האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

THE LOGIC OF BACKWARD INDUCTION

By

ITAI ARIELI and ROBERT J. AUMANN

Discussion Paper # 652 November 2013

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/

The Logic of Backward Induction*

Itai Arieli,† Robert J. Aumann‡

Abstract

The logic of backward induction (BI) in perfect information (PI) games has been intensely scrutinized for the past quarter century. A major development came in 2002, when P. Battigalli and M. Siniscalchi (BS) showed that an outcome of a PI game is consistent with common strong belief of utility maximization if and only if it is the BI outcome. Both BS's formulation, and their proof, are complex and deep. We show that the result continues to hold when utility maximization is replaced by a rationality condition that is even more compelling; more important, the formulation and proof become far more transparent, accessible, and self-contained.

1 Introduction

The logic of—and rationale for—backward induction (BI for short) in perfect information (PI) games has been intensely scrutinized for the past quarter century. A major development was the result of P. Battigalli and M. Siniscalchi [4, 2002] (BS), that an outcome of a PI game is consistent with common strong belief (*csb*) of utility maximization (*um*) if and only if (iff)

*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 21. Needless to say, these gentlemen are not responsible for errors, misstatements, or expressions of opinion herein.

†Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology. Email: iarieli@tx.technion.ac.il

‡Department of Mathematics and Center for the Study of Rationality, the Hebrew University of Jerusalem. Aumann thanks the Israel Science Foundation for research support. Email: raumann@math.huji.ac.il

it is the[1] BI outcome. Here "belief" means attribution of probability 1, and a player "strongly believes" a proposition if he believes it whenever it is logically possible. "Common" strong belief of a proposition means that it is true, all players strongly believe it, all strongly believe the foregoing, all strongly believe the foregoing, and so on ad infinitum. The proof of the BS result, and indeed even its precise formulation, are complex and deep.

The result presented here is similar in form and content to that of BS, but its formulation and proof are far more direct and transparent. The most visible difference is that we replace utility maximization by *action rationality* (simply *rationality* (r) for short); this means that a player never takes a *belief-dominated* action—one for which another action is available that he believes (with probability 1) would yield him a higher payoff, no matter what is done subsequently.[2] We then establish our

Main Theorem: *An outcome of a PI game is consistent with common strong belief of action rationality* ($csbr$) *iff it is a*[3] *BI outcome.*

In practice, action rationality is weaker than utility maximization*:* utility maximizers cannot take belief-dominated actions.[4] Thus our condition ($csbr$) appears weaker than that of BS ($csbum$), so the result that it implies BI appears stronger. Strictly speaking, that is not correct; neither result entails the other[5]. Nevertheless, it does seem reasonable to conclude that $csbr$ is at least as compelling as $csbum$, if not more so.

But our main interest lies not in the relative strengths of the two results, which when all is said and done, convey similar messages. Rather, what we provide here is a direct, uncomplicated, and accessible reformulation and proof of BS's landmark insight.

Specifically, our treatment differs from BS's in three important ways:
(1) **Proof Transparency:** BS first show that $csbum$ is equivalent to extensive form rationalizability (EFR), a task that is by no means easy. They then cite known results—of considerable depth—according to which EFR leads to the BI outcome. In contrast, in our treatment it is fairly clear that $csbr$ is equivalent to iterated elimination of strongly dominated actions,[6] which is shown in a few lines to yield the BI outcomes.

[1] In general, PI games do not have unique BI outcomes; BS use a genericity condition that does imply uniqueness.

[2] Even by that player himself.

[3] Our result is completely general—we use no genericity condition—so there may be several essentially different BI outcomes.

[4] The converse is false: Action rational players need not be utility maximizers.

[5] See Section 10.2.

[6] Henceforth called the *Pruning Process.*

(2) **Belief:** BS work with utility maximization, so must consider all probabilities from 0 to 1. In contrast, we use probability 1 belief only; simple, easily formulated properties of this concept suffice for our proof, obviating the need for the complex logical apparatus required by numerical probabilities.
(3) **Syntax:** The notion of common strong belief involves that of logical impossibility, whose natural formulation is syntactic. BS use semantics, perhaps because it is then easier to deal with numerical probabilities; but this requires using complex, indirect methods to formulate the notion of logical impossibility.

Two other differences are worth noting. One is in the matter of genericity. The genericity conditions that are required to show that *csbum* yields BI are by no means innocent; inter alia, they exclude chess, the queen of PI games. Indeed, *csbum* does *not* in general yield BI. As already noted, our treatment requires no genericity at all.

The second noteworthy difference is in the matter of strategies. Whereas *csbum* yields the BI *outcome,* it does not yield the BI *strategies*—those prescribing the BI action for every move of each player—even with genericity. In contrast, *csbr* does yield the BI strategies.

The plan of the paper is as follows: In Section 2 we formulate the result, and in Section 3 demonstrate it, carefully but still informally, assuming genericity; this may already satisfy many readers. The remainder of the paper has three parts, any or all of which may be skipped: Section 4 removes the genericity condition, Sections 5-8 comprise the formal development, and Sections 9-10 are devoted to discussion. 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 discusses the concept of action rationality. Section 10 reviews some of the literature, including BS.

Finally, our hats are off to BS. Finding an easier path to their discovery, far from detracting from its importance, only enhances it.

2 Informal Formulation

A PI game is defined by a *rooted tree* T. Terminal nodes are called *outcomes,* non-terminal nodes—*agents.* Each agent h chooses (or *plays)* an *action* a_h—an edge at h that does not lead back to the root. At each outcome, each agent has a *payoff.*[7]

[7]Several agents may act for one player; their payoffs are then his. See Section 9.

In parts of this and the next section[8] we assume *genericity*—that each agent 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 agent has beliefs[9] concerning actions played, and beliefs held, by other agents. Call an action a_h of agent 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,
and
r^1 : each agent believes the foregoing, unless it precludes reaching him,
and
r^2 : each agent believes the foregoing, unless it[10] precludes reaching him,
and
...
and
r^n : each agent believes the foregoing,[11] unless it precludes reaching him,
and so on ad infinitum.

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 agent h by that one of his sons' labels that yields him his highest payoff. 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.

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

[8] Specifically, Sections 2.2, 2.3, and 3.2.

[9] Informally, "belief" means attribution of probability 1. The actions and beliefs of an agent are conditional on his being reached: what he does and believes, if reached.

[10] the foregoing

[11] *All* the foregoing; i.e., r and r^1 and r^2 and ... and r^{n-1}.

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 agent, and the entire subsequent branch, and then iterating until no dominated actions remain.

3.2 Backward Induction $\leftrightarrow$ Pruning Process (Generic)

Lemma 1: *Pruning a dominated action a_h does not change h's label.*
Proof:[12] Label each action by the label of the node to which it leads. By definition of the BI process, 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 2: *Simultaneously pruning several dominated actions does not change the remaining agents' labels.*
Theorem A: *Generically, $BI(T)$ is the unique outcome of the fully pruned tree T_P.*
Proof: Repeated use of Corollary 2 shows that $BI(T)$ is an outcome of T_P. If T_P had more outcomes, one could first eliminate agents who have only one action, then prune some action of a preterminal[13] agent. ☺

3.3 Pruning Process $\leftrightarrow$ *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*[14] $r \wedge r^1 \wedge ... \wedge r^n$.
Demonstration[15]**:** Formally, this is proved in Section 8. Informally, note first that r excludes precisely what is pruned at stage 1 of the PP, no less

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

[13]One all of whose sons are outcomes.

[14]$\wedge$ means "and."

[15]The idea is very simple: When one spells out what $r \wedge r^1 \wedge ... \wedge r^n$ says, it turns out to be stage $(n+1)$ of the PP. The reader may prefer verifying this himself to reading the explicit demonstration.

and no more. "No less" follows from a dominated action being a fortiori belief-dominated; "no more," from r allowing an agent to believe that *all* agents subsequent to his action are reachable. So the result T^1 of stage 1 of the PP comprises precisely what is reachable under r. This is case $n = 0$ of the theorem.

Next, we examine what is reachable under $r \wedge r^1$. By r^1, every agent believes r, unless it precludes reaching him; as we have seen, this means that every agent in T^1 believes that the reachable agents are precisely those in T^1. So we can repeat the above argument, with T^1 substituted for T, and conclude that the result T^2 of stage 2 of the PP comprises precisely what is reachable under $r \wedge r^1$. This is case $n = 1$.

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 $r \wedge r^1 \wedge ... \wedge r^{n-1}$; we examine what is reachable under $r \wedge r^1 \wedge ... \wedge r^n$. By r^n, every agent believes $r \wedge r^1 \wedge ... \wedge r^{n-1}$, unless it precludes reaching him; by the induction hypothesis, this means that every agent in T^n believes that the reachable agents are precisely those in T^n. So we can repeat the argument for $n = 0$, with T^n substituted for T, and conclude that the result T^{n+1} of stage $n + 1$ of the PP comprises precisely what is reachable under $r \wedge r^1 \wedge ... \wedge r^n$, as asserted. ☺

Corollary 1: *An outcome survives the PP iff it is reachable under csbr.*
Demonstration: By Theorem B, an outcome is consistent with $r \wedge r^1 \wedge ... \wedge r^n$ iff it survives stage $n + 1$ of the PP. So it is consistent with *all* the r^n—i.e., with *csbr*—iff it survives *all* stages of the PP. ☺
Demonstration of the Generic Main Theorem: Combine Corollary 1 with Theorem A. ☺

4 Dispensing with Genericity

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

4.1 Backward Induction

As in the generic case, we use an inductive labelling process, but now each node h of the tree T is labelled with a *set* Z_h of outcomes.[16] Among the

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

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 $\min h' > \max h$.

Start the process by labelling each outcome z by $\{z\}$. Then, inductively label each agent 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.*[17]

The idea is that if all the sons of an agent h are leaves, 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:[18] *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. ☺

[17] 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.

[18] 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."

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 [5, 1980], Chapter 1), a sentence "holds" at each state in an arbitrary state space iff in the corresponding syntactic formalism, it is a *tautology,*[19] 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 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 indirect: to prove something, 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[20] 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

[19] 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, and in particular in this paper itself.

[20] See the next footnote.

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. Note, however, that 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 does not follow from the axioms.

In the present context, the syntactic formalism has an important advantage in addition to its transparency. This has to do with 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 we do in mathematics 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 become elements in more complex assertions, and these, in turn, become elements in still more complex assertions, and so on. Specifically, *csbr*—common strong belief of rationality—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 a *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 signifies that the sentence following it is a tautology; and whereas this operator is familiar in the metalanguage of logic, it is unusual[21] that it becomes part of the formal language itself, from which new assertions can be formed.

The tautology operator can be treated also within the semantic formalism, but it is considerably more awkward to do so, as we shall see in the discussion of BS in Section 10 below.

[21]Halpern and Lakemeyer [8, 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.

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 "agent 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 agent 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*[22] 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 agent 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 agents 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 agents H.

[22] 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.

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 agents h,
(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 agent chooses exactly one action. (4) and (5) represent classical modal belief axioms (see, e.g., [5, 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 agents 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[23] (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.

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 the appendix. ☺

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 $\neg tf$ 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

[23](6.2.1) stands for Formula 1 in Section 6.2. A similar convention will be used throughout, also for lemmas, corollaries, and so on.

formalism for our syntax.

For each agent h, denote by $\mathbf{A}_{\not\geq h}$ the set of profiles of actions of agents other than h and not preceding him. Define a *simple model*[24] 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 $\mathbf{A}_{\not\geq h}$.

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.* An agent 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 agent g preceding h must be the action a_g^h leading to h. There are no other a priori restrictions on h's beliefs; so we may describe them as a subset $\mathbf{C}(a_h)\times\mathbf{D}(a_h)$ of $\mathbf{A}$, where $\mathbf{C}(a_h)$ is a non-empty subset of $\mathbf{A}_{\not\geq h}$, and $\mathbf{D}(a_h) := \{a_h\} \times \bigtimes_{g:g\prec h}\{a_g^h\}$.

Example 1: $\mathbf{C}(a_h) := \bigtimes_{g:g\not\geq h} A_g$. Here the players have no non-trivial beliefs.

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) $\|\neg f\| := \mathbf{A}\backslash\|f\|$,
(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 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 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 3: *Every sentence that holds in some state of some simple model is consistent*[25]*.*

[24]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 25.

[25]In the language of formal logic, Corollary 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 24).

7 The Main Theorem: Formal Formulation

7.1 Rationality

Call an action of an agent *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[26] for h to play a'_h than a_h. Define

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

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 rationality obtains—i.e., that all agents 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 agents 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 for each agent h, either h believes g, or g is inconsistent 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 (written $sb^n g$) is defined inductively by

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

where $g^0 = g$ and $g^{m+1} = sb(g^0 \wedge g^1 \wedge ... \wedge g^m)$ for all m. It follows that

(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[27] *csbg*) comprises all[28] $sb^n g$ for all n.

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

[26] I.e., better for the player at h, given the actions at nodes after h.

[27] Note that *csbg* is not itself a sentence, but rather an infinite sequence of sentences.

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

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 it sounds convincing, is not entirely rigorous. We now complete the proof of the Main Theorem, by proving Theorem B rigorously.

Let T be the tree of an unrestricted[29] PI game, T^n the subtree that survives stage n of the PP. By (7.2.2), $r \wedge r^1 \wedge ... \wedge r^n = sb^n r$; 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, an agent h playing an action a_h believes in the maximum degree of iterated elimination of dominated actions (of all agents) 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 non-negative integer n and agent h, define a subset A_h^n of A_h inductively by

$A_h^0 := A_h$,

$$A_h^n := \begin{cases} \text{the set of } h\text{'s actions in } T^{n-1} \text{ that are undominated in } T^{n-1}, & \text{if } h \in T^{n-1}; \\ A_h^{n-1}, & \text{if } h \notin T^{n-1}. \end{cases}$$

Then define

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

note that the $\mathbf{A}^n$ are nested. For each action a_h, define

$$n(a_h) := \begin{cases} \text{the greatest } m \text{ for which } a_h \text{ is in } T^m \text{ and is undominated in } T^m; \\ 0, \text{ if there is no such } m. \end{cases}$$

Define a simple model $\mathbf{C}$ by

$\mathbf{C}(a_h) := \bigtimes_{g \neq h} A_g^{n(a_h)}$.

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 agents after h of a profile $\mathbf{a}_{\neq h}$ 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[30] 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

[29] Not necessarily generic.

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

Remark 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 2: *No action is belief-dominated in* $\mathbf{C}$, *unless it is dominated in* T.

Proof. 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) $\mathbf{C}(a_h) := \bigtimes_{g \neq h} A^n_g$. So if h plays a_h, he believes in $\mathbf{C}$ that agents g after h play actions in A^n_g. That implies that if a g after h is in T^{n-1}, then h believes in $\mathbf{C}$ that g plays an action in T^{n-1} that is undominated in T^{n-1}; i.e., an action in T^n. So h believes in $\mathbf{C}$ that agents in T^{n-1} after h play actions in T^n. So by (i), a_h is belief-undominated in $\mathbf{C}$. ☺

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 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) *an agent is in* T^{n+1} *iff reaching that agent is consistent with* $sb^n r$.

Remark: 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[31] with $sb^n r$. Then there is a profile $\mathbf{a}$ in $\mathbf{A}$, with the actions a_g of agents 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 agents g before h; and for other agents g, letting a_g be an arbitrary member of A^{n+1}_g. 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 $sb^0 r$. Indeed, since $\mathbf{c} \notin \mathbf{A}^1$, there is an h for which $c_h \notin A^1_h$. 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 $c_h \rightarrow (\bigwedge_{a'_h \in A_h} \neg b^h p(a'_h, c_h))$, 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, and by (7.3.1), $sb^0 r = r$, which entails r^h. So $\mathbf{c}^\wedge \wedge sb^0 r$ entails $c_h \wedge r^h$, which is inconsistent. So $\mathbf{c}^\wedge \wedge sb^0 r$ is inconsistent. ☺

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

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

($\mathbf{1}_n$): Let $\mathbf{c} \notin \mathbf{A}^{n+1}$; we will prove that $\mathbf{c}^\wedge$ is inconsistent with sb^nr. 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^nr, 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) $\vdash \bigvee_{T^n} \mathbf{a}^\wedge_{\succ h} \rightarrow p(d_h, c_h)$,

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 an agent 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), $\vdash sb^nr \rightarrow b^h(sb^{n-1}r)$ for all h in T^n. By (1_{n-1}), $\vdash sb^{n-1}r \rightarrow \bigvee_{\mathbf{a} \in \mathbf{A}^n} \mathbf{a}^\wedge_{\succ h}$; so by generalization (6.2.10) and (6.2.4),

(6) $\vdash sb^nr \rightarrow b^h(\bigvee_{\mathbf{a} \in \mathbf{A}^n} \mathbf{a}^\wedge)$.

It may be seen that $\vdash (b^hf \wedge b^hg) \rightarrow b^h(f \wedge g)$; so by (6), (4), and (6.2.4),

$\vdash sb^nr \rightarrow b^h(\bigvee_{T^n} \mathbf{a}^\wedge_{\succ h}) \rightarrow b^hp(d_h, c_h)$.

But by (7.2.2) and (7.1.1-3), $\vdash sb^nr \wedge \mathbf{c} \rightarrow r \wedge c_h \rightarrow \neg b^hp(d_h, c_h)$, so $sb^nr \wedge c_h$ entails a contradiction; i.e., $\mathbf{c}$ is inconsistent with sb^nr. ☺

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

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

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 $\mathbf{C}(c_h) \subset \bigtimes_{g \not\preceq h} A_g^n$, 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 an agent 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

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

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 $\mathbf{a} \in \mathbf{A}^{n+1}$. ☺

9 The Rationality Concept

The underlying rationality concept in this study is *action* rationality (r), which means that a player i never takes an action when another action is available that ensures[32] him a higher payoff. It is a "local" concept, focusing on the particular agent of i taking the action—the here and now—what is done at this moment. Indeed, "players" have no formal status in our model. If one wishes, one may define a "player" as a collection of agents, all with the same payoff at each outcome.

Contrast this with *strategic* rationality (sr), which means that a player never uses a *strategy* when another strategy is available that ensures him a higher payoff. The sr concept is "global:" It considers the entire tree—the long term—what will be done from now on.

On its face, sr sounds more natural. Why would a player take an action when another action is available that, while not by itself ensuring a higher payoff, does do so when combined with other actions that that player himself can take subsequently?

But r has a compelling immediacy that sr lacks. Overlooking an action that by itself ensures a higher payoff is a real blunder; avoiding such blunders is the least that one can ask of a rational player. The other side of that coin is that long-term strategies are often complex and opaque, and a wise player may well say, "I'll cross my bridges when I come to them."

More fundamentally, sr focusses on the players, r on the play. With sr, it is the *players* who are rational; with r, it is the *actions,* the *play*—the concept of "player" is secondary.

In practice, sr is stronger than r: strategically rational players cannot take belief-dominated actions. Though it does not follow[33] that $csbsr$ implies BI, $csbr$ does seem at least as compelling as $csbsr$, if not more so.

10 Battigalli-Sinischalchi and other Literature

The literature on the foundations of BI is much too large to survey here, so we confine ourselves to discussing directly relevant work. Foremost in this

[32] With probability 1; we use "sure" rather than the more cumbersome "almost sure."

[33] See the discussion at 10.2 below.

category is BS; we start by discussing it, and comparing it with the current study.

10.1 The BS Framework

BS use a semantic formalism, in which every state of the world consists of a profile of types, one for each player i; here a *type* of i consists of a strategy of i and a *conditional probability system* (CPS) over other players' types. Specifically, a CPS of i consists of a probability distribution over other players' types for each agent of i, representing that agent's beliefs; players update their beliefs in a Bayesian manner unless they are "surprised" (i.e., reach an agent to whom they previously attributed probability 0). The state spaces in BS are called *type spaces.*

10.2 BS's Rationality Concept

Rather than action rationality, BS use classical expected utility maximization (um), which means that a player never uses a strategy when a strategy with a higher expected payoff is available. This requirement is similar to strategic rationality, in that it selects strategies rather than actions; but it is even stronger, in that it rules out not only belief dominated strategies, but also those with a suboptimal expected payoff. A fortiori, it is practically speaking stronger than action rationality (see Section 9).

Normally, when a weaker requirement appears in the hypothesis of a theorem, the theorem itself is stronger. Nevertheless, our result is not stronger[34] than BS's. That is for several reasons. First, both BS and we show not only that csb of the respective forms of rationality entail BI, but also that they are consistent; and consistency of a weak requirement does not entail that of a stronger one. Second, the sb (strong belief) operator is non-monotonic: f may entail g, without[35] sbf entailing sbg. Finally, though the two results are conceptually analogous, the formal contexts are quite different, as we shall see presently; and if for that reason alone, no conclusions can be drawn in either direction.

[34]That is, the BS result does not follow directly from ours. Technically, a true theorem "follows" from any statement.

[35]That could come about in situations in which f is logically impossible, whereas g is logically possible; in that case strong belief of f is automatic, but you will strongly believe g only if you believe g, and that need not be so.

10.3 Transparency of Formulation

Our result is formulated syntactically, whereas BS formulate theirs semantically. In most applications, results that are formulated semantically may also be formulated syntactically, and vice versa; as we said above (Section 5), the choice is a matter of the trade-off between transparency and convenience in each particular application. In the present application, however, 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—intuitively, 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 (see Section 5) is not in the syntactic language itself, but only in 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 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 a real obstacle for them. They overcome it by working not with arbitrary state spaces, as is usual with 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 far from elementary; see [3, 1999]. But such an object does exist.

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 *csbum* "holds." Thus, BS *must* formulate their results using the complete type space, whose construction uses deep tools, and whose very conception is complex and deep. In contrast, our main theorem formulates

its conceptual content in a transparent and straightforward manner.[36]

Finally, it is worth mentioning that while probabilities are implicit in the notion of "belief" that underlies our approach, only the probability 1 (and so, implicitly, 0) plays any role. So we axiomatize the notion of belief—attribution of probability 1—fairly simply by itself, without reference to other probabilities. By contrast, since BS use expected payoffs, they must work with the whole gamut of numerical probabilities, a much more complex task.

10.4 Depth of Proof

Quite apart from BS's formulation, their *proof* is far deeper than ours. Like ours, their proof has two parts. In one—analogous to our Theorem B—they show that *csbum* is equivalent to extensive form rationalizability (EFR) a la Pearce [10, 1984] or Battigalli [2, 1997]; this is the deepest part of their proof. In the other—analogous to our Theorem A—they recall that EFR entails the BI outcome, as proved by Reny [11, 1992], who used heavy tools from algebraic topology developed by Kohlberg and Mertens [9, 1986] (see also Battigalli [2, 1997]). An "elementary" proof—which is nevertheless quite involved—of this result is obtained by applying to PI games a recent theorem of Chen and Micali [6, 2013] about general extensive games. Looking at PI games only, Heifetz and Perea [7, 2013] obtained a considerably simpler elementary proof that EFR entails BI; but it, too, is still fairly complex, certainly far more so than that of our Theorem A.

10.5 BI Strategies

Whereas in PI games,[37] EFR implies that the BI *outcome* is reached, it may be inconsistent with a player using his BI *strategy.* For an example, see Reny [11, 1992], p.637, Fig.3, where the *only* EFR strategy of Player 2 prescribes an action different from the BI action at his first move. Thus for some agents, *csbum* may actually be inconsistent with the BI action. In contrast, the BI action is always consistent with *csbr*, for all agents.

[36] 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, whose very formulation involves a whole lot of complex math, shows up as nothing but plain old 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.

[37] Even generic ones, as in Section 2.

10.6 Genericity

Our proof uses no genericity condition at all, whereas BS require the game to have "no relevant ties" (NRT) (see also [2, 1997]) ; i.e., that the last common ancestor of any two outcomes have different payoffs at those outcomes. NRT ensures a unique BI outcome. As already remarked, it is by no means innocent; inter alia, it excludes chess.

Without any genericity condition, *csbum* is still equivalent to EFR, and EFR still implies BI; but there may be BI outcomes that are not EFR. For an example, see Chen and Micali [6, 2013], p.149, Example 7. Heifetz and Perea [7, 2013] do assume NRT.

10.7 Strong Belief

Interestingly, backward induction looks forward, forward induction backward. In the labelling process defining BI (Section 2.2), each agent's label—conceptually, the outcome he expects—is determined by the *subsequent* agents' labels. So each agent looks *forward* only. On the other hand, in forward induction—as, say, in Kohlberg and Mertens [9, 1986]—agents look *backward,* to glean information about the future from what happened previously.

So one may think of *csbr* as based on a kind of negative forward induction reasoning. Basically, we look forward: optimize, assuming certain rationality conditions for all agents. But we also look backward: we ask, if those assumptions were correct, would the previous agents have made the choices that they did? If the answer is no, then we abandon the assumptions. So "strong belief" is a kind of forward induction: it looks backward, even if only to abandon an assumption.

An early harbinger of strong belief is Reny [12, 1993], who formulated the following two conditions as desirable for "a theory of games:"

(a) each player believes that his opponent is an expected utility maximizer, *so long as this is consistent with the history of play* (our italics), and

(b) both players believe (a), each believes that the other believes (a), etc.

Thus strong belief makes its maiden appearance in Reny's Condition (a). For some reason, he preferred not to apply the italicized caveat to Condition (b); it remained for Battigalli and Sinischalchi to do so.

10.8 Conclusion

Though much of this section has dealt with differences between BS's result and ours, we reiterate that when all is said and done, the two results convey

similar messages. Our contribution may be seen largely as a transparent and accessible reformulation and proof of BS's landmark insight.

11 Appendix: Proof of Theorem D

Lemma 1: *The list* $\mathfrak{B}$ *of all basic tautologies is coherent.*
Proof: Follows from Lemma (6.4.2) and Example (6.4.1). ☺

Now denote by 1 an arbitrary but fixed basic tautology (such as $a_h \vee \neg a_h$), 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 2: $\mathfrak{L}$ *is coherent, strongly closed, and includes all tautologies.*
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 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 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 $\vdash t(g \leftrightarrow g') \rightarrow (tg \leftrightarrow t(g'))$, 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 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 1, and $\mathfrak{L} \cap \chi = \mathfrak{B}$; so $\mathfrak{T} \cap \chi = \mathfrak{B}$. ☺

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

Proof: By Lemma 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 4, so $\mathfrak{T} = \{\ f\ | f' \in \mathfrak{B}\} = \mathfrak{L}$. ☺

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

References

[1] P. Battigalli, Strategic rationality orderings and the best rationalization principle, *Games and Economic Behavior* **13** (1996), 178-200.

[2] P. Battigalli, On rationalizability in extensive games, *Journal of Economic Theory* **74** (1997), 40-61.

[3] P. Battigalli and M. Siniscalchi, Hierarchies of conditional beliefs and interactive epistemology in dynamic games, *Journal of Economic Theory* **88** (1999), 188-230.

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

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

[6] J. Chen and S. Micali, The order independence of iterated dominance in extensive games, *Theoretical Economics* **8** (2013), 125–163.

[7] A. Heifetz and A. Perea, On the outcome equivalence of backward Induction and extensive form rationalizability, (2013), Unpublished.

[8] J. Y. Halpern and G. Lakemeyer, Multi-agent only knowing, *Journal of Logic and Computation* **11** (2001), 41-70.

[9] E. Kohlberg and J. F. Mertens, On strategic stability of equilibria, *Econometrica* **54** (1986), 1003–1037.

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

[11] P. Reny, Backward induction, normal form perfection and explicable equilibria, *Econometrica* **60** (1992), 627–649.

[12] P. Reny, Common belief and the theory of games with perfect information, *Journal of Economic Theory* **59** (1993), 257–274.