $f \in \mathbf{T}'$.[13]

[13]The reason for defining two different languages and the corresponding axiomatizations is technical; it simplifies our analysis.

2.4 Semantics

The notion of strong belief that plays a central role in our theorem depends crucially on that of consistency—that a formula does not contradict the axioms, that its negation cannot be proved. Proving consistency syntactically is a tricky matter; to prove a formula, one writes down a proof, but how does one show that something can*not* be proved? On the face of it, it would seem that one would have to write all possible proofs, and show that none of them end with the given formula.

To cope with this difficulty we present a friendly semantic formalism that represents a way to interpret the formal language. This will enable us to determine whether a given formula is consistent. We start by defining *models* for our language.

Definition 2.4. A *model* $M = \{\Omega, \mathbf{p}, (\mathcal{K}_i)_{i\in I}, ((B_i^h)_{h\in H_i})_{i\in I}\}$ for the syntax χ consists of,

1. A non-empty set Ω (the *states of the world,* or simply *states;*)
2. a function $\mathbf{p}$ from Ω to $\times_i P_i$ ($\mathbf{p}_i(\omega)$ is i's plan in state ω);
3. for each player i, a partition $\mathcal{K}_i$ of Ω (if ω is in an atom K of $\mathcal{K}_i$, then i *knows* that the true state is in K); and
4. for each player i, node h of i, and atom K of $\mathcal{K}_i$, a nonempty subset $B_i^h(K)$ of K (if ω is in $B^h(K)$, then i *believes* that the true state is in $B_i^h(K)$), where
5. if h and h' are nodes of i with $h \prec h'$, then $B_i^{h'}(K)$ is either included in $B_i^h(K)$, or disjoint from it; and

6. at every state ω in $B_i^h(K)$, the plans $\mathbf{p}_j(\omega)$ of player j other than i allow h.

To each formula f in the syntax, assign a subset $\|f\|$ of Ω, representing the set of states at which f *holds* (or is "true"); formally, $\|f\|$ is defined inductively over the "depth" of a formula as follows:

1. $\|p_i\| := \{\omega : \mathbf{p}_i(\omega) = p_i\}$;
2. $\|\neg f\| := \Omega \backslash \|f\|$;
3. $\|f \vee g\| := \|f\| \cup \|g\|$;
4. $\|b_i^h(f)\| := \cup\{K : B_i^h(K) \subset \|f\|\}$.

In these terms, Requirement 6 may be restated as 6'. $B_i^h(K) \subset \|h^o\|$.

An element of the set $\|f\|$ will be called a *model* of the formula f.

Lemma 2.2. Every formula $f \in \chi$, such that $\vdash_{\mathbf{AX}} f$, is true in every state of the world of every model.

Proof. See Theorem A.4 in the Appendix. □

Corollary 2.3. Every formula that is true in some state of the world, in some model, is consistent.

3 The Theorem

3.1 Rationality

Call a player *rational* if at every node allowed by his plan, he does not believe that he has a plan that yields him a higher payoff.[14] Formally, if p_i and q_i

[14] Like Aumann [2], we replace utility maximization by a weaker condition, namely, that the player does not believe that he has a better plan. Unlike Aumann, we demand this not at *every* node, but only at nodes allowed by the plan actually in use.

are different plans of player i, set $Q^h_{p_i}(q_i) = \bigvee\{p_{-i} \in P_{-i}(h) | u_i(q_i, p_{-i}) > u_i(p_i, p_{-i})\}$; in words, $Q^h_{p_i}(q_i)$ is the disjunction of opposition h-plans in $P_{-i}(h)$ for which q_i yields more than p_i to i (if there are no such p_{-i}, let $Q^h_{p_i}(q_i)$ be a contradiction). The formula that asserts that plan p_i is *rational* for i is then

$$r(p_i) := \bigwedge_{\{h | h \in H(p_i)\}} \bigwedge_{\{q_i \in P_i(h) | q_i \neq p_i\}} \neg b^h Q^h_{p_i}(q_i).$$

Define player i to be *rational* if he uses a rational plan, that is,

$$r_i := \bigwedge_{p_i \in P_i} (p_i \longrightarrow r(p_i)).$$

Remark. We do not claim that the above definition of "rationality" is the only reasonable one. We do however claim that if i is "rational" in any commonly accepted sense (such as utility maximization), then certainly r_i obtains.

The formula corresponding to all players being rational is

$$r := \bigwedge_i r_i.$$

3.2 Strong Belief

Say that i *strongly believes* a formula g (written $sb^i(g)$) if for each node h of i, either

(i) i believes g at h, or

(ii) g precludes h being reached (or equivalently, g is inconsistent with h).

In words, i continues to believe g no matter what happens, unless he reaches a node that is logically impossible under g. In symbols:

$$sb^i(g) = \bigwedge_{h \in H_i} [b^h(g) \vee t(\neg(h \wedge g))].$$

Say that g is *strongly believed* (or that there is *strong belief* of g, written $sb(g)$) if each player strongly believes g. *Mutual strong belief* of g of order n (written $sb^n(g)$) is defined inductively as $sb^{n-1}(g) \wedge sb(sb^{n-1}(g))$; that is, each iteration provides for the foregoing iteration and strong belief thereof (note that the strong belief operator does not commute with conjunction). *Common strong belief* of g comprises all the formulas $sb^n(g)$ for all n.

The main result of this paper states the following:

Theorem. Common strong belief of rationality is consistent and entails the unique backward induction outcome for every generic PI game.

4 Outline of the Proof of the Main Theorem

The proof has two parts. The first describes an elimination process culminating with the BI outcome. The second identifies the result of the $(k+1)$'th step of that process with $(sb)^k(r)$, i.e., k'th order strong belief of rationality.

Before each step of the elimination process, there is a set of *current* plans of each player; a node is *relevant* at that step if it is allowed by some profile of current plans. Before the first step, all plans are current. To go from one step to the next, retain only those plans p_i of player i that, at each relevant node h allowed by p_i, are not strictly dominated by an h-plan of i w.r.t. current opposition h-profiles.[15] The process clearly "ends" after finitely many steps, in the sense that the set of current plans does not change thereafter. We will show that there are plan profiles that survive the process, and all of them lead to the BI outcome (see Lemma 5.1).

This describes the first part of the proof. The second part demonstrates

[15] I.e., there is no h-plan q_i of i such that $u_i(q_i, p_{-i}) > u_i(p_i, p_{-i})$ for all opposition profiles p_{-i} consisting of relevant h-plans.

that a plan survives the $(k+1)$'th step of the process if and only if it is consistent with k'th order strong belief of rationality (see Lemma 5.2).

The proof of the second part is by induction. Suppose the lemma true up to and including k. To demonstrate "if," we show that if a plan p_i does not survive the $(k+1)$'th step, then it is inconsistent with k'th order strong belief of rationality. Indeed, in that case there is a relevant node h allowed by p_i such that p_i is not a best reply to any opposition h-profile of current plans. But then p_i cannot be rational under any possible beliefs of i at h that are consistent with k'th order strong belief of rationality.

To demonstrate "only if," we must show that any plan surviving the $(k+1)$'th step of the process is consistent with k'th order strong belief of rationality. Consistency is demonstrated semantically, that is, by building a model in which it holds.

5 Proof of the Main Theorem

Before proving our Main Theorem it is helpful to link between consistency in the language χ' with respect to $\mathbf{AX}'$ and consistency in χ with respect to $\mathbf{AX}$. The following result is essentially a restatement of the definition of strong belief and of a tautology in χ'.

Proposition 1. A formula $f \in \chi$ is consistent (or a tautology) with respect to $\mathbf{AX}$ iff it is consistent (or a tautology) with respect to $\mathbf{AX}'$. Moreover,

$$\vdash_{\mathbf{AX}'} sb^i(f) \leftrightarrow \bigwedge_{h \in H_i(f)} b_i^h(f),$$

where $H_i(f) = \{h \in H_i : \vdash_{\mathbf{AX}'} \neg t(\neg(h \wedge f))\}$.

Proposition 1 provides a straightforward inductive way to convert any formula of the form $sb^n(r)$ to a logically equivalent formula in χ, i.e., a formula not involving the modality t.

In the proof of the theorem we use a finite family of models that will be helpful in proving consistency for the formulas $sb^n(r)$. This family is a sub-family of the models introduced in Definition 2.4.

Definition 5.1. A model for the syntax χ is called *simple* if:

1. The set of states of the world is $\Omega = \prod_{i \in I} P_i$, where P_i is the set of plans for player i.

2. $\mathbf{p} : \Omega \to P$ is the identity map.

3. The atoms K of $\mathcal{K}_i$ are determined by $\mathbf{p}_i$; i.e., two states belong to the same atom of $\mathcal{K}_i$ iff they specify the same plan for i.

In a simple model, if I is the atom of Player i at whose states he plays p_i, and h is a node of i, we will sometimes write $B^h(p_i)$ instead of $B_i^h(I)$.

Consider the following inductively defined sequence of plans:
For every player i, define $P_i^0 = P_i$, $P_{-i}^0 = \Pi_{j \neq i} P_j^0$ and $P^0 = \Pi_j P_j^0$. For $n \geq 0$, assume P_i^n to be defined for every player i, and let H^n be those non-terminal nodes that are allowed (reachable) by profiles of plans from $P^n (:= \Pi_j P_j^n)$. Now define P_i^{n+1} as the set of all plans p_i satisfying the following requirements:

1. $p_i \in P_i^n$.

2. For every node $h \in H(p_i) \cap H^n$ and for every $q_i \in P_i(h)$, p_i is not strictly dominated by q_i with respect to P_{-i}^n in the subgame starting at h; i.e., there exists a $p_{-i} \in P_{-i}^n(h)$ ($P_{-i}^n(h)$ are those opposition plans in P_{-i}^n that allow h) for which $u_i(p_i, p_{-i}) \geq u_i(q_i, p_{-i})$.

In the terminology of the above "outline," P_i^n comprises i's "current" plans.

Lemma 5.1. All the P_i^n are nonempty. Moreover, for every generic PI game there exists an m such that $P_i^n = P_i^m$ for every player i and every $n > m$. And every profile of plans in P^m leads to the unique backward induction outcome.

Sketch of the proof. Consider the following inductively defined elimination process: For every i, $W_i^0 = P_i^0$, assume W_i^n is defined for every i, set W_i^{n+1} to be those plans in W_i^n that are not (weakly) dominated by any plan from P_i with respect to W_{-i}^n.[16]

We prove that $W^n = P^n$ for every $n \geq 0$. For $n = 0$ it trivially holds. Now assume $W^n = P^n$ and let $p_i \in W_i^n \setminus W_i^{n+1}$. So p_i is weakly dominated by some q_i with respect to $W_{-i}^n (= P_{-i}^n)$. So, there exists $h \in H^n$ that are allowed by both p_i and q_i for which p_i prescribe an action a, and q_i prescribe a different action, b. But since the game is generic, $h \in H^n$, and p_i is weakly dominated by q_i, one can deduce that q_i strongly dominates p_i in the subgame starting at h with respect to $P_{-i}^n(h)$. And so, $p_i \notin P_i^{n+1}$; therefore, $P_i^{n+1} \subset W_i^{n+1}$.

For the other direction, let $p_i \in P_i^{n+1} \setminus P_i^n$ by definition there exists a node $h \in H^n$ that is allowed by p_i and a plan $q_i \in P_i(h)$ such that q_i strongly dominates p_i at h. Define a plan for player i, l_i as follows:
For $h' \in H_i$, if $h' = h$ or h' follows h, set $l_i(h') = q_i(h')$; otherwise set $l_i(h) = p_i(h)$. The node h is allowed by P^n; therefore by the induction hypothesis it is allowed also by W^n and so $l_i(h)$ weakly dominates p_i with respect to W_{-i}^n. Therefore $W_i^{n+1} = P_i^{n+1}$.

[16] Alternatively, one can retain those plans that are not dominated by any plan from W_i^n with respect to W_{-i}^n. These processes are the same.

Clearly, $\forall n\ W^n \neq \emptyset$ and the existence of m is obvious. Moreover, in [4] Battigalli have proved that all the profiles in W^m lead to the unique BI outcome. $\square$

Lemma 5.2. For each n, a plan p_i of player i is consistent[17] with $sb^n(r)$ iff $p_i \in P_i^{n+1}$.

Proof: By induction on n. The induction hypothesis consists of two parts, *if* and *only if,* stated as follows:

"Only if": If a plan p_i of player i is not in P_i^{n+1}, then p_i is inconsistent with $sb^n(r)$ (where $sb^0(r) := r$).

"If": There exists a simple model $\mathcal{M}_n$ such that whenever $0 \leq k \leq n$, every profile of plans in P^{k+1}, when viewed as a point in Ω, is a model for $sb^k(r)$ (i.e., is in $\|sb^k(r)\|$).

n=0.

Only if: Suppose that $p_i \notin P_i^1$; we will show that p_i is not consistent with r. Since $p_i \notin P_i^1$, there is a node $h \in H(p_i)$, and a plan $q_i \in P_i(h)$ such that for every opposition h-plan p_{-i}, we have $u_i(p_i, p_{-i}) < u_i(q_i, p_{-i})$. That is, q_i is better for i than p_i, no matter what the opposition does; so p_i cannot be rational, no matter what i believes. Formally, by definition of rationality $\vdash_{\mathbf{AX}'} r \wedge p_i \rightarrow \neg b^h Q^h_{p_i}(q_i)$. But in this case, $Q^h_{p_i}(q_i) = h^o_i$, which contradicts axiom (3.6).

If: We construct a simple model $\mathcal{M}_0$, and show that r holds at every point in P^1 (which is nonempty by Lemma 5.1). So it will follow that $sb^0(r) = r$ is consistent.

For every player i, plan $p_i \in P_i^1$ and node $h \in H_i$, define $B^h(p_i)$ as the

[17]Since we may take $sb^n(r) \in \chi$ by proposition 1, consistency in $\mathbf{AX}$ and in $\mathbf{AX}'$ are the same.

set of *all* plan profiles that are consistent with h and p_i; i.e.,

$$B^h(p_i) := \{(p_i, p_{-i}) : \ p_{-i} \in P^0_{-i}(h)\} = \|h^o_i\|$$

Let $p \in P^1$; by definition of P^1_i we deduce that for every $h \in H(p_i)$ and for every $q_i \in P_i(h)$, there exists $p_{-i} \in P_{-i}(h)$ such that $u_i(p_i, p_{-i}) \geq u_i(q_i, p_{-i})$. Therefore $P_{-i}(h) \nsubseteq Q^h_{p_i}(q_i)$; i.e., for every i, every node h allowed by p_i, and every $q_i \in P_i(h)$ other than p_i, player i does not believe that q_i strictly dominates p_i. Therefore,

$$p \in \|\neg b^h(Q^h_{p_i}(q_i))\| \ \forall i \forall h \in H(p_i) \text{ and } \forall q_i \in P_i(h),$$

when p is viewed as a point in Ω. So $p \in ||r(p_i)||$ for all i, or alternatively, $||p \wedge r(p_i)|| \neq \varnothing$, where p is now viewed as a formula. But $\vdash_{\mathbf{AX'}} r \wedge p \leftrightarrow \wedge_{i \in I} r(p_i)$. Thus p is indeed consistent with r.

At this point we have the basis for the induction. Now assume the induction hypothesis for $n-1$; we prove it for n as follows:
Only if: Let p_i be a plan of player i that is not in P^{n+1}_i. If $p_i \notin P^n_i$, then by the induction hypothesis we are done. So we may take $p_i \in P^n_i \diagdown P^{n+1}_i$. Then for some node $h \in H(p_i) \cap H^n$ there exists a strategy $q_i \in P_i(h)$ such that $u_i(p_i, p_{-i}) < u_i(q_i, P_{-i})$ for every $p_{-i} \in P^n_{-i} \cap P_{-i}(h)$. By the induction hypothesis, the plans of players other than i that are consistent with $sb^{n-1}(r)$ are precisely those in P^n_{-i}, so $\vdash_{AX} sb^{n-1}(r) \rightarrow \bigvee_{p_{-i} \in P^n_{-i}} p_{-i}$. Again by the induction hypothesis, h is consistent with $sb^{n-1}(r)$, so by the definition of $sb^n(r)$,

$$\vdash_{\mathbf{AX'}} sb^n r \rightarrow b^h(sb^{n-1} r).$$

Therefore by axioms (3.1) and (4.2), $\vdash_{\mathbf{AX'}} sb^n(r) \rightarrow b^h[\bigvee_{p_{-i} \in P^n_{-i}} p_{-i}]$. By axioms (3.6), $\vdash_{\mathbf{AX'}} b^h h^o$ and so $\vdash_{\mathbf{AX'}} sb^n(r) \rightarrow b^h h^o$. Using axioms (3.1) and (4.2) again, one can show that $(b^h f \wedge b^h g) \rightarrow b^h(f \wedge g)$; therefore, $\vdash_{\mathbf{AX'}}$

$sb^n(r) \to b^h[\bigvee_{p_{-i} \in P^n_{-i}(h)} p_{-i}]$. Since q_i dominates p_i in the subtree starting at h w.r.t. $P^n_{-i}(h)$, we get $P^n_{-i}(h) \subseteq Q^h_{p_i}(q_i)$. But $\vdash_{\mathbf{AX'}} r(p_i) \to \neg b^h[\bigvee Q^h_{p_i}(q_i)]$; therefore $\vdash_{\mathbf{AX'}} r(p_i) \to \neg b^h[\bigvee_{p_{-i} \in P^n_{-i}(h)} p_{-i}]$. Since $\vdash_{\mathbf{AX'}} (sb^n(r) \wedge p_i) \to r(p_i)$, deduce that

$$\vdash_{\mathbf{AX'}} (p_i \wedge sb^n(r)) \to b^h[\bigvee_{p_{-i} \in P^n_{-i}(h)} p_{-i}] \wedge \neg b^h[\bigvee_{p_{-i} \in P^n_{-i}(h)} p_{-i}].$$

.

If: By the induction hypothesis, in the model $\mathcal{M}_{n-1}$, we have that for all $k < n$,

$$p \in P^{k+1} \Leftrightarrow p \in \|sb^k(r)\|.$$

For each player i and plan p_i, define the model $\mathcal{M}_n$ as follows:
If $p_i \notin P^{n+1}_i$, then $B^h(p_i)$ is as in $\mathcal{M}_{n-1}$.
If $p_i \in P^{n+1}_i$ and $h \in H_i$, if $h \notin H^n$, then $B^h(p_i)$ is as in $\mathcal{M}_{n-1}$. If $h \in H^n$, redefine

$$B^h(p_i) := \{(p_i, p_{-i}) : p_{-i} \in P^n_{-i}(h)\}.$$

We show that if $p \in P^{k+1}$, then $p \in \|sb^k(r)\|$ for all $k \leq n$.
Note that from the definition of P^{n+1}, for every $h \in H^n \cap H(p_i)$ and $q_i \in P_i(h)$, there exists an opposition h-plan $p_{-i} \in P^n_{-i}$ such that $u_i(p_i, p^h_{-i}) \geq u_i(q_i, p^h_{-i})$. So $P^n_{-i}(h) \setminus Q^h_{q_i}(p_i) \neq \emptyset$. Therefore $B^h(p_i) \not\subset \|Q^h_{q_i}(p_i)\|$, and so

$$p_i \in \|\neg b^h(Q^h_{p_i}(q_i))\| \ \forall i \forall h \in H^n(p_i) \text{ and } \forall q_i \in P_i(h). \tag{5.1}$$

By the induction hypothesis, 5.1 is true for all $h \in H(p_i)$. And so as in the case $n = 0$, we deduce that $p \in \|r\|$. Moreover, for every $k \leq n$ and node $h \in H^k \cap H(p_i)$, we have—by the induction hypothesis and by the reconstruction of $B^h(p_i)$—that $B^h(p_i) \subseteq \|sb^k\|$. And so inductively we have that $p \in sb^k(r)$ for every $k < n$. As for $k = n$, one has $sb(sb^{n-1}(r)) = \bigwedge_{h \in H^k} b^h(sb^{k-1}(r))$.

Since $sb^n(r) = sb^{n-1}(r) \wedge sb(sb^{n-1}(r))$, one gets $p \in sb^n(r)$. For every other p, the inductive construction entails $p \in P_i^{k+1} \Leftrightarrow p \in \|sb^k(r)\|$. So the lemma is proved.

Thus, a plan p_i of player i is consistent with $sb^0(r) \wedge \ldots \wedge sb^n(r)$ iff $p_i \in P_i^{n+1}$. By Lemma 5.1, P^m is nonempty, equal to P^n for all $n \geq m$, and every profile in P^m leads to the unique BI outcome. So, P^m is consistent with $sb^0(r) \wedge \ldots \wedge sb^{n-1}(r)$ for all $n \geq m$, and the Main Theorem is proved.

6 Discussion

6.1 Battigalli and Siniscalchi

In this subsection we would like to further relate to the connection between BS model and the presented language. Basically BS uses Harsanyi *type space,* in which every type of every player comprises a *conditional probability system* over the other players' strategies and types. A natural question to ask, in this context, is whether BS type space provides a model for our axiomatization.

There is a natural way to identify each formula in our language with a corresponding event in BS type space. But it turns out that the connection between our syntax and BS *universal type space* is much stronger. In fact one can prove that with an additional axiom (see axiom (3.8) in the Appendix) BS type space provide a canonical model for our axiomatization.[18] That is, every tautology in our language is valid in every state of the world, when it translated to BS type space, and vice versa, every formula that is valid in every state of the world is tautology with respect to our augmented axiomatization.[19]

[18]See the Appendix for a precise definition of canonical model.

[19]The paper does not include the formal proof of this assertion; it can be supplied by

6.2 General Extensive Games

Here we restrict the analysis to PI extensive-form games, but in fact it is equally valid for general finite extensive games with perfect recall. The way to adjust the framework for this case is fairly obvious. Again we use plans rather than strategies, except that now H_i is the collection of information sets of i; indeed, in the PI case it is identical to the set of histories of player i.

The axiomatization stays the same but here we have a belief with a probability one modality for every player's information set rather than for every node. However, it turns out that our definition of rationality is too weak to apply to general extensive games. In particular, in order to obtain BS's or Pearce's [9] extensive form rationalizability one needs a stronger definition of rationality.[20]

A Appendix

We present a class of models for our axiomatization, **AX**, that links the syntax to the semantics. The most preferable way would be to link the syntax to a class of models that characterize it by soundness and completeness relation. The way to do that would be by looking at the canonical model of the language with respect to the logic that our axiomatization defines.

We would first like to introduce some more terminology:
An axiom system is said to be *sound* for a language $\Im$ with respect to a class $\mathcal{C}$ of models if every tautology f is valid with respect to $\mathcal{C}$ i.e., valid in every

the author upon request.

[20] Again, a more detailed and formal approach for this case is beyond the scope of this paper and can be supplied by the author upon request.

model in $\mathcal{C}$. An axiom system is said to be *complete* for a language $\Im$ with respect to a class of models $\mathcal{C}$ if every valid formula f with respect to $\mathcal{C}$ is provable in the axiom system.

Throughout we fix an extensive form PI game G.

A.1 The Canonical Model

Definition A.1. A set of formulas Γ is *maximally consistent* with respect to **AX** if it satisfies the following two conditions:

a. Γ is consistent with respect to **AX**.

b. Γ is maximal with respect to that property.

It can be seen that maximal sets do exist[21] and satisfy the following properties:

1. Γ is logically closed i.e., closed under modus ponens (4.1).
2. Γ contains all the theorems of **AX**.
3. For every formula f, $f \in \Gamma$ or $\neg f \in \Gamma$.
4. For every formula f, g, $f \vee g \in \Gamma$ *iff* $f \in \Gamma$ or $g \in \Gamma$.
5. For every formula f, g, $f \wedge g \in \Gamma$ *iff* $f \in \Gamma$ and $g \in \Gamma$.
6. Every consistent set of formulas can be extended to a maximally consistent set.

Now let Ω be the set of all maximally consistent sets; we call the elements of Ω *states of the world*.

[21] See [8] or any other modal logic textbook.

Definition A.2. For each $\Gamma \in \Omega$ and non-terminal node $h \in H_i$, we define Γ_i/h to be the set of all formulas that player i *h-believes* in Γ. More precisely,

$$\Gamma_i/h = \{ \; g \mid b_i^h g \in \Gamma\}.$$

For every player i and non-terminal node $h \in H_i$, define the usual accessibility binary relation R_h^i over Ω as follows: let $\Gamma, \Lambda \in \Omega$, $\Gamma R_h^i \Lambda$ iff $\Gamma_i/h \subseteq \Lambda$. Let $B_i^h(\Gamma)$ be the set of all states of the world that player i considers possible at $h \in H_i$, that is,

$$B_i^h(\Gamma) = \{ \; \Lambda \in \Omega \mid \Gamma R_h^i \Lambda\}.$$

Proposition 2.

1. Γ_i/h is consistent (therefore $B_i^h(\Gamma) \neq \emptyset$).

2. Γ_i/h is closed under (4.1) and (4.2).

3. $g \in \Gamma_i/h$ for every g such that $\vdash_{\mathbf{AX}} g$.

   If $\Gamma R_h^i \Lambda$ for some $\Gamma, \Lambda \in \Omega$, then $\Gamma_i/h = \Lambda_i/h$.

Proof. Part 2 follows from positive introspection, while part 3 is straightforward from generalization. For part 1, assume by way of contradiction that Γ_i/h is not consistent. Then we have $g_1, \dots g_k \in \Gamma_i/h$ such that $\mathbf{AX} \vdash \neg(g_1 \wedge \dots \wedge g_k)$. By definition, $b_i^h g_1, \dots b^h g_k \in \Gamma$ and so from K we get $b_i^h(g_1 \wedge \dots \wedge g_k) \in \Gamma$ but from part 3 $b_i^h \neg(g_1 \wedge \dots \wedge g_k) \in \Gamma$, a contradiction to D.

As for part 4, let $\Gamma, \Lambda \in \Omega$ such that $\Gamma R_h^i \Lambda$. By definition $\Gamma_i/h \subseteq \Lambda$ and so if for some formula f, $b_i^h f \in \Gamma$, then $f \in \Lambda$. Note that if $b_i^h f \in \Gamma$, then from positive introspection (3.3), $b^h b_i^h f \in \Gamma$. We therefore deduce that if $b_i^h f \in \Gamma$ then $b_i^h f \in \Lambda$. And so $\Gamma_i/h \subseteq \lambda_i/h$. For the other direction assume on the contrary that $b_i^h f \in \Lambda \setminus \Gamma$, for some formula f. Then since Γ is a maximally consistent set $\neg b_i^h f \in \Gamma$. From negative introspection (3.4) $b_i^h \neg b_i^h f \in \Gamma$ and so since $\Gamma_i/h \subseteq \Lambda$ we get that $\neg b_i^h f \in \Lambda$, which contradicts the consistency of Λ. $\square$

Note that as a consequence of part 1 of the proposition, one gets in particular that for all $\Gamma \in \Omega$, $B_i^h(\Gamma) \neq \emptyset$.

For every $i \in I$, define $\mathbf{p}_i : \Omega \to P_i$ as follows: $\mathbf{p}_i(\Gamma) = p_i$ iff $p_i \in \Gamma$; note that $\mathbf{p}_i$ is well defined. We would like to define a partition $\mathcal{K}_i$ for every player i. Thus one can see the canonical model as a member of the class of models introduced in Definition 2.4. Let $\Gamma, \Lambda \in \Omega$. For every player i define an equivalence relation, $\sim_i$ as follows: $\Gamma \sim_i \Lambda$ if, for some node $h \in H_i$, $\Gamma_i/h = \Lambda_i/h$. The relation $\sim_i$ defines a partition $\mathcal{K}_i$ over Ω. One has to show that $B_i^h(\cdot)$ is measurable with respect to $\mathcal{K}_i$. That is, if $\Gamma \sim_i \Lambda$, then $B_i^h(\Gamma) = B_i^h(\Lambda)$. Assume that $\Gamma_i/h = \Lambda_i/h$ for some $h \in H_i$ and let $h' \in H_i$. If $f \in \Gamma_i/h'$ then $b^{h'} f \in \Gamma$ and from positive introspection (i.e., axiom (3.3)) $b_i^h b_i^{h'} f \in \Gamma$. It follows that $b_i^h b^{h'} f \in \Lambda$. Therefore $b_i^{h'} f \in \Lambda$ and $f \in \Lambda_i/h'$, and vice versa. Therefore $\Gamma_i/h' = \Lambda_i/h'$ for every node $h' \in H_i$, and so $B_i^{h'}(\Gamma) = B_i^{h'}(\Lambda)$ for every node $h' \in H_i$.

Now according to Definition 2.4 $M = \{\Omega, \mathbf{p}, (\mathcal{K}_i)_{i \in I}, ((B_i^h)_{h \in H_i})_{i \in I}\}$ defines a model for the language χ.

At this point we have define a model as a member of the class introduced in Definition 2.4. We need to show first that it indeed satisfies properties 1-6 stated in the definition.

Lemma A.1. The model M satisfies properties 1-6 stated in Definition 2.4.

Proof. For properties 1-3 we have nothing to prove. As for property 4 we have to show that for every partition element $K \in \mathcal{K}_i$ $B_i^h(K) \subset K$. Let $\Gamma \in K$; since $B_i^h(\cdot)$ is measurable with respect to $\mathcal{K}_i$ one has to show that

$\{\Lambda \mid \Gamma R^i_h \Lambda\} \subseteq K$ for every $h \in H_i$. But by part 4 of Proposition 2 $\Gamma R^i_h \Lambda$ entails $\Gamma_i/h = \Lambda_i/h$ and therefore $\Gamma \sim_i \Lambda$.

Property 5: Let $h, h' \in H_i$ with $h \prec h'$ and assume $\Lambda \in B_i^h(K) \cap B_i^{h'}(K) \neq \emptyset$. Since for every $\Gamma \in K$ one has $\Gamma_i/h' \subseteq \Lambda$, deduce that $h'^o \in \Lambda$. Also for every $\Gamma \in K$ one has $\Gamma_i/h \subseteq \Lambda$; therefore $\neg b_i^h h'^o \in \Gamma$. And so, if for some formula f, $b_i^h f \in \Gamma$, then because Γ is a maximally consistent set, $b_i^h f \wedge \neg b^h \neg h'^o \in \Gamma$; therefore by axioms (3.6) and (4.2) $b_i^{h'} f \in \Gamma$. We have shown that for every $\Gamma \in K$, $\Gamma_i/h \subseteq \Gamma_i/h'$ and so $B_i^{h'}(K) \subseteq B_i^h(K)$.

Property 6: For $\Lambda, \Gamma \in K$ if $\Gamma_i/h \subseteq \Lambda$ then $h^o \in \Lambda$. Therefore, if $\Lambda \in B_i^h(K)$, then $h^o \in \Lambda$ and so, if $p_j \notin P_j(h)$ for some player $j \neq i$, one has $\vdash_{\mathbf{AX}} p_j \to \neg h^o$. So, if $\Lambda \in B_i^h(K)$, $p_j \notin \Lambda$. □

Lemma A.2. For every $\Gamma \in \Omega$ and every formula f, $\Gamma \in \|f\|$ *iff* $f \in \Gamma$.

Proof. We will prove the lemma using induction on the depth of the formula. For formulas of depth zero the proof is immediate from the properties of maximal consistent sets and the truth assessment policy. We prove the proposition first for formulas of the form $f = b_i^h g$, where g is from depth $n-1 \geqslant 0$. The general case follows from the properties of maximally consistent sets and the truth assessment policy.

$\Leftarrow$: If $f \in \Gamma$, then by definition of Γ_i/h, $g \in \Gamma_i/h$; therefore $g \in \Lambda$ for every $\Lambda \in B_i^h(\Gamma)$ by the induction hypothesis $B_i^h(\Gamma) \subseteq \|g\|$; therefore $\Gamma \in \|f\|$.

$\Rightarrow$: If $\Gamma \in \|f\|$, then $B_i^h(\Gamma) \subseteq \|g\|$, so $g \in \Lambda$ for every Λ such that $\Gamma_i/h \subseteq \Gamma$; therefore $\Gamma_i/h \vdash_{\mathbf{AX}} g$ for otherwise we could have constructed a maximally consistent set Λ' such that $\Gamma_i/h \cup \{\neg g\} \subseteq \Lambda'$. But because Γ_i/h contains all the theorems of **AX** and is closed under 4.1 and 4.2 we get that $g \in \Gamma_i/h$ and therefore $f \in \Gamma$.

□

Recall that a canonical model $M = \{\Omega, \mathbf{p}, (\mathcal{K}_i)_{i\in I}, ((B_i^h)_{h\in H_i})_{i\in I}\}$ for χ with respect to **AX** is a model for which every **AX**-tautology formula is true in every $\omega \in \Omega$ and every **AX**-consistent formula f is valid (i.e., true in some $\omega \in \Omega$). Thus Lemma A.2 leads to the following immediate corollary:

Corollary A.3. $M = \{\Omega, \mathbf{p}, (\mathcal{K}_i)_{i\in I}, ((B_i^h)_{h\in H_i})_{i\in I}\}$ is a canonical model with respect to **AX**.

Denote the class of models for our language, was introduced in Definition 2.4, by $M(G)$. One has the following desired property of the $M(G)$:

Theorem A.4. The class of models $M(G)$ is sound and complete with respect to **AX**.

Proof. It is fairly easy to see that for each axiom schema f of the form (1), (2.1), (2.2), (3.1)-(3.7), one has $\|f\| = \Omega$ for every model $M \in M(G)$. Let f and g be formulas such that $\|f\| = \Omega$ and $\|f \rightarrow g\| = \Omega$; by our truth assessment policy one can deduce that $\|g\| = \Omega$, and also that for every $i \in I$ and $h \in H_i$, $\|b_i^h(f)\| = \Omega$. Therefore, if g is a tautology in **AX** (i.e., $\vdash_{\mathbf{AX}} f$), one can easily deduce using induction on the minimal proof length of f that $\|f\| = \Omega$ in every model $M \in M(G)$.

Completeness is a straightforward consequence of Corollary A.3. $\square$

The set $B_i^h(K)$ is the set of states that player i considers possible if the other player plays in accordance with node h, or, equivalently, one can think of $B_i^h(K)$ as a support of a probability measure formed by player i. From properties 5 and 6 one can see that the belief update is not strictly Bayesian, that is, whenever $B_i^{\hat{h}}(K) \cap \|\hat{h}\| \neq \emptyset$, $B_i^{\hat{h}}(K) \subseteq B_i^h(\omega) \cap \|\hat{h}^o\|$, rather than an equality. If we want strict Bayesian updating we must add the following axiom:

$$(3.8) \quad b_i^{\hat{h}} f \rightarrow b_i^h(f \vee \neg\hat{h}^o) \text{ where } h \prec \hat{h}.$$

Denote by $\mathbf{AX}^+$ the axiom system $\mathbf{AX}$ with the addition of (3.8) and by $M^+(G)$ the class of models which instead of property 5 of Definition 2.4 have the following property:

5' If h and h' are nodes of i with $h \prec h'$, then $B_i^{h'}(K)$ is either equal to $B_i^h(K) \cap \|h'\|$ or disjoint from it.

One then gets the following theorem:[22]

Proposition 3. $M^+(G)$ is sound and complete with respect to $\mathbf{AX}^+$.

In order to be able to define truth assessment in Ω for formulas in χ' we have to be able to interpret formulas of the form $t(f)$. For $\Gamma \in \Omega$ define

$$\Gamma \models t(f) \text{ iff } \|f\| = \Omega.$$

So $t(f)$ is true in Ω iff f is true in each $\Gamma \in \Omega$.

Proof of Lemma 2.1. Set,

$$\mathbf{T}' = \{f \in \chi' : \|f\| = \Omega\}.$$

$\mathbf{T}'$ obviously satisfies $\mathbf{T}' \cap \chi = \mathbf{T}$. The uniquness of $\mathbf{T}'$ follows by a simple induction over the construction of a formula in χ'. $\square$

References

[1] I. Arieli, Backward Induction and Common strong Belief of Rationality, *Texts in Logic and Games,* Vol.4: New Perspectives on Games and Interaction, Amsterdam University Press (2008).

[22] The proof for this proposition is omitted, and can be supplied by the author upon request.

[2] R. J. Aumann, Backward induction and common knowledge of rationality, *Games and Economic Behavior.* **8** (1995), 6-19.

[3] R. J. Aumann, Interactive epistemology. I. Knowledge. *International Journal of Game Theory.* (1999), 263-300.

[4] P. Battigalli, Strategic rationality orderings and the best rationalization principle, *Games Econ. Behav.* **13** (1996), 178-200.

[5] P. Battigalli, On rationalizability in extensive games, *J. Econ. Theory* **74** (1997), 40-61.

[6] P. Battigalli and M. Siniscalchi, Hierarchies of conditional beliefs and interactive epistemology in dynamic games *Journal of Economic Theory* (1999), 188-230.

[7] P. Battigalli and M. Siniscalchi, Strong belief and forward induction reasoning, *Journal of Economic Theory* **106** (2002), 356-391.

[8] B. F. Chellas Modal logic. An introduction. Cambridge University Press, Cambridge-New York, 1980.

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

[10] R. Stalnaker, Belief revision in games: forward and backward induction. Logic & foundation of the theory of games and decisions. *Mathematical Social Sciences* (1998), 31-56.