Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 159 (2015) 443–464

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

The logic of backward induction ☆

Itai Arieli [a], Robert J. Aumann [b,*,1]

[a] *Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Israel*

[b] *Department of Mathematics and Federmann Center for the Study of Rationality, the Hebrew University of Jerusalem, Israel*

Received 24 October 2013; final version received 30 June 2015; accepted 1 July 2015

Available online 13 July 2015

Dedicated to Moshe Arieli on the occasion of his 70th birthday

Abstract

Call a perfect information (PI) game *simple* if each player moves just once. Call a player *rational* if he never takes an action while believing, with probability 1, that a different action would yield him a higher payoff. Using syntactic logic, we show that an outcome of a simple PI game is consistent with common strong belief of rationality iff it is a backward induction outcome. The result also applies to general PI games in which a player's agents act independently, rendering forward inferences invalid.

JEL classification: C72; D83

Keywords: Backward induction; Common strong belief; Perfect information; Syntactic interactive epistemology; Strong belief

☆ The authors wish to express their deep gratitude to Adam Brandenburger; in the nineties of the last century, he was instrumental in formulating the underlying concepts, and he also contributed importantly to shaping the current presentation. We also thank Joe Halpern, who referred us to the work cited in Footnote 23. Finally, we are much indebted to the editor and the referees, whose comments brought about a very significant improvement of the paper. Needless to say, none of the above are responsible for errors, misstatements, or expressions of opinion herein.

* Corresponding author.

E-mail addresses: iarieli@tx.technion.ac.il (I. Arieli), raumann@math.huji.ac.il (R.J. Aumann).

[1] Aumann thanks the Israel Science Foundation (Grant number 1216/10) for research support.

http://dx.doi.org/10.1016/j.jet.2015.07.004

1. Introduction

The backward induction (BI) algorithm for perfect information (PI) games is based on the following reasoning: The last player, who must choose between outcomes of the game, chooses an action that maximizes his payoff; taking this as given, the previous player maximizes *his* payoff; and so on, until the beginning of the game is reached.

Though on its face convincing enough, this reasoning has for the past quarter century been discussed, scrutinized, analyzed intensely, and even rejected. Our goal here is to clarify the assumptions on the players' knowledge and rationality on which it rests.

Two preliminary observations: First, the above reasoning, and indeed the BI algorithm, is unchanged if each player i is split into several independent "agents," one for each of i's decision nodes, each with the same payoff as i. So[2] we restrict attention to *simple* PI games—those in which each player moves at just one node. Call a player *rational* if he has no action that he believes (with probability 1) yields him a higher payoff than the action he took.

Second, for the reasoning to work, more than just rationality is needed; roughly, the players must also ascribe rationality to each other. More precisely, we may take the players' rationality (r) to be *common knowledge* (*ck*); i.e., all players know that all are rational, all know that all know it, all know *that,* and so on. This implies that players at preterminal nodes[3] choose rationally; knowing that, players at prepreterminal[4] nodes choose rationally—i.e., the BI action; and so on. Formally, that *ckr* entails BI follows from a theorem of Aumann (1995).

Aumann's work was criticized because *ckr* involves a conceptual conundrum. A player i on the BI path must either continue on the BI path or go off. If he goes off, the player j who is reached is unreachable under *ckr*; and j knows this, as commonly known propositions are a fortiori known by all players. Since j bases his choice on what he thinks subsequent players will do, and his "knowledge" has been contradicted, it is not clear what he (j) will do. But i must base *his* choice on what *he* thinks j will do, so it is not clear what i should do, either. Specifically, it is not clear that the BI action is indeed rational for i—that he could not do better by leaving the BI path.

To avoid grappling with this conundrum,[5] we replace "knowledge" by "strong belief", where a player *strongly believes* a proposition p if he believes it unless it is logically inconsistent with his node being reached. *Common* strong belief of p signifies that p holds, all players strongly believe p, all strongly believe the foregoing, all strongly believe the foregoing, and so on. We then have our

Main Theorem. *An outcome of a simple PI game is consistent with common strong belief of rationality* (*csbr*) *iff it is a BI outcome.*

With *csbr*, the conundrum disappears. As before, suppose a player i on the BI path considers going off, say to j. When *ckr* obtains, j knows that he cannot be reached, as he knows *ckr*; so if he *is* reached, then his knowledge is contradicted, and we get the conundrum. Also when *csbr* obtains, j cannot be reached, by our theorem. But j does *not know* that he cannot be reached,

[2] Please see Section 9.1 for an explanation of this point.

[3] Those all of whose sons are outcomes.

[4] Those all of whose sons are terminal or preterminal.

[5] As have Binmore (1996) and Aumann (1996), inter alia.

as he does not know[6] *csbr*. He knows only those components of *csbr* that pertain to his beliefs and actions if reached,[7] and these are consistent.[8] Also i has beliefs about the actions of other players, including j; our theorem implies that with these beliefs, it is rational for him to stay on the BI path.

The theorem has both substantive and methodological interest. Substantively, it characterizes the BI outcome in simple PI games by *csbr*, which means that in general PI games, BI is characterized by common strong belief (*csb*) of the rationality of the *agents*. Previously, Battigalli and Siniscalchi (2002) (henceforth BS), who introduced the fundamental notion of strong belief, had characterized the BI outcome by *csb* of the rationality of the *players*. Ascribing rationality to *players* renders forward inferences relevant (Section 9.1.1). Thus what BS show is that reasoning with forward as well as backward inferences leads to the BI outcome; whereas we show that the BI outcome results when reasoning with backward inferences only (Section 9.1.2)—a task that, while not straightforward, is intrinsically simpler than BS's.

Methodologically, the theorem is formulated and proved in syntactic rather than semantic terms (Section 5). The notion of strong belief depends on that of logical impossibility. In the more widely used semantic methodology, formulating the notion of logical impossibility involves defining and constructing "complete" state spaces (Section 9.2). In the syntactic methodology, "logical impossibility" has its ordinary, everyday meaning.[9]

Several other features of our treatment are worth noting:

(1) **Belief:** Rather than expected payoff maximization, we use probability 1 belief only; easily formulated properties of this concept suffice, obviating the need for the logical apparatus required by numerical probabilities.

(2) **Genericity:** Much of the BI literature makes essential use of various genericity conditions (Section 10.3). Our treatment, too, is more transparent with genericity, but the result holds for all PI games, generic or not.

(3) **Strategies:** In simple PI games, *csbr* is always consistent not only with the BI outcome, but also with the BI *strategies*. There may also be other strategies with which it is consistent (Section 10.2).

The plan of the paper is as follows: Section 2 formulates the result, and Section 3 demonstrates it, carefully but still informally, assuming genericity; this may already satisfy some readers, who may skip any or all of the remainder. Section 4 removes the genericity condition. Sections 5–8 comprise the formal development: Section 5 discusses syntactic and semantic formalisms, the advantages of each, and the reasons for choosing the syntactic one here; Section 6 lays out the basic logical apparatus, which is used in Section 7 to formulate the result precisely, and in Section 8 to prove it rigorously. Section 9 is devoted to discussion: Section 9.1, of substance, and Section 9.2, of methodology. Finally, Section 10 reviews some of the literature, concentrating on BS. Appendix A proves a needed technical result outside the paper's mainstream.

[6] We (the analysts) know *csbr*; the players do not.

[7] Spelled out in Remark 2.1. We don't mean that j doesn't know anything else, but only that *csbr* does not call for him to know anything else.

[8] As *csbr* is consistent, which follows from our theorem.

[9] That is, a "logically impossible" proposition is one that involves a logical contradiction.

2. Informal formulation

A simple PI game is defined by a *rooted tree* T. Terminal nodes are called *outcomes,* non-terminal nodes—*players.* Each player h chooses (or *plays)* an *action* a_h—an edge at h that does not lead back to the root. At each outcome, each player has a *payoff.*

Parts of this and the next section[10] assume *genericity*—that each player has different payoffs at different outcomes; this renders the treatment significantly more transparent. In particular, BI then determines a unique outcome. The genericity assumption is removed in Section 4.

2.1. Actions, beliefs, rationality, and csbr

Assume that each player has beliefs concerning actions played, and beliefs held, by other players. Call an action a_h of player h *belief-dominated* if he has an action a'_h that he believes yields him a higher payoff than a_h. Call it *rational* if it is not belief-dominated. Say that *csbr* obtains if

r: only rational actions are played,[11]
and
r^1: each player believes r, unless it precludes reaching him,
and
r^2: each player believes the foregoing, unless it[12] precludes reaching him,
and
...
and
r^n: each player believes the foregoing,[13] unless it precludes reaching him,

and so on ad infinitum.

In terms of the actions and beliefs of the individual players, we have

Remark 2.1. *csbr* obtains iff each player plays a rational action, and for each n, believes[14] $r \wedge r^1 \wedge \ldots \wedge r^n$ when it does not preclude reaching him.

2.2. Backward induction (generic)

Label the nodes of the game tree T as follows: Label each outcome z by z. Proceeding by (backward) induction, label each player h by that one of his sons' labels that yields him the most. Denote the root's label by $BI(T)$, and call it the *BI outcome.*

2.3. The Main Theorem (generic)

Say that a node v is *reachable under* an assertion f if reaching v is consistent with f.

[10] Specifically, Sections 2.2, 2.3, and 3.2.
[11] The play and beliefs of a player are those that apply if he is reached.
[12] The foregoing.
[13] *All* the foregoing; i.e., r and r^1 and r^2 and ... and r^{n-1}.
[14] $\wedge$ means "and."

Generic Main Theorem. *An outcome of a generic simple PI game is reachable under csbr iff it is the BI outcome.*

3. Informal demonstration

The demonstration has two parts. The first describes a process of successively "pruning" branches of the game tree, and shows that only the BI outcome survives that process. The second identifies *csbr* with the pruning process.

3.1. The pruning process

Say that action a'_h *(strictly) dominates* action a_h if a'_h yields a higher payoff to h, no matter what is done by subsequent players. The *pruning process (PP)* proceeds by eliminating each dominated action of each player, and the entire subsequent branch, and then iterating until no dominated actions remain.

3.2. Pruning process → BI outcome (generic)

Lemma 3.1. *Pruning a dominated action a_h does not change h's label.*

Proof. [15]Label each action by the label of the node to which it leads. The label of an action depends only on the labels of nodes after that action, so the labels of actions other than a_h do not change when a_h is pruned. Since a_h is dominated, h's label is one of those other labels, so it does not change. □

Corollary 3.2. *Simultaneously pruning several dominated actions does not change the remaining players' labels.*

Theorem A. *Generically, $BI(T)$ is the unique outcome of the fully pruned tree T_P.*

Proof. Repeated use of Corollary 3.2 shows that $BI(T)$ is an outcome of T_P. If T_P had more outcomes, one could first eliminate players who have only one action, then prune some action of a preterminal player. □

3.3. Pruning process ↔ csbr

This part does *not* assume genericity; it is completely general.

Theorem B. *For all n, a node survives stage $n+1$ of the pruning process if and only if it is reachable under $r \wedge r^1 \wedge \ldots \wedge r^n$.*

Demonstration. [16]Formally, this is proved in Section 8. Informally, denote the result of stage n of the PP by T^n. Note first that r excludes precisely the dominated actions, no more and no less.

[15] The argument here is fully rigorous, so the word "proof" (rather than "demonstration") is in place.

[16] This is not an outline of the formal proof in Section 8, but a relatively brief argument showing informally "why" the theorem holds.

"No less" follows from a dominated action being a fortiori belief-dominated; "no more," from r allowing a player to believe that *all* players subsequent to his action could be reached. So T^1 comprises precisely what is reachable under r. This is case $n = 0$ of the theorem.

Next, we examine reachability under $r \wedge r^1$. As we have seen, r implies that only nodes in T^1 can be reached. So by r^1, all players in T^1 believe that only nodes in T^1 can be reached. Since by r, all players are also rational, it follows that actions in T^1 that are dominated in T^1 are excluded by $r \wedge r^1$. So $r \wedge r^1$ implies that only nodes in T^2 are reached. But under r, every node in T^1 is possible; so under r^1, each player h in T^1 may consider all subsequent nodes in T^1 possible. So $r \wedge r^1$ excludes *only* actions at nodes in T^1 that are dominated in T^1; so in T^1, that leaves precisely the nodes in T^2. The nodes outside of T^1 are already excluded by r, so T^2 is precisely what is reachable under $r \wedge r^1$. This is case $n = 1$ of the theorem.

For general n, we proceed by induction. Assume that the result T^n of stage n of the PP comprises precisely what is reachable under $r \wedge r^1 \wedge \ldots \wedge r^{n-1}$; we examine what is reachable under $r \wedge r^1 \wedge \ldots \wedge r^n$. By r^n, every player believes $r \wedge r^1 \wedge \ldots \wedge r^{n-1}$, unless it precludes reaching him; by the induction hypothesis, this means that every player in T^n believes that only players in T^n—and all[17] such players—can be reached. So as with $n = 1$, but with T^n instead of T^1, we conclude that the result T^{n+1} of stage $n + 1$ of the PP comprises precisely what is reachable under $r \wedge r^1 \wedge \ldots \wedge r^n$, as asserted. □

Corollary 3.3. *An outcome survives the PP iff it is reachable under csbr.*

Demonstration. By Theorem B, an outcome is consistent with $r \wedge r^1 \wedge \ldots \wedge r^n$ iff it survives stage $n + 1$ of the PP. So it is consistent with r and *all* the r^n—i.e., with *csbr*—iff it survives *all* stages of the PP. □

Demonstration of the Generic Main Theorem. Combine Corollary 3.3 with Theorem A. □

4. Dispensing with genericity

Without the genericity restriction, the BI process—as specified above (Section 2.2)—does not apply, since a player may have several sons whose labels are best for him. We here generalize the definition of BI to the unrestricted case (Section 4.1), show that BI and the PP still have the same result (Section 4.2), and finally, state and demonstrate the unrestricted Main Theorem (Section 4.3).

4.1. Backward induction

As in the generic case, we use an inductive labeling process, but now each node h of the tree T is labeled with a *set* Z_h of outcomes.[18] Among the outcomes in Z_h, denote the maximum and minimum payoffs to h's *father* by $\max h$ and $\min h$ respectively. Call a node h *inferior* if it has a brother h' with $\min h' > \max h$.

[17] Justifying the word "all" is the main burden of the rigorous proof (Section 8).

[18] The idea is that if h is reached, any outcome in Z_h may occur.

Start the process by labeling each outcome z by $\{z\}$. Then, inductively label each player by the union of the labels of his non-inferior sons. Denote the root's label by $BI(T)$, and call its members *BI outcomes.*[19]

The idea is that if all the sons of a player h are outcomes, then he chooses an action that maximizes his payoff; the resulting outcomes constitute the set Z_h. If h's father $\widehat{h}$ chooses h, then $\widehat{h}$ can count on h choosing *some* member of Z_h, but not *which* one. So if h is inferior—has a brother who is necessarily better for $\widehat{h}$ than h—then $\widehat{h}$ would certainly rather choose the brother, so his choosing h may be excluded. But $\widehat{h}$ might well choose any non-inferior son; this defines the label $Z_{\widehat{h}}$. The process continues similarly to $\widehat{h}$'s ancestors, until the root is reached.

4.2. *Backward induction ↔ pruning process*

Theorem C. $BI(T)$ is the set of outcomes of the fully pruned tree T_P.

Proof. Let $O(T_P)$ be the set of outcomes of T_P. That $BI(T) \subset O(T_P)$ follows as in Theorem A. Suppose $BI(T) \subsetneqq O(T_P)$. Let T' be a minimal subtree of T with $BI(T') \subsetneqq O(T_P)$; i.e., $O(T''_P) = BI(T'')$ for any proper subtree T'' of T'. Let $z \in O(T'_P) \backslash BI(T')$, let h' be the root of T', let $a_{h'}$ be the action leading to z, and let h'' be the son of h' to which $a_{h'}$ leads. Since z is not in $BI(T')$, it is eliminated by the BI process at h'; so h'' is inferior. So letting T'' be the tree with root h'', we conclude from $O(T''_P) = BI(T'')$ that $a_{h'}$ is dominated in T'_P, which contradicts T'_P being fully pruned. □

4.3. *The Main Theorem*

Main Theorem. [20]*An outcome of a PI game is reachable under csbr iff it is a BI outcome.*

Demonstration. Like that of the Generic Main Theorem, with Theorem C instead of Theorem A. □

5. Syntax and semantics

The Main Theorem belongs to an area of mathematical game theory called *interactive epistemology.* There are two parallel kinds of formalism in that area: the *semantic* and the *syntactic.* Semantic formalisms employ *state spaces;* each such space consists of a set of *states of the world* (or simply *states),* together with a structure representing the players' knowledge and beliefs (partitions, probability distributions, and the like). A particular state space 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, one establishes the assertion at each state in an arbitrary state space.

Syntactic formalisms are different; they work directly with sentences, rather than with states. There is a formal language, and there are axioms, inference rules, and formal proofs using the axioms and rules. In many contexts (see Chellas, 1980, Chapter 1), a sentence "holds" at each

[19] In generic games, this process reduces to that of Section 2.2, except that there $BI(T)$ is the unique BI outcome, and here it is the singleton consisting of that outcome.

[20] This differs from the Generic Main Theorem (Section 2.3) only in that the word "generic" is removed, and "a BI outcome" replaces "the BI outcome."

state in an arbitrary state space iff in the corresponding syntactic formalism, it is a *tautology,*[21] by which we mean that it follows logically from the axioms and inference rules.

Each kind of formalism has advantages. The main advantages of semantic formalisms are practical: they are easier to fathom, and easier to work with. They are also more widely used, especially in the economic literature, and so are more familiar. The main advantage of a syntactic formalism is conceptual: basically it says in plain words what it is that one wants to prove, and then proves it, logically, from explicit assumptions. To prove something in a semantic formalism, one must first restate it in the language of sets, and then establish it in an arbitrary state space. As Professor Dov Samet has put it (private communication), if you want to explain it to your barber, say it syntactically; there's no way he'll understand the semantic formulation.

There is, however, one important respect in which semantic formalisms are superior—one kind of task they can perform, that most syntactic formalisms cannot. Namely, they can prove consistency. In most[22] syntactic formalisms, one cannot show directly from the axioms that a sentence is consistent—that its negation is not a tautology. For that, one needs a *model* of the sentence—a state in a semantic state space at which the sentence in question "holds." Indeed, throughout mathematics, consistency proofs have traditionally used models, starting with the Bolyai–Lobachevsky proof that Euclid's parallel postulate does not follow from his other axioms—i.e., that its negation is consistent with those axioms.

In particular, the proof of our main theorem—that *csbr* is consistent and entails a BI outcome—intertwines syntactic with semantic methods. But, whereas the *proof* uses semantics, the *formulation* is purely syntactic. Indeed, the consistency of an assertion is intrinsically a syntactic notion: it means that the negation of the assertion is not a tautology.

In the present context, there is an additional aspect of the syntactic formalism that bears mentioning. This concerns the fundamental notion of "strong belief," which calls for the notion of "tautology" to play an important formal role *within* the statement of the result. Of course this paper, like all others in mathematics, is about tautologies; all theorems are tautologies—what mathematics does is to establish tautologies. But usually, the notion of "tautology" is not part of the statement of the result; the result is stated without involving the notion of tautology, and then one simply asserts and proves the statement.

Here the situation is different. Assertions that some specific statements are or are not tautologies are elements in more complex assertions; these, in turn, are elements in still more complex assertions, and so on. Specifically, strong belief of a statement means that it is believed unless it is logically impossible; i.e., unless its negation is a *tautology. Common* strong belief iterates this assertion, 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 signifies that the sentence following it is a tautology; and whereas this operator is familiar in the metalanguage of logic, it is unusual[23] that it becomes part of the object language—the formal syntax—from which new assertions can be formed.

---

[21] In formal logic, this is usually called a "theorem;" the word "tautology" is reserved for theorems of the propositional calculus. We prefer to reserve the term "theorem" for the more usual kind of theorem—the kind that appears in mathematical papers like this.

[22] See the next footnote.

[23] Halpern and Lakemeyer (2001) have published a model of this kind; they use an operator called *Val*, which is t in our language. With this operator and a few more axioms, they construct a formalism in which consistency can be proved syntactically.

With additional machinery, the tautology operator can be treated also within the semantic formalism, as we shall see in Section 9.2 below.

6. Framework

6.1. Syntax

Given a finite PI game, we construct a formal language. The building blocks are as follows:

- *Atomic sentences*. These have the form "player h chooses action a_h," denoted simply a_h.
- *Left parentheses* and *right parentheses*.
- *Connectives and operators of the propositional calculus*. As is known, it is sufficient to take "or" ($\vee$) and "not" ($\neg$) as primitives, and in terms of them to define "and" ($\wedge$) and "implies" ($\rightarrow$).
- *Belief modalities.* For each player h, there is a belief modality b_h. Informally, if g is a sentence (see below), then $b_h(g)$ means that if reached, h ascribes probability 1 to g. Verbally, we say "h believes g."
- A *tautology modality*, denoted t. Informally, if f is a sentence, $t(f)$ signifies that f is a tautology.

Define a *sentence*[24] as a finite string obtained by applying the following two rules, in some order, finitely often:

- Every atomic sentence is a sentence.
- If f and g are sentences, so are $(f) \vee (g)$, $\neg(f)$, $t(f)$, and $b_h(f)$, for every player h.

Henceforth, we may omit parentheses when the intended meaning is clear.

The set of all sentences for the game under consideration is called the *syntax* of that game, denoted χ'. Call a sentence f in χ' *basic* if it does not involve the modality t. The set of all basic sentences is called the *basic syntax,* denoted χ.

If h is a player and g a node, then $g \succ h$ (or $h \prec g$) means that g follows h in the game tree. The sentence "h is reached" is denoted simply h; i.e., $h := \bigwedge_g a_g^h$, where the conjunction is over all players g that precede h, and a_g^h is the action at g that leads to h. The set of all actions of h is denoted A_h, and the set of all players H.

6.2. Basic 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.

And, for all sentences f and g, and all players h,

[24] A.k.a. "formula" in the logic literature. The term "sentence," which, too, is used in the logic literature, seems conceptually more apt and indicative of this object's role.

(2) $\bigvee a_h$, the disjunction being over all actions a_h at h.
(3) $\neg(a_h \wedge a'_h)$, where a_h and a'_h are different actions at h.
(4) $b_h(f \to g) \to (b_h f \to b_h g)$.
(5) $b_h f \to \neg b_h \neg f$.
(6) $\neg b_h f \to b_h \neg b_h f$.
(7) $a_h \leftrightarrow b_h a_h$ for all actions a_h at h.
(8) $b_h h$.

Axioms (2) and (3) say that each player chooses exactly one action. (4) and (5) represent classical modal belief axioms (see, e.g., Chellas, 1980), with clear conceptual content. (6) is "negative introspection:" that if you do not believe something, then you believe that you do not believe it; it is known to entail "positive introspection," that if you believe something, then you believe that you believe it. (7) and (8) say that h believes that he chooses the action that he indeed chooses, and that he is reached.

A *list* $\mathfrak{L}$ is a set of sentences in χ'. It is *logically closed* if it satisfies *modus ponens*:

(9) $f \in \mathfrak{L}$ and $f \to g \in \mathfrak{L}$ implies $g \in \mathfrak{L}$;

epistemically closed if it satisfies *generalization*:

(10) $f \in \mathfrak{L}$ implies $b_h f \in \mathfrak{L}$ for all players h;

tautologically closed if

(11) $f \in \mathfrak{L}$ implies $tf \in \mathfrak{L}$;

closed if it satisfies (9) and (10); and *strongly closed* if it satisfies (9), (10) and (11). The *(strong) closure* of $\mathfrak{L}$ is the smallest (strongly) closed list that includes $\mathfrak{L}$. Conditions (9–11) are often called *inference rules*.

A sentence f is called a *basic tautology* if it is in the closure of the list of all basic sentences having one of the forms (1–8). The set of all basic tautologies is denoted $\mathfrak{B}$.

6.3. The logic of tautologies

In ordinary usage, a "tautology" is a statement that is necessarily true—simply by the laws of logic and grammar—and does not make a substantive assertion about the real world. A tautology in the syntax χ—i.e., a basic tautology—is a basic sentence that follows from the axioms: i.e., holds no matter what the players actually do and believe. Similarly in the full syntax χ', tautologies should hold no matter what the players actually do and believe. Formally, we define a *tautology* as a sentence in the strong closure of the following list:

(1) Axioms[25] (6.2.1)–(6.2.8);
(2) $\neg tf$, when f is a basic sentence that is not a basic tautology;
(3) $t(f \to g) \to (tf \to tg)$, for all sentences f and g.

[25] (6.2.1) stands for Axiom 1 in Section 6.2. A similar convention is used throughout, also for numbered formulas, definitions and assertions.

Denote the set of all tautologies by $\mathfrak{T}$. Write $\vdash f$ if $f \in \mathfrak{T}$. Call a list $\mathfrak{L}$ *coherent* if there is no f for which both f and $\neg f$ are in $\mathfrak{L}$.

Theorem D.

(4) $\mathfrak{T}$ is coherent,
(5) a basic sentence is a tautology iff it is a basic tautology; and
(6) for every sentence f there is a basic sentence f' with $\vdash f \leftrightarrow f'$;

Proof. See Appendix A. □

There is an important conceptual difference between tautologies, as just defined, and basic tautologies, defined in Section 6.2. A basic tautology f is *provable* from the axioms: it is possible to write a finite string of sentences ending with f, in which each sentence is either an axiom or follows from the previous sentences using the rules of modus ponens (9) and generalization (10). In contrast, a tautology as just defined need not be provable in this sense. The reason lies in (2) above, which says that if a basic sentence f is not a tautology, then it is a tautology that it is not a tautology. But proving $\neg t f$ would seem to require examining all the basic tautologies to see that f is not among them; and that is not a finite process. So while our concept of tautology is perfectly well-defined, it is not equivalent to provability.

We end this section by setting forth some terminology. A sentence f is *inconsistent* if its negation is a tautology; otherwise it is *consistent*. It *entails* g if $f \to g$ is a tautology. It is *consistent with* g if $f \wedge g$ is consistent. The sentences $f_1, f_2, \ldots$ 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.

6.4. Semantics

The notion of strong belief, which plays a central role in our theorem, depends crucially on that of consistency. Proving consistency of a sentence f—i.e., proving $\neg t(\neg f)$—is a tricky matter. As noted above, it would seem to require writing down all tautologies and checking that $\neg f$ is not among them—which seems impossible. To cope with this difficulty, we construct a semantic formalism for our syntax.

For each player h, denote by $\mathbf{A}_{\not=h}$ the set of profiles of actions of players other than h and not preceding him. Define a *simple model*[26] of the syntax χ as a mapping $\mathbf{C}$ that assigns to each action a_h of each player h, a non-empty subset $\mathbf{C}(a_h)$ of $\mathbf{A}_{\not=h}$.

Conceptually, a profile $\mathbf{a} := (a_h)_{h\in H}$ of actions constitutes a *state of the world;* the set $\mathbf{A} := \bigtimes_{h\in H} A_h$ of all such profiles is the *state space,* and subsets of $\mathbf{A}$ are *events.* A player h playing an action a_h has beliefs about the possible states. In each state $\mathbf{b}$ that he considers possible, the action b_h must be a_h, since he knows what he plays; moreover, the action b_g of each player g preceding h must be the action a_g^h leading to h. There are no other a priori restrictions on h's beliefs;

[26] We use the term "simple model" because in the literature, the term "model" has a wider meaning, of which our simple models constitute instances. See Footnote 28.

so we may describe them as a subset $\mathbf{C}(a_h) \times \mathbf{D}(a_h)$ of $\mathbf{A}$, where $\mathbf{D}(a_h) := \{a_h\} \times \bigtimes_{g:g \prec h} \{a_g^h\}$, and $\mathbf{C}(a_h)$ is a non-empty subset[27] of $\mathbf{A}_{\not\succeq h}$.

Example 6.1. $\mathbf{C}(a_h) := \bigtimes_{g:g \not\succeq h} A_g$. Here each player believes only that he has been reached; for players off the path to him, he considers any action possible.

Given a model $\mathbf{C}$ and a sentence f, denote by $\|f\|$ the *realization* of f in the model $\mathbf{C}$, i.e., the event that f *holds*. Formally, when f is basic, define $\|f\|$ inductively by:

(1) $\|a_h\| := \{\mathbf{b} \in \mathbf{A} : b_h = a_h\}$,
(2) $\|\neg f\| := \mathbf{A} \setminus \|f\|$,
(3) $\|f \vee g\| := \|f\| \cup \|g\|$, and
(4) $\|b_h(f)\| := \{\mathbf{a} \in \mathbf{A} : \mathbf{C}(a_h) \times \mathbf{D}(a_h) \subset \|f\|\}$;

it follows from Lemma 6.2 below that

(5) $\|f\| = \|g\|$ *whenever* $f \leftrightarrow g$ *is a basic tautology.*

When f is not basic, define

(6) $\|f\| := \|f'\|$, where f' is a basic sentence with $\vdash f \leftrightarrow f'$

(see (6.3.6)); by (5) and (6.3.5), $\|f\|$ does not depend on the choice of f'.

Lemma 6.2. *Every basic tautology holds in every state of every simple model.*

Proof. It suffices to show that each of the axioms (6.2.1–8) holds in every state of every simple model, and that the list of sentences holding in a given state of a given simple model is closed. These checks are standard. □

Corollary 6.3. *Every sentence that holds in some state of some simple model is consistent.*[28]

7. The Main Theorem: formal formulation

7.1. Rationality

Call an action of a player *rational* if he has no action that he believes would yield him a higher payoff. In symbols, if a_h and a'_h are actions, denote by $p(a'_h, a_h)$ the statement that it is better[29] for h to play a'_h than a_h. Define

[27] Though h believes that only nodes following a_h are possible, we must nevertheless consider beliefs regarding actions at nodes that follow other actions, in order to assess the rationality of choosing a_h. As for actions at nodes that do not follow h, these really play no substantive role; they are there for notational convenience only.

[28] In the language of formal logic, Corollary 6.3 says that the class of simple models is *sound.* We do *not* assert that it is *complete,* which is the converse of soundness—that every consistent sentence holds in some state of some simple model. Indeed, we don't need completeness, as what interests us is proving consistency. So to keep our formalism as simple as possible, we use a restricted class of models (see Footnote 26).

[29] I.e., better for h, given the actions of players after h.

(1) $r(a_h) := \bigwedge_{a'_h \in A_h} \neg b_h p(a'_h, a_h)$;

$r(a_h)$ says that the action a_h is rational. Next, define

(2) $r_h := \bigwedge_{a_h \in A_h} (a_h \to r(a_h))$;

r_h says that h is *rational*—that he plays only rational actions. Finally, define

(3) $r := \bigwedge_{h \in H} r_h$;

r says that action rationality obtains—i.e., that all players are rational.

On its face, it is not clear that $p(a'_h, a_h)$ is a sentence in the syntax. To see that it is, let $\mathbf{a}_{\succ h}$ be a profile of actions of players after h, and $\mathbf{a}^{\wedge}_{\succ h}$ their conjunction. Together with an action a'_h, the profile $\mathbf{a}_{\succ h}$ determines an outcome $z(\mathbf{a}_{\succ h}, a'_h)$, and so a payoff $u_h(z(\mathbf{a}_{\succ h}, a'_h))$ to h. Then $p(a'_h, a_h)$ is the disjunction of all those conjunctions $\mathbf{a}^{\wedge}_{\succ h}$ for which $u_h(z(\mathbf{a}_{\succ h}, a'_h)) > u_h(z(\mathbf{a}_{\succ h}, a_h))$; so it is indeed in the syntax, so r_h and r are also in the syntax.

7.2. Common strong belief and the Main Theorem

Say that a sentence g is *strongly believed* (written *sbg*) if each player h believes g whenever g is consistent with h being reached; i.e.,

(1) $sbg := \bigwedge_{h \in H} [b_h g \vee t\neg(h \wedge g)]$.

Mutual strong belief of g of order $n \geq 0$ (written $sb^n g$) is defined inductively by

(2) $sb^n g := g^0 \wedge g^1 \wedge \ldots \wedge g^n$,

where $g^0 := g$ and[30] $g^{m+1} = sb(g^0 \wedge g^1 \wedge \ldots \wedge g^m)$ for all m. So, for $n \geq 1$,

(3) $sb^n g = sb^{n-1} g \wedge sb(sb^{n-1} g)$;

thus each iteration provides for the previous iteration and strong belief thereof. *Common strong belief* of g (written[31] *csbg*) comprises all[32] $sb^n g$ for all n.

The Main Theorem is stated in Section 1, and restated in Section 4.2.

8. The Main Theorem: formal proof

One part of the demonstration of the Main Theorem—that the Pruning Process (PP) yields the BI outcomes (Theorem C)—is rigorous. But the other, which identifies stages of the PP with iterated strong belief of rationality (Theorem B), while sounding convincing, is not entirely rigorous. We now complete the proof of the Main Theorem by proving Theorem B rigorously.

[30] So $sb^n g = sb^{n-1} g \wedge sb(g \wedge sb^{n-1} g)$ for $n > 1$, which in turn yields, by induction, that all $sb^n g$ are consistent with all h. For $n = 1$, if g is consistent with h, then so is $b_h g$.

[31] *csbg* is not a sentence, but an infinite conjunction of sentences.

[32] Note that *csbg* asserts, inter alia, that g itself is actually true.

Let T be the tree of an unrestricted[33] PI game, T^n the subtree[34] that survives stage n of the PP. By (7.2.2), $r \wedge r^1 \wedge \ldots \wedge r^n = sb^n r$; thus Theorem B asserts that a node h is in T^{n+1} iff it is reachable under $sb^n r$—i.e., iff the sentence $h \wedge sb^n r$ is consistent. Consistency proofs use semantic models (Section 6.4). So, we define a simple model $\mathbf{C}$ of the syntax χ; in it, a player h playing an action a_h believes in the maximum degree of iterated elimination of dominated actions (of all players) that allows h and for which a_h is rational. We will show that in this model, the nodes that may be reached when $sb^n r$ holds are precisely those in T^{n+1}.

Let T' be a subtree of T; think of it as comprising actions as well as nodes. Say that an action a_h in T' is *dominated in* T' if there is an action a'_h in T' that is better for h than a_h no matter what is done subsequently, as long as all subsequent actions are in T'. For each positive integer n and player h, let m be the largest integer $\leq n-1$ with $h \in T^m$, and let A_h^n be the set of all actions of player h in T^m that are undominated in T^m. Then define

$$\mathbf{A}^n := \bigtimes_{h \in H} A_h^n;$$

note that the $\mathbf{A}^n$ are nested. For each undominated action a_h, let $n(a_h)$ be the greatest[35] m for which a_h is in T^m and is undominated in T^m. Define a simple model $\mathbf{C}$ by

$$\mathbf{C}(a_h) := \bigtimes_{g \not\succeq h} A_g^{n(a_h)}.$$

Given an action a_h and a profile $\mathbf{a}_{\succ h}$ of actions after h, say that h, *when playing* a_h, *believes in* $\mathbf{C}$ *that* $\mathbf{a}_{\succ h}$ *is possible,* if $\mathbf{a}_{\succ h}$ is the restriction to the players after h of a profile $\mathbf{a}_{\not\succeq h}$ in $\mathbf{C}(a_h)$. Call a_h *belief-dominated in* $\mathbf{C}$ if h has an action a'_h that yields him a better payoff[36] than a_h, for all $\mathbf{a}_{\succ h}$ that he believes in $\mathbf{C}$ are possible when playing a_h. Denote by $\|f\|$ the realization of f in the model $\mathbf{C}$. It may be verified that

Remark 8.1. A state $\mathbf{a}$ in the model $\mathbf{C}$ is in $\|r\|$ iff no action a_h is belief-dominated in $\mathbf{C}$.

Lemma 8.2. *No action is belief-dominated in* $\mathbf{C}$*, unless it is dominated in* T.

Proof. For brevity and clarity, in this proof the term "belief" will mean "belief in $\mathbf{C}$."

Let a_h be undominated in T. Thus the first line in the definition of $n(a_h)$ applies; so setting $n := n(a_h)$, we have (i) a_h is in T^n and is undominated in T^n, and (ii) $\mathbf{C}(a_h) := \bigtimes_{g \not\succeq h} A_g^n$. So if h plays a_h, he believes that players g after h play actions in A_g^n. We shall show (iii) a_h is not belief-dominated by any action b_h at h in T.

First we show (iv) a_h is not belief-dominated by any action b_h in T^n. Suppose it is. Then the player g immediately following b_h is also in T^n. Player h believes that if he would play b_h, then g would[37] play an action b_g in A_g^n, i.e., one that is undominated in T^{n-1}. So since $g \in T^n$, also $b_g \in T^n$. Applying the same argument repeatedly, we conclude that h believes that b_h would lead

[33] Not necessarily generic.

[34] A *subtree* T' of a tree T is obtained from T by eliminating one or more branches. Thus T and T' have the same root.

[35] There always is a greatest such m, since there are only finitely many T^n, as the PP is a finite process.

[36] I.e., $u_h(z(\mathbf{a}_{\succ h}, a'_h)) > u_h(z(\mathbf{a}_{\succ h}, a_h))$.

[37] Please see Footnote 27.

to an outcome in T^n. Conversely, it may be seen that if h would play b_h, then he believes that any outcome in T^n after b_h would be possible.

Since a_h is in T^n, it may be seen similarly that h believes that an outcome following a_h is possible iff it is in T^n. So by (i), b_h cannot belief-dominate a_h.

Next, we show that (v) a_h is not belief-dominated by any action b_h in T^m, for any $m \leq n$; this yields (iii). The proof is by (backward) induction on m. For $m = n$, we have just proved it. Assume it true for a given m, and let b_h be in T^{m-1}. If b_h is in T^m, then (v) is the induction hypothesis. So let b_h be in $T^{m-1} \setminus T^m$. Since $b_h \notin T^m$, it is dominated in T^{m-1}, say by c_h. If $c_h \notin T^m$, then it is dominated in T^{m-1} by an action c'_h; by the transitivity of domination, c'_h dominates b_h in T^{m-1}. Continuing in this way, we eventually reach an action d_h in T^m that dominates b_h in T^{m-1}. But h believes that players after b_h play actions in T^{m-1}; so since d_h dominates b_h in T^{m-1}, and b_h belief-dominates a_h, it follows that d_h belief-dominates a_h. But by (iv), that cannot be. □

If $\mathbf{a} \in \mathbf{A}$, denote by $\mathbf{a}^\wedge$ the conjunction $\bigwedge_{h \in H} a_h$ of the actions in $\mathbf{a}$.

Proposition 8.3. *For each $n \geq 0$,*

(1_n) *If $\mathbf{a}^\wedge$ is consistent with $sb^n r$, then $\mathbf{a} \in \mathbf{A}^{n+1}$; and*
(2_n) $\mathbf{A}^{n+1} = \|sb^n r\|$.

Together, (1_n) *and* (2_n) *yield*

(3_n) *a player is in T^{n+1} iff reaching that player is consistent with $sb^n r$.*

Remark 8.4. That (3_n) obtains for all n is precisely Theorem B, which is what is needed to complete the proof of the Main Theorem.

Proof. We first show that (1_n) and (2_n) yield (3_n). To show "if," let h be consistent[38] with $sb^n r$. Then there is a profile $\mathbf{a}$ in $\mathbf{A}$, with the actions a_g of players g before h leading to h, such that $\mathbf{a}^\wedge$ is consistent with $sb^n r$. So from (1_n) we get $\mathbf{a} \in \mathbf{A}^{n+1}$, and it follows that $h \in T^{n+1}$.

To show "only if," let $h \in T^{n+1}$. Define a profile $\mathbf{a}$ in $\mathbf{A}$ by letting a_g lead to h for players g before h; and for other players g, letting a_g be an arbitrary member of A_g^{n+1}. Then $\mathbf{a} \in \mathbf{A}^{n+1}$, so ($2_n$) yields $\mathbf{a} \in \|sb^n r\|$. So $\mathbf{a}^\wedge \wedge sb^n r$ holds in state $\mathbf{a}$ of the model $\mathbf{C}$, so is consistent. But $\mathbf{a}^\wedge$ entails h, so also $h \wedge sb^n r$ is consistent. □

We prove (1_n) and (2_n)—and so also (3_n)—by induction on n.

(1_0). Let $\mathbf{c} \notin \mathbf{A}^1$; we prove that $\mathbf{c}^\wedge$ is inconsistent with $sb^0 r$. Since $\mathbf{c} \notin \mathbf{A}^1$, there is an h for which $c_h \notin A_h^1$. So c_h is dominated, say by d_h. So $u_h(z(\mathbf{a}_{\succ h}, d_h)) > u_h(z(\mathbf{a}_{\succ h}, c_h))$ for all $\mathbf{a}_{\succ h}$, so $p(d_h, c_h)$ is a tautology. So by generalization (6.2.10), $b_h\, p(d_h, c_h)$ is a tautology. So $\bigwedge_{a'_h \in A_h} \neg b_h\, p(a'_h, c_h)$ is inconsistent. Now r_h entails $c_h \to (\bigwedge_{a'_h \in A_h} \neg b_h\, p(a'_h, c_h))$, so $c_h \wedge r_h$ entails $\bigwedge_{a'_h \in A_h} \neg b_h\, p(a'_h, c_h)$. So $c_h \wedge r_h$ is inconsistent. But $\mathbf{c}^\wedge$ entails c_h, so $\mathbf{c}^\wedge \wedge r$ entails $c_h \wedge r_h$, which is inconsistent. So $\mathbf{c}^\wedge \wedge sb^0 r = \mathbf{c}^\wedge \wedge r$ is inconsistent. □

[38] Recall that in the syntax, h means that the player h is reached (Section 6.1).

(2_0). Let $\mathbf{a} \in \mathbf{A}^1$; thus no a_h is dominated in T. So by Lemma 8.2, no a_h is belief-dominated in $\mathbf{C}$. So by Remark 8.1, $\mathbf{a}$ is in $\|r\|$, which by (7.2.2), $= \|sb^0 r\|$. So $\mathbf{A}^1 \subset \|sb^0 r\|$. For the opposite inclusion, let $\mathbf{a} \in \|sb^0 r\| = \|r\|$. Then $\mathbf{a}^\wedge \wedge sb^0 r$ holds in the model $\mathbf{C}$, so is consistent; so 1_0 yields $\mathbf{a} \in \mathbf{A}^1$. □

Now let $n > 0$, and assume (1_{n-1}) and (2_{n-1}) (and so also (3_{n-1})).

(1_n). Let $\mathbf{c} \notin \mathbf{A}^{n+1}$; we will prove that $\mathbf{c}^\wedge$ is inconsistent with $sb^n r$. If $\mathbf{c} \notin \mathbf{A}^n$, then by $(\mathbf{1}_{n-1})$, $\mathbf{c}^\wedge$ is inconsistent with $sb^{n-1} r$, so a fortiori with $sb^n r$, which by definition entails $sb^{n-1} r$. So we may assume $\mathbf{c} \in \mathbf{A}^n \backslash \mathbf{A}^{n+1}$. So $c_h \in A_h^n \backslash A_h^{n+1}$ for some h in T^n. So some action d_h in T^n dominates c_h in T^n; that is, d_h is better than c_h, no matter what the subsequent actions are, as long as they are in T^n. So

(4) $\vdash \bigvee_{T^n} \mathbf{a}_{\succ h}^\wedge \to p(d_h, c_h)$,

the disjunction being over all profiles $\mathbf{a}_{\succ h}$ of actions after h all of which are in T^n. By (3_{n-1}), reaching a player h is consistent with $sb^{n-1} r$ iff h is in T^n. So by (7.2.1),

(5) $sb(sb^{n-1} r) = \bigwedge_{h \in T^n} b_h(sb^{n-1} r)$.

So by (7.2.3), $\vdash sb^n r \to b_h(sb^{n-1} r)$ for all h in T^n. By (1_{n-1}), $\vdash sb^{n-1} r \to \bigvee_{\mathbf{a} \in \mathbf{A}^n} \mathbf{a}_{\succ h}^\wedge$; so by generalization (6.2.10) and (6.2.4),

(6) $\vdash sb^n r \to b_h(\bigvee_{\mathbf{a} \in \mathbf{A}^n} \mathbf{a}^\wedge)$.

It may be seen that $\vdash (b_h f \wedge b_h g) \to b_h(f \wedge g)$; so by (6), (4), and (6.2.4),

$\vdash sb^n r \to b_h(\bigvee_{T^n} \mathbf{a}_{\succ h}^\wedge) \to b_h p(d_h, c_h)$.

But by (7.2.2) and (7.1.1–3), $\vdash sb^n r \wedge \mathbf{c} \to r \wedge c_h \to \neg b_h p(d_h, c_h)$, so $sb^n r \wedge c_h$ entails a contradiction; i.e., $\mathbf{c}$ is inconsistent with $sb^n r$. □

(2_n). Let $\mathbf{c} \in \mathbf{A}^{n+1}$. By (7.2.3) and (5), $\vdash sb^{n-1} r \wedge \bigwedge_{h \in T^n} b_h(sb^{n-1} r) \leftrightarrow sb^n r$; so by $(\mathbf{2}_{n-1})$,

(7) $\mathbf{A}^n \cap \bigcap_{h \in T^n} \|(b_h(sb^{n-1} r))\| = \|sb^n(r)\|$.

First note that since $\mathbf{A}^{n+1} \subset \mathbf{A}^n$, it follows that

(8) $\mathbf{c} \in \mathbf{A}^n$.

Next, let $h \in T^n$. From $\mathbf{c} \in \mathbf{A}^{n+1}$ we get $c_h \in A_h^{n+1}$, so c_h is in T^n and is undominated in T^n. So $n(c_h) \geq n$. So $\mathbf{C}(c_h) \subset \bigtimes_{g \not\succeq h} A_g^n$, since the $\mathbf{A}^n$ are nested. Next, we show that $\mathbf{D}(c_h)$, which $:= \{c_h\} \times \bigtimes_{g : g \prec h} \{c_g^h\}$, is included in $\bigtimes_{g \preceq h} A_g^n$. The action c_h itself is in A_h^n; indeed it is in A_h^{n+1}, which A_h^n includes. When g precedes h, the action c_g^h leading to h is in T^n, and leads to a player in T^n; so $c_g^h \in A_g^n$. So indeed, $\mathbf{D}(c_h) \subset \bigtimes_{g \preceq h} A_g^n$. So $\mathbf{C}(a_h) \times \mathbf{D}(a_h) \subset \bigtimes_{g \in H} A_g^n = \mathbf{A}^n$. So $(\mathbf{2}_{n-1})$ and (6.4.4) yield

$$\begin{aligned} \mathbf{c} &\in \{\mathbf{a} \in \mathbf{A} : \mathbf{C}(a_h) \times \mathbf{D}(a_h) \subset \mathbf{A}^n\} \\ &= \{\mathbf{a} \in \mathbf{A} : \mathbf{C}(a_h) \times \mathbf{D}(a_h) \subset \|sb^{n-1} r\|\} = \left\| b_h(sb^{n-1} r) \right\|; \end{aligned}$$

together with (7) and (8), this yields $\mathbf{c} \in \|sb^n(r)\|$. So $\mathbf{A}^{n+1} \subset \|sb^n r\|$.

For the opposite inclusion, let $\mathbf{a} \in \|sb^n r\|$. Then $\mathbf{a}^\wedge \wedge sb^n r$ holds in the model $\mathbf{C}$, so is consistent; so 1_n yields $\mathbf{a} \in \mathbf{A}^{n+1}$. □

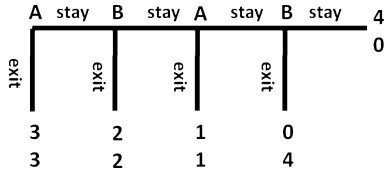

Fig. 1. Reny's game.

9. Discussion

9.1. Players and agents

9.1.1. Forward inferences

To understand the substantive contribution of this work, it is useful to relate it to the seminal work of BS (Battigalli and Siniscalchi, 2002). Like us, BS show that common strong belief (*csb*) of rationality yields the BI outcome.[39] Unlike us, they refer to general PI games, in which players may have several decision nodes; whereas we refer to *simple* PI games, in which each player has just one node. Equivalently, one may say that both BS and we consider general PI games G; but we assume *csb* of the *agents'* rationality,[40] whereas BS assume *csb* of the *players'* rationality.[41]

The game G of Fig. 1 (Reny, 1992) illustrates the distinction. First assume *csb* of the agents' rationality (r). By Remark 2.1, Bob_1 is rational and believes that all agents—in particular Ann_1—are rational; there are no further restrictions on his beliefs.[42] So he may well believe that Ann_2 exits, in which case he, too, should exit.

Now assume *csb* of the *players'* rationality. If Bob's first node is reached, he *can* impute rationality to Ann; but only if he believes that she believes that he will stay at his second node. In that case, his only rational strategy is to stay at his first node and exit at his second. Thus, if he is rational and imputes rationality to Ann, he should *stay* at his first node.[43]

That is a *forward* inference. At his first node, Bob asks, what can I infer from the past about the future? What did Ann think when she did what she did, and what can I infer about what she *will* do? So we're presuming that the player Ann chooses a strategy that dictates what both her agents do. Forward inferences depend on *players* choosing *strategies* rationally; they are impossible in simple games, where each player has just one node.

Note that *csb* of player rationality uses backward as well as forward inferences. After using a forward inference to decide what Ann_2 will do, Bob_1 reasons *backward* from Ann_2's inferred action to decide what he should do. Non-simple games involve both kinds of inference.

9.1.2. Backward inferences

Recall the BI reasoning, as set forth at the start of the paper: The last player, who must choose between outcomes, chooses an action that maximizes his payoff; taking this as given, the previous player maximizes *his* payoff; and so on, until the beginning of the game is reached.

[39] We do not formulate their definition or result precisely here.

[40] I.e., *csbr* in the simple game G' that is like G, except that the players in G' are the agents in G.

[41] Their definition of rationality is based on utility maximization rather than probability 1 belief.

[42] He can't believe $r \wedge r^1$, as Ann_1's rationality and her belief in Bob_2's rationality would require her to exit, which would precludes reaching him.

[43] We are reasoning here with pure strategies. Similar, but more complicated, considerations apply to mixed strategies.

These are backward inferences. Starting from the outcomes, we infer *back* to the actions of preterminal players; then *back* to prepreterminal players; and so on. No forward inferences are involved; indeed, the concept of "player" plays no role in BI—the agents act independently, as if the game were simple and they the players. Nothing in the BI reasoning connects the agents of a player to each other; *strategies*—as opposed to actions—play no role.

Basically, therefore, the BI reasoning applies only to simple games. By extension, it applies also to general PI games G in which the agents act independently. So to clarify the epistemic assumptions underlying BI, one replaces G by a simple game G' whose players are the agents in G; this rules out forward inferences. And then, *csbr* yields the BI outcome.

9.2. *Consistency and completeness*

To understand the methodological contribution of this work, it is, as with its substantive contribution, useful to relate it to BS (Battigalli and Siniscalchi, 2002).

Our result is formulated syntactically; BS's, semantically. In most applications, semantic results may also be formulated syntactically, and vice versa. In the present application, the matter is less straightforward.

To explain, we start by describing the usual relationship between semantic and syntactic formalisms more carefully. Each sentence in a syntactic formalism corresponds to a set in each semantic state space—conceptually, the set of states in that space 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 space); and a tautology in the syntax corresponds to the entire state space, since it must hold at each state. Conversely, if a sentence in the syntax corresponds in each arbitrary state space to the entire state space, then it is a tautology.

All that is well and good as long as the tautology operator t (Section 5) is not in the formal object language, but only in the metalanguage. As soon as t becomes part of the object language, 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 state space, since it refers simultaneously to *all* state spaces. For a sentence to be a tautology means that in *each* state space, the sentence corresponds to the entire space; and there is no way of saying that within a particular space.

BS work with semantics, so that is an obstacle for them. They overcome it by working not with arbitrary state spaces, as is usual in the semantic approach, but with one particular one, called the *complete type space;* this has the property that if in the complete type space, a sentence corresponds to the entire space, then in every semantic state space, that sentence corresponds to the entire space. Constructing the complete type space, and proving that it has the basic property just enunciated, is not elementary; see Battigalli and Siniscalchi (1999). But such an object does exist; indeed, has by now become familiar in epistemic game theory.

Now, this complete type space enables a valid semantic representation of sentences involving the tautology operator t. Namely, the tautology operator corresponds to a set operator that takes the entire complete type space to itself, and all its proper subsets to the empty set. Formally, then, BS prove that the BI outcome is reached at any element of the complete type space at which there is *csb* of player rationality.[44]

[44] See Footnote 41.

A reader has remarked that completeness of the type space requires all possible types to be present; this, he says, is an assumption on the players' reasoning that the syntactic analysis "hides," and that should be made explicit. But that is precisely the beauty of the syntactic analysis! With it, completeness shows up as nothing but plain logical reasoning. On the contrary, it is incomplete type spaces that make hidden implicit assumptions—namely, that certain types are *absent*—that go beyond plain logical reasoning.

10. Literature

10.1. Battigalli–Siniscalchi and related literature

The literature on the foundations of BI is far too large to survey here, so we will review only the most directly relevant work. Foremost is BS (Battigalli and Siniscalchi, 2002). Though in substance, their result is utterly different from ours—theirs deals with forward and backward inferences, ours with backward inferences only—there are interesting parallels in form. For one thing, the notion of strong belief, originated by BS, plays a central role also in our theorem. For another, the *formulations* of the results are roughly parallel: Our result characterizes *csb* of the agents' rationality (r); theirs, of the players' rationality[45] (pr). Finally, the proofs are analogous in grand structure, though BS's proof lies far deeper than ours—as indeed the result itself does.

Like ours, BS's proof has two parts. In one—analogous to our Theorem B—they show that *csbpr* is equivalent to extensive form rationalizability (EFR) a la Pearce (1984) or Battigalli (1997). In the other—analogous to our Theorem A—they recall that EFR entails the BI outcome, as proved by Reny (1992), who used tools from algebraic topology developed by Kohlberg and Mertens (1986) (see also Battigalli, 1997).[46] An "elementary" proof[47] of this result is obtained by applying to PI games a theorem of Chen and Micali (2013) about general extensive games. Looking at PI games only, Heifetz and Perea (2015) obtained an alternative elementary proof that EFR entails the BI outcome.

Battigalli and Friedenberg (2012) show that completeness is indispensable for the BS result. As already mentioned, completeness is implicit in the syntactic approach.

10.2. Other literature

Perhaps the first to provide an epistemic foundation for BI based on action rationality was Stalnaker (1998). Like us, he treats simple games only. But unlike our players, whose beliefs are predicated on being reached, his players have prior beliefs at the beginning of play, which are updated as play proceeds. He assumes common prior belief of rationality (*cbr*), and that the prior beliefs of each player about other players are in a sense independent. This implies that at each node v, there is *cbr* among the players at and after v; together with rationality, this yields BI.

In broad outline, our paper is conceptually similar. One would like to assume *cbr*; like with *ckr*, this leads to difficulties off the BI path. So at off-path nodes, one somehow weakens *cbr*. *Stalnaker's* players do so by saying, well, I see that the players before me did not conform to *cbr*. But by independence, that says nothing about the players after me; so among *them,* I retain *cbr*.

[45] See Footnote 41.

[46] Professor Marciano Siniscalchi has pointed out that in simple PI games, EFR is in fact equivalent to the Pruning Process. The proof is not entirely straightforward.

[47] One not using algebraic topology.

Our players say, for me, any deviation from *cbr* by some players indicates that others may be similar.[48] So I abandon *cbr*, maintaining only as much mutual belief of rationality as is possible at my node.

Note that Stalnaker gets not only the BI outcome, but also the BI *strategies,* and them *only.* In contrast, *csbr* allows the BI strategies, but not *only* them.[49] And *csbpr* may exclude the BI strategies altogether (Section 9.1.1).

Other approaches that determine the BI strategies include Aumann (1995), Asheim and Perea (2005), and Perea (2008).

Halpern and Pass (2009) define a semantic operator $\diamond$ that may be interpreted as consistency w.r.t. a given class of models; from it, one may derive a tautology operator w.r.t. that class. Though it is not a priori clear, it might be possible to base a semantic treatment parallel to our syntactic treatment on this operator, rather than on the complete type space.

10.3. Genericity

Though the conceptual insights of this paper are most easily seen under genericity, we in fact require no genericity at all. In contrast, BS require the game to have "no relevant ties" (NRT) (see also Battigalli, 1997); i.e., that the last common ancestor of any two outcomes have different payoffs at those outcomes. NRT ensures a unique BI outcome. The requirement has teeth; e.g., it excludes chess. Heifetz and Perea (2015) also use NRT.

Without any genericity condition, EFR still implies a BI outcome. But there are BI outcomes that are inconsistent with EFR; see Chen and Micali (2013), p. 149, Example 7.

Appendix A. Proof of Theorem D

Lemma A.1. *The list* $\mathfrak{B}$ *of all basic tautologies is coherent.*

Proof. Follows from Lemma 6.2 and Example 6.1. □

Let 1 denote an arbitrary but fixed basic tautology (like $a_h \vee \neg a_h$), and set $0 := \neg 1$. Define a mapping from sentences $f \in \chi'$ to basic sentences $f' \in \chi$ inductively as follows:

(1) $(a_h)' := a_h$ for every action a_h.
(2) $(f \vee g)' := f' \vee g'$.
(3) $(\neg f)' := \neg f'$.
(4) $(b_h f)' := b_h(f')$.
(5) $(tf)' := 1$ if $f' \in \mathfrak{B}$ and $(tf)' = 0$ if $f' \notin \mathfrak{B}$.

It may be seen that $f' = f$ for every basic sentence f. Now let

(6) $\mathfrak{L} = \{f : f' \in \mathfrak{B}\}$.

Lemma A.2. $\mathfrak{L}$ *is coherent, strongly closed, and includes all tautologies.*

[48] This is, after all, a forward inference of sorts. Lessons from the past are applied to the future, if only to abandon previously held beliefs. In contrast, Stalnaker's agents learn nothing from the past.

[49] Thus in Reny's game (Fig. 1), BI calls for Bob_1 to exit, whereas *csbr* allows him either to stay or to exit.

Proof. We first prove coherence. Suppose f and $\neg f$ are in $\mathfrak{L}$. Then by definition, f' and $(\neg f)'$ are basic tautologies. But $(\neg f)' = \neg f'$, so the list of basic tautologies is incoherent, contradicting Lemma A.1.

To see that $\mathfrak{L}$ is logically closed, assume $f \in \mathfrak{L}$ and $(f \to g) \in \mathfrak{L}$. Then by definition, $f' \in \mathfrak{B}$ and $(f \to g)' \in \mathfrak{B}$. But $(f \to g)' = f' \to g'$; so $g' \in \mathfrak{B}$, which entails $g \in \mathfrak{L}$. That $\mathfrak{L}$ is epistemically closed follows similarly from $\mathfrak{B}$ being epistemically closed. To see that $\mathfrak{L}$ is tautologically closed, let $f \in \mathfrak{L}$; then by definition, $f' \in \mathfrak{B}$, so $(t(f))' = 1$, so $t(f) \in \mathfrak{L}$.

That $\mathfrak{L}$ contains all sentences in (6.3.1) follows from k' being an axiom if k is.

That $\mathfrak{L}$ contains all sentences in (6.3.2) follows from the definition of $\mathfrak{L}$.

To see that $\mathfrak{L}$ contains all sentences in (6.3.3), let $k = t(f \to g) \to (t(f) \to t(g))$. If $(t(f \to g))' = 0$ then clearly k' is in $\mathfrak{B}$. If $(t(f \to g))' = 1$, then $f' \to g' \in \mathfrak{B}$; so since $\mathfrak{B}$ is logically closed, g' is in $\mathfrak{B}$ whenever f' is in $\mathfrak{B}$; so $(tf \to tg)'$ is in $\mathfrak{B}$. □

Lemma A.3. $\vdash f \leftrightarrow f'$ *for every sentence* f.

Proof. By induction. For basic f it is clear, since $f' = f$. Next, let $f = b_h(g)$ for some g. By the induction hypothesis, $\vdash g \leftrightarrow g'$, so since $\mathfrak{T}$ is epistemically closed, $\vdash b_h(g \leftrightarrow g')$. Now by axiom (6.2.4), $\vdash b_h(g \leftrightarrow g') \to (b_h g \leftrightarrow b_h(g'))$. So since $\mathfrak{T}$ is logically closed, $\vdash b_h(g) \leftrightarrow b_h(g')$; i.e., $\vdash f \leftrightarrow f'$.

If $f = tg$, then by the induction hypothesis,

(7) $\vdash g \leftrightarrow g'$.

So since $\mathfrak{T}$ is tautologically closed, $\vdash t(g \leftrightarrow g')$. But $\vdash t(g \leftrightarrow g') \to (tg \leftrightarrow t(g'))$, by (6.3.3). So since $\mathfrak{T}$ is logically closed,

(8) $\vdash tg \leftrightarrow t(g')$.

If $g' \in \mathfrak{B}$, then $\vdash g$, by (7); so $\vdash tg$, since $\mathfrak{T}$ is tautologically closed; by (5), $(tg)' = 1$, so $\vdash (tg)'$; so

(9) $\vdash tg \leftrightarrow (tg)'$.

If $g' \notin \mathfrak{B}$, then $\vdash \neg t(g')$, by (6.3.2); by (5), $(tg)' = 0$, so $\vdash \neg(tg)'$; so $\vdash t(g') \leftrightarrow (tg)'$; so by (8), we again get (9). So (9) holds in either case; i.e., $\vdash f \leftrightarrow f'$.

If $f = \neg g$ or $f = g \vee k$, the induction hypothesis easily yields $\vdash f \leftrightarrow f'$. □

Lemma A.4. $\mathfrak{T} \cap \chi = \mathfrak{B}$.

Proof. Since $\mathfrak{T}$ is closed and contains the basic axioms (6.2.1–8), it follows that $\mathfrak{B} \subset \mathfrak{T}$. But $\mathfrak{T} \subset \mathfrak{L}$ by Lemma A.1, and $\mathfrak{L} \cap \chi = \mathfrak{B}$; so $\mathfrak{T} \cap \chi = \mathfrak{B}$. □

Lemma A.5. $\mathfrak{T} = \mathfrak{L}$.

Proof. By Lemma A.3, a sentence f is in $\mathfrak{T}$ iff f' is in $\mathfrak{T}$. But f' is basic and $\mathfrak{T} \cap \chi = \mathfrak{B}$ by Lemma A.4, so $\mathfrak{T} = \{ f \mid f' \in \mathfrak{B}\} = \mathfrak{L}$. □

(6.3.4) now follows from Lemmas A.2 and A.4, (6.3.5) from Lemma A.4, and (6.3.6) from Lemma A.3. □

References

Asheim, G., Perea, A., 2005. Sequential and quasi-perfect rationalizability in extensive games. Games Econ. Behav. 53, 15–42.

Aumann, R.J., 1995. Backward induction and common knowledge of rationality. Games Econ. Behav. 8, 6–19.
Aumann, R.J., 1996. Reply to Binmore. Games Econ. Behav. 17, 138–146.
Battigalli, P., 1997. On rationalizability in extensive games. J. Econ. Theory 74, 40–61.
Battigalli, P., Friedenberg, A., 2012. Forward induction reasoning revisited. Theor. Econ. 7, 57–98.
Battigalli, P., Siniscalchi, M., 1999. Hierarchies of conditional beliefs and interactive epistemology in dynamic games. J. Econ. Theory 88, 188–230.
Battigalli, P., Siniscalchi, M., 2002. Strong belief and forward induction reasoning. J. Econ. Theory 106, 356–391 (BS).
Binmore, K., 1996. A note on backward induction. Games Econ. Behav. 17, 135–137.
Chellas, B.F., 1980. Modal Logic: An Introduction. Cambridge University Press, Cambridge–New York.
Chen, J., Micali, S., 2013. The order independence of iterated dominance in extensive games. Theor. Econ. 8, 125–163.
Halpern, J.Y., Lakemeyer, G., 2001. Multi-agent only knowing. J. Log. Comput. 11, 41–70.
Halpern, J.Y., Pass, R., 2009. A logical characterization of iterated admissibility. In: Theoretical Aspects of Rationality and Knowledge: Proc. Twelfth Conference. TARK 2009, pp. 146–155.
Heifetz, A., Perea, A., 2015. On the outcome equivalence of backward induction and extensive form rationalizability. Int. J. Game Theory 44, 37–59.
Kohlberg, E., Mertens, J.F., 1986. On strategic stability of equilibria. Econometrica 54, 1003–1037.
Pearce, D., 1984. Rationalizable strategic behavior and the problem of perfection. Econometrica 52, 1029–1050.
Perea, A., 2008. Minimal belief revision leads to backward induction. Math. Soc. Sci. 56, 1–26.
Reny, P., 1992. Backward induction, normal form perfection and explicable equilibria. Econometrica 60, 627–649.
Stalnaker, R., 1998. Belief revision in games: forward and backward induction. Math. Soc. Sci. 36, 31–56.