# Beliefs, Plans, and Perceived Intentions in Dynamic Games*

Pierpaolo Battigalli
Department of Decision Sciences and IGIER, Bocconi University
pierpaolo.battigalli@unibocconi.it

Nicodemo De Vito
Department of Decision Sciences, Bocconi University
nicodemo.devito@unibocconi.it

Draft of September 2018

**Abstract**

We adopt the epistemic framework of Battigalli and Siniscalchi (*J. Econ. Theory*, 1999) to model the distinction between a player's contingent behavior, which is part of the external state, and his plan, which is described by his beliefs about his own behavior. This allows us to distinguish between intentional and unintentional behavior, and to explicitly model how players' revise their beliefs about the intentions of others upon observing their actions. We illustrate our approach with detailed examples and with a new derivation of backward induction from epistemic conditions. Specifically, we prove that common full belief in optimal planning and in belief in continuation consistency imply the backward induction strategies and beliefs. We also present within our framework other relevant epistemic assumptions and relate them to similar ones studied in the previous literature.

KEYWORDS: Epistemic game theory, plans, perceived intentions, backward induction, forward induction.

*We thank Federico Bobbio, Roberto Corrao, Enrico De Magistris and Davide Ferri for careful proof-reading, and the partecipants to conference and seminar presentations at TARK 2017, EEA-ESEM 2018, Heidelberg and Northwestern University for useful comments. Pierpaolo Battigalli gratefully acknowledges the financial support of ERC, grant 324219.

# 1 Introduction

Players who reason strategically anticipate the moves of others under the assumption that they are rational and "sophisticated." In dynamic games, players have to understand past moves in order to predict future moves. Assumptions about how players would revise their beliefs upon observing unexpected moves are therefore paramount. According to forward-induction thinking, past moves are interpreted, if possible, as intentional choices carrying out strategically rational plans. According to backward-induction thinking, instead, past unexpected moves are interpreted as deviations from the strategically rational plans ascribed to opponents, but similar deviations are not expected to occur in the future, as in the trembling-hand story by Selten (1975).

A flexible theory of strategic reasoning in dynamic games should therefore allow for the distinction between plan and choice and should allow to model the perception of past moves by others as intentional or unintentional. Yet, most epistemic models for games conflate plan and contingent behavior, as they assume implicitly or explicitly that, at *every* state of the world, each player $i$ knows (or at least holds a correct belief about) his contingent behavior.[1] Since they do not have states where plans and behavior do not coincide, such models formally rule out the possibility that unexpected moves are interpreted as deviations from the plans ascribed to other players.

In this paper, we use the epistemic framework of Battigalli and Siniscalchi (1999a) to model how players change their perceptions about the intentions of others. Players hold (first-order) beliefs about the behavior of everybody, including themselves,[2] and plans are modeled as beliefs about own behavior. We illustrate the framework with examples and results about the behavioral implications of different assumptions about strategic reasoning. In particular, our main result provides epistemic conditions for the backward-induction strategies of generic games with perfect information. To do this, we use three main ingredients, which correspond to events in our framework:

- **optimal planning** ($OP$), which is the result of "folding-back" calculations given beliefs about the contingent behavior of others,

[1] See, for example, the surveys on epistemic game theory by Battigalli and Bonanno (1999), and Dekel and Siniscalchi (2015). To be precise, we consider **doxastic and epistemic** models of games. Yet, to be consistent with current use in game theory, we abuse the term "epistemic," which refers to the analysis of players' interactive knowledge, and extend it to encompass also the (doxastic) analysis of interactive beliefs.

[2] Of course, players hold higher-order beliefs as well.

- **consistency** ($C$), that is, coincidence between plan and contingent behavior, and
- **belief in continuation consistency** ($BCC$), that is, upon reaching any history $h$, each player believes that the opponents' behavior will be consistent with their plans starting from $h$, whether or not they were consistent in the past.

**Rationality** is given by the conjunction of optimal planning and consistency ($R = OP \cap C$). Much of the literature on epistemic game theory analyzes the behavioral implications of rationality and some versions of "common belief" in rationality. We instead take a different route and consider "common belief" in **doxastic** events, that is, events concerning how players think, not how they behave. Say that a player **fully believes** an event $E$ if he assigns probability 1 to $E$ conditional on *every* history $h$. Note that the assumption of full belief in doxastic events is not problematic because they cannot be falsified by the observation of behavior. With this, we show the following (Theorem 1):

*Common full belief in $OP \cap BCC$ implies the backward-induction plans and beliefs about others (if players are also consistent their behavior conforms to backward induction).*

We extend this result to cover all multistage games with observable actions: we show (Theorem 2) that the aforementioned assumptions imply that players use backwards rationalizable strategies (Penta 2015, Perea 2014), which coincide with the backward-induction strategies in generic games with perfect information. Furthermore, we prove that—in the universal type structure—rationality and common strong belief in rationality (Battigalli and Siniscalchi 2002) imply that players use strongly rationalizable strategies (Theorem 3).[3] Finally, we illustrate our approach showing that the same behavioral implications obtain under the following assumptions: Let $C^*$ denote the set of states where $C$ (consistency) holds and there is *common full belief of* $C$; with this, we prove that in a complete type structure strong rationalizability characterizes the behavioral implications of $OP \cap C^*$ (a subset of $R$) and common strong belief in $OP \cap C^*$ (Theorem 4). We argue that these assumptions are implicit in the framework of Battigalli and Siniscalchi (2002), where first-order beliefs concern only the behavior of co-players.

The epistemic analysis of backward induction dates back to Aumann (1995). Other articles with epistemic conditions for backward induction include Battigalli

[3]Our terminology is clarified and justified in Section 6. Here we just mention that strong rationalizability is often called "extensive-form rationalizability."

and Siniscalchi (2002), Bonanno (2013) and Perea (2014). We provide detailed comments on the related literature in Section 7. Here we just note that the formal language of the aforementioned papers does not allow to distinguish between plan and contingent behavior, hence it cannot express our key assumptions of consistency ($C$) and belief in continuation consistency ($BCC$).

The rest of the paper is structured as follows. Section 2 introduces the framework. Section 3 illustrates it and heuristically introduces the main ideas with two examples. Section 4 analyzes optimal planning, consistency and rationality. Section 5 contains the main result of this paper, that is, an epistemic characterization of backward-induction reasoning. Section 6 analyzes forward-induction reasoning. Finally, Section 7 discusses certain conceptual aspects and possible extensions of the analysis, and it comments on the related literature.

# 2 Framework

In this section we present the building blocks of our analysis: finite games with observable actions (subsection 2.1), systems of conditional probabilities (subsection 2.2) and type structures (subsection 2.3).

## 2.1 Finite games with observable actions

We focus on finite multistage games with perfect monitoring of past actions. Given some preliminaries about sequences and trees, we define these games and the external states describing players' contingent behavior.

### 2.1.1 Sequences and trees

Let $\mathbb{N}_0$ be the set of natural numbers including 0, that is, $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. Given an arbitrary nonempty set $X$, the set of all finite sequences of elements of $X$ is $X^{<\mathbb{N}_0} := \cup_{n\in\mathbb{N}_0} X^n$, where $X^0 := \{\varnothing\}$ and $\varnothing$ denotes the **empty sequence**. For all $\mathbf{x} \in X^{<\mathbb{N}_0}$ and $\mathbf{y} \in X^{<\mathbb{N}_0}$, $(\mathbf{x}, \mathbf{y})$ denotes the concatenation of $\mathbf{x}$ with $\mathbf{y}$. We write $\mathbf{x} \preceq \mathbf{x}'$ if $\mathbf{x}$ is a prefix of $\mathbf{x}'$, that is, $\mathbf{x}' = (\mathbf{x}, \mathbf{y})$ for some $\mathbf{y}$. Note that $(\varnothing, \mathbf{x}) = (\mathbf{x}, \varnothing) = \mathbf{x}$, hence $\varnothing \preceq \mathbf{x}$ and $\mathbf{x} \preceq \mathbf{x}$ for every $\mathbf{x} \in X^{<\mathbb{N}_0}$. We let $\prec$ denote the asymmetric part of $\preceq$.

A nonempty set $\mathbf{Y} \subseteq X^{<\mathbb{N}_0}$ is a **tree** if it is closed with respect to prefixes, that is, for every $\mathbf{x}' \in \mathbf{Y}$ and every prefix $\mathbf{x}$ of $\mathbf{x}'$, $\mathbf{x} \in \mathbf{Y}$; therefore, $\varnothing \in \mathbf{Y}$. For every tree $\mathbf{Y} \subseteq X^{<\mathbb{N}_0}$, we say that a sequence $\mathbf{x}$ is **terminal** in $\mathbf{Y}$ if $\mathbf{x} \prec \mathbf{x}'$ implies $\mathbf{x}' \notin \mathbf{Y}$ for all $\mathbf{x}' \in X^{<\mathbb{N}_0}$.

### 2.1.2 Games

A **finite game with observable actions** is a structure

$$\Gamma = \left\langle I, \bar{H}, (A_i, u_i)_{i \in I} \right\rangle$$

given by the following elements:

- $I$ is a finite set of **players**, and, for each $i \in I$, $A_i$ is a finite, nonempty set of potentially feasible **actions**.
- $\bar{H} \subseteq A^{<\mathbb{N}_0}$ is a finite tree of feasible **histories**, that is, $\bar{H}$ is a tree of sequences of elements of $A := \prod_{i \in I} A_i$ with distinguished root $\varnothing$. We let $Z$ denote the set of terminal histories, and $H := \bar{H} \backslash Z$ is the set of nonterminal histories.
- For each $h \in H$, the set of feasible action profiles

  $$A(h) := \left\{ a \in A : (h, a) \in \bar{H} \right\},$$

  is such that $A(h) = \prod_{i \in I} A_i(h)$, where $A_i(h)$ is the projection of $A(h)$ on $A_i$.
- For each $i \in I$, $u_i : Z \to \mathbb{R}$ is the payoff (utility) function for player $i$.

Since the restrictions of $\prec$ and $\preceq$ on $\bar{H}$ represent the strict and weak precedence relations between the histories/nodes of the game tree, we say that $h$ **(weakly) precedes** $h'$ if $(h \preceq h')$ $h \prec h'$; equivalently, we say that $h'$ **(weakly) follows** $h$ and write $(h' \succeq h)$ $h' \succ h$.

Player $i$ is **active** at history $h \in H$ if he has at least two feasible actions ($|A_i(h)| \geq 2$), and he is **inactive** otherwise (that is, if $|A_i(h)| = 1$).[4] There are simultaneous moves given $h$ if at least two players are active at $h$. If there is only one active player at each $h \in H$, we say that the game has **perfect information**.

### 2.1.3 External states and contingent behavior

For each $i \in I$, let $S_i := \prod_{h \in H} A_i(h)$ and $S := \prod_{i \in I} S_i$. An **external state** is a profile $s = (s_i)_{i \in I} \in S$, and each $s_i \in S_i$ is called **personal external state** of player $i$. The set of external states of players other than $i$ is $S_{-i} := \prod_{j \in I \backslash \{i\}} S_j$.[5]

[4] When $i$ is not active at $h \in H$, think of the unique element of $A_i(h)$ as the "action" of waiting one's turn to move.

[5] In keeping with standard game-theoretic notation, given any collection of sets $X_i$ $(i \in I)$, we let $X_{-i} := \prod_{j \neq i} X_j$ with typical element $x_{-i} \in X_{-i}$.

An external state $(s_i)_{i\in I} \in S$ is interpreted as an *objective description of players' contingent behavior*, which may or may *not* coincide with what players plan to do. Note that each $s_i \in S_i$ corresponds technically to a strategy of player $i$, but we avoid this terminology because we call "strategy" what player $i$ plans to do, which is part of his epistemic type (cf. Section 4).

Each external state $s = (s_i)_{i\in I} \in S$ induces a terminal history. Thus, we can define a **path function** $\zeta : S \to Z$ associating each external state with the corresponding terminal history. So, for each $h \in H$, we can define the set of external states inducing $h$:

$$S(h) := \{s \in S : h \prec \zeta(s)\}.$$

The projection

$$S_i(h) := \{s_i \in S_i : \exists s_{-i} \in S_{-i}, (s_i, s_{-i}) \in S(h)\}$$

is the set of external states of $i$ that do not prevent $h$ from being reached. Similarly, the projection

$$S_{-i}(h) := \{s_{-i} \in S_{-i} : \exists s_i \in S_i, (s_i, s_{-i}) \in S(h)\}$$

is the set of profiles of external states of players other than $i$ that do not prevent $h$ from being reached. Note that, in a game with observable actions,

$$S(h) = \prod_{i\in I} S_i(h)$$

for every $h \in H$.[6]

Finally,

$$U_i := u_i \circ \zeta : S \to \mathbb{R}$$

determines the payoff $U_i(s) = u_i(\zeta(s))$ of player $i$ as a function of the external state $s$.

## 2.2 Conditional beliefs

For every compact metrizable space $X$, we let $\Delta(X)$ denote the set of probability measures on the Borel subsets of $X$, called **events**. For every $\nu \in \Delta(X)$, the support of $\nu$ is denoted by $\text{supp}\nu$. The set $\Delta(X)$ is endowed with the *weak**-topology, so that $\Delta(X)$ becomes a compact metrizable space.

[6] In more general games, perfect recall implies the following factorization: $S(h_i) = S_i(h_i) \times S_{-i}(h_i)$ for each player $i$ and each information set $h_i$ of $i$.

We consider arrays of probability measures indexed by elements of a countable collection $\mathcal{C}$ of "conditioning events," i.e., $\mu := (\mu(\cdot|C))_{C\in\mathcal{C}} \in \Delta(X)^{\mathcal{C}}$.[7]

**Definition 1** *Let $X$ be a compact metrizable space and $\mathcal{C}$ be a countable family of* clopen *(i.e., both closed and open) and nonempty subsets of $X$. A* ***conditional probability system (CPS)*** *on $(X,\mathcal{C})$ is an array of probability measures $\mu := (\mu(\cdot|C))_{C\in\mathcal{C}}$ such that, for all $C, D \in \mathcal{C}$ and events $E$, $\mu(C|C) = 1$ and*

$$E \subseteq D \subseteq C \Rightarrow \mu(E|C) = \mu(E|D)\mu(D|C). \tag{2.1}$$

Condition (2.1) is the so-called **chain rule** of conditional probabilities and it can be written as follows: if $E \subseteq D \subseteq C$, then

$$\mu(D|C) > 0 \Rightarrow \mu(E|D) = \frac{\mu(E|C)}{\mu(D|C)}.$$

We write $\Delta^{\mathcal{C}}(X)$ for the set of CPSs on $(X,\mathcal{C})$. Under the stated assumptions, $\Delta^{\mathcal{C}}(X)$ is a compact metrizable space (see Lemma 1 in Battigalli and Siniscalchi 1999a).

Given compact metrizable spaces $X$ and $Y$, the set $X \times Y$ is endowed with the product topology. Let $\mathcal{C}$ be a countable collection of clopen subsets of $X$ such that $\emptyset \notin \mathcal{C}$. With a small abuse of notation, we write $\mathcal{C}\times Y$ for the corresponding collection of clopen "cylinders" in $X \times Y$, that is,

$$\mathcal{C} \times Y := \{C \subseteq X \times Y : \exists F \in \mathcal{C}, C = F \times Y\}.$$

For every probability measure $\nu \in \Delta(X \times Y)$, we let $\text{marg}_X \nu$ denote the marginal of $\nu$ on $X$. Now consider a CPS $\mu := (\mu(\cdot|C \times Y))_{C\in\mathcal{C}} \in \Delta^{\mathcal{C}\times Y}(X \times Y)$. Then the **marginal** of $\mu$ on $(X,\mathcal{C})$ is defined as the array of probability measures

$$\text{marg}_X \mu := (\text{marg}_X \mu(\cdot|C))_{C\in\mathcal{C}} \in [\Delta(X)]^{\mathcal{C}}.$$

It can be easily verified that $\text{marg}_X \mu$ is a CPS on $(X,\mathcal{C})$.

[7] For every pair of sets $P$ and $Q$, $Q^P$ denotes the collection of functions with domain $P$ and codomain $Q$. Thus, $\mu$ is a function from $\mathcal{C}$ to $\Delta(X)$. We write $\mu(\cdot|C)$ to stress the interpretation as a conditional probability given the conditioning event $C \in \mathcal{C}$.

## 2.3 Type structures

We represent a player's plan, or strategy, as a system of conditional beliefs about his own behavior. If a player holds conditional beliefs about his own behavior as well as other players', first-order beliefs are CPSs on $(S, \mathcal{S})$, where $\mathcal{S}$ is the common collection of conditioning events about behavior corresponding to nonterminal histories:

$$\mathcal{S} := \{F \subseteq S : \exists h \in H, F = S(h)\}.$$

For any $i \in I$, let $T_{-i}$ denote the set of possible "types" of the other players, that is, the set of their possible "ways to think." Then the conditioning event for $i$ corresponding to history $h \in H$ is $S(h) \times T_{-i}$; thus, a CPS for $i$ is an array of probability measures $\mu_i := (\mu_i(\cdot|S(h) \times T_{-i}))_{h \in H}$ that satisfies the chain rule and $\mu_i(S(h) \times T_{-i}|S(h) \times T_{-i}) = 1$ for each $h \in H$.

**Definition 2** *A $\Gamma$-based **type structure** is a tuple*

$$\mathcal{T} = \left(S, H, (T_i, \beta_i)_{i \in I}\right)$$

*such that, for every $i \in I$,*
*(a) the **type set** $T_i$ is a* compact metrizable *space,*
*(b) the **belief map** $\beta_i : T_i \to \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ is* continuous.

*A **personal state** of player $i$ is a pair $(s_i, t_i) \in S_i \times T_i$. A **state of the world** is a profile $(s_i, t_i)_{i \in I} \in \prod_{i \in I}(S_i \times T_i)$.*

To ease notation, we will often write $\beta_{i,h}(t_i)$ to denote the beliefs of type $t_i$ conditional on history $h$, that is,

$$\beta_{i,h}(t_i)(\cdot) := \beta_i(t_i)(\cdot|S(h) \times T_{-i}).$$

A type structure provides an implicit representation of the higher-order beliefs of the players. Specifically, each type $t_i$ in a type structure induces a corresponding hierarchy of conditional beliefs satisfying an intuitive coherence condition. Battigalli and Siniscalchi (1999a) show that a **canonical** type structure can always be constructed by letting the set of types of each $i$ be the collection of all possible hierarchies of CPSs that satisfy coherence and common full belief in coherence.[8] More precisely, each type $t_i$ in the canonical structure is an infinite hierarchy of CPSs, i.e., $t_i = (\mu_i^n)_{n \in \mathbb{N}}$

[8] Loosely speaking, this means that lower-order beliefs are the marginals of higher-order beliefs and there is common belief of this conditional on each history—see Section 5 for a formal definition of "full belief."

where $\mu_i^1$ is the first-order belief, a CPS on $(S, \mathcal{S})$, $\mu_i^2$ is the second-order belief, a CPS on $\left(S \times \left[\Delta^{\mathcal{S}}(S)\right]^{I\setminus\{i\}}, \mathcal{S} \times \left[\Delta^{\mathcal{S}}(S)\right]^{I\setminus\{i\}}\right)$ whose marginal is $\mu_i^1$, and so on. Denoting by $T_i^*$ the set of player $i$'s belief hierarchies satisfying coherence and common full belief in coherence, there is a canonical homeomorphism $\beta_i^* : T_i^* \to \Delta^{\mathcal{S}\times T_{-i}^*}\left(S \times T_{-i}^*\right)$ that determines, for each type (hierarchy) $t_i$, a CPS $\beta_i^*(t_i)$ on the set of external states and the set of belief hierarchies of the co-players. Moreover, such canonical type structure turns out to be "universal," or "terminal" in the sense that every other type structure can be mapped into it in a unique belief-preserving way.[9] Hence, each type structure is hierarchy-equivalent to a substructure of the canonical one.

With this in mind, we consider in the next section two illustrative examples with type structures that are "small," but nonetheless sufficiently rich for the purposes of our epistemic analysis; that is, the essential epistemic features would not change if we considered the corresponding belief hierarchies with the backdrop of the canonical structure.

It is worthwhile to compare the notion of type structure as per Definition 2 to type structures that only describe players' beliefs about the behavior and beliefs of other players. We refer to the latter type structures as "standard," since they are widely used in epistemic game theory.[10] A $\Gamma$-based **standard type structure** is a tuple $\mathcal{T} = \left(H, (S_{-i}, T_i, \beta_i)_{i\in I}\right)$ where, as in Definition 2, $T_i$ is a compact metrizable space of player $i$'s types, but each belief map is a (continuous) function $\beta_i : T_i \to \Delta^{\mathcal{S}_{-i}\times T_{-i}}(S_{-i} \times T_{-i})$, where $\mathcal{S}_{-i}$ denotes the collection of conditioning events about the behavior of player $i$'s opponents, i.e., $\mathcal{S}_{-i} := \{F \subseteq S_{-i} : \exists h \in H, F = S_{-i}(h)\}$.

The epistemic approach *via* standard type structures has the advantage of providing a parsimonious description of beliefs that can in principle be elicited by observing choices of side bets.[11] Furthermore, the approach is adequate for the analysis of expected utility maximizing players in dynamic games.[12]

However, we argue that in the analysis of dynamic games there are conceptual advantages in introducing players' beliefs about their own behavior.[13] Such beliefs explicitly represent how a player expects to choose at later histories, which guides the player's current choice. Also, they allow to formally distinguish between the

[9] In the terminology of Mertens and Zamir (1985), one can say that every type structure is a "belief-closed substructure" of the canonical type structure.

[10] See Definition 12.23 in Dekel and Siniscalchi (2015).

[11] Under the assumption that players choose rationally complemented by a strong invariance assumption; see Siniscalchi (2016).

[12] See the monographs by Battigalli et al. (2017) and Perea (2012), and the comprehensive survey by Dekel and Siniscalchi (2015).

[13] We will maintain the implicit assumption that players are introspective, hence know their own way to think, and that this is commonly believed at every history.

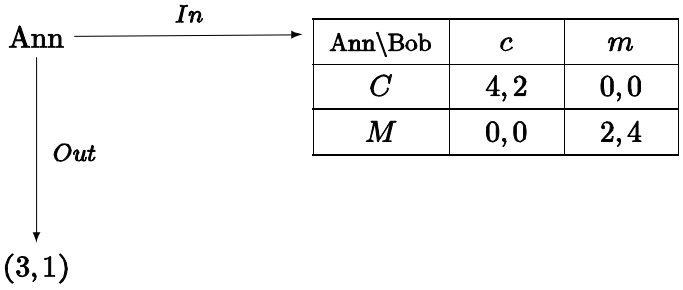
description of the contingent behavior of a player, which is what co-players ultimately care about, and what this player *plans* to do and achieve, that is, his *intentions*. Of course, intentions do not affect payoffs, but thinking about the intentions of co-players helps interpret their past observed actions and predict their future actions, e.g., by forward or backward-induction reasoning.[14] By contrast, when we use standard type structures, we implicitly assume that the personal external states $s_i$ ($i \in I$) in every state of the world $(s_i, t_i)_{i \in I}$ simultaneously represent players' contingent behavior and their plans. Since this is true for every state, it is implicitly assumed that it is transparent (i.e., true and commonly believed also at every history) that players execute their plans and that evidence about behavior is (regarded as) evidence about intentions.

# 3 Two illustrative examples

In this section we illustrate the framework and informally introduce the building blocks of our analysis by means of examples based on two well known games.

## 3.1 Perceived intentions in the Battle of Sexes with Outside Option

Consider the game depicted in Figure 3.1 ("Battle of Sexes with Outside Option," BoSOO) between two players, Ann ($a$) and Bob ($b$). If Ann does not choose the outside option, Ann and Bob play a simultaneous-moves game in which they have to choose between a concert with music by Chopin or Mozart.




| Ann\Bob | $c$ | $m$ |
|---|---|---|
| $C$ | 4, 2 | 0, 0 |
| $M$ | 0, 0 | 2, 4 |

Figure 3.1: The BoSOO game.

[14] Furthermore, the theory of psychological games allows intentions, or beliefs about intentions to affect players' utility. See Battigalli and Dufwenberg (2009), and Battigalli et al. (2018).

The set of nonterminal histories is $H = \{\varnothing, (In)\}$, while the sets of personal external states of each player are[15]

$$S_a = \{In.C, In.M, Out.C, Out.M\}\,,\; S_b = \{c, m\}\,.$$

This game has two pure subgame perfect equilibria, $(In.C, c)$ and $(Out.M, m)$, where only the former conforms to the standard forward-induction story. We now exhibit a type structure with types corresponding to both equilibria, where each type is consistent with a kind of backward-induction condition. For each player $i \in \{a, b\}$, let $T_i = \{t_i^1, t_i^2\}$; the belief maps are shown in Table 1.

| $\beta_i$ | $\varnothing$ | $(In)$ |
|---|---|---|
| $t_i^1$ | $\left(\left(In.C, c\right), t_{-i}^1\right), 1$ | $\left(\left(In.C, c\right), t_{-i}^1\right), 1$ |
| $t_i^2$ | $\left(\left(Out.M, m\right), t_{-i}^2\right), 1$ | $\left(\left(In.M, m\right), t_{-i}^2\right), 1$ |

**Table 1**: Type structure for BoSOO game.

To understand the description of the type structure in Table 1, consider for instance the beliefs of Ann's type $t_a^1$ conditional on the empty sequence $\varnothing$, that is, $\beta_{a,\varnothing}\left(t_a^1\right)\left(\left\{\left(\left(In.C, c\right), t_b^1\right)\right\}\right) = 1$.

At both states of the world

$$\left(s^1, t^1\right) = \left(\left(In.C, t_a^1\right), \left(c, t_b^1\right)\right) \text{ and } \left(s^2, t^2\right) = \left(\left(Out.M, t_a^2\right), \left(m, t_b^2\right)\right)$$

players **"plan optimally"** in the following sense: each player plans to take, at each history where she or he is active, the best action given her or his (conditional) belief, and this yields a dynamically optimal plan. For example, type $t_a^2$ of Ann predicts that—if the proper subgame were reached—Bob would choose $m$ and she would choose $M$; given her conditional belief, $M$ is the expected utility maximizing action; thus, in a sense, Ann is planning to behave optimally in the subgame. Given her prediction about what would happen in the subgame, Ann of type $t_a^2$ plans to stay out of it, that is, she is initially certain that she is going to choose $Out$. Overall, the plan of type $t_a^2$ is $Out.M$ and—given $t_a^2$'s beliefs about Bob—it satisfies a **folding-back** property that can be informally stated for general multi-stage games as follows:

*Actions planned for the last stage are best replies to the last-stage conditional beliefs about the other player; given the last-stage predictions, actions planned for the second-to-last stage are best replies to the second-to-last-stage conditional beliefs, and so on.*

[15] We write $X.Y$ for the personal external state of Ann that describes action $X$ at history $\varnothing$ and action $Y$ at history $(In)$.

Thus, we say that Ann plans optimally at state $(s^2, t^2)$. By itself, this is not enough to deem Ann rational at $(s^2, t^2)$: we say that a player is **rational** at a state $(s, t)$ if she plans optimally *and* her contingent behavior, as objectively described by $s_i$, corresponds to her plan. In other words, we view the inconsistency between plan and behavior as a form of irrationality. For example, at any state $((In.C, t_a^2), (s_b, t_b^2))$ $(s_b \in \{c, m\})$ Ann is irrational because—although type $t_a^2$ satisfies optimal planning (that is, the folding-back property)—behavior $In.C$ is different from $t_a^2$'s plan $Out.M$.

We say that player $i$

- **strongly believes** event $E$ if $i$ assigns probability 1 to $E$ conditional on each history $h$ that does not contradict $E$;[16]
- **fully believes** event $E$ if $i$ assigns probability 1 to $E$ conditional on each history $h$.[17]

At state $(s^1, t^1) = ((In.C, t_a^1), (c, t_b^1))$, Bob's belief conditional on $(In)$ about Ann's plan is that she did what she planned to do, that she intends to continue with the same plan $In.C$, and that she will actually behave as planned; that is, Bob believes in Ann's rationality also in the subgame. Given the interactive beliefs at $(s^1, t^1)$ conditional on $(In)$, one can see that there is common belief in rationality also in the subgame, which implies that there is **rationality and common strong belief in rationality** (RCSBR) at state $(s^1, t^1)$.

Consider now state $(s^2, t^2) = ((Out.M, t_a^2), (m, t_b^2))$. Upon observing $In$, Bob could think that Ann's personal state is $(In.C, t_a^1)$, thus maintaining his belief in Ann's rationality. Instead, at $(s^2, t^2)$ and conditional on $(In)$, Bob maintains his belief that Ann's type is $t_a^2$, hence, that her plan was $Out.M$ and she did not follow through. Thus, Bob does not strongly believe that Ann is rational. However, Bob also believes that—despite her initial deviation—Ann is going to follow her plan in the subgame. In other words, Ann's initial deviation from the plan she was supposed to hold is not interpreted as evidence that her intentions are different, but rather as a "mistake," and such mistake is not deemed as evidence that further "mistakes" are likely. Given the behavior and interactive beliefs at $(s^2, t^2)$, there cannot be common full belief in rationality, but there is common full belief that players plan optimally (although deviations from the hypothesized plans would be acknowledged ex post). Furthermore, conditional on each history, players believe

[16] See the formal definitions in Section 6.

[17] See the formal definition in Section 5. Note that it is impossible to fully believe an event $E$ if $E$ implies that some history $h \in H$ cannot be reached. In this case, the event "$i$ fully believes $E$" is empty, but it is still well defined.

that everybody's behavior will be consistent with plan from that point onward, and there is common belief in such "**belief in continuation consistency**." We view this as an epistemic representation of backward-induction thinking, as the following example further illustrates.

## 3.2 Forward and backward-induction reasoning in a perfect information game

Consider the game with perfect information depicted in Figure 3.2 between Ann ($a$) and Bob ($b$).[18]

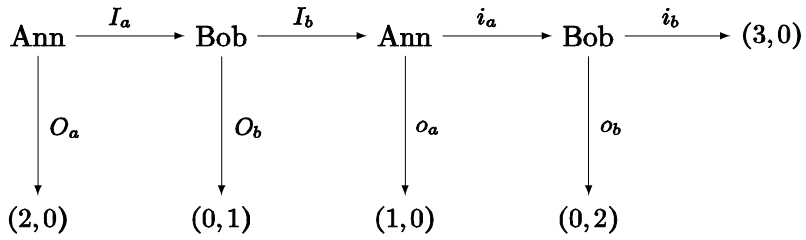



Figure 3.2: A game with perfect information.

The set of nonterminal histories is

$$H = \{\varnothing, (I_a), (I_a, I_b), (I_a, I_b, i_a)\},$$

while the sets of personal external states of each player are

$$\begin{aligned} S_a &= \{I_a.i_a, I_a.o_a, O_a.i_a, O_a.o_a\}, \\ S_b &= \{I_b.i_b, I_b.o_b, O_b.i_b, O_b.o_b\}. \end{aligned}$$

As is well known, strong rationalizability (Pearce 1984, Battigalli 1997)[19] and backward induction yield the same path, $(O_a)$, but have very different off-path behavioral implications: for Bob, the unique strongly rationalizable behavior is $I_b.o_b$,

[18] Cf. Reny 1992, Figure 3.

[19] The solution concept of strong rationalizability is also known as "extensive-form rationalizability." We find such terminology ambiguous and hence we avoid it, because this solution concept refers to just one out of several meaningful versions of rationalizability for extensive-form games. We find it semantically and conceptually appropriate to use "strong" for this version of rationalizability in light of its epistemic foundation, which is based on the notion of strong belief. See Section 6.

while backward induction yields $O_b.o_b$. We can formally interpret the difference as the result of different hypotheses about how players revise their beliefs about the plans, or intentions, of co-players. We consider a type structure with types corresponding to forward-induction reasoning (fi), or backward-induction reasoning (bi), plus a "simpleton" type ($*$) of Ann who plans optimally, but holds naively optimistic beliefs about Bob. Each belief map is as shown in Table 2.

| $\beta_a$ | $\varnothing$ | $(I_a)$ | $(I_a, I_b)$ | $(I_a, I_b, i_a)$ |
|---|---|---|---|---|
| $t_a^{\text{fi}}$ | $\left((O_a.o_a, I_b.o_b), t_b^{\text{fi}}\right), 1$ | $\left((I_a.o_a, I_b.o_b), t_b^{\text{fi}}\right), 1$ | $\left((I_a.o_a, I_b.o_b), t_b^{\text{fi}}\right), 1$ | $\left((I_a.i_a, I_b.o_b), t_b^{\text{fi}}\right), 1$ |
| $t_a^{\text{bi}}$ | $\left((O_a.o_a, O_b.o_b), t_b^{\text{bi}}\right), 1$ | $\left((I_a.o_a, O_b.o_b), t_b^{\text{bi}}\right), 1$ | $\left((I_a.o_a, I_b.o_b), t_b^{\text{bi}}\right), 1$ | $\left((I_a.i_a, I_b.o_b), t_b^{\text{bi}}\right), 1$ |
| $t_a^*$ | $((I_a.i_a, I_b.i_b), \cdot), 1$ | $((I_a.i_a, I_b.i_b), \cdot), 1$ | $((I_a.i_a, I_b.i_b), \cdot), 1$ | $((I_a.i_a, I_b.i_b), \cdot), 1$ |

| $\beta_b$ | $\varnothing$ | $(I_a)$ | $(I_a, I_b)$ | $(I_a, I_b, i_a)$ |
|---|---|---|---|---|
| $t_b^{\text{fi}}$ | $\left((O_a.o_a, I_b.o_b), t_a^{\text{fi}}\right), 1$ | $((I_a.i_a, I_b.o_b), t_a^*), 1$ | $((I_a.i_a, I_b.o_b), t_a^*), 1$ | $((I_a.i_a, I_b.o_b), t_a^*), 1$ |
| $t_b^{\text{bi}}$ | $\left((O_a.o_a, O_b.o_b), t_a^{\text{bi}}\right), 1$ | $\left((I_a.o_a, O_b.o_b), t_a^{\text{bi}}\right), 1$ | $\left((I_a.o_a, I_b.o_b), t_a^{\text{bi}}\right), 1$ | $\left((I_a.i_a, I_b.o_b), t_a^{\text{bi}}\right), 1$ |

**Table 2**: Type structure for the game of Figure 3.2.

We now explain in detail the features of the type structure.

**Ann** Type $t_a^{\text{fi}}$ has always the same beliefs about Bob: Bob's type is $t_b^{\text{fi}}$, he plans $I_b.o_b$, and he is going to execute his plan. The plan of $t_a^{\text{fi}}$ is $O_a.o_a$ in the following sense: conditional on each history where she is active, $t_a^{\text{fi}}$ assigns probability one to the corresponding action in $O_a.o_a$ (of course, given history $(I_a, I_b)$, Ann of type $t_a^{\text{fi}}$ must acknowledge that she deviated from her plan at the root). Given this, the plan of $t_a^{\text{fi}}$ is folding-back optimal. See Figure 3.3, where marked arcs of Ann represent her planned actions, marked arcs of Bob represent expected actions, the number in parentheses above each node of Bob represents Ann's expected payoffs conditional on reaching it, and the type of Bob in square brackets above each node of Ann represents Ann's conditional higher-order beliefs.

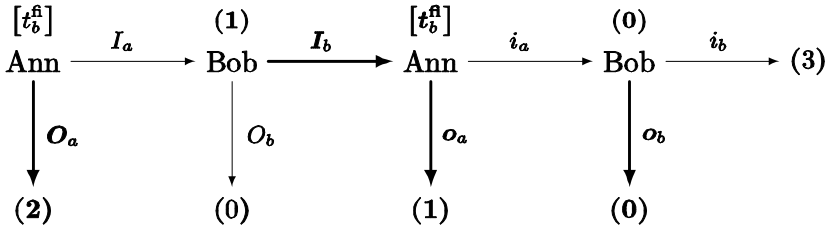



Figure 3.3: Plan and beliefs of type $t_a^{\text{fi}}$ of *Ann*.

Type $t_a^*$ of Ann is a "simpleton" who always believes that Bob plays $I_b.i_b$ and whose plan is $I_a.i_a$. (The higher-order beliefs of such type are irrelevant for the example, hence the dot in Table 2.) Given this, the folding-back optimal plan of $t_a^*$ is indeed $I_a.i_a$. See Figure 3.4.

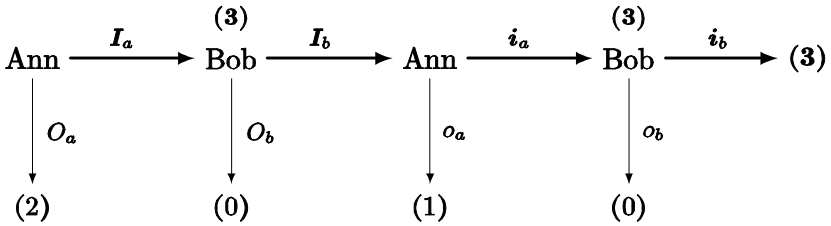



Figure 3.4: Plan and (first-order) beliefs of $t_a^*$.

Type $t_a^{\text{bi}}$ of Ann conforms to backward induction. Specifically, first-order beliefs yield the backward-induction pair $(O_a.o_a, O_b.o_b)$, higher-order beliefs are always concentrated on the backward-induction type of Bob.

**Bob** Type $t_b^{\text{fi}}$ plans $I_b.o_b$, believes at the beginning of the game that Ann's type is $t_a^{\text{fi}}$ and that she plays according to her plan $O_a.o_a$; upon observing action $I_a$, Bob of type $t_b^{\text{fi}}$ would believe that Ann's type is the simpleton $t_a^*$, and that she is playing $I_a.i_a$ as planned by $t_a^*$. See Figure 3.5, where the type of Ann on top of the root represents the initial higher-order belief of $t_b^{\text{bi}}$, and types above nodes of Bob represent his conditional higher-order beliefs.

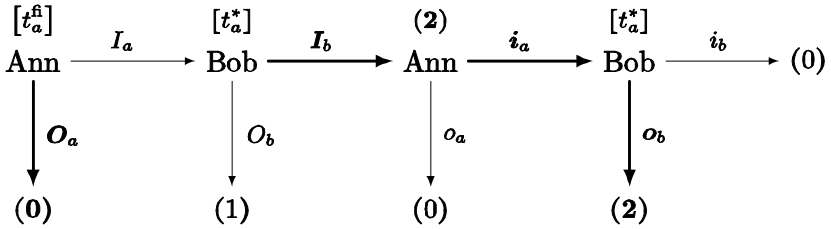



Figure 3.5: Plan and beliefs of type $t_b^{\text{fi}}$ of Bob.

Finally, it is immediate to check that type $t_b^{\text{bi}}$ of Bob conforms to backward induction.

**Rationality** Recall that rationality within a type structure is characterized by folding-back optimality of the subjective plan *and* consistency between subjective plan and objective behavior. This implies that if player $i$ believes in the rationality of co-player $-i$ conditional on observing history $h$, then $i$ also believes that each previous move of $-i$ in $h$ was made on purpose, in other words, that it was intentional. We can verify that a player is rational at each personal state of the extended type structure of Table 2 where she or he behaves as planned. In particular, Ann is rational at each $(s_a, t_a) \in \left\{\left(O_a.o_a, t_a^{\text{fi}}\right), \left(O_a.o_a, t_a^{\text{bi}}\right), \left(I_a.i_a, t_a^*\right)\right\}$, and Bob is rational at each $(s_b, t_b) \in \left\{\left(I_b.o_b, t_b^{\text{fi}}\right), \left(O_b.o_b, t_b^{\text{bi}}\right)\right\}$.

**Forward induction: Strong belief in optimal planning and consistency** With this, we can further verify that there is (intuitively) RCSBR at state

$$\left(\left(O_a.o_a, t_a^{\text{fi}}\right), \left(I_b.o_b, t_b^{\text{fi}}\right)\right),$$

that is, at this state players reason by forward induction. Specifically, upon observing the initially unexpected move $I_a$, type $t_b^{\text{fi}}$ keeps believing that Ann is rational, hence that action $I_a$ was intentional, although motivated by the rather naive beliefs of type $t_a^*$.

**Backward induction: Belief in continuation consistency** At state

$$\left(\left(O_a.o_a, t_a^{\text{bi}}\right), \left(O_b.o_b, , t_b^{\text{bi}}\right)\right)$$

Bob does not strongly believe in Ann's rationality; hence, RCSBR does not hold. Yet, there is something that players hold on to at this state: they always believe in (folding-back) optimal planning, although this means they would give up their belief in consistency between plan and behavior upon observing unexpected moves. Indeed, since each type $t_i^{\text{bi}}$ $(i = a, b)$ plans optimally and fully believes that the co-player's type is $t_{-i}^{\text{bi}}$, there is common full belief in optimal planning. On top of this, there is something else these types hold on to: although they interpret unexpected moves as unintentional mistakes, they expect that, in the continuation game, behavior will be consistent with plan. Call this epistemic event "**belief in continuation consistency**," or BCC. Then, at state $\left(\left(O_a.o_a, t_a^{\text{bi}}\right), \left(O_b.o_b, , t_b^{\text{bi}}\right)\right)$ there is BCC and also common full belief in BCC. To sum up, at this state the following epistemic hypotheses hold: (a) players are rational, i.e., they plan optimally and behavior is consistent with plan, (b) there is BCC, and (c) there is common full belief in optimal planning and BCC. We claim that this is an accurate epistemic representation of backward-induction reasoning. We provide a formal motivation for this claim in

Section 5, where we show that—in each finite, perfect-information game without relevant ties—epistemic hypotheses (a)-(c) yield the backward-induction behavior and beliefs.

# 4 Beliefs, plans and intentions

We first introduce a natural independence assumption that cleanly separates between plans and beliefs about others (4.1), next we analyze optimal planning (4.2), and finally we define rationality as the conjunction of optimal planning and consistency between plan and behavior (4.3).

## 4.1 Independence

For every $\Gamma$-based type structure $\mathcal{T}$ and every type of a player, viz. $t_i$, let

$$\beta_{i,i}(t_i) := \left(\text{marg}_{S_i}\beta_i(t_i)(\cdot|S_i(h))\right)_{h\in H}$$

and

$$\beta_{i,-i}(t_i) := \left(\text{marg}_{S_{-i}\times T_{-i}}\beta_i(t_i)(\cdot|S_{-i}(h)\times T_{-i})\right)_{h\in H}$$

respectively denote the marginal belief systems of $t_i$ about $i$'s own behavior and about the co-players $-i$.

**Definition 3** *We say that type $t_i$ in a $\Gamma$-based type structure $\mathcal{T}$ satisfies **independence** if, for all $h, h' \in H$,*

$$S_{-i}(h) = S_{-i}(h') \Rightarrow \beta_{i,-i}(t_i)(\cdot|S_{-i}(h)\times T_{-i}) = \beta_{i,-i}(t_i)(\cdot|S_{-i}(h')\times T_{-i}),$$

$$S_i(h) = S_i(h') \Rightarrow \beta_{i,i}(t_i)(\cdot|S_i(h)) = \beta_{i,i}(t_i)(\cdot|S_i(h')), \tag{4.1}$$

*and*

$$\beta_i(t_i)(\cdot|S(h)\times T_{-i}) = \beta_{i,i}(t_i)(\cdot|S_i(h))\times\beta_{i,-i}(t_i)(\cdot|S_{-i}(h)\times T_{-i}). \tag{4.2}$$

In words, $\beta_i(t_i)$ is the "product" of two marginal CPSs, one about $i$ himself and one about $-i$.[20]

[20] Condition (4.1) implies a weaker form of (4.2): if $h \prec h'$, then

$$\beta_i(t_i)(S(h')|S(h)\times T_{-i}) = \beta_{i,i}(t_i)(S_i(h')|S_i(h))\times\beta_{i,-i}(t_i)(S_{-i}(h')\times T_{-i}|S_{-i}(h)\times T_{-i}).$$

Note that from $\beta_{i,i}(t_i)$ and $\beta_{i,-i}(t_i)$ we can derive a **plan**

$$\sigma_{t_i,i} \in \prod_{h \in H} \Delta\left(A_i(h)\right),$$

which is—technically—a behavioral strategy (see Kuhn 1953), and a system of possibly correlated measures

$$\sigma_{t_i,-i} \in \prod_{h \in H} \Delta\left(A_{-i}(h)\right),$$

again a behavioral strategy if $-i$ is just one player. Formally, for all $h \in H$, $a_i \in A_i(h)$, and $a_{-i} \in A_{-i}(h)$,

$$\begin{aligned} \sigma_{t_i,i}(a_i|h) &: = \beta_{i,i}(t_i)\left(S_i(h,a_i)\,|S_i(h)\right), \\ \sigma_{t_i,-i}(a_{-i}|h) &: = \beta_{i,-i}(t_i)\left(S_{-i}(h,a_{-i}) \times T_{-i}|S_{-i}(h) \times T_{-i}\right), \end{aligned}$$

where $S_i(h,a_i) := \{s_i \in S_i(h) : s_i(h) = a_i\}$ is the set of personal external states of $i$ consistent with $h$ and choosing $a_i$ given $h$, and $S_{-i}(h,a_{-i}) := \prod_{j \neq i} S_j(h,a_j)$.

**Remark 1** *If $t_i$ satisfies independence, then*

$$\mathrm{marg}_S \beta_i(t_i)\left(S(h,a)\,|S(h)\right) = \sigma_{t_i,i}(a_i|h) \times \sigma_{t_i,-i}(a_{-i}|h)$$

*for all $h \in H$ and $a = (a_i, a_{-i}) \in A(h)$.*

We take independence to be a precondition for the rationality of player $i$. Refer back to the type structure in Table 2. The key feature of types $t_a^{\mathrm{fi}}$ and $t_a^{\mathrm{bi}}$ is that Ann's beliefs about the type $t_b$ and contingent behavior $s_b$ of Bob are independent of what Ann does, and in particular do not depend on whether Ann deviated or not from her plan. Indeed type $t_a^{\mathrm{fi}}$ (resp. $t_a^{\mathrm{bi}}$) of Ann initially plans to go out immediately and believes that Bob's personal state is $\left(I_b.o_b, t_b^{\mathrm{fi}}\right)$ (resp. $\left(O_b.o_b, t_b^{\mathrm{bi}}\right)$); upon observing a deviation to $I_a$ from her own plan, Ann keeps the same belief about Bob.

Next we define the other ingredients of the definition of rationality in this paper.

## 4.2 Optimal planning

For every $\Gamma$-based type structure $\mathcal{T}$, the expected payoff of type $t_i$ conditional on reaching history $h' \in \bar{H}$ is

$$V_{t_i}(h') := \sum_{s \in S(h')} U_i(s)\,\mathrm{marg}_S \beta_i(t_i)\left(s|S(h')\right)$$

(in particular, $V_{t_i}(z) = u_i(z)$ for each $z \in Z$). With this, the value of taking action $a_i \in A_i(h)$ conditional on $h \in H$ for a type $t_i$ that satisfies *independence* can be meaningfully defined as follows:

$$V_{t_i}(h, a_i) := \sum_{a_{-i} \in A_{-i}(h)} \sigma_{t_i,-i}(a_{-i}|h) V_{t_i}(h, (a_i, a_{-i})).$$

**Definition 4** *Type $t_i$ in a $\Gamma$-based type structure $\mathcal{T}$* ***plans optimally*** *if it satisfies independence and*

$$\text{supp}\sigma_{t_i,i}(\cdot|h) \subseteq \arg\max_{a_i \in A_i(h)} V_{t_i}(h, a_i)$$

*for all $h \in H$.*

In other words, we say that a type satisfying independence plans optimally if his plan has the one-shot-deviation (OSD) property. The set of types in $\mathcal{T}$ of player $i$ that satisfy optimal planning is denoted by

$$\overline{OP}_i := \left\{ t_i \in T_i : (4.1)\text{-}(4.2) \text{ hold}; \forall h \in H, \text{supp}\sigma_{t_i,i}(\cdot|h) \subseteq \arg\max_{a_i \in A_i(h)} V_{t_i}(h, a_i) \right\}.$$

The corresponding optimal-planning event about $i$ is $OP_i := S_i \times \overline{OP}_i$. We define $OP := \prod_{i \in I} OP_i$, and we can call $OP_i$ and $OP$ "events" because they are closed, hence Borel.

**Remark 2** *$\overline{OP}_i$ and $OP_i$ are closed.*

This is a shortcut to define optimality of a plan as the result of folding-back optimization, as it is well known that the latter is equivalent to the OSD property in every finite-horizon decision problem. Intuitively, if $h$ is a "pre-terminal" history, that is, $(h, a) \in Z$ for every $a \in A(h)$, then the OSD property implies the same maximization at $h$ as folding-back optimality; thus, $V_{t_i}(h) = V^*_{t_i}(h)$, where $V^*_{t_i}(h)$ denotes the value of reaching $h$ obtained by folding back. By backward recursion one can then prove that $V_{t_i}(h, a_i) = V^*_{t_i}(h, a_i)$ and $V_{t_i}(h) = V^*_{t_i}(h)$ for each $h \in H$ and $a_i \in A_i(h)$.

The following dynamic programming result is standard.

**Remark 3** *Fix a type $t_i$ that satisfies independence; $t_i$ plans optimally if and only if*

$$\text{supp}\beta_{i,i}(t_i)(\cdot|S_i(h)) \subseteq \arg\max_{s_i \in S_i(h)} \sum_{s_{-i} \in S_{-i}(h)} U_i(s_i, s_{-i}) \text{marg}_{S_{-i}} \beta_{i,-i}(t_i)(s_{-i}|S_{-i}(h))$$

*for all $h \in H$.*

## 4.3 Consistency and rationality

Recall that a personal state of player $i$ in a $\Gamma$-based type structure $\mathcal{T}$ is a pair $(s_i, t_i)$ that contains two possibly distinct descriptions of the "strategy" of player $i$: $s_i$ is interpreted as an *objective* description of $i$'s contingent behavior, that is, what other players have to predict in order to assess the likely consequences of their actions; $\sigma_{t_i,i}$—derived from $\beta_{i,i}(t_i)$—is the *subjective plan* of $i$. A consistent player behaves as planned; a rational player plans optimally and is consistent:

**Definition 5** *Player $i$ is **consistent from** history $h$ at personal state $(s_i, t_i)$ of a $\Gamma$-based type structure $\mathcal{T}$ if $s_i$ and $\sigma_{t_i,i}$ coincide on the subgame with root $h$, that is, $\sigma_{t_i,i}(s_i(h')|h') = 1$ for all $h' \in H$ with $h \preceq h'$; player $i$ is **consistent** at $(s_i, t_i)$ if he is consistent from the empty history $\varnothing$; player $i$ is **rational** at $(s_i, t_i)$ if he is consistent at $(s_i, t_i)$ and type $t_i$ plans optimally.*

To ease notation, for each $h \in H$, let

$$H(h) := \{h' \in H : h \preceq h'\}$$

denote the set of nonterminal histories that weakly follow $h$. For every $\Gamma$-based type structure $\mathcal{T}$, the sets of personal states where $i$ is consistent from $h$, consistent, and rational are respectively denoted by

$$\begin{aligned} C_i^{\succeq h} &: = \{(s_i, t_i) \in S_i \times T_i : \forall h' \in H(h), \sigma_{t_i,i}(s_i(h')|h') = 1\}, \\ C_i &: = C_i^{\succeq \varnothing}, \\ R_i &: = C_i \cap OP_i. \end{aligned}$$

Also these sets are events about $i$, because they are closed, hence Borel.

**Remark 4** *$C_i^{\succeq h}$ $(h \in H)$ and $R_i$ are closed.*

We define the set of all states of the world where each player is consistent as

$$C := \prod_{i \in I} C_i;$$

by Remark 4, $C$ is a Borel subset of $\prod_{i \in I}(S_i \times T_i)$.

For example, in the type structure of Section 3 for the game of Figure 3.2, all types plan optimally, and so the players are rational at all personal states at which they are consistent and irrational at the other states. Furthermore, all types of Bob believe at the beginning of the game that Ann is consistent (and rational). But there

is a key difference in epistemic attitudes conditional on the unexpected move $I_a$ of Ann: forward-induction type $t_b^{\text{fi}}$ of Bob would keep believing that Ann is consistent also if he observed $I_a$, hence $t_b^{\text{fi}}$ must change belief about the plan of Ann conditional on $I_a$; backward-induction type $t_b^{\text{bi}}$ instead would keep the initial belief in Ann's plan to go out and would think—upon observing $I_a$—that Ann is not (globally) consistent and yet she will be consistent from $h$.

Some remarks on the notion of rationality are in order. First note that the notion of rationality considered here is richer and stronger than the notion of rationality usually adopted in epistemic game theory. It is richer because here we distinguish between plan and objective behavior, and the requirement that they coincide is part of the rationality conditions. It is stronger because, if $i$ is rational at $(s_i, t_i)$, then $s_i$ is optimal given $\beta_{i,-i}(t_i)$ conditional on *every* history $h$, not only those consistent with $s_i$ itself. There are two related reasons for this stronger requirement. First, here we take the perspective that players can only (irreversibly) choose actions, rather than strategies; therefore, the conceptually primary notion of optimization must concern the choice of actions at different histories, and a dynamically optimal plan must satisfy such "action optimality" at *every* history of $i$, otherwise early choices of $i$ may be based on the prediction that $i$ himself would choose irrationally in some future contingency. Second, we interpret optimality as the result of folding-back planning: when $i$ is considering what action he would choose, should history $h$ occur, he has already determined his contingent plan for histories following $h$, but not yet for those preceding $h$.

Finally, note that our notion of consistency requires that players hold deterministic plans. It makes sense to consider a weaker notion of consistency whereby $s_i$ is in the "support" of $\sigma_{t_i,i}$, that is, $\sigma_{t_i,i}(s_i(h)|h) > 0$ for all $h \in H$.[21] But this generalization would not change the substance of our results.

# 5 Backward-induction reasoning: neglect of perceived deviations

In this section we present epistemic assumptions that—we claim—capture faithfully the spirit of backward-induction (henceforth BI) reasoning. We first show that these assumptions yield the BI plans and beliefs in perfect-information games without relevant ties. Next we generalize the result to games with observable actions, showing

[21] This means that $s_i$ is in the support of the mixed strategy that corresponds to $\sigma_{t_i,i}$ according to Kuhn's (1953) transformation.

that they yield a solution concept called "backwards rationalizability" (cf. Perea 2014, Penta 2015).[22]

Although our notion of type structure allows us to represent subjective plans and consistency between plan and behavior, we focus on what each player believes about other players. Specifically, for any player $i \in I$, event $E_{-i} \subseteq S_{-i} \times T_{-i}$, and history $h \in H$, we let

$$\mathrm{B}_{i,h}(E_{-i}) := S_i \times \left\{ t_i \in T_i : \beta_{i,h}(t_i)(S_i \times E_{-i}) = 1 \right\}$$

denote the event that $i$ **believes** $E_{-i}$ **given** $h$. Thus,

$$\mathrm{B}_i(E_{-i}) := \bigcap_{h \in H} \mathrm{B}_{i,h}(E_{-i})$$

denotes the event that $i$ **fully believes** $E_{-i}$. Note that these belief operators satisfy conjunction and monotonicity. Furthermore, if $E_{-i}$ is closed then $\mathrm{B}_{i,h}(E_{-i})$ and $\mathrm{B}_i(E_{-i})$ are closed as well. We let $\mathrm{B}(\cdot)$ denote the **mutual full belief** operator, that is, $\mathrm{B}(E) := \prod_{i \in I} \mathrm{B}_i(E_{-i})$ for each event $E := \prod_{i \in I} E_i$; as standard, $\mathrm{B}^m = \mathrm{B} \circ \mathrm{B}^{m-1}$ denotes the $m$-th iteration ($m \in \mathbb{N}$) of the selfmap B, that is,

$$\mathrm{B}^m(E) := \mathrm{B}\left(\mathrm{B}^{m-1}(E)\right),$$

where $\mathrm{B}^0(E) := E$ by convention.

Our representation of BI reasoning is based on the following epistemic assumption: each player $i$ **believes in** the **continuation consistency** of the other players, that is, for each history $h \in H$, $i$ would believe $C_{-i}^{\succeq h} := \prod_{j \neq i} C_j^{\succeq h}$ upon observing $h$. The corresponding events are

$$\begin{aligned} BCC_i &:= \bigcap_{h \in H} \mathrm{B}_{i,h}\left(C_{-i}^{\succeq h}\right), \\ BCC &:= \prod_{i \in I} BCC_i. \end{aligned}$$

In a sense, a player who believes in continuation consistency stubbornly neglects the past: no evidence of deviations from what he believes to be the plans of co-players makes him doubt that in the future they will follow such plans, as in the "trembling-hand" story by Selten (1975). Since each $C_{-i}^{\succeq h}$ is a product of closed sets, hence itself closed, $BCC_i$ is closed as well.

With this, define recursively the following epistemic events:

[22] A version of our result about BI in games with perfect information can be obtained as a corollary of the theorem on backwards rationalizability. But we present it first as a separate result (with the proof in the main text) because it is simpler and it allows to better appreciate our framework.

- $OP_i^1 := OP_i \cap BCC_i$,
- $OP_i^{m+1} := OP_i^m \cap \mathrm{B}_i\left(OP_{-i}^m\right)$, where $OP_{-i}^m := \prod_{j \neq i} OP_j^m$.

For each $m \in \mathbb{N}$, we define the set $OP^m \subseteq \prod_{i \in I}(S_i \times T_i)$ in the usual way, that is, $OP^m := \prod_{i \in I} OP_i^m$. Note that each $OP_i^1$ is closed ($i \in I$); furthermore, if $OP_{-i}^m$ is closed, then $\mathrm{B}_i\left(OP_{-i}^m\right)$ and $OP_i^{m+1} = OP_i^m \cap \mathrm{B}_i\left(OP_{-i}^m\right)$ are closed. It follows by induction that $\left(OP_i^m\right)_{m \in \mathbb{N}}$ is a well defined decreasing sequence of closed sets.

**Remark 5** *For each $m \in \mathbb{N}$,*

$$\begin{aligned} OP^{m+1} &= (OP \cap BCC) \cap \bigcap_{k=1}^{m} \mathrm{B}^k (OP \cap BCC) \\ &= \left(OP \cap \bigcap_{k=1}^{m} \mathrm{B}^k (OP)\right) \cap \left(BCC \cap \bigcap_{k=1}^{m} \mathrm{B}^k (BCC)\right). \end{aligned}$$

## 5.1 Backward induction in games with perfect information

Consider a game $\Gamma$ that can be solved by BI and let $s^{\mathrm{bi}}$ denote its **BI external state**, that is, the outcome of the BI algorithm. We claim that optimal planning, belief in continuation consistency, and common belief in both imply that players believe, conditional on each $h \in H$, that everybody will play according to $s^{\mathrm{bi}}$ in the subgame with root $h$. To simplify the exposition, we focus on games with *perfect information* (PI games) and without relevant ties, but the result can be extended to other BI-solvable games, such as finitely repeated Prisoners' Dilemmas. Recall that a PI game $\Gamma$ is **without relevant ties** if for all $z, z' \in Z$ and all $i \in I$, if $z \neq z'$ and $i$ is the active player at the last common predecessor of $z$ and $z'$, then $u_i(z) \neq u_i(z')$. The game of Figure 3.2 is an instance of a PI game without relevant ties.

We first note that the aforementioned epistemic assumptions can be satisfied in every game with a pure subgame perfect equilibrium, hence in every BI-solvable game. For each $s_i \in S_i$ and $h \in H$, let $s_i^h$ denote the minimal modification of $s_i$ that makes $h$ reachable.[23]

**Remark 6** *For every finite game $\Gamma$ with observable actions, if there is a pure subgame perfect equilibrium $\bar{s}$ then there exists a $\Gamma$-based type structure $\mathcal{T}$ such that*

[23] Note that, for any pair of related histories $h' \prec h$, there is a unique action profile $\alpha(h', h) = (\alpha_i(h', h))_{i \in I} \in A(h')$ such that $(h', \alpha(h', h)) \preceq h$. With this, given $s_i \in S_i$ and history $h \in H$, $s_i^h$ is defined as the personal external state that coincides with $s_i$ at every history $h'$ that does not precede $h$ and takes action $\alpha_i(h', h)$ at every $h' \prec h$.

$OP^{\infty} := \cap_{m\in\mathbb{N}} OP^{m} \neq \emptyset$ *and* $C \cap OP^{\infty} \neq \emptyset$. *To see this, consider the following type structure: the type set of each player is a singleton, that is,* $T_i := \{\bar{t}_i\}$ *for each* $i \in I$*; furthermore, each belief map is such that* $\beta_{i,h}(\bar{t}_i)\left(\left\{\left(\bar{s}_j^h\right)_{j\in I}\right\} \times T_{-i}\right) = 1$ *for every* $h \in H$. *It is immediate to check that in this type structure* $OP^{\infty} = S \times \{\bar{t}\}$ *and* $C \cap OP^{\infty} = \{(\bar{s}, \bar{t})\}$.

In BI-solvable games with perfect information, the number of steps of the BI algorithm necessary to obtain belief in the BI continuation behavior in a subgame with root $h$ is given by the **height** of $h$, $L(h) := \max_{z\in Z, z\succ h} \ell(z) - \ell(h)$, where $\ell(\cdot)$ denotes the length of a sequence. To state the following result it is convenient to let $\sigma_{t_i}(a|h) := \beta_i(t_i)(S(h,a) \times T_{-i} | S(h) \times T_{-i})$ denote the probability assigned by type $t_i$ to action profile $a \in A(h)$ conditional on $h$. In a PI game, this is just the probability assigned by $t_i$ to $a_{\iota(h)}$, the action of the only player $\iota(h)$ who is **active at** $h$.

**Lemma 1** *Fix a finite PI game* $\Gamma$ *without relevant ties and a* $\Gamma$*-based type structure* $\mathcal{T}$. *For each history* $h \in H$ *and each personal state* $\left(s_{\iota(h)}, t_{\iota(h)}\right) \in C_{\iota(h)} \cap OP_{\iota(h)}^{L(h)}$ *of the player who is active at* $h$*, this player believes that the BI contingent behavior will be followed in the subgame with root* $h$ *and, furthermore, his behavior conforms to BI in the same subgame, that is,* $\sigma_{t_{\iota(h)}}(s^{\mathrm{bi}}(h')|h') = 1$ *and* $s_{\iota(h)}(h') = s_{\iota(h)}^{\mathrm{bi}}(h')$ *for each* $h' \in H(h)$.

**Proof**. Let

$$T_i^{\mathrm{bi},1}(h) := \left\{t_i \in T_i : \forall h' \in H(h), \sigma_{t_i}(s^{\mathrm{bi}}(h')|h') = 1\right\}$$

denote the set of types of $i$ whose first-order beliefs conform to BI in the subgame with root $h$, and let

$$\left[s_i^{\mathrm{bi}}\right]^{\succeq h} := \left\{s_i \in S_i : \forall h' \in H(h), s_i(h') = s_i^{\mathrm{bi}}(h')\right\}$$

denote the set of external states of $i$ that coincide with $s_i^{\mathrm{bi}}$ on $H(h)$. First note that, for every $h \in H$ and $t_{\iota(h)} \in T_{\iota(h)}^{\mathrm{bi},1}(h)$,

$$\arg \max_{a_{\iota(h)} \in A_{\iota(h)}(h)} V_{t_{\iota(h)}}\left(h, a_{\iota(h)}\right) = s_{\iota(h)}^{\mathrm{bi}}(h),$$

because of perfect information, no relevant ties and $t_{\iota(h)}$'s belief in the BI continuation after every action. We prove by induction on the height of history $h$ that

$$\begin{aligned} OP_{\iota(h)}^{L(h)} &\subseteq S_{\iota(h)} \times T_{\iota(h)}^{\mathrm{bi},1}(h), \\ C_{\iota(h)}^{\succeq h} \cap OP_{\iota(h)}^{L(h)} &\subseteq [s_{\iota(h)}^{\mathrm{bi}}]^{\succeq h} \times T_{\iota(h)}, \end{aligned}$$

for each $h \in H$.

*Basis step.* Suppose that $L(h) = 1$. Then, $H(h) = \{h\}$, $OP^{L(h)}_{\iota(h)} = OP_{\iota(h)} \cap BCC_{\iota(h)}$ and $BCC_{\iota(h)}$ puts no restriction on beliefs about future moves. Thus,

$$\begin{aligned} OP^{L(h)}_{\iota(h)} &= OP^{1}_{\iota(h)} \\ &= OP_{\iota(h)} \cap BCC_{\iota(h)} \\ &\subseteq S_{\iota(h)} \times \left\{ t_{\iota(h)} \in T_{\iota(h)} : \mathrm{supp}\sigma_{t_{\iota(h)},\iota(h)}(\cdot|h) \subseteq \arg \max_{a_{\iota(h)} \in A_{\iota(h)}(h)} V_{t_{\iota(h)}}\left(h, a_{\iota(h)}\right) \right\} \\ &= S_{\iota(h)} \times \left\{ t_{\iota(h)} \in T_{\iota(h)} : \sigma_{t_{\iota(h)}}(s^{\mathrm{bi}}(h)|h) = 1 \right\} \\ &= S_{\iota(h)} \times T^{\mathrm{bi},1}_{\iota(h)}(h), \end{aligned}$$

and

$$C^{\succeq h}_{\iota(h)} \cap OP^{L(h)}_{\iota(h)} \subseteq [s^{\mathrm{bi}}_{\iota(h)}]^{\succeq h} \times T^{\mathrm{bi},1}_{\iota(h)}(h).$$

*Inductive step.* Fix an integer $k$ with $1 \leq k < L(\varnothing)$. Suppose by way of induction that, for every history $h'$ with $L(h') \leq k$,

$$\begin{aligned} OP^{L(h')}_{\iota(h')} &\subseteq S_{\iota(h')} \times T^{\mathrm{bi},1}_{\iota(h')}\left(h'\right), \\ C^{\succeq h'}_{\iota(h')} \cap OP^{L(h')}_{\iota(h')} &\subseteq [s^{\mathrm{bi}}_{\iota(h')}]^{\succeq h'} \times T_{\iota(h')}. \end{aligned}$$

Let $L(h) = k + 1$. Note that, by definition of the sequences $\left(OP^m_j\right)_{m\in\mathbb{N}}$ $(j \in I)$,

$$\begin{aligned} OP^{L(h)}_{\iota(h)} &= OP^{k+1}_{\iota(h)} \\ &= OP^{k}_{\iota(h)} \cap \mathrm{B}_{\iota(h)}(OP^{k}_{-\iota(h)}) \\ &= OP^{k}_{\iota(h)} \cap BCC_{\iota(h)} \cap \mathrm{B}_{\iota(h)}(OP^{k}_{-\iota(h)}), \end{aligned}$$

where the latter equality holds because, by definition, $OP^k_j \subseteq BCC_j$ for each $j$ and $k$.

Next note that $OP^k_{\iota(h')} \subseteq OP^{L(h')}_{\iota(h')}$ for every $h' \succ h$, because $\left(OP^m_j\right)_{m\in\mathbb{N}}$ is a nested sequence of subsets for each $j$, and $L(h') \leq k = L(h) - 1$ by assumption. By definition of $BCC_{\iota(h)}$ and of full belief, by monotonicity, and by the inductive hypothesis, $BCC_{\iota(h)} \cap \mathrm{B}_{\iota(h)}(OP^k_{-\iota(h)})$ implies that $\iota(h)$ expects his co-players to take

the BI actions at future histories, which must have height $k$ or less:

$$\begin{aligned}
& BCC_{\iota(h)} \cap \mathrm{B}_{\iota(h)}\left(OP^{k}_{-\iota(h)}\right) \\
\subseteq\ & \mathrm{B}_{\iota(h)}\left(\left(\prod_{h' \succ h : \iota(h') \neq \iota(h)} C^{\succeq h'}_{\iota(h')} \cap OP^{L(h')}_{\iota(h')}\right) \times S_{-\iota(h)\iota(h')} \times T_{-\iota(h)\iota(h')}\right) \\
\subseteq\ & \mathrm{B}_{\iota(h)}\left(\left(\prod_{h' \succ h : \iota(h') \neq \iota(h)} \left([s^{\mathrm{bi}}_{\iota(h')}]^{\succeq h'} \times T_{\iota(h')}\right)\right) \times S_{-\iota(h)\iota(h')} \times T_{-\iota(h)\iota(h')}\right) \\
\subseteq\ & S_{\iota(h)} \times \left\{t_{\iota(h)} : \forall h' \in H(h), \iota(h') \neq \iota(h) \Rightarrow \sigma_{t_{\iota(h)}}(s^{\mathrm{bi}}(h')|h') = 1\right\},
\end{aligned}$$

where $-ij$ denotes $I \backslash \{i,j\}$. With this, (folding-back) optimal planning of $\iota(h)$ implies that he plans to choose the BI action at $h$ and every $h' \succ h$ with $\iota(h') = \iota(h)$:

$$\begin{aligned}
OP^{L(h)}_{\iota(h)} \subseteq\ & OP_{\iota(h)} \cap \left(S_{\iota(h)} \times \left\{t_{\iota(h)} : \forall h' \in H(h), \iota(h') \neq \iota(h) \Rightarrow \sigma_{t_{\iota(h)}}(s^{\mathrm{bi}}(h')|h') = 1\right\}\right) \\
\subseteq\ & S_{\iota(h)} \times \left\{t_{\iota(h)} : \forall h' \in H(h), \sigma_{t_{\iota(h)}}(s^{\mathrm{bi}}(h')|h') = 1\right\} \\
=\ & S_{\iota(h)} \times T^{\mathrm{bi},1}_{\iota(h)}(h).
\end{aligned}$$

Adding consistency from $h$, we get that $\iota(h)$ would indeed take the BI action at each history in the subgame with root $h$:

$$C^{\succeq h}_{\iota(h)} \cap OP^{L(h)}_{\iota(h)} \subseteq [s^{\mathrm{bi}}_{\iota(h)}]^{\succeq h} \times T_{\iota(h)}.$$

■

Say that player $i$ has the **backward-induction plan** at personal state $(s_i, t_i)$ if he plans to follow the BI contingent behavior. This gives the epistemic event

$$BIP_i := \left\{(s_i, t_i) : \forall h \in H, \sigma_{t_i,i}(s^{\mathrm{bi}}_i(h)|h) = 1\right\}.$$

We let $BIP := \prod_{i \in I} BIP_i$ denote the set of all states of the world in which each player has the backward-induction plan at his personal state.

**Corollary 1** *Fix a finite PI game $\Gamma$ without relevant ties and a $\Gamma$-based type structure $\mathcal{T}$. Then $OP_i^{L(\varnothing)} \subseteq BIP_i$ and $OP_i^{L(\varnothing)} \cap C_i \subseteq \{s^{\mathrm{bi}}_i\} \times T_i$ for every $i \in I$.*

**Proof.** Let $H^1_i$ denote the set of histories where player $i$ is active for the first time. Then $\{s^{\mathrm{bi}}_i\} = \bigcap_{h \in H^1_i} [s^{\mathrm{bi}}_i]^{\succeq h}$ and $BIP_i \subseteq \bigcap_{h \in H^1_i} S_i \times T^{\mathrm{bi},1}_i(h)$. Also, $L(\varnothing) \geq L(h)$,

$C_i = C_i^{\succeq \varnothing} \subseteq C_i^{\succeq h} = C_{\iota(h)}^{\succeq h}$ and $OP_i^{L(\varnothing)} \subseteq OP_i^{L(h)} = OP_{\iota(h)}^{L(h)}$ for every $h \in H_i^1$. Therefore, Lemma 1 implies

$$\begin{aligned} OP_i^{L(\varnothing)} \cap C_i &\subseteq \bigcap_{h \in H_i^1} OP_i^{L(h)} \cap C_i^{\succeq h} \\ &\subseteq \bigcap_{h \in H_i^1} [s_i^{\mathrm{bi}}]^{\succeq h} \times T_i^{\mathrm{bi},1}(h) \\ &\subseteq \left(\{s_i^{\mathrm{bi}}\} \times T_i\right) \cap BIP_i. \end{aligned}$$

■

We say that an event $E := \prod_{i \in I} E_i$ is **transparent** at state $(s,t)$ if $(s,t) \in E$ and there is common full belief in $E$ at $(s,t)$; thus, the set of states where $E$ is transparent is

$$E \cap \bigcap_{m \in \mathbb{N}} \mathrm{B}^m(E) = \bigcap_{n \in \mathbb{N}_0} \mathrm{B}^n(E).$$

**Corollary 2** *Fix a finite PI game $\Gamma$ without relevant ties and a $\Gamma$-based type structure $\mathcal{T}$. Then consistency and transparency of optimal planning and of belief in continuation consistency imply BI behavior:*

$$C \cap \bigcap_{n \in \mathbb{N}_0} \mathrm{B}^n(OP \cap BCC) \subseteq \{s^{\mathrm{bi}}\} \times T.$$

**Proof.** By Remark 5,

$$\begin{aligned} & C \cap \bigcap_{m \in \mathbb{N}_0} \mathrm{B}^m(OP \cap BCC) \\ \subseteq\ & C \cap \bigcap_{n=0}^{L(\varnothing)} \mathrm{B}^n(OP \cap BCC) = C \cap OP^{L(\varnothing)}. \end{aligned}$$

By Corollary 1,

$$C \cap OP^{L(\varnothing)} \subseteq \{s^{\mathrm{bi}}\} \times T.$$

■

The foregoing analysis yields our main result about BI reasoning.

**Theorem 1** *Fix a finite PI game $\Gamma$ without relevant ties and a $\Gamma$-based type structure $\mathcal{T}$. Then transparency of optimal planning and of belief in continuation consistency implies transparency of BI planning:*

$$\bigcap_{n\in\mathbb{N}_0} \mathrm{B}^n\left(OP \cap BCC\right) \subseteq \bigcap_{n\in\mathbb{N}_0} \mathrm{B}^n\left(BIP\right).$$

**Proof.** The mutual full belief operator $\mathrm{B}\left(\cdot\right)$ satisfies conjunction and (as a consequence) monotonicity. Therefore, one can show by standard arguments that, for all $m \in \mathbb{N}_0$ and events $E, F$,

$$\bigcap_{k=0}^{m} \mathrm{B}^k(E) \subseteq F \Rightarrow \bigcap_{n\in\mathbb{N}_0} \mathrm{B}^n\left(E\right) \subseteq \bigcap_{n\in\mathbb{N}_0} \mathrm{B}^n\left(F\right).$$

Remark 5 and Corollary 1 imply that

$$\bigcap_{k=0}^{L(\varnothing)-1} \mathrm{B}^k\left(OP \cap BCC\right) = OP^{L(\varnothing)} \subseteq BIP.$$

Therefore,

$$\bigcap_{n\in\mathbb{N}_0} \mathrm{B}^n\left(OP \cap BCC\right) \subseteq \bigcap_{n\in\mathbb{N}_0} \mathrm{B}^n\left(BIP\right).$$

■

## 5.2 Backwards rationalizability

We now show how the foregoing analysis on BI reasoning can be extended to the general class of finite multistage games with perfect monitoring of past actions. Specifically, we show that the behavioral implications of the aforementioned epistemic assumptions are characterized by the solution concept of backwards rationalizability (cf. Penta 2015, Perea 2014), which we introduce next.

Let $\mathcal{Q}$ be the collection of all subsets of $S$ with the form $Q = \prod_{i\in I} Q_i$, where $Q_i \subseteq S_i$ for every $i$. For every $h \in H$, let $\chi^h : \mathcal{Q} \to \mathcal{Q}$ be the operator defined as follows: for all $Q \in \mathcal{Q}$,

$$\begin{aligned} \chi_i^h\left(Q_i\right) &: = \left\{s_i \in S_i\left(h\right) : \exists \bar{s}_i \in Q_i, \forall h' \in H\left(h\right), s_i\left(h'\right) = \bar{s}_i\left(h'\right)\right\}, \\ \chi^h\left(Q\right) &: = \prod_{i\in I} \chi_i^h\left(Q_i\right). \end{aligned}$$

In words, each $\chi_i^h(Q_i)$ is the set of all $s_i \in S_i(h)$ whose continuation in subgame with root $h$ coincides with those in $Q_i$. Note that $\chi^{\varnothing}(Q) = Q$, and $\chi^h(S) = S(h)$ for all $h \in H$.

For every CPS $\mu_i$ on $(S_{-i}, \mathcal{S}_{-i})$, we let $\rho_i(\mu_i)$ denote the set of all **sequential best replies** to $\mu_i$, that is,

$$\rho_i(\mu_i) := \left\{ s_i \in S_i : \forall h \in H, s_i^h \in \arg\max_{r_i \in S_i(h)} \mathbb{E}_{\mu_i}[U_i(r_i, \cdot)|h] \right\},$$

where $\mathbb{E}_{\mu_i}[U_i(r_i, \cdot)|h]$ denotes the expected payoff of $r_i$ conditional on $h$ given CPS $\mu_i$.[24]

**Definition 6** *Consider the following procedure.*

**(Step 0)** *For every $i \in I$, let $\hat{S}_i^0 := S_i$. Also, let $\hat{S}_{-i}^0 := \prod_{j \neq i} \hat{S}_j^0$ and $\hat{S}^0 := \prod_{i \in I} \hat{S}_i^0$.*

**(Step $n > 0$)** *For every $i \in I$ and every $s_i \in S_i$, let $s_i \in \hat{S}_i^n$ if and only if there exists $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that*

1. *$s_i \in \rho_i(\mu_i)$;*
2. *$\mu_i\left(\chi_{-i}^h\left(\hat{S}_{-i}^{n-1}\right)|S_{-i}(h)\right) = 1$ for every $h \in H$.*

*Also, let $\hat{S}_{-i}^n := \prod_{j \neq i} \hat{S}_j^n$ and $\hat{S}^n := \prod_{i \in I} \hat{S}_i^n$.*

*Finally, let $\hat{S}^\infty := \cap_{n \in \mathbb{N}_0} \hat{S}^n$. The profiles in $\hat{S}^\infty$ are called **backwards rationalizable**.*

One can show by standard arguments that backwards rationalizability is a non-empty solution procedure:

**Remark 7** *$\hat{S}^\infty \neq \emptyset$.*

We illustrate the above iterative procedure by means of the PI game of Figure 3.2. At the first step, we have

$$\hat{S}^1 = \{I_a.i_a, O_a.i_a, O_a.o_a\} \times \{I_b.o_b, O_b.o_b\}.$$

For Ann, we rule out $I_a.o_a$ because Ann plans to choose action $o_a$ only if her conditional belief at history $(I_a, I_b)$ assigns sufficiently low probability to $I_b.i_b$ and $O_b.i_b$;

[24] Recall that $s_i^h$ denotes the minimal modification of $s_i$ that makes $h$ reachable.

given this conditional belief, the optimal action for Ann at history $\varnothing$ is $O_a$. For Bob, we rule out both $I_b.i_b$ and $O_b.i_b$ because action $i_b$ is not optimal at history $(I_a, I_b, i_a)$.

One can verify that at the second step

$$\hat{S}^2 = \{O_a.o_a\} \times \{I_b.o_b, O_b.o_b\},$$

and the procedure ends at the third step, i.e.,

$$\hat{S}^\infty = \hat{S}^3 = \{O_a.o_a\} \times \{O_b.o_b\}.$$

In Appendix B, we show how the solution concept of backwards rationalizability can be given a characterization in terms of the so-called "backwards procedure" (Penta 2015), which is an extension of the BI algorithm to the general class of finite multistage games with observable past actions. Specifically, the "backwards procedure" coincides with the BI algorithm in PI games without relevant ties.

We now show that backwards rationalizability characterizes the behavioral implications of consistency and common full belief of optimal planning and belief in continuation consistency in the canonical type structure.

We say that a $\Gamma$-based type structure $\mathcal{T}$ is **complete** if each belief map $\beta_i$ is onto. As is well known, the canonical type structure is complete (see Proposition 2 in Battigalli and Siniscalchi 1999a).[25] With this, we can now state the main result of this section.[26]

**Theorem 2** *Fix a finite game $\Gamma$ and a $\Gamma$-based* complete *type structure $\mathcal{T}$. Then,*
*(i) for every $n \in \mathbb{N}$, $\text{proj}_S(OP^n \cap C) = \hat{S}^n$;*
*(ii) $\text{proj}_S(OP^\infty \cap C) = \hat{S}^\infty$.*

The proof of Theorem 2 is in Appendix A.

# 6 Forward-induction reasoning: rationalization of past moves

In Section 3, we have informally claimed that forward-induction reasoning can be modelled by the epistemic assumption of strong belief in rationality, that is, strong belief in consistency and optimal planning. Here we make this claim precise.

[25] Although the canonical type structure is the conceptual backdrop of our epistemic analysis, we do not need to use it explicitly for the statements of the formal results.

[26] For any pair of sets $X$ and $Y$, we let $\text{proj}_X$ denote the canonical projection map from $X \times Y$ onto $X$.

A player strongly believes a nonempty event $E$ if he is certain of $E$ at all histories consistent with $E$. Formally, fix a game $\Gamma$ and an associated $\Gamma$-based type structure $\mathcal{T}$. For every $i \in I$ and event $E_{-i} \subseteq S_{-i} \times T_{-i}$, let

$$\mathrm{SB}_i(E_{-i}) := \bigcap_{h \in H : S_{-i}(h) \times T_{-i} \cap E_{-i} \neq \emptyset} \mathrm{B}_{i,h}(E_{-i})$$

denote the event that $i$ **strongly believes** $E_{-i}$. With this, rationality and common strong belief in rationality (RCSBR) can be defined as follows. For each $i \in I$, let $R_i^1 := R_i$ (recall that $R_i$ is the set of personal states $(s_i, t_i)$ where $i$ is consistent and $t_i$ plans optimally: $R_i = C_i \cap OP_i$); for each $n \in \mathbb{N}$, define $R_i^{n+1}$ recursively as follows:

$$R_i^{n+1} := R_i^n \cap \mathrm{SB}_i\left(R_{-i}^n\right),$$

where $R_{-i}^n := \prod_{j \neq i} R_j^n$. An easy induction argument shows that each set $R_i^n$ is closed in $S_i \times T_i$, hence Borel.[27] The set of states consistent with RCSBR is therefore defined as

$$R^\infty := \prod_{i \in I} \bigcap_{n \in \mathbb{N}} R_i^n.$$

**Definition 7** *Consider the following procedure.*

**(Step 0)** *For every $i \in I$, let $S_i^0 := S_i$. Also, let $S_{-i}^0 := \prod_{j \neq i} S_j$ and $S^0 := S$.*

**(Step $n > 0$)** *For every $i \in I$ and every $s_i \in S_i$, let $s_i \in S_i^n$ if and only if there exists $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that*

1. *$s_i \in \rho_i(\mu_i)$;*
2. *for every $m \in \{0, ..., n-1\}$ and $h \in H$,*

$$S_{-i}^m \cap S_{-i}(h) \neq \emptyset \Rightarrow \mu_i\left(S_{-i}^m | S_{-i}(h)\right) = 1.$$

*Also, let $S_{-i}^n := \prod_{j \neq i} S_j^n$ and $S^n := \prod_{i \in I} S_i^n$.*

*Finally, let $S^\infty := \cap_{n \in \mathbb{N}} S^n$. The external states in $S^\infty$ are called **strongly rationalizable**.*

[27] Recall that if $E_{-i} \subseteq S_{-i} \times T_{-i}$ is closed, so is $\mathrm{B}_{i,h}(E_{-i})$. By finiteness of the game, $\mathrm{SB}_i(E_{-i})$ is a finite intersection of closed sets, hence it is closed. Using this fact and Remark 4, it follows by induction that each set $R_i^n$ is closed.

As for backwards rationalizability, one can show by standard arguments that strong rationalizability is a nonempty solution procedure:

**Remark 8** $S^{\infty} \neq \emptyset$.

In Section 7 we will compare strong rationalizability as per Definition 7 to the extensive-form rationalizability concept put forward by Pearce (1984) and further analyzed by Battigalli (1996, 1997).

The following result states that strong rationalizability characterizes the behavioral implications of RCSBR.

**Theorem 3** *Fix a finite game* $\Gamma$ *and a* $\Gamma$*-based* complete *type structure* $\mathcal{T}$*. Then,*
*(i) for every* $n \in \mathbb{N}$, $\text{proj}_S \prod_{i \in I} R_i^n = S^n$*;*
*(ii)* $\text{proj}_S R^{\infty} = S^{\infty}$.

The proof of Theorem 3 is omitted, since it is very similar to the proof of Theorem 4 below.

Comparing Theorem 3 and the characterization result of Battigalli and Siniscalchi (2002, Proposition 6) we see that there is a key difference in the definition of RCSBR: Battigalli and Siniscalchi (2002) use standard type structures, so they implicitly assume that the personal external states simultaneously represent players' contingent behavior and their plans. In the current framework such implicit assumption can be naturally interpreted as follows: the players execute their plans, and this is commonly believed at every history; formally, *event* $C$ (*consistency*) *is transparent*. We support this interpretation by showing that strong rationalizability characterizes also the behavioral implications of (a) optimal planning and transparency of consistency, and (b) common strong belief in (a).

Formally, for each player $i \in I$,

$$\mathrm{B}_i\left(C_{-i}\right) := \bigcap_{h \in H} \mathrm{B}_{i,h}\left(C_{-i}\right)$$

is the event that $i$ fully believes $C_{-i}$; and

$$\mathrm{B}\left(C\right) := \prod_{i \in I} \mathrm{B}_i\left(C_{-i}\right)$$

is the set of states consistent with mutual full belief in consistency. So,

$$C^* := \bigcap_{m \in \mathbb{N}_0} \mathrm{B}^m\left(C\right)$$

is the set of states where there is **transparency of consistency**, that is, consistency holds and there is common full belief in it. Note that each $\mathrm{B}^m(C)$ is closed. So $(\mathrm{B}^m(C))_{m\in\mathbb{N}_0}$ is a decreasing sequence of closed sets and $C^*$ is closed as well. Moreover,

$$C^* = \prod_{i\in I} C_i^*,$$

where $C_i^* := \mathrm{proj}_{S_i\times T_i} C^*$.

For each player $i \in I$, let

$$R_i^{*,1} := C_i^* \cap OP_i,$$

and, for each $n \in \mathbb{N}$, define $R_i^{*,n+1}$ recursively by

$$R_i^{*,n+1} := R_i^{*,n} \cap \mathrm{SB}_i\left(R_{-i}^{*,n}\right),$$

where $R_{-i}^{*,n} := \prod_{j\neq i} R_j^{*,n}$. A standard induction argument shows that each set $R_i^{*,n}$ is closed in $S_i \times T_i$, hence Borel. We let

$$\begin{aligned} R_i^{*,\infty} &:= \bigcap_{n\in\mathbb{N}_0} R_i^{*,n}, \\ R^{*,\infty} &:= \prod_{i\in I} R_i^{*,\infty}. \end{aligned}$$

Therefore $R^{*,\infty}$ is the set of states consistent with optimal planning and transparency of consistency, and common strong belief thereof.

**Theorem 4** *Fix a finite game* $\Gamma$ *and a* $\Gamma$*-based* complete *type structure* $\mathcal{T}$*. Then,*
*(i) for every* $n \in \mathbb{N}$, $\mathrm{proj}_S \prod_{i\in I} R_i^{*,n} = S^n$*;*
*(ii)* $\mathrm{proj}_S R^{*,\infty} = S^\infty$*.*

The proof of Theorem 4 is in Appendix A.

# 7 Discussion

In this section we consider alternative solution concepts and epistemic assumptions, we discuss an extension of our framework, and we compare our work with the closest related literature. A note on terminology: throughout the discussion, the word "strategy" will be used in its technical meaning as referred to both plans and contingent behavior.

**Forward induction and solution concepts** It can be shown that strong rationalizability is behaviorally equivalent to (the correlated version of) the extensive-form rationalizability concept put forward by Pearce (1984) and clarified by Battigalli (1996, 1997).[28] Specifically, let

$$H_i(s_i) := \{h \in H : s_i \in S_i(h)\}$$

denote the set on nonterminal histories made reachable by strategy $s_i$. We say that $s_i'$ and $s_i''$ are **behaviorally equivalent** if $H_i(s_i') = H_i(s_i'')$ and $s_i'(h) = s_i''(h)$ for each $h \in H_i(s_i')$. Kuhn (1953) shows that $s_i'$ and $s_i''$ are behaviorally equivalent if and only if they are realization equivalent, that is, $\zeta(s_i', s_{-i}) = \zeta(s_i'', s_{-i})$ for all $s_{-i}$, which means they induce the same consequences and are observationally indistinguishable. A class of behaviorally equivalent strategies is called "plan of action" by Rubinstein (1991).[29] Essentially, Pearce's solution concept replaces the best reply correspondence $\rho_i(\cdot)$ with the following weaker version:

$$\bar{\rho}_i(\mu_i) = \left\{ s_i \in S_i : \forall h \in H_i(s_i), s_i^h \in \arg\max_{r_i \in S_i(h)} \mathbb{E}_{\mu_i}[U_i(r_i, \cdot) | h] \right\}.$$

Let $\left(\bar{S}_i^n\right)_{i \in I, n \in \mathbb{N}}$ denote the solution procedure obtained by replacing $\rho_i(\cdot)$ with $\bar{\rho}_i(\cdot)$ in Definition 7. Much of the literature on epistemic game theory and rationalizability for games in extensive form (including Battigalli and Siniscalchi 2002) refers to this solution concept. Yet, it is well known that, for every player $i$, belief $\mu_i$ and strategy $\bar{s}_i$, we have that $\bar{s}_i \in \bar{\rho}_i(\mu_i)$ if and only if there is some behaviorally equivalent strategy $s_i$ such that $s_i \in \rho_i(\mu_i)$. Thus, a straightforward induction argument shows that, for all $i$, $n$, and $\bar{s}_i$, we have that $\bar{s}_i \in \bar{S}_i^n$ if and only if there is some behaviorally equivalent $s_i \in S_i^n$.[30]

[28] In Section 3.2 we explained why we avoid the "extensive-form rationalizability" terminology. Note also that, for $n$-person games, the literature following Pearce (1984) mostly focused on the "correlated" version, that can be characterized by iterated conditional dominance (Shimoji and Watson 1998). Furthermore, Battigalli (1996) criticizes Pearce's "uncorrelated" solution concept because his condition of independence of beliefs across opponents is flawed. Battigalli (1996) proposes an alternative definition of independent rationalizability for which Battigalli and Siniscalchi (1999b) provide an epistemic justification.

[29] Rubinstein (1991) considers a notion of "forward planning" according to which it is not necessary to plan what to do after deviations from one's own plan. Much of the epistemic literature on dynamic games relies (implicitly or explicitly) on this notion of forward planning. Instead we consider a folding-back interpretation that makes sense of planning for *every* contingency, including own deviations.

[30] This is noticed, for example, by Battigalli et al. (2013) and Heifetz and Perea (2015).

According to Definition 7, the unique strongly rationalizable pair in the BoSOO game of Figure 3.1 is $(In.C, c)$, and the unique strongly rationalizable pair in the PI game of Figure 3.2 is $(O_a.o_a, I_b.o_b)$. In the second game, the earlier definition due to Pearce also allows for strategy $O_a.i_a$; instead, we rule out this strategy because, if Ann believes that the strategy of Bob is $I_b.o_b$, then her folding-back optimal plan is $O_a.o_a$, that is, $O_a.o_a \in \bar{S}_a^{\infty}$. These differences are immaterial because $O_a.o_a$ and $O_a.i_a$ are behaviorally equivalent, hence also realization equivalent.

Let $z^{\text{bi}}$ denote the BI path of any finite PI game $\Gamma$ without relevant ties. If $\mathcal{T}$ is a $\Gamma$-based complete type structure $\mathcal{T}$, then

$$\zeta\left(\text{proj}_S R^{\infty}\right) = \zeta\left(\text{proj}_S R^{*,\infty}\right) = \left\{z^{\text{bi}}\right\};$$

that is, forward-induction reasoning yields the BI path. This result is the analogue of Proposition 8 in Battigalli and Siniscalchi (2002) and it follows from Theorems 3 and 4, the equivalence between $(S_i^n)_{i\in I, n\in\mathbb{N}}$ and $\left(\bar{S}_i^n\right)_{i\in I, n\in\mathbb{N}}$, and Theorem 4 in Battigalli (1997).[31]

**Common initial belief in rationality** The notion of initial, or weak rationalizability (Battigalli 2003) is an extension to games with observable actions of a solution concept put forward and analyzed by Ben Porath (1997) for games with perfect information. This solution concept is weaker than strong and backwards rationalizability because it allows a player to believe anything about the co-players if he is surprised. For example, in the BoSOO game of Figure 3.1 only strategy $In.M$ is deleted. In PI games without relevant ties, initial rationalizability is behaviorally equivalent to one round of elimination of weakly dominated strategies followed by the iterated deletion of strictly dominated strategies (see Ben Porath 1997). Such equivalence holds generically in games with observable actions.[32]

Say that player $i$ **initially believes** event $E$ if $i$ assigns probability 1 to $E$ at the beginning of the game. Using arguments similar to those in the proof of Theorem 4, it can be shown—as an analogue of Theorem 3—that the behavioral implications of rationality and common initial belief in rationality are characterized by initial rationalizability (cf. Battigalli and Siniscalchi 2007). A similar result holds for "optimal planning and transparency of consistency" and common initial belief thereof.

[31] Like a related proof by Reny (1992), Battigalli's proof relies on properties of stable sets. Heifetz and Perea (2015), and Perea (2018) provide more transparent proofs.

[32] Fix a strategy $s_i$. There is no first-order CPS $\mu_i$ such that $s_i \in \bar{\rho}_i(\mu_i)$ if and only if $s_i$ is strictly dominated conditional on reaching some $h \in H_i(s_i)$. The latter condition implies that $s_i$ is weakly dominated, and the converse fails only for a negligible set of payoff functions $u_i$. See Shimoji (2004) and Shimoji and Watson (1998).

**Extension to dynamically inconsistent and belief-dependent preferences** Our perspective on rationality and the ensuing epistemic approach can be extended to cover dynamically inconsistent preferences due, for example, to non-exponential discounting (Frederick et al. 2002), or some versions of ambiguity aversion (Marinacci 2015), and belief-dependent preferences, which in turn may be dynamically inconsistent when preferences over outcomes depend on one's own plan (Battigalli and Dufwenberg 2009).[33] Given beliefs about other players (or nature), sophisticated planning is an intra-personal equilibrium condition expressed by the OSD property, which in this case is not equivalent to sequential optimality.[34] Rationality is given by the conjunction of sophisticated planning and consistency between plan and behavior. With this, the epistemic assumptions analyzed in this paper can be applied to a much wider set of interactive situations.

Compared to the traditional multi-self approach to games with dynamically inconsistent preferences, we bring a different perspective. The traditional approach does not really distinguish between the "selves" at different nodes of different players, or the same player: preferences may differ in both cases, but belief systems are presumed to be the same (barring asymmetric information, as we do here); thus an *inter*-personal (e.g., sequential) equilibrium is assumed. We instead only maintain that each player is introspective and sophisticated, which justifies *intra*-personal equilibrium as a starting point. But we do not assume that players know each other as they know themselves. Therefore, inter-personal equilibrium can only be a conclusion of the analysis that holds under special circumstances (e.g., games with complete and perfect information) and epistemic assumptions (e.g., versions of "common belief in rationality").

**Related literature** Starting with the seminal contribution of Aumann (1995), various epistemic justifications for BI behavior have been offered in the literature (see the review by Perea 2007).[35] Here we outline the differences between our epistemic

[33] In the context of dynamic games, see—for example—Battigalli et al. (2017) on ambiguity aversion, and Battigalli et al. (2018) on the role of emotions and belief-dependent preferences. Note that own-plan dependence of preferences over outcomes may require non-deterministic plans to satisfy the OSD property given beliefs about others.

[34] The plan of a sophisticated player with dynamically inconsistent preferences may be "sophisticated" and yet not "optimal" in an obvious sense, because the OSD principle fails. Hence, in this case it is better to talk about sophisticated, rather than optimal planning.

[35] The survey by Perea (2007) restricts attention to sufficient epistemic conditions for the BI *behavior*. By contrast, the result in Battigalli and Siniscalchi (2002) pertains to the BI *path*. Arieli and Aumann (2015) adopt a syntactic approach to provide epistemic conditions for BI behavior in PI games where each player moves at most once. A similar result not relying on strong belief,

conditions for BI behavior (Theorem 1 and Theorem 2) and those that appear to be conceptually closest, namely Baltag et al. (2009) and Perea (2014). Baltag et al. (2009) use a dynamic epistemic-logic formalism related to, but different from, the one we have adopted in this paper. Their approach is based on the framework of the so called "plausibility models" (see van Benthem 2007), which can be seen as an extension of standard knowledge spaces to take into account the dynamics of beliefs and knowledge. They use this formalism to capture a future-oriented concept of rationality, called "dynamic rationality": at any stage of the game, the rationality of a player depends *only* on his current beliefs and knowledge; so a player can be dynamically rational at history $h$ even if he has made "irrational" moves at some history $h' \prec h$. This is somewhat similar to our event that player $i$ is consistent from $h$ ($C^{\succeq h}$) and plans optimally ($OP$). As the authors show, dynamic rationality is a coarsening of Aumann's (1995) concept of "substantive rationality" in a belief-revision context;[36] then they use the notion of "stable belief" to show that dynamic rationality and common knowledge of stable belief in dynamic rationality entails BI behavior in generic PI games.

Perea (2014) defines "common belief in future rationality" within a standard type-structure formalism (i.e., without players' beliefs about their own behavior), and he shows that its behavioral implications are characterized by a version of backwards rationalizability which is weaker than ours (Definition 6). As in all standard type structures, and differently from our framework, it is implicitly assumed in Perea (2014) that the personal external states $s_i$ ($i \in I$) simultaneously represent players' contingent behavior and their plans. In particular, the personal external states are defined as "plans of action," that is, classes of behaviorally equivalent strategies (Rubinstein 1991), hence, maximization is required only at histories consistent with the given plan $s_i$, that is, $h \in H_i(s_i)$. Perea's version of backwards rationalizability is based on best reply correspondence $\bar{\rho}_i(\cdot)$ rather than $\rho_i(\cdot)$; with this, it can be shown that a strategy $\bar{s}_i$ is backwards rationalizable in Perea's sense if and only if it is behaviorally equivalent to some $s_i$ that is backwards rationalizable in our sense. Specifically, in PI games without relevant ties, backwards rationalizability *à la* Perea yields the set of profiles $(s_i)_{i \in I}$ such that each $s_i$ is behaviorally equivalent to $s_i^{\text{bi}}$: for

but rather on epistemic independence is obtained within standard type structures by Battigalli and Siniscalchi (1999b).

[36] As is well known (see, for instance, Halpern 2001), Aumann's framework is "static" in the sense that it does not allow the players to revise their beliefs about co-players' behavior when doing hypothetical reasoning. Aumann defines "substantive rationality" in terms of knowledge, and shows that common knowledge of "substantive rationality" yields BI. Samet (2013) shows that common (probability 1) belief of "substantive rationality" yields BI, provided that "substantive rationality" is defined in doxastic terms, that is, in terms of belief.

instance, in the game of Figure 3.2, both $(O_a.o_a, O_b.o_b)$ and $(O_a.i_a, O_b.o_b)$ are backwards rationalizable in Perea's sense. This shows that Perea's sufficient conditions for BI and backwards rationalizability (Theorems 6.1 and 4.3 in Perea 2014) are somewhat different from ours,[37] although they are similar in spirit and have equivalent implications: indeed, our representation of BI reasoning in BI-solvable games yields precisely the unique profile $s^{\text{bi}}$.

Like us, Battigalli et al. (2013) model plans as beliefs about own behavior, but in their framework—differently from us—the set of external states is $Z$, i.e., the set of complete paths. While in our framework the external personal state of a player is (technically) also a strategy, in their framework the only mathematical objects that correspond to (behavior) strategies are players' systems of conditional beliefs about their own actions. Furthermore, Battigalli et al. (2013) focus only on RCSBR[38] in PI games, proving a result analogous to Theorem 3. We conjecture that we could reformulate our analysis having $Z$ as the set of external states. Battigalli et al. (2018) make steps in this direction while also allowing for belief-dependent preferences.

[37] Asheim (2002) and Asheim and Perea (2005) also provide epistemic analyses of BI, but use formalisms different from the one in Perea (2014). In Asheim (2002) and Asheim and Perea (2005) type structures do not include players' beliefs about their own behavior, and—more importantly—beliefs are represented by "Conditional Lexicographic Probability Systems" (Blume et al. 1991), rather than CPSs. As in Perea (2014), they obtain sufficient epistemic conditions for BI "plans of actions," rather than the BI strategies.

[38] As defined in their different framework.

# Appendix A: Proofs of Theorems 2 and 4

We first record the following result that will be useful for the proofs of Theorems 2 and 4.

**Lemma 2** *Let $X$ and $Y$ be compact metrizable spaces. If $(E^m)_{m=1}^{\infty}$ is a decreasing sequence of nonempty, closed subsets of $X \times Y$, then*

$$\text{proj}_X \cap_{m=1}^{\infty} E^m = \cap_{m=1}^{\infty} \text{proj}_X E^m.$$

**Proof.** The inclusion $\subseteq$ is obvious. For the other direction, let $x \in \cap_{m=1}^{\infty} \text{proj}_X E^m$. For each $m$, let $E_x^m := \{y \in Y : (x,y) \in E^m\}$. So, we need to establish the existence of some $y \in Y$ such that $y \in \cap_{m=1}^{\infty} E_x^m$, that is, $\cap_{m=1}^{\infty} E_x^m \neq \emptyset$. This will imply the thesis. First note that each $E_x^m$ is a nonempty closed subset of $Y$, so compact. Specifically, non-emptiness of each $E_x^m$ follows from the fact that $x \in \cap_{m=1}^{\infty} \text{proj}_X E^m$. Moreover, $(E_x^m)_{m=1}^{\infty}$ is a decreasing sequence of sets; therefore, by the finite intersection property, $\cap_{m=1}^{\infty} E_x^m \neq \emptyset$. ■

## Proof of Theorem 2

For the proof of the theorem, we find it convenient to introduce further notation and preliminary results.

Fix a finite game $\Gamma$. For a given $h \in H$, let

$$\rho_i^{\succ h}(\mu_i) := \left\{ s_i \in S_i : \forall h' \in H(h), s_i^{h'} \in \arg\max_{r_i \in S_i(h')} \mathbb{E}_{\mu_i}\left[U_i(r_i,.)\,|h'\right] \right\}.$$

Note that $\rho_i^{\succ \varnothing}(\mu_i) = \rho_i(\mu_i)$.

**Lemma 3** *Fix $h \in H$ and a CPS $\mu_i$ on $(S_{-i}, \mathcal{S}_{-i})$.*

*(i) If $s_i \in \rho_i(\mu_i)$, then $s_i \in \rho_i^{\succ h}(\mu_i)$.*

*(ii) If $s_i \in \rho_i^{\succ h}(\mu_i)$, then there exists $\bar{s}_i \in S_i$ such that $\bar{s}_i \in \rho_i(\mu_i)$ and $s_i(h') = \bar{s}_i(h')$ for all $h' \in H(h)$.*

**Proof.** Part (i) is immediate. Part (ii) follows from standard dynamic programming results. ■

**Lemma 4** *For every $i \in I$, $h \in H$ and $n \in \mathbb{N}$,*

$$\chi_i^h\left(\hat{S}_i^n\right) = \left\{ \begin{array}{ll} s_i \in S_i(h): & \exists \mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i}), \\ & 1)\ s_i \in \rho_i^{\succ h}(\mu_i), \\ & 2)\ \forall h' \in H, \mu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^{n-1}\right)|S_{-i}(h')\right) = 1 \end{array} \right\}.$$

**Proof.** Let $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$. Then, by definition, there exists $\bar{s}_i \in \hat{S}_i^n$ such that $s_i(h') = \bar{s}_i(h')$ for all $h' \in H(h)$. Hence $\bar{s}_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ satisfying $\mu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^{n-1}\right)|S_{-i}(h')\right) = 1$ for all $h' \in H$. Part (i) of Lemma 3 implies that $\bar{s}_i \in \rho_i^{\succeq h}(\mu_i)$, and since $\bar{s}_i$ coincides with $s_i$ at all histories weakly following $h$, we have $s_i \in \rho_i^{\succeq h}(\mu_i)$.

For the other direction, let $s_i \in S_i(h)$ such that $s_i \in \rho_i^{\succeq h}(\mu_i)$ for some $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ satisfying $\mu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^{n-1}\right)|S_{-i}(h')\right) = 1$ for all $h' \in H$. Part (ii) of Lemma 3 yields the existence of $\bar{s}_i \in S_i$ such that $\bar{s}_i \in \rho_i(\mu_i)$ and $s_i(h') = \bar{s}_i(h')$ for all $h' \in H(h)$. By Definition 6, we have $\bar{s}_i \in \hat{S}_i^n$. Hence $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$. ■

The proof of Theorem 2 relies on Lemma 5 below. To formally state and prove Lemma 5, we need some additional notation. Fix a finite game $\Gamma$ and a $\Gamma$-based type structure $\mathcal{T}$. For each $i \in I$, let

$$OP_i^0 := S_i \times T_i.$$

The sets $OP^0$ and $OP_{-i}^0$ are defined in the usual way, that is, $OP^0 := \prod_{i \in I} OP_i^0$ and $OP_{-i}^0 := \prod_{j \neq i} OP_j^0$.

**Lemma 5** *Fix a finite game $\Gamma$ and a $\Gamma$-based type structure $\mathcal{T}$. The following statements hold:*

*(i) for all $n \in \mathbb{N}_0$ and $h \in H$,*

$$\chi^h\left(\mathrm{proj}_S\left(OP^n \cap C\right)\right) \subseteq \chi^h\left(\hat{S}^n\right);$$

*(ii) if $\mathcal{T}$ is complete, then, for all $n \in \mathbb{N}_0$ and $h \in H$,*

$$\chi^h\left(\mathrm{proj}_S\left(OP^n \cap C\right)\right) = \chi^h\left(\hat{S}^n\right).$$

**Proof.** We first prove the following preliminary result:

**Claim 1** *Fix $n \in \mathbb{N}_0$ and $h \in H$. Then*

$$\forall i \in I, \chi_i^h\left(\mathrm{proj}_{S_i}\left(OP_i^n \cap C_i\right)\right) \subseteq \mathrm{proj}_{S_i}\left(OP_i^n \cap C_i^{\succeq h}\right) \cap S_i(h).$$

*Proof of Claim 1.* First note that $C_i \subseteq C_i^{\succeq h}$ and $\chi_i^h\left(\text{proj}_{S_i}\left(OP_i^n \cap C_i\right)\right) \subseteq S_i(h)$ for each $i \in I$. Consequently, if $OP_i^n \cap C_i$ or $OP_i^n \cap C_i^{\succeq h}$ are empty, then the result is immediate. So in what follows we will assume that $OP_i^n \cap C_i$ is nonempty. Let $s_i \in \chi_i^h\left(\text{proj}_{S_i}\left(OP_i^n \cap C_i\right)\right)$. Then $s_i \in S_i(h)$, and so we only need to show the existence of $t_i \in T_i$ such that $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq h}$; this will imply $s_i \in \text{proj}_{S_i}\left(OP_i^n \cap C_i^{\succeq h}\right) \cap S_i(h)$, as required. By definition, there exists $\bar{s}_i \in \text{proj}_{S_i}\left(OP_i^n \cap C_i\right)$ such that $s_i(h') = \bar{s}_i(h')$ for every $h' \succeq h$. Hence $(\bar{s}_i, t_i) \in OP_i^n \cap C_i$ for some $t_i \in T_i$. Optimal planning and consistency at $(\bar{s}_i, t_i)$ entails that $\bar{s}_i \in \rho_i(\nu_i)$, where $\nu_i$ denotes the marginal of $\beta_i(t_i)$ on $(S_{-i}, \mathcal{S}_{-i})$. Part (i) of Lemma 3 implies that $\bar{s}_i \in \rho_i^{\succeq h}(\nu_i)$, and since $\bar{s}_i$ and $s_i$ coincide at every history weakly following $h$, we obtain $s_i \in \rho_i^{\succeq h}(\nu_i)$. Therefore $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq h}$. $\square$

Part (i): We prove the following claim:

$$\forall i \in I, \forall h \in H, \forall n \in \mathbb{N}_0, \text{proj}_{S_i}\left(OP_i^n \cap C_i^{\succeq h}\right) \cap S_i(h) \subseteq \chi_i^h\left(\hat{S}_i^n\right).$$

Then Claim 1 will give the result. The proof is by induction on $n \in \mathbb{N}_0$.

*Basis step.* Note that, for every $i \in I$ and $h \in H$,

$$\begin{aligned} \text{proj}_{S_i}\left(OP_i^0 \cap C_i^{\succeq h}\right) \cap S_i(h) &= \text{proj}_{S_i}\left(C_i^{\succeq h}\right) \cap S_i(h) \\ &\subseteq S_i(h) \\ &= \chi_i^h\left(\hat{S}_i^0\right), \end{aligned}$$

so the result follows immediately.

*Inductive step.* Assume that the result is true for each $m = 0, ..., n$. We show that it is also true for $m = n + 1$.

Fix any $i \in I$ and $h \in H$ arbitrarily. Let $s_i \in \text{proj}_{S_i}\left(OP_i^{n+1} \cap C_i^{\succeq h}\right) \cap S_i(h)$, so that $(s_i, t_i) \in OP_i^{n+1} \cap C_i^{\succeq h}$ for some $t_i \in T_i$. Since $OP_i^{n+1} \subseteq OP_i^n$, it follows that $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq h}$, and so, by the induction hypothesis, $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$. By Lemma 4, we get that $s_i \in \rho_i^{\succeq h}(\nu_i)$, where $\nu_i$ denotes the marginal of $\beta_i(t_i)$ on $(S_{-i}, \mathcal{S}_{-i})$. So, in order to show that $s_i \in \chi^h\left(\hat{S}_i^{n+1}\right)$, it is enough to show (by Lemma 4) that $\nu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^n\right) | S_{-i}(h')\right) = 1$ for every $h' \in H$.

To this end, first note that $(s_i, t_i) \in OP_i^{n+1}$ implies $(s_i, t_i) \in \text{B}_i\left(OP_{-i}^n\right) = \cap_{h' \in H} \text{B}_{i,h'}\left(OP_{-i}^n\right)$. Note also that $(s_i, t_i) \in BCC_i = \cap_{h' \in H} \text{B}_{i,h'}\left(C_{-i}^{\succeq h'}\right)$; hence, by

the conjunction property of the operator $\mathrm{B}_{i,h'}(\cdot)$, it follows that, for each $h' \in H$, $(s_i, t_i) \in \mathrm{B}_{i,h'}\left(OP^n_{-i} \cap C^{\succeq h'}_{-i}\right)$. Using this fact, we get that, for all $h' \in H$,

$$\begin{aligned}
\nu_i\left(\chi^{h'}_{-i}\left(\hat{S}^n_{-i}\right)|S_{-i}(h')\right) &\geq \nu_i\left(\mathrm{proj}_{S_{-i}}\left(OP^n_{-i} \cap C^{\succeq h'}_{-i}\right) \cap S_{-i}(h')\,|S_{-i}(h')\right) \\
&= \nu_i\left(\mathrm{proj}_{S_{-i}}\left(OP^n_{-i} \cap C^{\succeq h'}_{-i}\right)|S_{-i}(h')\right) \\
&= \mathrm{marg}_{S_{-i}}\beta_{i,h'}(t_i)\left(\mathrm{proj}_{S_{-i}}\left(OP^n_{-i} \cap C^{\succeq h'}_{-i}\right)\right) \\
&= \beta_{i,h'}(t_i)\left(\mathrm{proj}^{-1}_{S_{-i}}\left(\mathrm{proj}_{S_{-i}}\left(OP^n_{-i} \cap C^{\succeq h'}_{-i}\right)\right)\right) \\
&\geq \beta_{i,h'}(t_i)\left(S_i \times \left(OP^n_{-i} \cap C^{\succeq h'}_{-i}\right)\right) \\
&= 1,
\end{aligned}$$

where the first inequality follows from the induction hypothesis, the first equality follows from basic properties of a CPS,[39] the second and third equalities follow by definition, the second inequality follows from a trivial fact about the inverse images of functions, and the last equality follows from the conjunction property of the operator $\mathrm{B}_{i,h'}(\cdot)$. This shows that $\nu_i$ satisfies the required properties. Since $i \in I$ and $h \in H$ are arbitrary, the conclusion follows.

Part (ii): Let $\mathcal{T}$ be complete. First note that

$$\forall i \in I, S_i = \mathrm{proj}_{S_i}(C_i). \tag{7.1}$$

To see this, let $s_i \in S_i$, and consider the CPS $\mu_{s_i} \in \Delta^{\mathcal{S}\times T_{-i}}(S \times T_{-i})$ defined as follows: fix an arbitrary $\mu_{s_i,-i} \in \Delta^{\mathcal{S}_{-i}\times T_{-i}}(S_{-i} \times T_{-i})$, and, for all $h \in H$, let

$$\mu_{s_i}(\cdot|S(h) \times T_{-i}) := \mu_{i,i}(\cdot|S_i(h)) \times \mu_{s_i,-i}(\cdot|S_{-i}(h) \times T_{-i}),$$

where $\mu_{i,i}$ is the CPS on $(S_i, \mathcal{S}_i)$ that satisfies $\mu_{i,i}\left(s^h_i|S_i(h)\right) = 1$ for all $h \in H$. By completeness, there exists $t_{s_i} \in T_i$ such that $\beta_i(t_{s_i}) = \mu_{s_i}$. Then $(s_i, t_{s_i}) \in C_i$ because, for all $h \in H$,

$$\begin{aligned}
\sigma_{t_{s_i},i}(s_i(h)|h) &\geq \beta_{i,i}(t_{s_i})\left(s^h_i|S_i(h)\right) \\
&= \mu_{i,i}\left(s^h_i|S_i(h)\right) \\
&= 1.
\end{aligned}$$

Therefore $s_i \in \mathrm{proj}_{S_i}(C_i)$. The inclusion $\mathrm{proj}_{S_i}(C_i) \subseteq S_i$ is obvious.

[39] Let $\mu$ be a CPS on $(S, \mathcal{S})$, and fix a conditioning event $C \in \mathcal{S}$. Then $\mu(C|C) = 1$ implies $\mu(E \cap C|C) = \mu(E|C)$ for every event $E \subseteq S$.

We now prove the following claim: for every $i \in I$, $h \in H$ and $n \in \mathbb{N}_0$,

$$\chi_i^h\left(\hat{S}_i^n\right) = \chi_i^h\left(\text{proj}_{S_i}\left(OP_i^n \cap C_i\right)\right)$$

and there exists a map $\varphi_i^n : S_i \to S_i \times T_i$ such that $\varphi_i^n\left(\chi_i^h\left(\hat{S}_i^n\right)\right) \subseteq OP_i^n \cap C_i^{\succeq h}$.

This will imply the thesis. The proof is by induction on $n \in \mathbb{N}_0$.

*Basis step.* Fix any $i \in I$ and $h \in H$ arbitrarily. Note that, by (7.1),

$$\text{proj}_{S_i}\left(OP_i^0 \cap C_i\right) = \text{proj}_{S_i}\left(C_i\right) = S_i = \hat{S}_i^0,$$

and so $\chi_i^h\left(\hat{S}_i^0\right) = \chi_i^h\left(\text{proj}_{S_i}\left(OP_i^0 \cap C_i\right)\right)$. It also follows from (7.1) that, for each $s_i \in S_i$, there exists $t_{s_i} \in T_i$ such that $(s_i, t_{s_i}) \in C_i$. So, for every $s_i \in S_i$, we choose and fix some $t_{s_i}$ satisfying $(s_i, t_{s_i}) \in C_i$, and we define the map

$$\begin{array}{rccc} \varphi_i^0 : & S_i & \to & S_i \times T_i \\ & s_i & \mapsto & \left(s_i, t_{s_i}\right). \end{array}$$

This map satisfies the required properties, in that if $s_i \in \chi_i^h\left(\hat{S}_i^0\right) = S_i\left(h\right)$, then

$$\varphi_i^0\left(s_i\right) \in C_i = OP_i^0 \cap C_i \subseteq OP_i^0 \cap C_i^{\succeq h}.$$

Since $i \in I$ and $h \in H$ are arbitrary, this concludes the proof of the basis step.

*Inductive step.* Assume that the result is true for each $m = 0, ..., n$. We show that it is also true for $m = n + 1$. Towards this end, we first record the following implication of the induction hypothesis: for all $i \in I$ and $h \in H$,

$$\varphi_i^n\left(\chi_i^h\left(\hat{S}_i^n\right)\right) \subseteq OP_i^n \cap C_i^{\succeq h} \Rightarrow \chi_i^h\left(\hat{S}_i^n\right) \subseteq \left(\varphi_i^n\right)^{-1}\left(OP_i^n \cap C_i^{\succeq h}\right).$$

Fix $i \in I$ and $h \in H$ arbitrarily. Part (i) gives that $\chi_i^h\left(\text{proj}_{S_i}\left(OP_i^{n+1} \cap C_i\right)\right) \subseteq \chi_i^h\left(\hat{S}_i^{n+1}\right)$. For the converse, let $s_i \in \chi_i^h\left(\hat{S}_i^{n+1}\right)$. So, there exists $\bar{s}_i \in \hat{S}_i^{n+1}$ such that $s_i\left(h'\right) = \bar{s}_i\left(h'\right)$ for every $h' \succeq h$. Moreover, there exists $\nu_{\bar{s}_i} \in \Delta^{\mathcal{S}_{-i}}\left(S_{-i}\right)$ such that $\bar{s}_i \in \rho_i(\nu_{\bar{s}_i})$ and $\nu_{\bar{s}_i}\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^n\right)|S_{-i}\left(h'\right)\right) = 1$ for every $h' \in H$. Consider the CPS $\mu_{_{\bar{s}_i},-i} \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}\left(S_{-i} \times T_{-i}\right)$ defined as follows: for all events $E_{-i} \subseteq S_{-i} \times T_{-i}$ and $h' \in H$,

$$\mu_{_{\bar{s}_i},-i}\left(E_{-i}|S_{-i}\left(h'\right) \times T_{-i}\right) := \nu_{\bar{s}_i}\left(\left(\varphi_{-i}^n\right)^{-1}\left(E_{-i}\right)|S_{-i}\left(h'\right)\right).$$

Note that this is a well-defined CPS on $(S_{-i} \times T_{-i}, \mathcal{S}_{-i} \times T_{-i})$ whose marginal on $(S_{-i}, \mathcal{S}_{-i})$ is $\nu_{\bar{s}_i}$. Note also that, for all $h' \in H$,

$$\begin{aligned}
\mu_{\bar{s}_i,-i}\left(OP_{-i}^n \cap C_{-i}^{\succeq h'} | S_{-i}(h') \times T_{-i}\right) &= \nu_{\bar{s}_i}\left(\left(\varphi_{-i}^n\right)^{-1}\left(OP_{-i}^n \cap C_{-i}^{\succeq h'}\right) | S_{-i}(h')\right) \\
&\geq \nu_{\bar{s}_i}\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^n\right) | S_{-i}(h')\right) \\
&= 1,
\end{aligned}$$

where the inequality follows from the implication of the induction hypothesis.

So we get that, for all $h' \in H$,

$$\mu_{\bar{s}_i,-i}\left(OP_{-i}^n | S_{-i}(h') \times T_{-i}\right) = \mu_{\bar{s}_i,-i}\left(C_{-i}^{\succeq h'} | S_{-i}(h') \times T_{-i}\right) = 1; \tag{7.2}$$

we will make use of this fact below.

Let $\mu_{i,i}$ be the CPS on $(S_i, \mathcal{S}_i)$ that satisfies $\mu_{i,i}\left(\bar{s}_i^{h'} | S_i(h')\right) = 1$ for every $h' \in H$. Consider the CPS $\mu_{\bar{s}_i} \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h' \in H$,

$$\mu_{\bar{s}_i}(\cdot | S(h') \times T_{-i}) := \mu_{i,i}(\cdot | S_i(h')) \times \mu_{\bar{s}_i,-i}(\cdot | S_{-i}(h') \times T_{-i}).$$

By completeness, there exists $t_i \in T_i$ such that and $\beta_i(t_i) = \mu_{\bar{s}_i}$. We now show that

$$(\bar{s}_i, t_i) \in OP_i^{n+1} \cap C_i = OP_i \cap BCC_i \cap \bigcap_{l=0}^{n} \mathrm{B}_i\left(OP_{-i}^l\right) \cap C_i.$$

To this end, first note that, by inspection of the definition of $\mu_{\bar{s}_i}$, type $t_i$ satisfies independence; moreover, type $t_i$ plans optimally because, for all $h' \in H$,

$$\begin{aligned}
\mathrm{supp}\beta_{i,i}(t_i)(\cdot | S_i(h')) &= \mathrm{supp}\mu_{i,i}(\cdot | S_i(h')) \\
&= \left\{\bar{s}_i^{h'}\right\} \\
&\subseteq \arg\max_{r_i \in S_i(h')} \sum_{s_{-i} \in S_{-i}(h')} U_i(r_i, s_{-i}) \mathrm{marg}_{S_{-i}} \beta_{i,-i}(t_i)(s_{-i} | S_{-i}(h')) \\
&= \arg\max_{r_i \in S_i(h')} \sum_{s_{-i} \in S_{-i}(h')} U_i(r_i, s_{-i}) \mathrm{marg}_{S_{-i}} \mu_{\bar{s}_i,-i}(s_{-i} | S_{-i}(h')) \\
&= \arg\max_{r_i \in S_i(h')} \sum_{s_{-i} \in S_{-i}(h')} U_i(r_i, s_{-i}) \nu_{\bar{s}_i}(s_{-i} | S_{-i}(h')).
\end{aligned}$$

Hence $(\bar{s}_i, t_i) \in OP_i$. Furthermore, $(\bar{s}_i, t_i) \in C_i$ because for all $h' \in H$,

$$\begin{aligned}
\sigma_{t_i,i}(\bar{s}_i(h') | h') &\geq \beta_{i,i}(t_i)\left(\bar{s}_i^{h'} | S_i(h')\right) \\
&= \mu_{i,i}\left(\bar{s}_i^{h'} | S_i(h')\right) \\
&= 1.
\end{aligned}$$

We now check that $(\bar{s}_i, t_i) \in BCC_i = \cap_{h' \in H} \mathrm{B}_{i,h'} \left( C_{-i}^{\succeq h'} \right)$: for every $h' \in H$,

$$\begin{aligned} \beta_{i,h'}(t_i) \left( S_i \times C_{-i}^{\succeq h'} \right) &= \beta_i(t_i) \left( S_i \times C_{-i}^{\succeq h'} | S(h') \times T_{-i} \right) \\ &= \mu_{i,i}(S_i | S_i(h')) \times \mu_{\bar{s}_i,-i} \left( C_{-i}^{\succeq h'} | S_{-i}(h') \times T_{-i} \right) \\ &= 1, \end{aligned}$$

where the third equality follows from (7.2) and from the definition of $\mu_{i,i}$. It remains to show that $(\bar{s}_i, t_i) \in \mathrm{B}_i \left( OP_{-i}^n \right) = \cap_{h' \in H} \mathrm{B}_{i,h'} \left( OP_{-i}^n \right)$; since the sequence $\left( OP_{-i}^l \right)_{l=0,1,\ldots,n}$ is decreasing, monotonicity of the operator $\mathrm{B}_i(\cdot)$ will imply $(\bar{s}_i, t_i) \in \cap_{l=0}^n \mathrm{B}_i \left( OP_{-i}^l \right)$. Using again (7.2), we get that, for all $h' \in H$,

$$\begin{aligned} \beta_{i,h'}(t_i) \left( S_i \times OP_{-i}^n \right) &= \beta_i(t_i) \left( S_i \times OP_{-i}^n | S(h') \times T_{-i} \right) \\ &= \mu_{i,i}(S_i | S_i(h')) \times \mu_{\bar{s}_i,-i} \left( OP_{-i}^n | S_{-i}(h') \times T_{-i} \right) \\ &= 1. \end{aligned}$$

This concludes the proof that $(\bar{s}_i, t_i) \in OP_i^{n+1} \cap C_i$.

It follows that $\bar{s}_i \in \mathrm{proj}_{S_i} \left( OP_i^{n+1} \cap C_i \right)$, and so $s_i \in \chi_i^h \left( \mathrm{proj}_{S_i} \left( OP_i^{n+1} \cap C_i \right) \right)$. Hence we have shown that

$$\chi_i^h \left( \hat{S}_i^{n+1} \right) = \chi_i^h \left( \mathrm{proj}_{S_i} \left( OP_i^{n+1} \cap C_i \right) \right).$$

Then, part (i) and Claim 1 yield

$$\chi_i^h \left( \hat{S}_i^{n+1} \right) = \mathrm{proj}_{S_i} \left( OP_i^{n+1} \cap C_i^{\succeq h} \right) \cap S_i(h). \tag{7.3}$$

To conclude the proof of the inductive step, we show the existence of a map $\varphi_i^{n+1} : S_i \to S_i \times T_i$ such that

$$\varphi_i^{n+1} \left( \chi_i^h \left( \hat{S}_i^{n+1} \right) \right) \subseteq OP_i^{n+1} \cap C_i^{\succeq h}.$$

By (7.3), it follows that if $s_i \in \chi_i^h \left( \hat{S}_i^{n+1} \right)$, then there exists $t_{s_i} \in T_i$ such that $(s_i, t_{s_i}) \in OP_i^{n+1} \cap C_i^{\succeq h}$. Therefore, for each $s_i \in \chi_i^h \left( \hat{S}_i^{n+1} \right)$, we choose and fix some $t_{s_i}$ satisfying the above condition. We also fix an arbitrary $\bar{t}_i \in T_i$, and we define the map $\varphi_i^{n+1} : S_i \to S_i \times T_i$ as follows:

$$\varphi_i^{n+1}(s_i) = \begin{cases} (s_i, t_{s_i}), & \text{if } s_i \in \chi_i^h \left( \hat{S}_i^{n+1} \right), \\ (s_i, \bar{t}_i), & \text{if } s_i \in S_i \backslash \chi_i^h \left( \hat{S}_i^{n+1} \right). \end{cases}$$

This map satisfies the required properties. Since $i \in I$ and $h \in H$ are arbitrary, the proof of the inductive step is complete. ■

We can now provide the proof of Theorem 2.

**Proof of Theorem 2**. Part (i) follows from Lemma 5. As far as part (ii) is concerned, first note that $(OP^n \cap C)_{n\in\mathbb{N}}$ is a decreasing sequence of compact sets. Part (i) implies that $OP^n \cap C \neq \emptyset$ for every $n \in \mathbb{N}$. Hence, by the finite intersection property, $OP^\infty \cap C \neq \emptyset$. By part (i) and Lemma 2, $\text{proj}_S (OP^\infty \cap C) = \hat{S}^\infty$, as required. ■

## Proof of Theorem 4

For the proof of Theorem 4, we first record an abstract result (Lemma 6) for CPSs.

Let $X$ and $Y$ be compact metrizable spaces, and fix a CPS $\mu := (\mu(\cdot|C \times Y))_{C\in\mathcal{C}} \in \Delta^{\mathcal{C}\times Y}(X \times Y)$. We say that $\mu$ **strongly believes** a nonempty event $E \subseteq X \times Y$ if, for every $C \in \mathcal{C}$,

$$E \cap (C \times Y) \neq \emptyset \Rightarrow \mu(E|C \times Y) = 1.$$

We say that $\mu$ strongly believes a sequence of nonempty events $(E_0, ..., E_n)$ in $X \times Y$ if, for each $m = 0, ..., n$, $\mu$ strongly believes $E_m$. We say that $\mu$ **fully believes** a nonempty event $E \subseteq X \times Y$ if $\mu(E|C \times Y) = 1$ for every $C \in \mathcal{C}$.

**Lemma 6** *Fix a finite decreasing sequence of closed events $(E_0, ..., E_n)$ in $X \times Y$.*

*(i) If $\mu \in \Delta^{\mathcal{C}\times Y}(X \times Y)$ strongly believes $(E_m)_{m=0}^n$, then $\text{marg}_X \mu$ strongly believes $(\text{proj}_X E_m)_{m=0}^n$.*

*(ii) Let $\nu \in \Delta^{\mathcal{C}}(X)$. If $\nu$ fully believes $\text{proj}_X E_0$ and strongly believes $(\text{proj}_X E_m)_{m=1}^n$, then there exists $\mu \in \Delta^{\mathcal{C}\times Y}(X \times Y)$ such that (a) $\mu$ fully believes $E_0$, (b) $\mu$ strongly believes $(E_m)_{m=1}^n$, and (c) $\text{marg}_X \mu = \nu$.*

**Proof.** Part (i) follows from the marginalization property of strong belief (see Battigalli and Friedenberg 2012). The proof of part (ii) follows the same lines as those of Lemma 3 in Battigalli and Tebaldi (2018). ■

We also need the following facts pertaining to the epistemic events of interest.

**Lemma 7** *Fix a finite game $\Gamma$ and a $\Gamma$-based type structure $\mathcal{T}$. Then $(s_i, t_i) \in C_i^*$ if and only if $(s_i, t_i) \in C_i$ and $\beta_{i,-i}(t_i)$ fully believes $C_{-i}^* := \prod_{j\neq i} C_j^*$.*

**Proof**. Note that, by definition and the conjunction property of $\mathrm{B}(\cdot)$, we have

$$\begin{aligned} C^* &= \cap_{m\in\mathbb{N}_0}\mathrm{B}^m(C) \\ &= \mathrm{B}^0(C)\cap(\cap_{m\in\mathbb{N}}\mathrm{B}^m(C)) \\ &= C\cap(\cap_{m\in\mathbb{N}_0}\mathrm{B}(\mathrm{B}^m(C))) \\ &= C\cap\mathrm{B}(\cap_{m\in\mathbb{N}_0}(\mathrm{B}^m(C))) \\ &= C\cap\mathrm{B}(C^*). \end{aligned}$$

So the statement immediately follows. ■

**Lemma 8** *Fix a finite game $\Gamma$ and a $\Gamma$-based type structure $\mathcal{T}$. If $\mathcal{T}$ is complete, then, for every $i\in I$ and $h\in H$,*

$$C_i^*\cap(S_i(h)\times T_i)\neq\emptyset.$$

**Proof**. Note that, for all $m\in\mathbb{N}$,

$$\mathrm{B}^m(C)=\prod_{i\in I}\mathrm{B}_i\left(\mathrm{proj}_{S_{-i}\times T_{-i}}\mathrm{B}^{m-1}(C)\right).$$

We show by induction on $m\in\mathbb{N}_0$ that, for each $i\in I$ and $h\in H$,

$$\left(\mathrm{proj}_{S_i\times T_i}\mathrm{B}^m(C)\right)\cap(S_i(h)\times T_i)\neq\emptyset.$$

Since $C^*:=\bigcap_{m\in\mathbb{N}_0}\mathrm{B}^m(C)$, this will imply the thesis.

*Basis step.* Fix $i\in I$ and $h\in H$. Let $s_i\in S_i(h)$, and consider the CPS $\mu_i\in\Delta^{\mathcal{S}\times T_{-i}}(S\times T_{-i})$ defined as follows: pick any $\mu_{i,-i}\in\Delta^{\mathcal{S}_{-i}\times T_{-i}}(S_{-i}\times T_{-i})$, and, for all $h'\in H$, let

$$\mu_i(\cdot|S(h')\times T_{-i}):=\mu_{i,i}(\cdot|S_i(h'))\times\mu_{i,-i}(\cdot|S_{-i}(h')\times T_{-i}),$$

where $\mu_{i,i}$ is the CPS on $(S_i,\mathcal{S}_i)$ that satisfies $\mu_{i,i}\left(s_i^{h'}|S_i(h)\right)=1$ for all $h'\in H$. By completeness, there exists $t_i\in T_i$ such that $\beta_i(t_i)=\mu_i$. Then $(s_i,t_i)\in C_i$ because, for all $h'\in H$,

$$\begin{aligned} \sigma_{t_i,i}(s_i(h')|h') &\geq \beta_{i,i}(t_{s_i})\left(s_i^{h'}|S_i(h')\right) \\ &= \mu_{i,i}\left(s_i^{h'}|S_i(h')\right) \\ &= 1. \end{aligned}$$

Hence $\text{proj}_{S_i \times T_i} \text{B}^0(C) \cap (S_i(h) \times T_i) = C_i \cap (S_i(h) \times T_i) \neq \emptyset$. As $i$ and $h$ are arbitrary, the proof of the basis step is complete.

*Inductive step.* Assume that the result is true for each $m = 0, ..., n$. We show that it is also true for $m = n + 1$.

Fix $i \in I$ and $h \in H$. Let $s_i \in S_i(h)$. By the induction hypothesis, the (closed) set $\text{proj}_{S_{-i} \times T_{-i}} \text{B}^n(C)$ is nonempty, and for all $h' \in H$,

$$\text{proj}_{S_{-i} \times T_{-i}} \text{B}^n(C) \cap (S_{-i}(h') \times T_{-i}) \neq \emptyset.$$

So there exists $\mu_{i,-i} \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}(S_{-i} \times T_{-i})$ such that $\mu_{i,-i}$ fully believes event $\text{proj}_{S_{-i} \times T_{-i}} \text{B}^n(C)$. With this, consider the CPS $\mu_i \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h' \in H$,

$$\mu_i(\cdot | S(h') \times T_{-i}) := \mu_{i,i}(\cdot | S_i(h')) \times \mu_{i,-i}(\cdot | S_{-i}(h') \times T_{-i}),$$

where $\mu_{i,i}$ is the CPS on $(S_i, \mathcal{S}_i)$ that satisfies $\mu_{i,i}\left(s_i^{h'} | S_i(h')\right) = 1$ for all $h' \in H$. By completeness, there exists $t_i \in T_i$ such that $\beta_i(t_i) = \mu_i$. The same argument as in the basis step yields $(s_i, t_i) \in C_i$. Moreover $(s_i, t_i) \in \text{B}_i\left(\text{proj}_{S_{-i} \times T_{-i}} \text{B}^n(C)\right)$, because $\beta_{i,-i}(t_i) = \mu_{i,-i}$. It follows that $(s_i, t_i) \in \text{proj}_{S_i \times T_i} \text{B}^{n+1}(C) \cap (S_i(h) \times T_i)$. Since $i$ and $h$ are arbitrary, the conclusion follows. ■

**Remark 9** *Fix a finite game $\Gamma$ and a $\Gamma$-based type structure $\mathcal{T}$. Then, for each $i \in I$ and $n > 1$,*

$$R_i^{*,n+1} = R_i^{*,1} \cap \left( \bigcap_{m=1}^{n} \text{SB}_i\left(R_{-i}^{*,m}\right) \right).$$

With this, we are ready to provide the proof of Theorem 4.

**Proof of Theorem 4**. Part (i): First note that, by Lemma 8, $C_i^* \neq \emptyset$ for each $i \in I$. Moreover

$$\forall i \in I, S_i = \text{proj}_{S_i}(C_i^*). \tag{7.4}$$

The inclusion $\text{proj}_{S_i}(C_i^*) \subseteq S_i$ is obvious. Conversely, let $s_i \in S_i$. By Lemma 8, there exists $\mu_{s_i,-i} \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}(S_{-i} \times T_{-i})$ such that $\mu_{s_i,-i}$ fully believes $C_{-i}^*$. So consider the CPS $\mu_{s_i} \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h \in H$, let

$$\mu_{s_i}(\cdot | S(h) \times T_{-i}) := \mu_{i,i}(\cdot | S_i(h)) \times \mu_{s_i,-i}(\cdot | S_{-i}(h) \times T_{-i}),$$

where $\mu_{i,i}$ is the CPS on $(S_i, \mathcal{S}_i)$ that satisfies $\mu_{i,i}\left(s_i^h | S_i(h)\right) = 1$ for all $h \in H$. By completeness, there exists $t_{s_i} \in T_i$ such that $\beta_i(t_{s_i}) = \mu_{s_i}$. Then $(s_i, t_{s_i}) \in C_i$ because,

for all $h \in H$,

$$\begin{aligned}\sigma_{t_{s_i},i}\left(s_i\left(h\right)|h\right) &\geq \beta_{i,i}\left(t_{s_i}\right)\left(s_i^h|S_i\left(h\right)\right)\\ &= \mu_{i,i}\left(s_i^h|S_i\left(h\right)\right)\\ &= 1.\end{aligned}$$

Since $\beta_{i,-i}\left(t_i\right) := \mu_{s_i,-i}$ fully believes $C_{-i}^*$, Lemma 7 yields $\left(s_i, t_{s_i}\right) \in C_i^*$. Therefore $s_i \in \text{proj}_{S_i}\left(C_i^*\right)$.

We now prove the following claim:

$$\forall i \in I, \forall n \in \mathbb{N}, \text{proj}_{S_i} R_i^{*,n} = S_i^n.$$

The proof is by induction on $n \in \mathbb{N}$.

*Basis step.* Let $s_i \in \text{proj}_{S_i} R_i^{*,1}$, so that $\left(s_i, t_i\right) \in R_i^{*,1} := C_i^* \cap OP_i$ for some $t_i \in T_i$. Transparency of consistency at $\left(s_i, t_i\right)$ and optimal planning implies that $s_i$ satisfies the OSD property given $\text{marg}_{S_{-i}}\beta_i\left(t_i\right)$; so the OSD principle (Remark 3) implies that $s_i \in \rho\left(\text{marg}_{S_{-i}}\beta_i\left(t_i\right)\right)$. Thus $s_i \in S_i^1$.

Conversely, let $s_i \in S_i^1$. By definition, there exists $\nu_i \in \Delta^{\mathcal{S}_{-i}}\left(S_{-i}\right)$ such that $s_i \in \rho\left(\nu_i\right)$. Part (ii) of Lemma 6 yields the existence of $\mu_{i,-i} \in \Delta^{\mathcal{S}_{-i}\times T_{-i}}\left(S_{-i} \times T_{-i}\right)$ such that $\mu_{i,-i}$ fully believes $C_{-i}^*$ and $\text{marg}_{S_{-i}}\mu_{i,-i} = \nu_i$. Let $\mu_{i,i}$ be the CPS on $\left(S_i, \mathcal{S}_i\right)$ that satisfies $\mu_{i,i}\left(s_i^h|S_i\left(h\right)\right) = 1$ for each $h \in H$. Consider the CPS $\mu_i \in \Delta^{\mathcal{S}\times T_{-i}}\left(S \times T_{-i}\right)$ defined as follows: for all $h \in H$,

$$\mu_i\left(\cdot|S\left(h\right) \times T_{-i}\right) := \mu_{i,i}\left(\cdot|S_i\left(h\right)\right) \times \mu_{i,-i}\left(\cdot|S_{-i}\left(h\right) \times T_{-i}\right).$$

Since $\beta_i$ is surjective, there exists $t_i \in T_i$ such that $\beta_i\left(t_i\right) = \mu_i$. We now show that $\left(s_i, t_i\right) \in C_i^* \cap OP_i$. Player $i$ is consistent at $\left(s_i, t_i\right)$, because

$$\begin{aligned}\sigma_{t_i,i}\left(s_i\left(h\right)|h\right) &\geq \beta_{i,i}\left(t_i\right)\left(s_i^h|S_i\left(h\right)\right)\\ &= \text{marg}_{S_i}\beta_i\left(t_i\right)\left(s_i^h|S_i(h)\right)\\ &= \mu_{i,i}\left(s_i^h|S_i\left(h\right)\right)\\ &= 1\end{aligned}$$

for all $h \in H$. Since $\mu_{i,-i}$ fully believes $C_{-i}^*$, Lemma 7 yields $\left(s_i, t_i\right) \in C_i^*$. By inspection of the definition of $\mu_i$, we see that type $t_i$ satisfies independence; moreover, type $t_i$ plans optimally because, for all $h \in H$,

$$\begin{aligned}\text{supp}\beta_{i,i}\left(t_i\right)\left(\cdot|S_i\left(h\right)\right) &= \left\{s_i^h\right\}\\ &\subseteq \arg\max_{r_i \in S_i(h)} \sum_{s_{-i} \in S_{-i}(h)} U_i\left(r_i, s_{-i}\right) \text{marg}_{S_{-i}}\beta_{i,-i}\left(t_i\right)\left(s_{-i}|S_{-i}\left(h\right)\right)\\ &= \arg\max_{r_i \in S_i(h)} \sum_{s_{-i} \in S_{-i}(h)} U_i\left(r_i, s_{-i}\right) \nu_i\left(s_{-i}|S_{-i}\left(h\right)\right).\end{aligned}$$

Hence $(s_i, t_i) \in OP_i$.

*Inductive step.* Assume that the result is true for each $m \leq n$. We show that it is also true for each $m \leq n+1$.

Let $s_i \in \text{proj}_{S_i} R_i^{*,n+1}$, so that $(s_i, t_i) \in R_i^{*,n+1}$ for some $t_i \in T_i$. Then, by Remark 9, $(s_i, t_i) \in R_i^{*,1} \cap \left(\cap_{m \leq n} \text{SB}_i\left(R_{-i}^{*,m}\right)\right)$. Transparency of consistency at $(s_i, t_i)$ and optimal planning implies that $s_i$ satisfies the OSD property given $\nu_i := \text{marg}_{S_{-i}} \beta_i(t_i)$; so the OSD principle (Remark 3) implies that $s_i \in \rho(\nu_i)$. Part (i) of Lemma 6 entails that $\nu_i$ strongly believes $\left(\text{proj}_{S_{-i}} R_{-i}^{*,m}\right)_{m=1}^{n}$, hence, by the induction hypothesis, $\nu_i$ strongly believes $\left(S_{-i}^{m}\right)_{m=1}^{n}$; that is, Condition 2 in the recursive step of Definition 7 is satisfied. Thus $s_i \in S_i^{n+1}$.

Conversely, let $s_i \in S_i^{n+1}$. By definition, there exists $\nu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho(\nu_i)$ and $\nu_i$ strongly believes $\left(S_{-i}^{m}\right)_{m=1}^{n}$. By the induction hypothesis, $\nu_i$ strongly believes $\left(\text{proj}_{S_{-i}} R_{-i}^{*,m}\right)_{m=1}^{n}$. Moreover, $\nu_i$ fully believes $S_{-i}$ by definition, and so, by (7.4), $\nu_i$ fully believes $\text{proj}_{S_{-i}} C_{-i}^{*}$. Hence part (ii) of Lemma 6 yields the existence of $\mu_{i,-i} \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}(S_{-i} \times T_{-i})$ such that

(a) $\mu_{i,-i}$ strongly believes $\left(R_{-i}^{*,m}\right)_{m=1}^{n}$,

(b) $\mu_{i,-i}$ fully believes $C_{-i}^{*}$, and

(c) $\text{marg}_{S_{-i}} \mu_{i,-i} = \nu_i$.

Let $\mu_{i,i}$ be the CPS on $(S_i, \mathcal{S}_i)$ that satisfies $\mu_{i,i}\left(s_i^h | S_i(h)\right) = 1$ for each $h \in H$. Consider the CPS $\mu_i \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h \in H$,

$$\mu_i(\cdot | S(h) \times T_{-i}) := \mu_{i,i}(\cdot | S_i(h)) \times \mu_{i,-i}(\cdot | S_{-i}(h) \times T_{-i}).$$

Since $\beta_i$ is surjective, there exists $t_i \in T_i$ such that $\beta_i(t_i) = \mu_i$. It remains to show that $(s_i, t_i) \in R_i^{*,n+1}$. By Remark 9, this is equivalent to show that $(s_i, t_i) \in R_i^{*,1} \cap \left(\cap_{m \leq n} \text{SB}_i\left(R_{-i}^{*,m}\right)\right)$. Since $\beta_{i,-i}(t_i) = \mu_{i,-i}$, it immediately follows that $(s_i, t_i) \in \cap_{m \leq n} \text{SB}_i\left(R_{-i}^{*,m}\right)$. The proof that $(s_i, t_i) \in R_i^{*,1}$ is the same as that of the basis step. Therefore $(s_i, t_i) \in R_i^{*,n+1}$.

Part (ii): Note that $\left(\prod_{i \in I} R_i^{*,n}\right)_{n \in \mathbb{N}_0}$ is a decreasing sequence of compact sets. Part (i) implies that $\prod_{i \in I} R_i^{*,n} \neq \emptyset$ for every $n \in \mathbb{N}$. Hence, by the finite intersection property, $R^{*,\infty} \neq \emptyset$. By part (i) and Lemma 2, it follows that $\text{proj}_S R^{*,\infty} = S^{\infty}$, as required. ■

# Appendix B: An algorithmic characterization of backwards rationalizability

Penta (2015) shows that backwards rationalizability can be given an algorithmic characterization by a procedure, called "backwards procedure," which is a generalization of the BI algorithm to a wide class of games. In what follows, we will introduce formally the "backwards procedure" and show its equivalence with the solution concept of backwards rationalizability in Definition 6. Towards this end, we need additional notations and definitions. To ease language, in this appendix we slightly modify our terminology. Since the elements $s_i$ that we call "personal external states of $i$" mathematically correspond to the strategies of player $i$, even though they do not represent the plan in $i$'s mind, we call them "objective strategies."

Fix a game $\Gamma$. The set of objective sub-strategies of player $i$ in the sub-tree with root $h \in H$ is denoted by $S_i^{\succeq h}$, that is,

$$S_i^{\succeq h} := \prod_{h' \in H(h)} A_i(h').$$

A generic element of $S_i^{\succeq h}$ is denoted by $s_i^{\succeq h}$. For each $h \in H$, the objective sub-strategy induced by $s_i^{\succeq h} \in S_i^{\succeq h}$ in the sub-tree with root $\bar{h} \succeq h$ is denoted by

$$(s_i^{\succeq h} | \bar{h}) := (s_i(h'))_{h' \in H(\bar{h})} \in S_i^{\succeq \bar{h}}.$$

Recall that $L(h)$ denotes the height of the sub-tree starting at $h \in \bar{H}$, that is, $L(h) := \max_{z \in Z(h)} \ell(z) - \ell(h)$, where $\ell(h)$ denotes the length of $h$. For convenience, we let $K := L(\varnothing)$ denote the "height of the game."

We also find it convenient to use the following notation: for every $k \in \{1, ..., K\}$, let

$$H^k := \{h \in H : L(h) = k\}.$$

Next, fix some $k > 1$. For each $h \in H^k$, let

$$H^{k-1}(h) := \left\{h' \in H^{k-1} : h' \succ h\right\}.$$

Recall that

$$U_i := u_i \circ \zeta : S \to \mathbb{R}$$

is the utility of player $i$ as a function of the external state. Following Penta (2015), we define (objective) strategic-form payoff functions for continuations from a given history: for each $h \in H$ and each $s \in S$, let $U_i(s|h) := u_i(\zeta(s|h))$, where $\zeta(s|h)$ denotes the terminal history induced by profile $s$ from history $h$.

Finally, for each $\nu_i \in \Delta\left(S_{-i}^{\succeq h}\right)$, let

$$BR_i^h(\nu_i) := \arg\max_{s_i^{\succeq h} \in S_i^{\succeq h}} \sum_{s_{-i}^{\succeq h} \in S_{-i}^{\succeq h}} U_i\left(s_i^{\succeq h}, s_{-i}^{\succeq h} | h\right) \nu_i\left(s_{-i}^{\succeq h}\right).$$

If $h = \varnothing$, we simply write $BR_i(\nu_i)$ instead of $BR_i^{\varnothing}(\nu_i)$.

We can formally introduce the "backwards procedure," which starts by considering first all histories of height 1, and then proceeding recursively for all histories of height $k > 1$.

**Definition 8** *Consider the following procedure.*

$(k = 1)$ *For every $i \in I$ and every $h \in H^1$, let*

$$\begin{aligned} P_i^{1,0}(h) &:= S_i^{\succeq h}, \\ P_{-i}^{1,0}(h) &:= \prod_{j \neq i} S_j^{\succeq h}, \end{aligned}$$

*and, for all $n \in \mathbb{N}$,*

$$\begin{aligned} P_i^{1,n}(h) &:= \left\{ s_i^{\succeq h} \in P_i^{1,n-1}(h) : \exists \nu_i \in \Delta\left(P_{-i}^{1,n-1}(h)\right), s_i^{\succeq h} \in BR_i^h(\nu_i) \right\}, \\ P_{-i}^{1,n}(h) &:= \prod_{j \neq i} P_j^{1,n}(h). \end{aligned}$$

*Also, for every $i \in I$ and every $h \in H^1$, let*

$$\begin{aligned} P_i^{1,\infty}(h) &:= \cap_{n \in \mathbb{N}_0} P_i^{1,n}(h), \\ P_{-i}^{1,\infty}(h) &:= \prod_{j \neq i} P_j^{1,\infty}(h). \end{aligned}$$

$(k > 1)$ *For every $i \in I$ and every $h \in H^k$, let*

$$\begin{aligned} P_i^{k,0}(h) &:= \left\{ s_i^{\succeq h} \in S_i^{\succeq h} : \forall h' \in H^{k-1}(h), (s_i^{\succeq h} | h') \in P_i^{k-1,\infty}(h') \right\}, \\ P_{-i}^{k,0}(h) &:= \prod_{j \neq i} P_j^{k,0}(h), \end{aligned}$$

*and, for all $n \in \mathbb{N}$,*

$$\begin{aligned} P_i^{k,n}(h) &:= \left\{ s_i^{\succeq h} \in P_i^{k,n-1}(h) : \exists \nu_i \in \Delta\left(P_{-i}^{k,n-1}(h)\right), s_i^{\succeq h} \in BR_i^h(\nu_i) \right\}, \\ P_{-i}^{k,n}(h) &:= \prod_{j \neq i} P_j^{k,n}(h). \end{aligned}$$

*Also, for every $i \in I$ and every $h \in H^k$, let*

$$\begin{aligned} P_i^{k,\infty}(h) &:= \cap_{n\in\mathbb{N}_0} P_i^{k,n}(h), \\ P_{-i}^{k,\infty}(h) &:= \prod_{j\neq i} P_j^{k,\infty}(h). \end{aligned}$$

*We say that $s \in S$ survives the backwards procedure if $s \in P^{K,\infty}(\varnothing) := \prod_{i\in I} P_i^{K,\infty}(\varnothing)$.*

The main result of this section is the following:

**Proposition 1** *Fix a finite game with observable actions $\Gamma$. Then $\hat{S}^\infty = P^{K,\infty}(\varnothing)$.*

To prove the result, some further notation and auxiliary results are needed.

For each $i \in I$ and $h \in H$, let $\pi_i^h : S_i \to S_i^{\succeq h}$ be the projection map that associates each $s_i \in S_i$ with the induced objective sub-strategy in the sub-tree with root $h$, that is, $\pi_i^h(s_i) = (s_i|h)$. Clearly, each map $\pi_i^h : S_i \to S_i^{\succeq h}$ is onto. Moreover, for every $i \in I$ and $h \in H$, we let $\pi_{-i}^h : S_{-i} \to S_{-i}^{\succeq h}$ denote the "product" of the maps $\pi_j^h$ $(j \neq i)$, that is, $\pi_{-i}^h(s_{-i}) = \left(\pi_j^h(s_j)\right)_{j\neq i}$.

We record an ancillary result that will be used in the proof below.

**Lemma 9** *Fix $i \in I$, $h \in H$ and a nonempty set $Q_i \subseteq S_i$. Then, for all $h' \in H(h)$, the following hold:*
*(i) $\pi_i^{h'}\left(\chi_i^h(Q_i)\right) = \pi_i^{h'}(Q_i)$;*
*(ii) $\chi_i^h(Q_i) \cap S_i(h') \subseteq \chi_i^{h'}(Q_i)$.*

**Proof**: Begin with part (i). Let $s_i^{\succeq h'} \in \pi_i^{h'}\left(\chi_i^h(Q_i)\right)$. We need to show the existence of $\bar{s}_i \in Q_i$ such that $(\bar{s}_i|h') = s_i^{\succeq h'}$. By definition, there exists $s_i \in \chi_i^h(Q_i)$ such that $(s_i|h') = s_i^{\succeq h'}$. Moreover, since $s_i \in \chi_i^h(Q_i)$, there exists $\bar{s}_i \in Q_i$ such that $s_i(h'') = \bar{s}_i(h'')$ for all $h'' \in H(h)$. Since $h' \in H(h)$, we obtain $(\bar{s}_i|h') = s_i^{\succeq h'}$, as required.

For the converse, let $s_i^{\succeq h'} \in \pi_i^{h'}(Q_i)$. We show the existence of $s_i \in \chi_i^h(Q_i)$ such that $(s_i|h') = s_i^{\succeq h'}$. By definition, there exists $\bar{s}_i \in Q_i$ such that $(\bar{s}_i|h') = s_i^{\succeq h'}$. Pick any $s_i \in \chi_i^h(Q_i)$ such that $s_i(h'') = \bar{s}_i(h'')$ for all $h'' \in H(h)$. Since $h' \in H(h)$, we get $(\bar{s}_i|h') = (s_i|h')$, and so $(s_i|h') = s_i^{\succeq h'}$.

To show part (ii), pick any $s_i \in S_i(h')$ such that $s_i \in \chi_i^h(Q_i)$. We show that $s_i \in \chi_i^{h'}(Q_i)$, i.e., there exists $\bar{s}_i \in Q_i$ such that $s_i(h'') = \bar{s}_i(h'')$ for all $h'' \in H(h')$. To this end, note that, since $s_i \in \chi_i^h(Q_i)$, it follows that there exists $\bar{s}_i \in Q_i$ such that

$s_i(h'') = \bar{s}_i(h'')$ for all $h'' \in H(h)$. But $h' \succeq h$, so this implies that $s_i(h'') = \bar{s}_i(h'')$ for all $h'' \succeq h'$. Therefore $s_i \in \chi_i^{h'}(Q_i)$. ■

**Proof of Proposition 1**: We first show that $\hat{S}^\infty \subseteq P^{K,\infty}(\varnothing)$. Specifically, we will prove the following claim:

$$\forall i \in I, \forall k \in \{1, ..., K\}, \forall h \in H^k, s_i \in \hat{S}_i^\infty \Rightarrow (s_i|h) \in P_i^{k,\infty}(h).$$

The proof is by induction on the height of histories.

(Step $k = 1$) Fix any $h \in H^1$. We prove that, for all $i \in I$ and $n \in \mathbb{N}_0$, if $s_i \in \hat{S}_i^\infty$ then $(s_i|h) \in P_i^{1,n}(h)$. The proof is by induction on $n \in \mathbb{N}_0$. For $n = 0$ the result is immediate. Then suppose that the result is true for each $m = 0, ..., n$. We show that it is true for $m = n + 1$. Pick any $s_i \in \hat{S}_i^\infty$, so that there exists a CPS $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\mu_i)$ and $\mu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^\infty\right)|S_{-i}(h')\right) = 1$ for all $h' \in H$. Note that part (i) of Lemma 9 yields $\pi_{-i}^h\left(\chi_{-i}^h\left(\hat{S}_{-i}^\infty\right)\right) = \pi_{-i}^h\left(\hat{S}_{-i}^\infty\right)$; hence it follows from the induction hypothesis that $\pi_{-i}^h(s_{-i}) \in P_{-i}^{1,n}(h)$ provided that $s_{-i} \in \hat{S}_{-i}^\infty$. We can therefore define a probability measure $\nu_i \in \Delta\left(P_{-i}^{1,n}(h)\right)$ as follows: for all $s_{-i}^{\succeq h} \in S_{-i}^{\succeq h}$,

$$\nu_i\left(s_{-i}^{\succeq h}\right) := \mu_i\left(\left(\pi_{-i}^h\right)^{-1}\left(s_{-i}^{\succeq h}\right)|S_{-i}(h)\right).$$

In other words, $\nu_i$ is the image measure of $\mu_i(.|S_{-i}(h))$ on $S_{-i}^{\succeq h}$ under the map $\pi_{-i}^h : S_{-i} \to S_{-i}^{\succeq h}$. The conclusion that $(s_i|h) \in BR_i^h(\nu_i)$ follows from the fact that $s_i \in \rho_i(\mu_i)$ and $(s_i^h|h) = (s_i|h)$. Hence $(s_i|h) \in P_i^{1,n+1}(h)$.

(Step $k > 1$) Suppose that the statement has been proved to hold for all histories of height $l = 1, ..., k-1$. Fix any $h \in H^k$. We show that, for all $i \in I$ and $n \in \mathbb{N}_0$, if $s_i \in \hat{S}_i^\infty$ then $(s_i|h) \in P_i^{k,n}(h)$. The argument proceeds by induction on $n \in \mathbb{N}_0$.

($n = 0$) Pick any $s_i \in \hat{S}_i^\infty$. By the induction hypothesis on the height of histories, it follows that $(s_i|h') \in P_i^{k-1,\infty}(h')$ for all $h' \in H^{k-1}(h)$. Hence, by definition, $(s_i|h) \in P_i^{k,0}(h)$.

($n \geq 0$) Suppose that the result is true for each $m = 0, ..., n$. We show that it is true for $m = n + 1$. The argument proceeds in the same way as in step $k = 1$. Let $s_i \in \hat{S}_i^\infty$. There is a CPS $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\mu_i)$ and $\mu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^\infty\right)|S_{-i}(h')\right) = 1$ for all $h' \in H$. Using again part (i) of Lemma 9 and the induction hypothesis, we obtain that $\pi_{-i}^h(s_{-i}) \in P_{-i}^{k,n}(h)$ for all $s_{-i} \in \hat{S}_{-i}^\infty$. We can define a probability measure $\nu_i \in \Delta\left(P_{-i}^{k,n}(h)\right)$ as the image measure of $\mu_i(.|S_{-i}(h))$ on $S_{-i}^{\succeq h}$ under the map $\pi_{-i}^h : S_{-i} \to S_{-i}^{\succeq h}$. Hence, the same argument as above entails that $(s_i|h) \in BR_i^h(\nu_i)$.

We now show that $P^{K,\infty}(\varnothing) \subseteq \hat{S}^{\infty}$. We do this by showing that $P^{K,\infty}(\varnothing) \subseteq \hat{S}^{n}$ for all $n \in \mathbb{N}_0$. For $n = 0$, the result follows from the fact that $\hat{S}^0 = S$. Then suppose that the result is true for each $m = 0, ..., n$. We prove the result for $m = n + 1$.

We first record a consequence of the induction hypothesis.

**Claim 2** *For every $i \in I$, $k \in \{1, ..., K\}$ and $h \in H^k$,*

$$P_i^{k,\infty}(h) = \pi_i^h\left(P_i^{K,\infty}(\varnothing)\right) \subseteq \pi_i^h\left(\hat{S}_i^n\right) = \pi_i^h\left(\chi_i^h\left(\hat{S}_i^n\right)\right)$$

*Proof of Claim 2.* The first equality follows from the definition of the backwards procedure, while the set inclusion follows from the induction hypothesis. Part (i) of Lemma 9 yields the last equality. □

We make use of this result to construct, for each $h \in H$, a profile of maps $\left(\varphi_i^h : S_i^{\succeq h} \to S_i\right)_{i \in I}$ satisfying some desirable properties.

Fix $i \in I$ and $h \in H$. There exists $k \in \{1, ..., K\}$ such that $h \in H^k$. Claim 2 yields, for each $s_i^{\succeq h} \in P_i^{k,\infty}(h)$, the existence of $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$ such that $\pi_i^h(s_i) = s_i^{\succeq h}$. Hence, for every $s_i^{\succeq h} \in P_i^{k,\infty}(h)$, we choose and fix some $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$ such that $\pi_i^h(s_i) = s_i^{\succeq h}$, we also choose an arbitrary $s_i^0 \in S_i$, and we define the map $\varphi_i^h : S_i^{\succeq h} \to S_i$ as follows:

$$\varphi_i^h\left(s_i^{\succeq h}\right) = \begin{cases} s_i, & \text{if } s_i^{\succeq h} \in P_i^{k,\infty}(h), \\ s_i^0, & \text{otherwise.} \end{cases}$$

By construction, each map $\varphi_i^h$ satisfies $\varphi_i^h\left(P_i^{k,\infty}(h)\right) \subseteq \chi_i^h\left(\hat{S}_i^n\right)$, which in turn implies

$$P_i^{k,\infty}(h) \subseteq \left(\varphi_i^h\right)^{-1}\left(\chi_i^h\left(\hat{S}_i^n\right)\right). \tag{7.5}$$

For every $i \in I$ and $h \in H$, we let $\varphi_{-i}^h : S_{-i}^{\succeq h} \to S_{-i}$ denote the "product" of the maps $\varphi_j^h$ $(j \neq i)$, that is, $\varphi_{-i}^h\left(s_{-i}^{\succeq h}\right) := \left(\varphi_j^h\left(s_j^{\succeq h}\right)\right)_{j \neq i}$.

Having done these preparations, we are now ready to provide the proof of the inductive step.

Let $s_i \in P_i^{K,\infty}(\varnothing)$. We show the existence of a CPS $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\mu_i)$ and $\mu_i\left(\chi_{-i}^h\left(\hat{S}_{-i}^n\right) | S_{-i}(h)\right) = 1$ for all $h \in H$. For every $k \in \{1, ..., K\}$ and every $h \in H^k$, there exists $\nu_i^h \in \Delta\left(P_{-i}^{k,\infty}(h)\right)$ such that $s_i^{\succeq h} \in BR_i^h\left(\nu_i^h\right)$. We

carefully select some of these probability measures to construct a CPS $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ that satisfies the required properties. The construction goes as follows.

For all $h \in H$ such that $\nu_i^{\varnothing}(S_{-i}(h)) > 0$, and for all $E_{-i} \subseteq S_{-i}$, let

$$\mu_i(E_{-i}|S_{-i}(h)) := \frac{\nu_i^{\varnothing}(E_{-i} \cap S_{-i}(h))}{\nu_i^{\varnothing}(S_{-i}(h))}.$$

Next, consider some $h' = (h, a) \in H^k$ $(k \neq K)$ such that $\nu_i^{\varnothing}(S_{-i}(h)) > 0$ and $\nu_i^{\varnothing}(S_{-i}(h')) = 0$. In this case, for all $E_{-i} \subseteq S_{-i}$, let

$$\mu_i(E_{-i}|S_{-i}(h')) := \nu_i^{h'}\left(\left(\varphi_{-i}^{h'}\right)^{-1}(E_{-i})\right),$$

and, for all $h'' \succ h'$ such that $\nu_i^{h'}\left(\left(\varphi_{-i}^{h'}\right)^{-1}(S_{-i}(h''))\right) > 0$, let

$$\mu_i(E_{-i}|S_{-i}(h'')) := \frac{\nu_i^{h'}\left(\left(\varphi_{-i}^{h'}\right)^{-1}(E_{-i} \cap S_{-i}(h''))\right)}{\nu_i^{h'}\left(\left(\varphi_{-i}^{h'}\right)^{-1}(S_{-i}(h''))\right)}.$$

For all other histories, we proceed as above, in order to obtain an array of probability measures $\mu_i = (\mu_i(.|S_{-i}(h)))_{h \in H}$ such that the chain rule holds; hence $\mu_i$ is a well-defined CPS on $(S_{-i}, \mathcal{S}_{-i})$.

We now show that $\mu_i\left(\chi_{-i}^h\left(\hat{S}_{-i}^n\right)|S_{-i}(h)\right) = 1$ for all $h \in H$. To this end, let $h' \in H$. There exists a unique $k' \in \{1, ..., K\}$ such that $h' \in H^{k'}$. By construction of $\mu_i$, there exists $h \in H$ such that $h' \in H(h)$ (hence $h \in H^k$ where $k \geq k'$) and such that

$$\mu_i(\cdot|S_{-i}(h)) = \nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}(\cdot)\right),$$

and $\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)>0$. We get that

$$\begin{aligned}
\mu_i\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^n\right)|S_{-i}\left(h'\right)\right) &= \frac{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^n\right)\cap S_{-i}\left(h'\right)\right)\right)}{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)} \\
&\geq \frac{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(\chi_{-i}^{h}\left(\hat{S}_{-i}^n\right)\cap S_{-i}\left(h'\right)\cap S_{-i}\left(h'\right)\right)\right)}{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)} \\
&= \frac{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(\chi_{-i}^{h}\left(\hat{S}_{-i}^n\right)\cap S_{-i}\left(h'\right)\right)\right)}{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)} \\
&= \frac{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(\chi_{-i}^{h}\left(\hat{S}_{-i}^n\right)\right)\cap\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)}{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)} \\
&= \frac{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)}{\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)} \\
&= 1,
\end{aligned}$$

where the first equality is by definition, the inequality follows from part (ii) of Lemma 9, the second and third equalities are obvious, while the fourth equality follows from the following fact: since $\nu_i^h\in\Delta\left(P_{-i}^{k,\infty}\left(h\right)\right)$, it follows from (7.5) that

$$\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(\chi_{-i}^h\left(\hat{S}_{-i}^n\right)\right)\right)=1;$$

using the fact that $\nu_i^h\left(\left(\varphi_{-i}^h\right)^{-1}\left(S_{-i}\left(h'\right)\right)\right)>0$, the fourth equality follows.[40]

Finally, the conclusion $s_i\in\rho_i\left(\mu_i\right)$ is immediate by construction of $\mu_i$. Hence $s_i\in\hat{S}_i^{n+1}$, as required. ■

[40]Specifically, the conclusion follows from the following, simple exercise in probability theory. Fix a finite probability space $\left(\Omega,2^\Omega,\mu\right)$. If $E$ and $F$ are nonempty events of $\Omega$ such that $\mu\left(E\right)=1$ and $\mu\left(F\right)>0$, then $E\cap F\neq\emptyset$ and $\mu\left(E\cap F\right)=\mu\left(F\right)$.

# References


[1] Arieli, I., and R.J. Aumann (2015): "The Logic of Backward Induction," *Journal of Economic Theory*, 159, 443-464.

[2] Asheim, G.B. (2002): "On the Epistemic Foundation for Backward Induction," *Mathematical Social Sciences*, 44, 121-144.

[3] Asheim, G.B., and A. Perea (2005): "Sequential and Quasi-perfect Rationalizability in Extensive Games," *Games and Economic Behavior,* 53, 15-42.

[4] Aumann, R.J. (1995): "Backward Induction and Common Knowledge of Rationality," *Games and Economic Behavior*, 8, 6-19.

[5] Baltag, A., S. Smets, and J.A. Zvesper (2009): "Keep 'Hoping' for Rationality: A Solution to the Backward Induction Paradox," *Synthese,* 169, 301-333.

[6] Battigalli, P. (1996): "Strategic Rationality Orderings and the Best Rationalization Principle," *Games and Economic Behavior,* 13, 178-200.

[7] Battigalli, P. (1997): "On Rationalizability in Extensive Games," *Journal of Economic Theory*, 74, 40-61.

[8] Battigalli, P. (2003): "Rationalizability in Infinite, Dynamic Games of Incomplete Information," *Research in Economics,* 57, 1-38.

[9] Battigalli, P., and G. Bonanno (1999): "Recent Results on Belief, Knowledge and the Epistemic Foundations of Game Theory," *Research in Economics,* 53, 149-226.

[10] Battigalli, P., and M. Dufwenberg (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1-35.

[11] Battigalli, P., and A. Friedenberg (2012): "Forward Induction Reasoning Revisited," *Theoretical Economics*, 7, 57-98.

[12] Battigalli, P., and M. Siniscalchi (1999a): "Hierarchies of Conditional Beliefs and Interactive Epistemology in Dynamic Games," *Journal of Economic Theory*, 88, 188-230.

[13] Battigalli, P., and M. Siniscalchi (1999b): "Interactive Beliefs, Epistemic Independence and Rationalizability," *Research in Economics,* 53, 243-246.

[14] Battigalli, P., and M. Siniscalchi (2002): "Strong Belief and Forward Induction Reasoning," *Journal of Economic Theory*, 106, 356-391.

[15] Battigalli, P., and M. Siniscalchi (2007): "Interactive Epistemology in Games with Payoff Uncertainty," *Research in Economics*, 61, 165-184.

[16] Battigalli, P., and P. Tebaldi (2018): "Interactive Epistemology in Simple Dynamic Games with a Continuum of Strategies," *Economic Theory*, https://doi.org/10.1007/s00199-018-1142-8.

[17] Battigalli, P., R. Corrao, and F. Sanna (2018): "Epistemic Game Theory Without Type Structures. An Application to Psychological Games," mimeo, Bocconi University.

[18] Battigalli, P., A. Di Tillio, and D. Samet (2013): "Strategies and Interactive Beliefs in Dynamic Games," in *Advances in Economics and Econometrics*, ed. by D. Acemoglu, M. Arellano, and E. Dekel. Cambridge, UK: Cambridge University Press, 391-422.

[19] Battigalli, P., M. Siniscalchi, and A. Friedenberg (2017): *Epistemic Game Theory: Reasoning about Strategic Uncertainty*, MIT Press, in preparation.

[20] Battigalli, P., E. Catonini, G. Lanzani, and M. Marinacci (2017): "Ambiguity Attitudes and Self-Confirming Equilibrium in Sequential Games," IGIER working paper 607.

[21] Ben Porath, E. (1997): "Rationality, Nash Equilibrium, and Backward Induction in Perfect Information Games," *Review of Economic Studies*, 64, 23-46.

[22] Blume, L., A. Brandenburger, and E. Dekel (1991): "Lexicographic Probabilities and Choice Under Uncertainty," *Econometrica,* 59, 61-79.

[23] Bonanno, G. (2013): "A Dynamic Epistemic Characterization of Backward Induction Without Counterfactuals," *Games and Economic Behavior,* 78, 31–45.

[24] Dekel, E., and M. Siniscalchi (2015): "Epistemic Game Theory," in *Handbook of Game Theory with Economic Applications, Volume 4*, ed. by P. Young and S. Zamir. Amsterdam: North-Holland, 619-702.

[25] Frederick, S., G. Loewenstein, and T. O'Donoghue (2002) "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature,* 40*, 351-401.*

[26] Halpern, J.Y. (2001): "Substantive Rationality and Backward Induction," *Games and Economic Behavior*, 37, 425-435.

[27] Heifetz, A., and A. Perea (2015): "On the Outcome Equivalence of Backward Induction and Extensive Form Rationalizability," *International Journal of Game Theory*, 44, 37-59.

[28] Kuhn, H.W. (1953): "Extensive Games and the Problem of Information," in *Contributions to the Theory of Games II*, ed. by H.W. Kuhn and A.W. Tucker. Princeton: Princeton University Press, 193-216.

[29] Marinacci, M. (2015): "Model Uncertainty," *Journal of the European Economic Association*, 13, 998-1076.

[30] Mertens, J.F., and S. Zamir (1985): "Formulation of Bayesian Analysis for Games With Incomplete Information," *International Journal of Game Theory*, 14, 1-29.

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

[32] Penta, A. (2015): "Robust Dynamic Implementation," *Journal of Economic Theory,* 160, 280-316.

[33] Perea, A. (2007): "Epistemic Foundations for Backward Induction: An Overview," in *Texts in Logic and Games, Volume 1*, ed. by J. van Benthem, D. Gabbay and B. Löwe. London: Amsterdam University Press, 159-193.

[34] Perea A. (2012): *Epistemic Game Theory: Reasoning and Choice*, CUP Press.

[35] Perea, A. (2014): "Belief in the Opponents' Future Rationality," *Games and Economic Behavior,* 83, 231-254.

[36] Perea, A. (2018): "Why Forward Induction Leads to the Backward Induction Outcome: A New Proof for Battigalli's Theorem," *Games and Economic Behavior,* 110, 120-138.

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

[38] Rubinstein, A. (1991): "Comments on the Interpretation of Game Theory," *Econometrica*, 59, 909-904.

[39] Samet, D. (2013): "Common Belief of Rationality in Games of Perfect Information," *Games and Economic Behavior*, 79, 192-200.

[40] Selten, R. (1975): "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25-55.

[41] Shimoji, M. (2004): "On the Equivalence of Weak Dominance and Sequential Best Response," *Games and Economic Behavior*, 48, 385-402.

[42] Shimoji, M., and J. Watson (1998): "Conditional Dominance, Rationalizability, and Game Forms," *Journal of Economic Theory,* 83, 161-195.

[43] Siniscalchi, M. (2016): "Structural Rationality in Dynamic Games," mimeo, Northwestern University.

[44] Van Benthem, J. (2007): "Dynamic Logic for Belief Revision," *Journal of Applied Non-Classical Logics*, 17, 129-155.