INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Beliefs, Plans, and Perceived Intentions in Dynamic Games

Pierpaolo Battigalli, Nicodemo De Vito

Working Paper n. 629

This Version: June 2021

IGIER – Università Bocconi, Via Guglielmo Röntgen 1, 20136 Milano –Italy
http://www.igier.unibocconi.it

Beliefs, Plans, and Perceived Intentions in Dynamic Games*

Pierpaolo Battigalli
Department of Decision Sciences and IGIER, Bocconi University

Nicodemo De Vito
Department of Decision Sciences, Bocconi University

June 18, 2021

Abstract

We adopt the epistemic framework of Battigalli and Siniscalchi (*J. Econ. Theory* 88:188-230, 1999) to model the distinction between a player's behavior at each node, 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. Rational players plan optimally and their behavior is consistent with their plans. We illustrate our approach with detailed examples and some results. We prove that optimal planning, belief in continuation consistency and common full belief in both imply the backward induction strategies and beliefs in games with perfect information and no relevant ties. More generally, we present within our framework relevant epistemic assumptions about backward and forward-induction reasoning, and relate them to similar ones studied in the previous literature.

*We thank three anonymous referees, Geir B. Asheim, Andrés Perea, and Marciano Siniscalchi for comments and suggestions on a previous draft of this paper, Federico Bobbio, Roberto Corrao, Carlo Cusumano, Enrico De Magistris, Francesco Fabbri, Davide Ferri, Nicolò Generoso, Shuige Liu, Julien Manili, and David Ruiz for careful proof-reading, and the participants to conference and seminar presentations at TARK 2017, EEA-ESEM 2018, ESWC 2020, University of Copenhagen, Heidelberg University, Northwestern University, and Queen Mary College for useful comments. Pierpaolo Battigalli gratefully acknowledges the financial support of ERC, grant 324219.

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

JEL: C72, C73.

1 Introduction

Players who reason strategically anticipate the moves of others under the assumption that they are rational and "sophisticated," i.e., that co-players reason strategically in some well specified sense. In games where some moves are sequential, henceforth **dynamic games**, players have to interpret 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 other players, but similar deviations are not expected to occur in the future, as in the trembling-hand story by Selten (1975).

The *overarching principle* of this paper is that a flexible framework to model strategic reasoning in dynamic games should *allow for the formal 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 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 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 instead the epistemic framework of Battigalli and Siniscalchi (1999), which allows us to *decouple plan and behavior*. With this, we can model how players change their perceptions about the intentions of others, e.g., by assuming that upon observing an unexpected action of a co-player they think that he deviated from his plan, or—alternatively—that he must be implementing a plan different from the one originally ascribed to him. Players hold (first-order) beliefs about the behavior

[1] See, for example, the surveys on epistemic game theory by Battigalli and Bonanno (1999), or 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.

of everybody, including themselves,[2] and plans are modeled as beliefs about own behavior. We use the framework to analyze examples and derive results about the behavioral implications of different assumptions on strategic reasoning. Our first main result provides epistemic conditions for the backward-induction strategies and beliefs. 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 behavior of others for every possible contingency,
- **consistency** (C), that is, coincidence between plan and behavior, and
- **belief in continuation consistency** (BCC), that is, upon observing any history h, each player believes that the co-players' behavior will be consistent with their plans *starting from* h, whether or not they had been 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. To model backward-induction reasoning 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 in the game. With this, we show the following (Theorem 1):

In games with perfect information and no relevant ties, correct common full belief in $OP \cap BCC$ implies the backward-induction plans and beliefs about others; if players are also consistent, then they are rational and 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 games with perfect information and no relevant ties.

Moving on to the analysis of forward-induction reasoning, we consider the assumptions of **rationality and common strong belief in rationality** (RCSBR,

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

cf. Battigalli and Siniscalchi 2002): players (1) **strongly believe** (i.e., believe whenever possible) that the co-players are rational, (2) strongly believe that—on top of being rational—the co-players also strongly believe the others are rational, and more generally strongly believe in the highest order of rationality and mutual strong belief in rationality consistent with observed behavior. We prove that—in the universal type structure—the behavioral implications of RCSBR are characterized by strong rationalizability (Theorem 3).[3] We further 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 in* C; with this, we prove that in the universal 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). Like us, Battigalli and Siniscalchi (2002) provide an epistemic justification of strong rationalizability. Differently from us, they use a framework whereby players only hold beliefs about other players and there is no separate description of plans and behavior. Our latter result shows that we can interpret their framework as implicitly assuming that players are consistent and there is common full belief in consistency.

To sum up, backwards rationalizability characterizes the behavioral implications of our epistemic assumptions on backward-induction reasoning, while strong rationalizability characterizes the behavioral implications (in the universal type structure) of RCSBR, which represents forward-induction reasoning. These solution concepts are different. In some games they even select disjoint sets of strategy profiles (see Section 3.2 and the discussion in Section 7). In both cases co-players are initially expected to behave as planned, and such differences follow from how backward and forward-induction reasoning shape players' perceptions of co-players' intentions upon observing unexpected actions: backward-induction reasoning explains such actions as one-off deviations from plans, forward-induction reasoning explains them as executing unexpected rationalizable plans.

While these results illustrate the expressive power of the adopted framework, we propose that its usefulness goes well beyond them. For example, decoupling plans from behavior is essential to correctly model players who intrinsically care about the intentions of co-players as in the psychological-game models of reciprocity, guilt aversion, and anger (see Battigalli and Dufwenberg 2020 and references therein).

Related literature The epistemic analysis of backward induction dates back to Aumann (1995). Other articles with epistemic conditions for either the backward-

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

induction strategies, or the backward-induction path include Asheim (2002), Battigalli and Siniscalchi (2002), Asheim and Perea (2005), Bonanno (2013, 2014), and Perea (2014). The aforementioned backwards rationalizability solution concept was independently put forward by Penta (2015) and Perea (2014) as an extension of the backward induction solution to games with imperfect information. The epistemic analysis of forward-induction reasoning is due to Battigalli and Siniscalchi (2002) and is revisited by Battigalli and Friedenberg (2012), who drop the assumption that the backdrop type structure is universal. Since we provide epistemic justifications of backwards rationalizability and strong rationalizability, our paper is most related to those of Perea (2014) and Battigalli and Siniscalchi (2002). Note that *the formal language of all the aforementioned papers does not allow to decouple plans from behavior,* which is instead essential to express the epistemic assumptions analyzed here. We briefly explained above how we relate to Battigalli and Siniscalchi (2002). For the class of games analyzed here, we can describe the main result by Perea (2014) as stating that backwards rationalizable behavior follows from the assumption that, conditional on the occurrence of every non-terminal history, players would be rational in the ensuing subgame and this would be commonly believed. Perea informally mentions an interpretation in terms of "mistakes." Since we can decouple plans from behavior, we make this interpretation precise: "mistakes" are deviations from planned behavior, the assumption of common full belief in (i) optimal planning and (ii) belief in continuation consistency implies that unexpected actions are explicitly interpreted by other players as one-off deviations from plans. We provide more detailed comments on the related literature in Section 7.

Structure of the paper 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 our 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 in detail on the related literature. Most of the proofs are collected in the Appendix.

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

For the sake of notational simplicity, we focus on finite multistage games with perfect monitoring of past actions.[4] Given some preliminaries about sequences and trees, we define these games and the external states describing players' 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, of sequences of action profiles $a \in A := \prod_{i\in I} A_i$. We let Z denote the set of terminal histories, and $H := \bar{H} \backslash Z$ is the set of non-terminal 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.

[4] Our framework and techniques can be extended to games with perfect recall.

- For each $i \in I$, $u_i : Z \to \mathbb{R}$ is the payoff (utility) function for player i.

The intended interpretation of Γ is that, as the game unfolds, each player is informed of the sequence of action profiles that has just occurred. Indeed, we assume more: as soon as a history h occurs it becomes common knowledge that h has occurred.

Since the restrictions of $\prec$ and $\preceq$ to $\bar{H}$ 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 $|A_i(h)| = 1$).[5] 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 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 $S_{-i} := \prod_{j \in I \setminus \{i\}} S_j$.[6] An external state $(s_i)_{i \in I} \in S$ is interpreted as an *objective description of players' behavior* conditional on the occurrence of each non-terminal history, which may or may *not* coincide with what players plan to do.[7] For this reason we often call "behavior" the external states. 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: the first element is $s(\varnothing)$, if $s(\varnothing) \notin Z$ the second element is $s((s(\varnothing)))$, and so on. Thus, we can determine a **path function** $\zeta : S \to Z$ and, for each $h \in H$, the set of external states inducing h:

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

[5] 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.

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

[7] In other words, an external state is a specification of the truth values of all the behavioral subjunctive conditionals of the form "if h occurred, a_i would be chosen" (with $h \in H$, $i \in I$ and $a_i \in A_i(h)$).

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 allow h (that is, do not prevent the realization of h). 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 allow h. Note that, in a game with observable actions,

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

for every $h \in H$.[8]

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 ν is denoted by suppν. The set $\Delta(X)$ is endowed with the topology of weak convergence, so that $\Delta(X)$ becomes a compact metrizable space.

We consider arrays of probability measures indexed by elements of a countable collection $\mathcal{C}$ of "conditioning events," i.e., $\mu := (\mu(\cdot|C))_{C \in \mathcal{C}} \in \Delta(X)^{\mathcal{C}}$ (see Renyi 1955).[9]

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*

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

[8] 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.

[9] For every pair of sets P and Q, Q^P denotes the set of functions with domain P and codomain Q. Thus, μ 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 $C \in \mathcal{C}$.

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 1999).

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 ν 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 μ on $(X, \mathcal{C})$ is defined as the array of probability measures

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

It can be verified that $\text{marg}_X \mu$ is a CPS on $(X, \mathcal{C})$. Thus, for every $C \in \mathcal{C}$, it makes sense to write $\text{marg}_X \mu(\cdot|C)$ instead of $\text{marg}_X \mu(\cdot|C \times Y)$.

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 non-terminal 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 $S(h) \times T_{-i}$;[10] 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 such that $\mu_i(S(h) \times T_{-i}|S(h) \times T_{-i}) = 1$ for each $h \in H$.

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

Definition 2 *A* Γ*-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* $(s_i, t_i)_{i \in I} \in \prod_{i \in I}(S_i \times T_i)$*.*
A type structure is ***complete*** *if, for every* $i \in I$*, the belief map* β_i *is* onto *(surjective).*

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 first-order and 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, where $\left(\text{marg}_S \beta_{i,h}(t_i)\right)_{h \in H}$ represents the first-order beliefs, that is, beliefs about behavior. Battigalli and Siniscalchi (1999) 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.[11] Such canonical type structure is complete. Furthermore, it is "universal," or "terminal" in the sense that every other type structure can be mapped into it in a unique belief-preserving way. 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.[12] A Γ-based **standard type structure** is a tuple

[11] 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.

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

$\mathcal{T} = \left(H, (S_{-i}, T_i, \beta_i)_{i \in I}\right)$ where, as in Definition 2, each T_i is a compact metrizable space of player i's types, and 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 players different from i, 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.[13] Furthermore, the approach is adequate for the analysis of expected utility maximizing players in dynamic games.[14]

However, we argue that in the analysis of dynamic games there are conceptual advantages in introducing players' beliefs about their own behavior. 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 description of the 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.[15] 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' 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 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

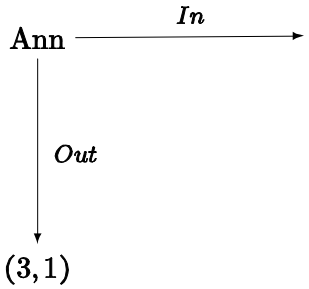
Consider the game depicted in Figure 3.1 ("Battle of Sexes with Outside Option," BoSOO) between two players, Ann (a) and Bob (b). If Ann does not choose the

[13] Under the assumption that players choose rationally complemented by a strong invariance assumption; see Siniscalchi (2020).

[14] See the monograph by Perea (2012), and the survey by Dekel and Siniscalchi (2015).

[15] 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. (2020).

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.

The set of non-terminal histories is $H = \{\varnothing, (In)\}$, while the sets of personal external states of each player are[16]

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

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.

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

Table 1: Type structure for the 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, $\beta_{a,\varnothing}\left(t_a^1\right)\left(\{((In.C, c), t_b^1)\}\right) = 1$.

At both states of the world

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

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

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 players; 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.

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 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 $\left(\left(In.C, t_a^2\right), \left(s_b, t_b^2\right)\right)$ $(s_b \in \{c, m\})$ 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;[17]
- **fully believes** event E if i assigns probability 1 to E conditional on each history h.[18]

At state $(s^1, t^1) = \left(\left(In.C, t_a^1\right), \left(c, t_b^1\right)\right)$, 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

[17] See the formal definitions in Section 6.

[18] 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.

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 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).[19]

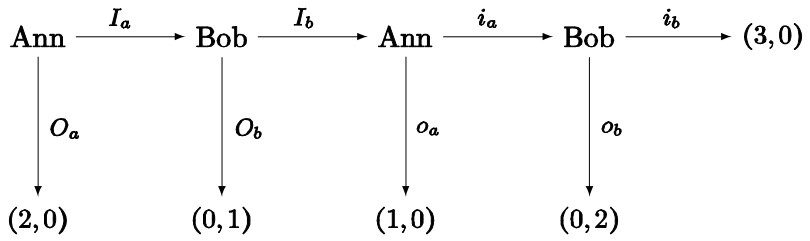


Figure 3.2: A game with perfect information.

[19] Cf. Reny 1992, Figure 3.

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)[20] 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 $I_b.o_b$, 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.

β_a	$\varnothing$	(I_a)	(I_a, I_b)	(I_a, I_b, i_a)
t_a^{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^{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^*	$((I_a.i_a, I_b.i_b), \cdot), 1$	$((I_a.i_a, I_b.i_b), \cdot), 1$	$((I_a.i_a, I_b.i_b), \cdot), 1$	$((I_a.i_a, I_b.i_b), \cdot), 1$

β_b	$\varnothing$	(I_a)	(I_a, I_b)	(I_a, I_b, i_a)
t_b^{fi}	$\left((O_a.o_a, I_b.o_b), t_a^{\text{fi}}\right), 1$	$\left((I_a.i_a, I_b.o_b), t_a^*\right), 1$	$\left((I_a.i_a, I_b.o_b), t_a^*\right), 1$	$\left((I_a.i_a, I_b.o_b), t_a^*\right), 1$
t_b^{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^{fi} has always the same beliefs about Bob: Bob's type is t_b^{fi}, he plans $I_b.o_b$, and he is going to execute his plan. The plan of t_a^{fi} is $O_a.o_a$ in the following sense: conditional on each history where she is active, t_a^{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^{fi}

[20]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.

must acknowledge that she deviated from her plan at the root). Given this, the plan of t_a^{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.

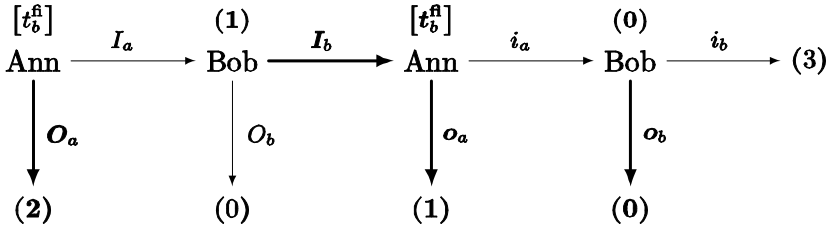


Figure 3.3: Plan and beliefs of type t_a^{fi} of *Ann*.

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.

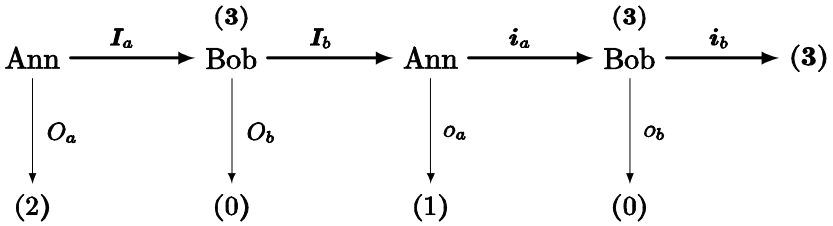


Figure 3.4: Plan and (first-order) beliefs of t_a^*.

Type t_a^{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^{fi} plans $I_b.o_b$, believes at the beginning of the game that Ann's type is t_a^{fi} and that she plays according to her plan $O_a.o_a$; upon observing action I_a, Bob of type t_b^{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^{bi}, and types above nodes of Bob represent his conditional higher-order beliefs.

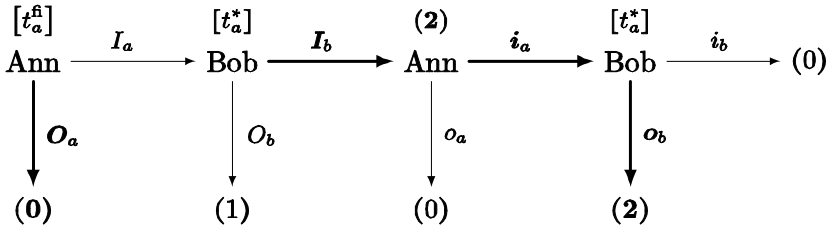


Figure 3.5: Plan and beliefs of type t_b^{fi} of Bob.

Finally, it is immediate to check that type t_b^{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^{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^{bi} $(i = a, b)$ plans optimally and fully believes that the co-player's type is t_{-i}^{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 (subsection 4.1), next we analyze optimal planning (subsection 4.2), and finally we define rationality as the conjunction of optimal planning and consistency between plan and behavior (subsection 4.3).

4.1 Independence

Recall that $\mathcal{S}_{-i} \times T_{-i}$ is the collection of observable events about i's co-players. Similarly, let $\mathcal{S}_i := \left\{F \in 2^{S_i} : \exists h \in H, F = S_i(h)\right\}$ denote the collection of observable events about i's behavior.

Definition 3 *We say that type t_i in a Γ-based type structure $\mathcal{T}$ satisfies **independence** if there exist two CPSs $\beta_{i,i}(t_i) \in \Delta^{\mathcal{S}_i}(S_i)$ and $\beta_{i,-i}(t_i) \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}(S_{-i} \times T_{-i})$ such that*

$$\forall h \in H, \beta_i(t_i)(\cdot | S(h) \times T_{-i}) = \beta_{i,i}(t_i)(\cdot | S_i(h)) \times \beta_{i,-i}(t_i)(\cdot | S_{-i}(h) \times T_{-i}). \quad (4.1)$$

In words, $\beta_i(t_i)$ is the "product" of two independent marginal CPSs: $\beta_{i,i}(t_i)$ is a CPS about i himself, and $\beta_{i,-i}(t_i)$ is a CPS about the co-players. To better understand this independence condition, note that two distinct histories h and h' reveal the same information about the behavior of others—that is, $S_{-i}(h) = S_{-i}(h')$—if they differ only because of actions taken by i. Condition (4.1) implies that i's beliefs about how other players behave and think (as a function of what they observe) conditional on such histories must be the same; hence, they must be independent of i's behavior. A similar condition applies to i's beliefs about his own behavior: if $h \neq h'$ and $S_i(h) = S_i(h')$, then these histories differ only because of actions taken by co-players, and we require that such differences do not affect i's predictions about his own behavior.[21]

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

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

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

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

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

$$\sigma_{t_i,i}(a_i|h) := \beta_{i,i}(t_i)(S_i(h,a_i)|S_i(h)),$$

$$\sigma_{t_i,-i}(a_{-i}|h) := \beta_{i,-i}(t_i)(S_{-i}(h,a_{-i}) \times T_{-i}|S_{-i}(h) \times T_{-i}),$$

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*

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

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^{fi} and t_a^{bi} is that Ann's beliefs about the type t_b and behavior s_b of Bob are independent of what Ann

[21] Of course, i's predictions about the actions he is going to choose may depend on the observed behavior of others, as such dependence may well be part of his plan.

does, and in particular do not depend on whether Ann deviated or not from her plan. Indeed type t_a^{fi} (resp. t_a^{bi}) of Ann initially plans to go out immediately and believes that Bob's personal state is $\left(I_b.o_b, t_b^{\text{fi}}\right)$ (resp. $\left(O_b.o_b, t_b^{\text{bi}}\right)$); 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 Γ-based type structure $\mathcal{T}$, the expected payoff of type t_i conditional on observing history $h' \in H$ is

$$V_{t_i}(h') := \sum_{s \in S(h')} U_i(s) \operatorname{marg}_S \beta_i(t_i)(s|S(h')).$$

For notational convenience also let $V_{t_i}(z) := u_i(z)$ 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 Γ-based type structure $\mathcal{T}$ **plans optimally** if it satisfies independence and*

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

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. Let T_i^* denote the set of i's types that satisfy independence. With this, the set of types satisfying optimal planning is

$$\overline{OP}_i := \left\{ t_i \in T_i^* : \forall h \in H, \operatorname{supp}\sigma_{t_i,i}(\cdot|h) \subseteq \arg\max_{a_i \in A_i(h)} V_{t_i}(h, a_i) \right\}.$$

The corresponding optimal-planning event for 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 sets.

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 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—which relies on the chain rule and the independence condition (4.1)—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 $h \in H$.

This is a version of the *OSD Principle*: the plan of a type t_i that satisfies independence and—as per Definition 4—the OSD property must also be sequentially optimal given the beliefs of t_i about the co-players. Intuitively, independence implies that t_i's system of beliefs about co-players satisfies the chain rule; hence, t_i does not change his mind about the relative probabilities of any two s'_{-i}, s''_{-i} unless observed behavior rules one of them out. This implies that t_i's preferences are dynamically consistent, which is the key condition for the equivalence between the OSD property and sequential optimality.

Conversely, if independence fails, then t_i may change (and—by introspection—expect to change) his mind about the behavior of co-players, and this may prevent the existence of a sequentially optimal plan. Suppose, for example, that t_a's marginal beliefs about Bob in the BoSOO game of Figure 3.1 are such that

$$\beta_a(t_a)(S_a \times \{c\} \times T_b | S_a \times S_b \times T_b) > 3/4,$$

$$\beta_a(t_a)(\{In.C, In.M\} \times \{c\} \times T_b | \{In.C, In.M\} \times S_b \times T_b) < 1/3.$$

Then the folding-back (OSD) plan is $Out.M$, because Ann predicts that she would choose M in the subgame and thus prefers to opt Out. But, at the root, she would like to commit to $In.C$ if she only could. Thus, she has no plan satisfying sequential optimality.

To sum up, our preferred interpretation of independence and optimal planning is the following. By strategic reasoning, player i forms a system of beliefs about the

co-players. Such beliefs—which satisfy the chain rule, hence form a CPS—give rise to a subjective decision tree. With this, player i devises a strategy by folding-back planning. The chain rule ensures that such strategy is sequentially optimal in the subjective decision tree. Furthermore, by construction the resulting beliefs about *everybody* satisfy independence between i and $-i$.

4.3 Consistency and rationality

Recall that a personal state of player i in a Γ-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 behavior at each history $h \in H$, that is, what other players have to predict in order to assess the likely consequences of their actions; focusing on types t_i that satisfy independence, $\sigma_{t_i,i}$—derived from $\beta_{i,i}(t_i)$—is the *subjective plan* of i. A consistent player would behave as planned at each non-terminal history; 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 Γ-based type structure $\mathcal{T}$ if (t_i satisfies independence and) 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 non-terminal histories that weakly follow h. For every Γ-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 (concerning i), because they are closed, hence Borel sets.

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^{fi} of Bob would keep believing that Ann is consistent also if he observed I_a, hence t_b^{fi} must change belief about the plan of Ann conditional on I_a; backward-induction type t_b^{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 history (I_a).

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 $h \in H$.[22] But this generalization would not change the substance of our results. Intuitively, it can be shown by dynamic programming arguments that, given beliefs about others, the value of

[22] 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.

any given history is the same for every plan (deterministic or not) satisfying optimal planning. This is the key condition we need for the essence of our results. However, considering non-deterministic plans makes the analysis more complex.[23]

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 that they yield a solution concept called "backwards rationalizability" (cf. Perea 2014, Penta 2015).[24]

Our notion of type structure allows us to represent subjective plans as beliefs about *own* behavior. As we explained, a key assumption about such beliefs, consistency, is that players correctly predict their own behavior, which is a prerequisite of rationality. Our analysis of strategic thinking focuses instead 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

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

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

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

denotes the event that i **fully believes** E_{-i}.[25] Note that these belief operators satisfy conjunction and monotonicity. Furthermore, if E_{-i} is closed then each $\mathrm{B}_{i,h}\left(E_{-i}\right)$ $(h \in H)$ and $\mathrm{B}_i\left(E_{-i}\right)$ are closed as well. We let $\mathrm{B}\left(\cdot\right)$ denote the **mutual full belief**

[23] For example, in our analysis all the relevant events are closed, hence compact; thus, we can apply the finite intersection property. In the more general case, we just have Borel measurability and the proof of existence results (non-emptiness) is less straightforward.

[24] 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.

[25] Battigalli and Siniscalchi (1999) define, for any fixed collection $\mathcal{F}$ of histories, the $\mathrm{B}_{i,\mathcal{F}}$ belief operator, where $\mathrm{B}_{i,\mathcal{F}}\left(E_{-i}\right)$ means that i would believe E_{-i} with probability one conditional on h for every $h \in \mathcal{F}$. The operators $\mathrm{B}_{i,h}$ and B_i used here are special cases.

operator, that is, $\mathrm{B}(E) := \prod_{i\in I} \mathrm{B}_i(E_{-i})$ for each Cartesian 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 self-map B, that is,

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

where $\mathrm{B}^0(E) := E$ by convention (i.e., B^0 is the identity map on the collection of Cartesian events). We say that an event $E = \prod_{i\in I} E_i$ is **transparent** at state (s,t) if $(s,t) \in E$ (i.e., E is the case) and there is common full belief in E at (s,t); thus, the set of states where E is transparent is

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

Our representation of BI reasoning is based on the following key 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 may "stubbornly neglect the past": he may observe deviations from the plans he ascribes to the co-players, and yet no evidence of such deviations makes him doubt that in the future they will follow their plans, as in the "trembling-hand" story by Selten (1975).

Note that, for each $h \in H$, $C_{-i}^{\succeq h}$ is a product of closed sets, hence it is closed, which implies that $\mathrm{B}_{i,h}\left(C_{-i}^{\succeq}\right)$ is closed as well. Therefore, BCC_i is closed. With this, define recursively the following epistemic events:

- $OP_i^1 := OP_i \cap BCC_i$,
- $OP_i^{m+1} := OP_i^m \cap \mathrm{B}_i\left(OP_{-i}^m\right)$, 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, for each $i \in I$, OP_i^1 is closed; 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 $\left(OP^m\right)_{m\in\mathbb{N}}$ is a well defined decreasing sequence of closed sets; thus, it makes sense to define $OP^\infty := \cap_{m\in\mathbb{N}} OP^m$.

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

$$OP^{m+1} = \bigcap_{k=0}^{m} \mathrm{B}^k (OP \cap BCC) = \left(\bigcap_{k=0}^{m} \mathrm{B}^k (OP) \right) \cap \left(\bigcap_{k=0}^{m} \mathrm{B}^k (BCC) \right)$$

and

$$OP^{\infty} = \bigcap_{m \in \mathbb{N}_0} \mathrm{B}^m (OP \cap BCC) = \left(\bigcap_{m \in \mathbb{N}_0} \mathrm{B}^m (OP) \right) \cap \left(\bigcap_{m \in \mathbb{N}_0} \mathrm{B}^m (BCC) \right).$$

In words, OP^{∞} is the event that optimal planning and belief in continuation consistency are transparent.

5.1 Backward induction

Consider a game Γ that can be solved by BI and let s^{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 full belief in both imply that players believe, conditional on each $h \in H$, that everybody will play according to s^{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 Γ 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 finite game.

Remark 6 *For every finite game* Γ *with observable actions there exists a* Γ*-based type structure* $\mathcal{T}$ *such that* $OP^{\infty} \neq \emptyset$ *and* $C \cap OP^{\infty} \neq \emptyset$*.*

Formally, this follows from Theorem 2 below and the observation that backwards rationalizability is a nonempty solution procedure. The result can be easily understood in the special case when Γ has a pure subgame perfect equilibrium $\bar{s}$. Consider the following type structure: The type set of each player is a singleton, that is, $T_i := \{\bar{t}_i\}$ for each $i \in I$. For each $s_i \in S_i$ and $h \in H$, let s_i^h denote the minimal modification of s_i allowing h.[26] With this, each belief map is such that

[26] 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 $h' \prec h$.

$\beta_{i,h}(\bar{t}_i)\left(\left\{\left(\bar{s}_j^h\right)_{j\in I}\right\}\times T_{-i}\right)=1$ 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, as every other player has a forced action (that is, to "wait").

Lemma 1 *Fix a finite PI game Γ without relevant ties and a Γ-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 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^{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

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

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,

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

and

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

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

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

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

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

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, $BCC_{\iota(h)} \cap \text{B}_{\iota(h)}(OP^{k}_{-\iota(h)})$ implies that $\iota(h)$ expects his co-players to take the BI actions at future histories, which must have height k or less. Formally, let $I(h) := \{i \in I : \exists h' \in H(h), i = \iota(h')\}$ denote the set of players who are active at some history of the subgame with root h. Note that, for every player i who is *not*

active at h ($i \neq \iota(h)$), $C_i^{\succeq h} = \cap_{h' \succ h} C_i^{\succeq h'}$. With this,

$$\begin{aligned}
& BCC_{\iota(h)} \cap \mathrm{B}_{\iota(h)}\left(OP^k_{-\iota(h)}\right) \\
\subseteq\ & \mathrm{B}_{\iota(h),h}\left(\left(\prod_{i \in I(h)\setminus\{\iota(h)\}} C_i^{\succeq h} \cap OP_i^k\right) \times \left(\prod_{j \in I\setminus(I(h)\cup\{\iota(h)\})} S_j \times T_j\right)\right) \\
\subseteq\ & \mathrm{B}_{\iota(h),h}\left(\left(\prod_{i \in I(h)\setminus\{\iota(h)\}} \left([s_i^{\mathrm{bi}}]^{\succeq h} \times T_i\right)\right) \times \left(\prod_{j \in I\setminus(I(h)\cup\{\iota(h)\})} S_j \times T_j\right)\right) \\
\subseteq\ & S_{\iota(h)} \times \left\{t_{\iota(h)} : \forall h' \in H(h), \iota(h') \neq \iota(h) \Rightarrow \sigma_{t_{\iota(h)}}(s^{\mathrm{bi}}(h')|h') = 1\right\},
\end{aligned}$$

where the second inclusion follows from the inductive hypothesis (besides monotonicity of $\mathrm{B}_{\iota(h),h}$). Thus, 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 behavior. This gives the epistemic event

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

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 Γ without relevant ties and a Γ-based type structure $\mathcal{T}$. Then $OP_i^{L(\varnothing)} \subseteq BIP_i$ and $OP_i^{L(\varnothing)} \cap C_i \subseteq \{s_i^{\mathrm{bi}}\} \times T_i$ for every $i \in I$.*

Proof. Let H_i^1 denote the set of histories where player i is active for the first time. Then $\{s_i^{\text{bi}}\} = \cap_{h \in H_i^1} [s_i^{\text{bi}}]^{\succeq h}$ and $\cap_{h \in H_i^1} S_i \times T_i^{\text{bi},1}(h) \subseteq BIP_i$. 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

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

■

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

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

Proof. By Remark 5,

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

By Corollary 1,

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

■

The previous results allow us to derive the implications of correct common full belief that (i) players plan optimally given their beliefs about others (OP) and (ii) believe in continuation consistency (BCC): in PI games without relevant ties, these assumptions imply that players plan according to backward induction (BIP) and, furthermore, there is common belief of this at every history. Thus, in particular, every player always believes that each co-player has the BI plan and is going to execute it in the future, regardless of past deviations from such plan. Hence, players neglect perceived deviations.

Theorem 1 *Fix a finite PI game Γ without relevant ties and a Γ-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 "backwards rationalizability" solution concept (cf. Penta 2015, Perea 2014), which we introduce next.

Let $\mathcal{Q}$ be the collection of all the Cartesian subsets $Q = \prod_{i\in I} Q_i$, 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 $Q \in \mathcal{Q}$,

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

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

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

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

where $\mathbb{E}_{\mu_i}\left[U_i(r_i, \cdot) | h\right]$ denotes the expected payoff of r_i given $\mu_i(\cdot | S_{-i}(h))$.[27]

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 external states in $\hat{S}^\infty$ are called **backwards rationalizable**.*

Note, part 2 of the recursive step requires that, at each $h \in H$, the CPS about $-i$ justifying s_i assign probability 1 to the continuations of behaviors from $\hat{S}_{-i}^{n-1}$ even if h is inconsistent with $\hat{S}_{-i}^{n-1}$, that is, even if $\hat{S}_{-i}^{n-1} \cap S_{-i}(h) = \emptyset$. One can show by standard arguments that backwards rationalizability is a nonempty solution concept:

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

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

[27] Also, recall that s_i^h denotes the minimal modification of s_i allowing h.

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 $I_b.i_b$, specifically $\mu_a\left(I_b.i_b \middle| \{I_b.I_b, I_b.o_b\}\right) \leq 1/3$; by the chain rule, such conditional belief implies that Ann assigns a low probability to $I_b.i_b$ also at the root, that is, $\mu_a\left(I_b.i_b | S_b\right) \leq 1/3$; thus, the optimal action for Ann at the root 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).

With this, one can verify that at the second step

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

In particular, both $I_a.i_a$ and $O_a.i_a$ are ruled out because action i_a is not justifiable at (I_a, I_b) as Ann believes that Bob would choose o_b at (I_a, I_b, i_a). The algorithm stops at the third step:

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

Intuitively, Bob is initially certain that Ann's plan is $O_a.o_a$ and she will implement it. Upon observing I_a, Bob would interpret this as a deviation from Ann's plan (instead of evidence that Ann's plan is different) and yet would believe that further deviations will not occur. Formally, upon observing I_a Bob would believe that Ann's behavior is described by the only element of $\chi_a^{(I_a)}\left(\hat{S}_a^2\right) = \{I_a.o_a\}$.

We can show (see Appendix B.1) that 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 class of finite multistage games with observable actions. Specifically, the "backwards procedure" coincides with the BI algorithm in perfect information games without relevant ties.[28] This implies that in every PI game with no relevant ties $\hat{S}^\infty = \left\{s^{\text{bi}}\right\}$ (cf. Perea 2014). Instead, in games with simultaneous actions at some histories, backwards rationalizability may be very permissive. Consider the BoSOO game depicted in Figure 3.1. At the first step, we have

$$\hat{S}^1 = \{Out.C, Out.M, In.C\} \times \{c, m\}.$$

Indeed, the algorithm just rules out $In.M$ for Ann, which is strictly dominated, whereas both c and m can be justified as best replies of Bob to some belief about Ann in the subgame. The algorithm stops at the first step: $\hat{S}^\infty = \hat{S}^1$. The reason is that Bob may believe that Ann's plan is $Out.M$. Upon observing In, he thinks Ann made a mistake in implementation, but is still going to continue as planned and choose M in the subgame (cf. Subsection 3.1).

[28] The two procedures yield the same output $\hat{S}^\infty$, but the backwards procedure may be "slower." For example, in the game of Figure 3.2 BI does not delete any s_a in the first step.

The following theorem shows that backwards rationalizability characterizes the behavioral implications of consistency and transparency of optimal planning and of belief in continuation consistency. Note that, formally, the behavioral implications of any epistemic event $E \subseteq S \times T$ are represented by the image of E under the canonical projection from $S \times T$ onto S, that is,

$$\text{proj}_S E := \{s \in S : \exists t \in T, (s,t) \in E\}.$$

Also, recall that a type structure is complete if the belief maps are onto (surjective), as in the canonical type structure where the belief maps are homeomorphisms.

Theorem 2 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Then,*
(i) for every $n \in \mathbb{N}$, $\text{proj}_S (OP^n \cap C) \subseteq \hat{S}^n$;
(ii) $\text{proj}_S (OP^\infty \cap C) \subseteq \hat{S}^\infty$.
Furthermore, if structure $\mathcal{T}$ is complete *these weak inclusions hold as equalities.*

6 Forward-induction reasoning: rationalization of past moves

In Section 3, we have informally claimed that the basic forward-induction reasoning step can be modelled by the epistemic assumption of strong belief in rationality, that is, strong belief in consistency and optimal planning. Furthermore, the whole forward-induction reasoning process is represented by "rationality and common strong belief in rationality" and its behavioral implications are characterized by the "strong rationalizability" solution concept. Here we make these informal claims precise.

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

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

denote the event that i **strongly believes** E_{-i}. Note that, unlike the conditional belief operators $\text{B}_{i,h}$ $(h \in H)$ and the full belief operator B_i, the strong belief operator SB_i is *not monotone*. The reason is the following: if $E_{-i} \subseteq F_{-i}$, then the collection of histories consistent with E_{-i} is included in the collection of histories consistent with F_{-i}; if the inclusion is strict, then $\text{SB}_i (F_{-i})$ requires that i believe F_{-i} conditional

on more histories than $\mathrm{SB}_i(E_{-i})$ requires i to believe E_{-i}.[29] With this, **rationality and common strong belief in rationality** (RCSBR) can be recursively defined as follows.[30] 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 $n \in \mathbb{N}$, let

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

where $R_{-i}^n := \prod_{j \neq i} R_j^n$. A standard inductive argument shows that each set R_i^n is closed in $S_i \times T_i$.[31] 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 $S^\infty := \cap_{n \in \mathbb{N}_0} S^n$. The external states in S^∞ are called **strongly rationalizable**.*

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

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

[29] This explains the difference between the structure of the theorems of this section, which are stated only for complete type structures, and those of Section 5 (cf. Battigalli and Siniscalchi 2002, and Battigalli and Friedenberg 2012).

[30] In Section 7 we compare our analysis of RCSBR with that of Battigalli and Siniscalchi (2002).

[31] Recall that if $E_{-i} \subseteq S_{-i} \times T_{-i}$ is closed, so is $\mathrm{B}_{i,h}(E_{-i})$. Thus, $\mathrm{SB}_i(E_{-i})$ is an 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.

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 (1997). The following result states that strong rationalizability characterizes the behavioral implications of RCSBR.

Theorem 3 *Fix a finite game* Γ *and a* Γ*-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) $\text{proj}_S R^\infty = S^\infty$.

Let us consider the first two steps of the strong rationalizability solution procedure and the epistemic assumptions that justify such steps. To simplify the exposition, we focus on two-person games. According to Theorem 3, step 1 says that behavior (personal external state) s_i can be justified as a sequential best reply to at least one first-order CPS if and only if there is a type t_i such that $(s_i, t_i) \in R_i := C_i \cap OP_i$, that is, the plan of type t_i agrees with behavior s_i and t_i plans optimally. Step 2 assumes that, on top of being rational, player i strongly believes in the rationality of his co-player $j = -i$. Now, consider a history h that contradicts j's rationality, that is, $(S_j(h) \times T_j) \cap R_j = \emptyset$. Then, despite his strong belief in j's rationality, player i cannot hold on to the assumption that j is rational; yet, he can still hold on to either C_j (j is implementing his plan), or OP_j (j plans optimally). In other words, player i can give "epistemic priority" to either C_j or OP_j, but the assumption of strong belief in rationality is silent on such epistemic priority. Next we present a modified theory of forward-induction reasoning whereby it is transparent that players always give epistemic priority to consistency between co-players' plans and behavior. It turns out that its behavioral implications are again characterized by strong rationalizability; hence, they are the same as those of RCSBR.

Let $C^* := \cap_{m \in \mathbb{N}_0} \text{B}^m(C)$ denote the set of states where there is **transparency of consistency**, that is, consistency holds and there is common full belief in it. One can show by induction that each $\text{B}^m(C)$ ($m \in \mathbb{N}_0$) is a Cartesian product of closed sets, which implies that the same holds for the intersection, that is, $C^* = \prod_{i \in I} C_i^*$, where each $C_i^* := \text{proj}_{S_i \times T_i} C^*$ is closed. With this, for each player $i \in I$, define recursively the following events:

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

and, for each $n \in \mathbb{N}$,

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

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

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

represents optimal planning and transparency of consistency, and common strong belief thereof.

Theorem 4 *Fix a finite game* Γ *and a* Γ*-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) $\text{proj}_S R^{*,\infty} = S^\infty$*.*

As anticipated in the Introduction, we interpret this theorem as an explicit statement of epistemic assumptions that are implicitly maintained in the analysis of forward-induction reasoning by means of standard type structures, whereby players' beliefs concern only the co-players. In such analysis, any element of S_i simultaneously represents a plan of i (which must be a sequential best reply to his first-order beliefs, if i is rational) and also an external state of i, that is, an objective description of how i would behave at any history where he is active. Because of this necessary coincidence, whenever a (non-terminal) history h occurs, the co-players infer that i must be implementing a strategy in $S_i(h)$, which is the premise to give bite to the assumption of strong belief in rationality. Theorem 4 replaces the implicit assumption of necessary coincidence of plan and behavior with the explicit assumption that such coincidence (i.e., consistency) is transparent.

7 Discussion

In this section we consider alternative solution concepts and epistemic assumptions, we discuss extensions 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 the notion of strong rationalizability defined here is behaviorally equivalent to the "extensive-form rationalizability" concept put forward by Pearce (1984) and clarified by Battigalli (1997).[32] Specifically, let

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

[32] In Section 3.2, we explained why we avoid the "extensive-form rationalizability" terminology. Note also that complete equivalence holds for two-person games (without chance moves). 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).

denote the set of non-terminal histories allowed 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). For example, the plan of Ann to go out at the root of the PI game of Figure 3.2 corresponds to the equivalence class $\{O_a.o_a, O_a.i_a\}$. In two-person games without chance moves, Pearce's solution concept is like the strong rationalizability procedure $(S^n)_{n\in\mathbb{N}}$ of Definition 7 with the best-reply correspondence $\rho_i(\cdot)$ replaced by the following weaker version: for every first-order belief (CPS) $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$,

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

Correspondence $\bar{\rho}_i$ captures a notion of **forward planning** and does not distinguish between strategies in the same equivalence class. Consider, for example, Ann in the PI game of Figure 3.2. She can plan to go out (O_a), which requires no further planned choice, or to go in (I_a), which requires further contingent planning in case Bob also goes in (I_b); suppose the planned contingent choice is to continue again (i_a) to have the possibility of getting 3 utils. The ex ante value of the first plan (O_a) is 2 and the ex ante value of the second plan $(I_a.i_a)$ is $3\mu_a(I_b.i_b)$; since the ex ante value of plan $I_a.o_a$ is $\mu_a(\{I_b.i_b, I_b.o_b\}) < 2$, Ann wants to implement plan $I_a.i_a$ if $\mu_a(I_b.i_b) > 2/3$, and O_a if $\mu_a(I_b.i_b) < 2/3$.

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 dynamic games (including Battigalli and Siniscalchi 2002) refers to this solution concept. Yet, it is well known that, for every player i, belief μ_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, an inductive 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 $s_i \in S_i^n$.[33] To illustrate, in the BoSOO game of Figure 3.1 $S^\infty = \bar{S}^\infty = \{In.C\} \times \{c\}$, whereas in the PI game of Figure 3.2 $S^\infty = \{O_a.o_a\} \times \{I_b.o_b\}$ and $\bar{S}^\infty = \{O_a.o_a, O_a.i_a\} \times \{I_b.o_b\}$. Our notion of strong rationalizability rules out $O_a.i_a$ because, if Ann believes that the behavior of Bob is $I_b.o_b$, then her folding-back optimal plan is $O_a.o_a$. With forward planning, we just get the equivalence class $\{O_a.o_a, O_a.i_a\}$, that is, the plan of going out at the root.

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

Path predictions of forward and backward induction Let z^{bi} denote the backward-induction path of any finite perfect-information (PI) game Γ without relevant ties. In the universal type structure, we have

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

that is, forward-induction reasoning—like backward-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 behavioral equivalence between $\left(S_i^n\right)_{i\in I, n\in\mathbb{N}}$ and $\left(\bar{S}_i^n\right)_{i\in I, n\in\mathbb{N}}$, and Theorem 4 in Battigalli (1997).[34] This shows that, while backward- and forward-induction reasoning may yield very different implications about plans (as in the game of Figure 3.2), in PI games without relevant ties they yield the same path. Such partial consistency between forward and backward induction can be extended to games with imperfect information: building on work by Chen and Micali (2012), one can show that the path implications of RCSBR (in the universal structure) refine the path implications of C (consistency) and OP^{∞} (common full belief in $OP \cap BCC$): $\zeta\left(\text{proj}_S R^{\infty}\right) \subseteq \zeta\left(\text{proj}_S\left(C \cap OP^{\infty}\right)\right)$. The BoSOO example shows that in non-PI games the inclusion may be strict.

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.[35]

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

[34] 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.

[35] Fix a strategy s_i. There is no first-order CPS μ_i such that $s_i \in \bar{\rho}_i\left(\mu_i\right)$ if and only if s_i is strictly dominated conditional on reaching some $h \in H_i\left(s_i\right)$. 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).

rationalizability (cf. Battigalli and Siniscalchi 2007). A similar result holds for optimal planning, transparency of consistency, and common initial belief in optimal planning.

Furthermore, in Appendix B.2 we provide an alternative epistemic justification of initial rationalizability which is closer to the one provided for backwards rationalizability. Say that player i **initially believes in** the **consistency** of the other players if i believes $C_{-i} := \prod_{j \neq i} C_j$ at the beginning of the game (of course, the assumption of initial belief in consistency is weaker than BCC). We show in Appendix B.2 that initial rationalizability characterizes the behavioral implications of consistency and transparency of optimal planning and of initial belief in consistency.

Introspective beliefs In this paper, we did not explicitly represent the beliefs of the players about their own beliefs, because we implicitly assumed that (a) players are fully introspective, so that they know their own way to think, and (b) that this is commonly believed at every history. To better understand this implicit assumption, let us introduce a terminological and conceptual distinction. For any history h, let $[h]$ denote the event that h occurs. Consider two related histories $h_0 \prec h_1$ (so that $[h_1] \subseteq [h_0]$) and the corresponding conditional beliefs of a player $\mu(\cdot|[h_0])$ and $\mu(\cdot|[h_1])$. We say that the player **updates** his belief from h_0 to h_1 if $\mu([h_1]|[h_0]) > 0$ and

$$\mu(E|[h_1]) = \frac{\mu(E|[h_0])}{\mu([h_1]|[h_0])}$$

for every event $E \subseteq [h_1]$; otherwise, if $\mu([h_1]|[h_0]) = 0$, we say that the player **revises** his belief from h_0 to h_1. Note that the chain rule

$$\mu(E|[h_1])\,\mu([h_1]|[h_0]) = \mu(E|[h_0])$$

(for every event $E \subseteq [h_1]$) holds in both cases, although trivially in the latter. We (like the related papers we cite) implicitly assume that players know both how they update—i.e., according to the rule of conditional probability—and how they revise. But we conjecture that the behavioral predictions of our theory can be preserved by dropping the assumption that they necessarily know how they revise. Thus, players may be uncertain about their own type. In such modified theory, we envision a rational player of type t_i who plans—w.l.o.g., deterministically—starting with his first-order belief $\mu^1_{t_i,\varnothing} \in \Delta(S_{-i})$ over the behaviors s_{-i} of others, and who performs folding-back planning from the longest histories h such that $\mu^1_{t_i,\varnothing}(S_{-i}(h)) > 0$. This yields a **partial plan** (or partial strategy) with domain $H_i\left(\mu^1_{t_i,\varnothing}\right)$, the tree of histories he deems possible. At any h such that $\mu^1_{t_i,\varnothing}(S_{-i}(h)) = 0$, but the immediate

predecessor is instead deemed possible, a similar folding-back planning would start all over again with the revised belief $\mu^1_{t_i,h}$. Thus, the *type of a* (rational) *player specifies a full plan.* Since rationality also requires consistency, the contingent behavior of such player satisfies the same sequential optimality property adopted here, which is what matters when we assume that other players believe in his rationality. But the plan (strategy) in the mind of the player at any history h may be partial, because he may be unable to anticipate how he would revise his beliefs upon observing unexpected moves. In such a modified theory the external state of a player and the deterministic plan he has in mind, say, at the beginning of the game, would be mathematical objects of different kinds, that is, elements of $\prod_{h\in H} A_i(h)$ and $\prod_{h\in H_i(\mu^1_{t_i,\varnothing})} A_i(h)$ respectively.

This approach raises the following issue. Consider two distinct histories $h' = (h,(a'_i,a_{-i}))$ and $h'' = (h,(a''_i,a_{-i}))$. Since h' and h'' reveal the same behavior about i's co-players, $S_{-i}(h') = S_{-i}(h'')$. Suppose that $\mu^1_{t_i,\varnothing}(S_{-i}(h)) > 0$ and $\mu^1_{t_i,\varnothing}(S_{-i}(h')) = 0$, which implies $\mu^1_{t_i,\varnothing}(S_{-i}(h'')) = 0$. Under the implicit assumption that i is fully introspective, we took for granted that the revised belief of t_i given these two histories h' and h'' would be the same, a natural form of independence following from the fact that the first-order CPS of t_i is defined for the collection of conditioning events $\mathcal{S}_{-i} = \left\{F_{-i} \in 2^{S_{-i}} : \exists h \in H, S_{-i}(h) = F_{-i}\right\}$. But if player i is only partially introspective in the aforementioned sense, this requirement is less compelling. Thus, we may want to allow for the possibility that $\mu^1_{t_i,h'} \neq \mu^1_{t_i,h''}$ in the previous case, i.e., that belief revision depends on i's behavior. This yields a weaker form of chain rule, whereby we only compare conditional beliefs at histories related by precedence. For example, in the previous case, $h \prec h'$ and $h \prec h''$, which implies $S_{-i}(h') \subseteq S_{-i}(h)$ and $S_{-i}(h'') \subseteq S_{-i}(h)$; thus, for all $E_{-i} \subseteq S_{-i}(h') = S_{-i}(h'')$, we must have

$$\begin{aligned}
\mu^1_{t_i,h}(E_{-i}) &= \mu^1_{t_i,h'}(E_{-i})\,\mu^1_{t_i,h}(S_{-i}(h')), \\
\mu^1_{t_i,h}(E_{-i}) &= \mu^1_{t_i,h''}(E_{-i})\,\mu^1_{t_i,h}(S_{-i}(h'')),
\end{aligned}$$

but this does not imply $\mu^1_{t_i,h'}(E_{-i}) = \mu^1_{t_i,h''}(E_{-i})$ because we may have $\mu^1_{t_i,h'}(E_{-i}) = \mu^1_{t_i,h''}(E_{-i}) = 0$.

Let us call **forward CPS** an array of first-order conditional beliefs $\left(\mu_{i,h}\right)_{h\in H} \in \prod_{h\in H} \Delta(S_{-i}(h))$ such that the chain rule applies for all $h,h' \in H$ with $h \prec h'$, and similarly for higher-order beliefs. We can rephrase what we just said as follows: *under the assumption of partial introspection, belief systems should be forward CPSs, but not necessarily CPSs.* Let us note that much of the literature on rationalizability in sequential games (somewhat implicitly) adopts this weaker form of belief system; see,

for example, Pearce (1984), Battigalli (1997), and Perea (2014). We can show that assuming forward CPSs rather than CPSs does not affect behavioral implications in the class of games analyzed in the main text, that is, multistage games with observable actions. Yet, so far we have not been able to extend this equivalence result to backwards rationalizability in games with imperfectly observable actions and perfect recall. This extension is discussed below.

General games with perfect recall Although the assumption of observable actions simplifies our analysis, in Appendix B.3 we extend our results about backward-induction reasoning by just assuming perfect recall.[36] The key is to be able to state and keep the assumption of belief in co-players' consistency starting from any particular history/node even though information sets are not singletons. We achieve this noting that only the nodes in an information set with strictly positive conditional probability matter for the analysis, and we can determine a well defined belief conditional on each one of those nodes. With this, BCC is defined by requiring that each player believe in the co-players' consistency starting from every node x he deems possible conditional on the information set containing x. This approach allows us to consider all games with perfect recall, not only those with ordered information sets, as in Perea (2014). Also, our approach makes the details of our backward rationalizability solution concept different from the one analyzed by Perea. So, we have a partially new solution concept. It is worth noting that we can prove the characterization theorem for this more general class of games by assuming that belief systems are *forward CPSs*, as in Perea (2014). Whether the analysis can be reformulated in terms of CPSs is an open question.

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), as well as belief-dependent preferences, which in turn may be dynamically inconsistent when preferences over outcomes depend on one's own plan (Battigalli and Dufwenberg 2009).[37] Given beliefs about other players (or nature), sophisticated

[36]The analysis of forward-induction reasoning in games with perfect recall already exists in the literature. We additionally have to account for the decoupling of plans and behavior, which does not create conceptual or technical difficulties.

[37]In the context of dynamic games, see—for example—Battigalli et al. (2019b) on ambiguity aversion, and Battigalli et al. (2019a) on the role of emotions and belief-dependent preferences. Note that own-plan dependence of preferences over outcomes may require non-deterministic plans

planning is an intra-personal equilibrium condition expressed by the OSD property, which in this case is not equivalent to sequential optimality.[38] 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").

Further comments on the 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).[39] Here we outline the differences between our epistemic conditions for BI (Theorem 1 and Theorem 2) and those that appear to be conceptually closest, namely Asheim (2002), Asheim and Perea (2005), Baltag et al. (2009), Bach and Heilmann (2011) and Perea (2014). Finally, we briefly comment on other papers where players' plans are modeled as beliefs.

In Asheim (2002) and Asheim and Perea (2005) type structures do not include players' beliefs about their own behavior, and beliefs are represented by lexicographic (conditional) probability systems, rather than CPSs. They obtain sufficient epistemic conditions for BI strategies based on an approach somewhat similar to ours. In particular, they assume common "certain belief" (analogous to full belief) of

to satisfy the OSD property given beliefs about others.

[38] 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.

[39] 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.

events that—absent additional "consistency" conditions—have no implication about behavior.

Baltag et al. (2009) use a dynamic epistemic-logic formalism related to, but different from, the formalism of this paper. Their approach is based on the framework of the so called "plausibility models," 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 each 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;[40] 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.

Bach and Heilmann (2011) consider generic PI games in which each player comprises a reasoning agent and a set of game agents ("selves"), each of them corresponding to a unique decision node. In their framework, the reasoning agent plans before the game, while each game agent is responsible for the actual choice at his node. A game agent is said to be high-connected if he acts in compliance to the plan of the reasoning agent. Bach and Heilmann define a condition called "forward belief in future-high-connectedness" and use it to obtain sufficient conditions for backward induction. Their approach is somewhat similar in spirit to ours, but their epistemic framework is very different, which makes a precise comparison difficult. In particular, we just rely on the resources of epistemic structures *à la* Battigalli and Siniscalchi (1999), which have a constructive foundation by means of hierarchies of conditional beliefs, whereas Bach and Heilmann use a notion of type structure that features beliefs about others and also posits a map from types to initial plans.

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 analyses based on

[40] 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.

standard type structures, it is implicitly assumed by Perea (2014) that the personal external states s_i ($i \in I$) simultaneously represent players' behavior and their plans. In particular, the personal external states are defined as classes of behaviorally equivalent strategies, hence, maximization is required only at histories h consistent with the given plan s_i. Perea's version of backwards rationalizability is based on best-reply correspondence $\bar{\rho}_i(\cdot)$ rather than $\rho_i(\cdot)$ (cf. our previous discussion of strong rationalizability). 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^{bi}: for 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.

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 setting the external personal state of a player is (technically) also a strategy, in theirs the only mathematical objects corresponding to (behavior) strategies are players' systems of conditional beliefs about their own actions. This has the advantage of preventing confusion between behavior (external state) and strategies (in the mind of players). But there is a price to pay: the language is not rich enough to express beliefs about behavioral subjunctive conditionals, such as "Ann believes that if she chose In Bob would choose C" (see the BoSOO game of Figure 3.1). Battigalli et al. (2013) and related works replace such beliefs with conditional beliefs about behavior, which is analogous to the transformation from mixed to behavior strategies of Kuhn (1953). On balance, we find both approaches worth pursuing. We conjecture that we could reformulate our analysis having Z as the set of external states. Battigalli et al. (2020) take steps in this direction while also allowing for belief-dependent preferences. Another difference with Battigalli et al. (2013) is that they focus only on forward induction and RCSBR in PI games, proving a result analogous to Theorem 3; by contrast, we consider more general games and analyze backward- as well as forward-induction reasoning.

Appendix A: Proofs omitted from the main text

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*

$$\mathrm{proj}_X \bigcap_{m=1}^{\infty} E^m = \bigcap_{m=1}^{\infty} \mathrm{proj}_X E^m.$$

Proof. The inclusion $\subseteq$ is obvious. For the other direction, let $x \in \cap_{m=1}^{\infty} \mathrm{proj}_X E^m$. 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, hence compact. Specifically, non-emptiness of each E_x^m follows from the fact that $x \in \cap_{m=1}^{\infty} \mathrm{proj}_X E^m$. Moreover, $(E_x^m)_{m=1}^{\infty}$ is a decreasing sequence of sets; therefore, by the finite intersection property of compact sets, $\cap_{m=1}^{\infty} E_x^m \neq \emptyset$. ■

Furthermore, the following notation will be used throughout the proofs of Theorems 2 and 4. For any $x \in X$, we let δ_x denote the Dirac measure supported by x. With this, given $s_i \in S_i$, we define

$$\delta_{s_i}^* := \left(\delta_{s_i^h}\right)_{h \in H} \in \left(\Delta\left(S_i\right)\right)^H$$

as the array of Dirac measures supported by s_i^h (the minimal modification of s_i allowing h); that is, $\delta_{s_i}^*$ is such that $\delta_{s_i}^*\left(s_i^h | S_i\left(h\right)\right) = 1$ for every $h \in H$. First, it can be verified that such array is $\mathcal{S}_i$-measurable, because $S_i\left(h'\right) = S_i\left(h''\right)$ implies $s_i^{h'} = s_i^{h''}$. This can be verified for pairs of histories $h' = (h, a')$ and $h'' = (h, a'')$ with $a_i' = a_i''$, taking into account that, for the other cases in which $h' \neq h''$ and $S_i\left(h'\right) = S_i\left(h''\right)$, player i has a forced move ("wait") at nodes following the "bifurcation" (longest common prefix) and preceding either h' or h''. Thus, it makes sense to write $\delta_{s_i}^* \in \left(\Delta\left(S_i\right)\right)^{\mathcal{S}_i}$. Also, it can be verified that $\delta_{s_i}^*$ satisfies the chain rule for pairs $h, h' \in H$ such that $h \prec h'$ (note, $h \prec h'$ implies $S_i\left(h\right) \supseteq S_i\left(h'\right)$). With this, it can be shown that $\delta_{s_i}^*$ is a CPS: $\delta_{s_i}^* \in \Delta^{\mathcal{S}_i}\left(S_i\right)$. We omit the details.

Backward-induction reasoning

For the proof of Theorem 2, we find it convenient to introduce further notation and preliminary results.

Fix a finite game Γ. Given $\bar{h} \in H$ and $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$, let

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

The proof of the following remark is immediate by inspection of the definition.

Remark 9 *Fix $\bar{h} \in H$ and $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$.*

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

(ii) If $s_i \in \rho_i^{\succeq \bar{h}}(\mu_i)$, 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 $h \in H\left(\bar{h}\right)$.

Lemma 3 *For every $i \in I$, $h \in H$ and $n \in \mathbb{N}$,*

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

Proof. Let $s_i \in \chi_i^h(\hat{S}_i^n)$. 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 $\mu_i(\chi_{-i}^{h'}(\hat{S}_{-i}^{n-1}) | S_{-i}(h')) = 1$ for all $h' \in H$. Remark 9.(i) entails 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, pick any $s_i \in S_i(h)$ such that $s_i \in \rho_i^{\succeq h}(\mu_i)$ for some $\mu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ satisfying $\mu_i(\chi_{-i}^{h'}(\hat{S}_{-i}^{n-1}) | S_{-i}(h')) = 1$ for all $h' \in H$. By Remark 9.(ii), 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 $h' \in H(h)$. By Definition 6, $\bar{s}_i \in \hat{S}_i^n$. Hence $s_i \in \chi_i^h(\hat{S}_i^n)$. ■

Lemma 4 *Fix $i \in I$, $\bar{h} \in H$ and a nonempty set $Q_i \subseteq S_i$. Then:*

(i) for all $h \in H(\bar{h})$,

$$\chi_i^{\bar{h}}(Q_i) \cap S_i(h) \subseteq \chi_i^h(Q_i);$$

(ii) for all $h', h'' \in H(\bar{h})$ such that $h' = \left(\bar{h}, a'\right)$, $h'' = \left(\bar{h}, a''\right)$ and $a_i' = a_i''$,

$$\chi_i^{h'}(Q_i) \cap \chi_i^{h''}(Q_i) \neq \emptyset.$$

Proof. Part (i): The conclusion is immediate if $\chi_i^{\bar{h}}(Q_i) \cap S_i(h)$ is empty. So assume that this set is nonempty. Pick any $s_i \in \chi_i^{\bar{h}}(Q_i) \cap S_i(h)$. Then, by definition of $\chi_i^{\bar{h}}(Q_i)$, there exists $\bar{s}_i \in Q_i$ such that $s_i(h') = \bar{s}_i(h')$ for all $h' \in H\left(\bar{h}\right)$. Since

$h \in H\left(\bar{h}\right)$, we have $s_i(h') = \bar{s}_i(h')$ for all $h' \in H(h)$. Since $s_i \in S_i(h)$, it follows from the definition of $\chi_i^h(Q_i)$ that $s_i \in \chi_i^h(Q_i)$.

Part (ii): Note that $S_i(h') = S_i(h'')$ because the two histories reveal the same behavior of i. Hence,

$$\chi_i^{\bar{h}}(Q_i) \cap S_i(h') = \chi_i^{\bar{h}}(Q_i) \cap S_i(h'') \subseteq \chi_i^{h'}(Q_i) \cap \chi_i^{h''}(Q_i),$$

where the inclusion follows from part (i). Pick any $s_i \in \chi_i^{\bar{h}}(Q_i)$. Then there exists $\bar{s}_i \in Q_i$ such that $s_i(h) = \bar{s}_i(h)$ for all $h \in H\left(\bar{h}\right)$. Since $h', h'' \in H(\bar{h})$, we have $s_i(h) = \bar{s}_i(h)$ for all $h \in H(h')$ and $s_i(h) = \bar{s}_i(h)$ for all $h \in H(h'')$. Next consider $\hat{s}_i \in S_i(h') = S_i(h'')$ defined as follows: $\hat{s}_i(h) = s_i(h)$ for each $h \in H \backslash \left\{\bar{h}\right\}$, and $\hat{s}_i\left(\bar{h}\right) = a_i' = a_i''$. So we have $\hat{s}_i(h) = \bar{s}_i(h)$ for all $h \in H(h')$ and for all $h \in H(h'')$, hence $\hat{s}_i \in \chi_i^{h'}(Q_i) \cap \chi_i^{h''}(Q_i)$. ■

The proof of Theorem 2 relies on Lemma 5 and Lemma 6 below. To ease the statements and proofs, let $OP_i^0 := S_i \times T_i$ for each $i \in I$. The sets OP^0 and OP_{-i}^0 are defined in the obvious way: $OP^0 := \prod_{i \in I} OP_i^0$ and $OP_{-i}^0 := \prod_{j \neq i} OP_j^0$.

Lemma 5 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Then, for all $n \in \mathbb{N}_0$ and $h \in H$,*

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

Proof. We first prove the following auxiliary 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 $\chi_i^h(\mathrm{proj}_{S_i}(OP_i^n \cap C_i)) \subseteq S_i(h)$ 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. Pick any $s_i \in \chi_i^h(\mathrm{proj}_{S_i}(OP_i^n \cap C_i))$. 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 $s_i \in \mathrm{proj}_{S_i}(OP_i^n \cap C_i^{\succeq h}) \cap S_i(h)$, as required. By definition of $\chi_i^h(\cdot)$, there exists $\bar{s}_i \in \mathrm{proj}_{S_i}(OP_i^n \cap C_i)$ such that $s_i(h') = \bar{s}_i(h')$ for every $h' \in H(h)$. 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 $\nu_i := \mathrm{marg}_{S_{-i}} \beta_i(t_i)$. Remark 9.(i) implies that $\bar{s}_i \in \rho_i^{\succeq h}(\nu_i)$, and since $\bar{s}_i(h') = s_i(h')$ for every $h' \in H(h)$, we obtain $s_i \in \rho_i^{\succeq h}(\nu_i)$. Therefore $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq h}$. □

We now prove the following claim:

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

With this, the result follows from Claim 1. The proof is by induction on $n \in \mathbb{N}_0$.

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

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

so the result follows immediately.

Inductive step. Assume that the result is true for $n \geq 0$. We show that it is also true for $n+1$.

Fix $i \in I$ and $\bar{h} \in H$ arbitrarily. Pick any $s_i \in \operatorname{proj}_{S_i}(OP_i^{n+1} \cap C_i^{\succeq \bar{h}}) \cap S_i\left(\bar{h}\right)$, so that $(s_i, t_i) \in OP_i^{n+1} \cap C_i^{\succeq \bar{h}}$ for some $t_i \in T_i$. Since $OP_i^{n+1} \subseteq OP_i^n$, it follows that $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq \bar{h}}$, and so, by the inductive hypothesis, $s_i \in \chi_i^{\bar{h}}(\hat{S}_i^n)$. By Remark 9.(i), $s_i \in \rho_i^{\succeq \bar{h}}(\nu_i)$, where $\nu_i := \operatorname{marg}_{S_{-i}}\beta_i(t_i)$. So, in order to show that $s_i \in \chi_i^{\bar{h}}(\hat{S}_i^{n+1})$, it is enough to show (by Lemma 3) that $\nu_i(\chi_{-i}^h(\hat{S}_{-i}^n)|S_{-i}(h)) = 1$ for every $h \in H$.

To this end, first note that $(s_i, t_i) \in OP_i^{n+1}$ implies $(s_i, t_i) \in \mathrm{B}_i\left(OP_{-i}^n\right) := \cap_{h \in H}\mathrm{B}_{i,h}\left(OP_{-i}^n\right)$. Note also that $(s_i, t_i) \in BCC_i := \cap_{h \in H}\mathrm{B}_{i,h}(C_{-i}^{\succeq h})$; hence, by the conjunction property of the operator $\mathrm{B}_{i,h}(\cdot)$, it follows that, for each $h \in H$, $(s_i, t_i) \in \mathrm{B}_{i,h}\left(OP_{-i}^n \cap C_{-i}^{\succeq h}\right) = \mathrm{B}_{i,h'}\left(\prod_{j \neq i}\left(OP_j^n \cap C_j^{\succeq h}\right)\right)$. Using this fact, we obtain, for all $h \in H$,

$$\begin{aligned}
\nu_i\left(\chi_{-i}^h\left(\hat{S}_{-i}^n\right)|S_{-i}(h)\right) &\geq \nu_i\left(\prod_{j \neq i}\operatorname{proj}_{S_j}\left(OP_j^n \cap C_j^{\succeq h}\right) \cap S_{-i}(h)\,|S_{-i}(h)\right) \\
&= \nu_i\left(\prod_{j \neq i}\operatorname{proj}_{S_j}\left(OP_j^n \cap C_j^{\succeq h}\right)|S_{-i}(h)\right) \\
&= \operatorname{marg}_{S_{-i}}\beta_{i,h}(t_i)\left(\prod_{j \neq i}\operatorname{proj}_{S_j}\left(OP_j^n \cap C_j^{\succeq h}\right)\right) \\
&= \beta_{i,h}(t_i)\left(S_i \times \prod_{j \neq i}\left(\operatorname{proj}_{S_j}\left(OP_j^n \cap C_j^{\succeq h}\right) \times T_j\right)\right) \\
&\geq \beta_{i,h}(t_i)\left(S_i \times \prod_{j \neq i}\left(OP_j^n \cap C_j^{\succeq h}\right)\right) = 1,
\end{aligned}$$

where the first inequality follows from the inductive hypothesis, the first equality follows from basic properties of a CPS,[41] the second and third equalities follow by definition, and the second inequality is immediate. This shows that ν_i satisfies the required properties. Since $i \in I$ and $\bar{h} \in H$ are arbitrary, the conclusion follows. ■

Lemma 6 *Fix a finite game* Γ *and a complete* Γ*-based type structure* $\mathcal{T}$*. Then, for each* $n \in \mathbb{N}_0$*,*

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

Proof. First note that

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

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

$$\mu_{s_i} \left(\cdot | S \left(h \right) \times T_{-i} \right) := \delta^*_{s_i} \left(\cdot | S_i \left(h \right) \right) \times \mu_{s_i,-i} \left(\cdot | S_{-i} \left(h \right) \times T_{-i} \right).$$

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 t_{s_i} satisfies independence and, for all $h \in H$,

$$\sigma_{t_{s_i},i} \left(s_i \left(h \right) | h \right) \geq \beta_{i,i} \left(t_{s_i} \right) \left(s_i^h | S_i \left(h \right) \right) = \delta^*_{s_i} \left(s_i^h | S_i \left(h \right) \right) = 1,$$

where the inequality holds because $s_i^h \in S_i \left(h, s_i \left(h \right) \right)$. Therefore $s_i \in \mathrm{proj}_{S_i} \left(C_i \right)$.

We now prove the result by induction on $n \in \mathbb{N}_0$.

Basis step. By (7.1), it follows that, for each $i \in I$,

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

and the result is immediate.

Inductive step. Assume that the result is true for $n \geq 0$. To show that it is also true for $n+1$, we need some preliminary definitions. First note that, by the inductive hypothesis, for every $i \in I$ and every $s_i \in \hat{S}_i^n$, there exists $t_{s_i} \in T_i$ such that $(s_i, t_{s_i}) \in OP_i^n \cap C_i$. So, for every $i \in I$ and $s_i \in \hat{S}_i^n$, we choose and fix some t_{s_i} satisfying the above condition, and we let τ_i denote the map that associates each $s_i \in \hat{S}_i^n$ with the corresponding type $\tau_i \left(s_i \right) = t_{s_i}$. For each $i \in I$, we let $\tau_{-i} := \left(\tau_j \right)_{j \neq i} : \hat{S}_{-i}^n \to T_{-i}$.

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

Furthermore, for all $i \in I$, $h \in H$ and $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$, we choose and fix some $\bar{s}_i \in \hat{S}_i^n$ such that $s_i(h') = \bar{s}_i(h')$ for every $h' \in H(h)$. We let $\hat{\varphi}_i^h : \chi_i^h\left(\hat{S}_i^n\right) \to \hat{S}_i^n$ denote the map that associates each $s_i \in \chi_i^h\left(\hat{S}_i^n\right)$ with the corresponding $\hat{\varphi}_i^h(s_i) = \bar{s}_i$. In particular, for each $i \in I$, and for all $h, h', h'' \in H$ such that $h' = (h, a')$, $h'' = (h, a'')$ and $a_i' = a_i''$, we require that $\hat{\varphi}_i^{h'}(s_i) = \hat{\varphi}_i^{h''}(s_i)$ for all $s_i \in \chi_i^{h'}\left(\hat{S}_i^n\right) \cap \chi_i^{h''}\left(\hat{S}_i^n\right)$ (by Lemma 4.(ii), $\chi_i^{h'}\left(\hat{S}_i^n\right) \cap \chi_i^{h''}\left(\hat{S}_i^n\right)$ is nonempty). Note that if $h = \varnothing$, then $\hat{\varphi}_i^h$ is the identity map, because $\chi_i^{\varnothing}\left(\hat{S}_i^n\right) = \hat{S}_i^n$.

With this, we recursively construct, for all $i \in I$ and $h \in H$, a map $\varphi_i^h : \chi_i^h(\hat{S}_i^n) \to \hat{S}_i^n$ that satisfies some desirable properties. The construction is based on the height of the histories, starting from the root. For each $h \in H \setminus \{\varnothing\}$, let $p(h)$ denote the immediate (strict) predecessor of h.

Fix a player $i \in I$. For $h = \varnothing$, let $\varphi_i^{\varnothing} := \hat{\varphi}_i^{\varnothing}$. Next, suppose that φ_i^h has been defined for all histories h with height $k \leq L(\varnothing)$. With this, for all $h \in H$ with $L(h) = k - 1$, we define

$$\varphi_i^h(s_i) = \begin{cases} \varphi_i^{p(h)}(s_i), & \text{if } s_i \in \chi_i^{p(h)}\left(\hat{S}_i^n\right) \cap S_i(h), \\ \hat{\varphi}_i^h(s_i), & \text{otherwise.} \end{cases}$$

The maps φ_i^h $(h \in H)$ satisfy the following property:

Claim 2 *Fix $i \in I$. For all $h, h' \in H$ such that $h \preceq h'$, and for all $s_i \in \chi_i^{h'}\left(\hat{S}_i^n\right)$,*

$$s_i \in \chi_i^h\left(\hat{S}_i^n\right) \cap S_i(h') \Rightarrow \varphi_i^h(s_i) = \varphi_i^{h'}(s_i).$$

Proof of Claim 2. Fix $h, h' \in H$ such that $h \preceq h'$. Fix also $s_i \in \chi_i^{h'}\left(\hat{S}_i^n\right)$ such that $s_i \in \chi_i^h\left(\hat{S}_i^n\right) \cap S_i(h')$. Since

$$\chi_i^h\left(\hat{S}_i^n\right) \cap S_i(h') \subseteq \chi_i^h\left(\hat{S}_i^n\right) \cap S_i(p(h')),$$

we have $s_i \in \chi_i^h\left(\hat{S}_i^n\right) \cap S_i(p(h'))$. Then, by Lemma 4.(i), $s_i \in \chi_i^{p(h')}\left(\hat{S}_i^n\right)$. Therefore $s_i \in \chi_i^{p(h')}\left(\hat{S}_i^n\right) \cap S_i(h')$ yields $\varphi_i^{p(h')}(s_i) = \varphi_i^{h'}(s_i)$. Next, if $h = p(h')$, we are done. Otherwise, set $\bar{h} = p(h')$. Repeating the above argument (with h' replaced by $\bar{h}$),

we obtain $\varphi_i^{p(\bar{h})}(s_i) = \varphi_i^{\bar{h}}(s_i) = \varphi_i^{h'}(s_i)$. Proceeding this way, it follows by induction that $\varphi_i^h(s_i) = \varphi_i^{h'}(s_i)$. $\square$

We will make use of Claim 2 below. For all $i \in I$ and $h \in H$, we let $\varphi_{-i}^h := (\varphi_j^h)_{j \neq i} : \chi_{-i}^h(\hat{S}_{-i}^n) \to \hat{S}_{-i}^n$.

We now provide the proof of the inductive step. Fix a player $i \in I$. We prove that $\hat{S}_i^{n+1} \subseteq \text{proj}_{S_i}\left(OP_i^{n+1} \cap C_i\right)$. Pick any $s_i \in \hat{S}_i^{n+1}$. We must show the existence of a type $t_i \in T_i$ such that $(s_i, t_i) \in OP_i^{n+1} \cap C_i$. Since $s_i \in \hat{S}_i^{n+1}$, there exists $\nu_{s_i} \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\nu_{s_i})$ and $\nu_{s_i}(\chi_{-i}^h(\hat{S}_{-i}^n)|S_{-i}(h)) = 1$ for every $h \in H$. Consider now an array of probability measures $\mu_{s_i,-i} := (\mu_{s_i,-i}(\cdot|S_{-i}(h) \times T_{-i}))_{h \in H}$ satisfying the following property: for all $h \in H$ and $s_{-i} \in \chi_{-i}^h(\hat{S}_{-i}^n)$,

$$\mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^h(s_{-i})\right)\right)|S_{-i}(h) \times T_{-i}\right) = \nu_{s_i}(s_{-i}|S_{-i}(h)),$$

where, by construction, $\varphi_{-i}^h(s_{-i}) \in \hat{S}_{-i}^n$ is such that $s_{-i}(h') = \varphi_{-i}^h(s_{-i})(h')$ for every $h' \in H(h)$. In words, conditional on every $h \in H$, measure $\mu_{s_i,-i}(\cdot|S_{-i}(h) \times T_{-i})$ is concentrated on the (finite) set of profiles $\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^h(s_{-i})\right)\right)$ such that the plan of each type $\tau_j(\varphi_j^h(s_j))$ $(j \neq i)$ coincides with s_j for every h' weakly following h.

Note that the marginal of $\mu_{s_i,-i}$ on $(S_{-i}, \mathcal{S}_{-i})$ is ν_{s_i}. We now claim that $\mu_{s_i,-i}$ is a CPS on $(S_{-i} \times T_{-i}, \mathcal{S}_{-i} \times T_{-i})$. To this end, fix any $h, \bar{h} \in H$ such that $\bar{h} \prec h$, which implies $S_{-i}(h) \subseteq S_{-i}\left(\bar{h}\right)$. Consider a profile $\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{\bar{h}}(s_{-i})\right)\right) \in S_{-i} \times T_{-i}$ where $s_{-i} \in \chi_{-i}^{\bar{h}}\left(\hat{S}_{-i}^n\right)$. Suppose further that $\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{\bar{h}}(s_{-i})\right)\right) \in S_{-i}(h) \times T_{-i}$. Then $s_{-i} \in \chi_{-i}^{\bar{h}}\left(\hat{S}_{-i}^n\right) \cap S_{-i}(h)$, hence, by Claim 2, $\varphi_{-i}^{\bar{h}}(s_{-i}) = \varphi_{-i}^h(s_{-i})$. Therefore,

$$\begin{aligned}
& \mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{\bar{h}}(s_{-i})\right)\right)|S_{-i}(h) \times T_{-i}\right) \mu_{s_i,-i}\left(S_{-i}(h) \times T_{-i}\left|S_{-i}\left(\bar{h}\right) \times T_{-i}\right.\right) \\
= \; & \mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^h(s_{-i})\right)\right)|S_{-i}(h) \times T_{-i}\right) \mu_{s_i,-i}\left(S_{-i}(h) \times T_{-i}\left|S_{-i}\left(\bar{h}\right) \times T_{-i}\right.\right) \\
= \; & \nu_{s_i}(s_{-i}|S_{-i}(h)) \, \nu_{s_i}\left(S_{-i}(h)\,|S_{-i}\left(\bar{h}\right)\right) \\
= \; & \nu_{s_i}\left(s_{-i}|S_{-i}\left(\bar{h}\right)\right) \\
= \; & \mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{\bar{h}}(s_{-i})\right)\right)|S_{-i}\left(\bar{h}\right) \times T_{-i}\right),
\end{aligned}$$

where the third equality holds because ν_{s_i} is a CPS. Next, fix a history $\bar{h}$. Consider distinct histories $h', h'' \in H$ that differ only for the last action of player i: $h' = \left(\bar{h}, a'\right)$, $h'' = \left(\bar{h}, a''\right)$ and $a'_{-i} = a''_{-i}$. Then $S_{-i}(h') = S_{-i}(h'')$, and, by Lemma 4.(ii) and Claim 2, it follows that, for all $s_{-i} \in \chi_{-i}^{\bar{h}}\left(\hat{S}_{-i}^n\right) \cap S_{-i}(h') = \chi_{-i}^{\bar{h}}\left(\hat{S}_{-i}^n\right) \cap S_{-i}(h'')$,

$$\varphi_{-i}^{h'}(s_{-i}) = \varphi_{-i}^{h''}(s_{-i}).$$

This conclusion also holds for all $s_{-i} \in \chi_{-i}^{h'}\left(\hat{S}_{-i}^{n}\right) \cap \chi_{-i}^{h''}\left(\hat{S}_{-i}^{n}\right)$, because if $s_{-i} \notin \chi_{-i}^{\bar{h}}\left(\hat{S}_{-i}^{n}\right) \cap S_{-i}(h')$, then $\varphi_{-i}^{h'}(s_{-i}) = \hat{\varphi}_{-i}^{h'}(s_{-i}) = \hat{\varphi}_{-i}^{h''}(s_{-i}) = \varphi_{-i}^{h''}(s_{-i})$ by construction. Note that $\nu_{s_i}(\cdot|S_{-i}(h')) = \nu_{s_i}(\cdot|S_{-i}(h''))$, and

$$\text{supp}\nu_{s_i}(\cdot|S_{-i}(h')) = \text{supp}\nu_{s_i}(\cdot|S_{-i}(h'')) \subseteq \chi_{-i}^{h'}\left(\hat{S}_{-i}^{n}\right) \cap \chi_{-i}^{h''}\left(\hat{S}_{-i}^{n}\right),$$

since $\nu_{s_i}\left(\chi_{-i}^{h'}\left(\hat{S}_{-i}^{n}\right)|S_{-i}(h')\right) = \nu_{s_i}\left(\chi_{-i}^{h''}\left(\hat{S}_{-i}^{n}\right)|S_{-i}(h'')\right) = 1$ by assumption. It follows that for all $s_{-i} \in \chi_{-i}^{h'}\left(\hat{S}_{-i}^{n}\right)$,

$$\begin{aligned}
& \mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{h'}(s_{-i})\right)\right)|S_{-i}(h') \times T_{-i}\right) = \nu_{s_i}(s_{-i}|S_{-i}(h')) \\
= \; & \nu_{s_i}(s_{-i}|S_{-i}(h'')) = \mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{h''}(s_{-i})\right)\right)|S_{-i}(h'') \times T_{-i}\right) \\
= \; & \mu_{s_i,-i}\left(\left(s_{-i}, \tau_{-i}\left(\varphi_{-i}^{h'}(s_{-i})\right)\right)|S_{-i}(h'') \times T_{-i}\right),
\end{aligned}$$

where the first equality is by definition, the second equality is immediate, the third equality is by definition, and the last equality holds because $\varphi_{-i}^{h'} = \varphi_{-i}^{h''}$. Thus, $\mu_{s_i,-i}$ satisfies an "own-action independence" property (besides the chain rule for histories $h, \bar{h} \in H$ such that $\bar{h} \prec h$); with this, one can show that $\mu_{s_i,-i}$ is a CPS.[42]

Next, consider the CPS $\mu_{s_i} \in \Delta^{\mathcal{S}\times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h \in H$,

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

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

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

First note that, by inspection of the definition of μ_{s_i}, type t_i satisfies independence; moreover, type t_i plans optimally (see Remark 3) because, for all $h \in H$,

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

Hence $(s_i, t_i) \in OP_i$. Furthermore, $(s_i, t_i) \in C_i$: for all $h \in H$,

$$\sigma_{t_i,i}(s_i(h)|h) \geq \beta_{i,i}(t_i)\left(s_i^h|S_i(h)\right) = \delta_{s_i}^{*}\left(s_i^h|S_i(h)\right) = 1,$$

[42] See the analysis of the Dirac array $\delta_{s_i}^{*}$, where the roles of i and $-i$ are reversed.

where the inequality holds because $s_i^h \in S_i(h, s_i(h))$.

Next, observe the following fact: fix $\bar{h} \in H$ arbitrarily, and consider a profile $(s_{-i}, \tau_{-i}(\varphi_{-i}^{\bar{h}}(s_{-i}))) \in S_{-i} \times T_{-i}$ such that $s_{-i} \in \chi_{-i}^{\bar{h}}(\hat{S}_{-i}^n)$. Then we have $(s_{-i}, \tau_{-i}(\varphi_{-i}^{\bar{h}}(s_{-i}))) \in OP_{-i}^n \cap C_{-i}^{\succeq \bar{h}}$, since, by definition of $\varphi_{-i}^{\bar{h}}$, $s_{-i}(h) = \varphi_{-i}^{\bar{h}}(s_{-i})(h)$ for every $h \in H(\bar{h})$. Hence, by definition of $\mu_{s_i,-i}$,

$$\forall h \in H, \mu_{s_i,-i}\left(OP_{-i}^n \cap C_{-i}^{\succeq h} | S_{-i}h) \times T_{-i}\right) = 1; \tag{7.2}$$

we use this fact to show that $(s_i, t_i) \in BCC_i \cap \mathrm{B}_i\left(OP_{-i}^n\right)$.

We first check that $(s_i, t_i) \in BCC_i := \cap_{h \in H} \mathrm{B}_{i,h}\left(C_{-i}^{\succeq h}\right)$. For every $h \in H$, we have

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

where the third equality follows from (7.2); hence $(s_i, t_i) \in BCC_i$. Next, we check that $(s_i, t_i) \in \mathrm{B}_i\left(OP_{-i}^n\right) := \cap_{h \in H} \mathrm{B}_{i,h}\left(OP_{-i}^n\right)$; since the sequence $\left(OP_{-i}^m\right)_{m=0}^n$ is decreasing, monotonicity of operator $\mathrm{B}_i(\cdot)$ implies $(s_i, t_i) \in \cap_{m=0}^n \mathrm{B}_i\left(OP_{-i}^l\right)$. By (7.2), it follows that, for all $h \in H$,

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

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

We therefore have $s_i \in \mathrm{proj}_{S_i}\left(OP_i^{n+1} \cap C_i\right)$; since s_i is arbitrary, we can claim that $\hat{S}_i^{n+1} \subseteq \mathrm{proj}_{S_i}\left(OP_i^{n+1} \cap C_i\right)$. The converse follows from Lemma 5 (with $h = \varnothing$). This concludes the proof of the inductive step. ■

We can now provide the proof of Theorem 2.

Proof of Theorem 2. Parts (i)-(ii) of the main statement of the theorem immediately follow from Lemma 5 (with $h = \varnothing$). The statement about complete type structures immediately follows from Lemma 6 as long as we consider finitely many steps, i.e., $n \in \mathbb{N}$. To see that it holds also in the limit, first note that

$\text{proj}_S(OP^n \cap C) = \hat{S}^n \neq \emptyset$ for every $n \in \mathbb{N}$, hence $OP^n \cap C \neq \emptyset$ for every $n \in \mathbb{N}$. Then

$$\text{proj}_S(OP^\infty \cap C) = \text{proj}_S \bigcap_{n=1}^{\infty} (OP^n \cap C) = \bigcap_{n=1}^{\infty} \text{proj}_S(OP^n \cap C) = \bigcap_{n=1}^{\infty} \hat{S}^n,$$

where the first equality holds by definition, the second follows from Lemma 2 (as $(OP^n \cap C)_{n=1}^{\infty}$ is a decreasing sequence of closed and nonempty sets), and the third follows from Lemma 6. ■

Forward-induction reasoning

We postpone the proof of Theorem 3, because we will show that the epistemic assumptions featured there (RCSBR) have the same behavioral implications as the epistemic assumptions of Theorem 4. With this, Theorem 3 follows from Theorem 4.

We first record some preliminary results.

Remark 10 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Then, for each $i \in I$ and $n > 1$,*

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

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

Lemma 7 *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 ν 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) μ fully believes E_0, (b) μ 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) is very similar to the proof of Lemma 3 in Battigalli and Tebaldi (2019). ■

To prove the following result, we find it convenient to define $R^0 := S \times T$ and $R^{*,0} := S \times T$ in a Γ-based type structure $\mathcal{T}$.

Lemma 8 *Fix a finite game* Γ *and a* Γ*-based* complete *type structure* $\mathcal{T}$*. Then, for every* $n \in \mathbb{N}_0$*,*
(i) $\prod_{i\in I} R_i^{*,n} \subseteq \prod_{i\in I} R_i^n$*, and*
(ii) $\operatorname{proj}_S \prod_{i\in I} R_i^n = \operatorname{proj}_S \prod_{i\in I} R_i^{*,n}$.

Proof. The proof is by induction on $n \in \mathbb{N}_0$. The basis step ($n = 0$) trivially holds by definition.

Suppose by way of induction that (IH) $\prod_{i\in I} R_i^{*,m} \subseteq \prod_{i\in I} R_i^m$ and $\operatorname{proj}_S \prod_{i\in I} R_i^m = \operatorname{proj}_S \prod_{i\in I} R_i^{*,m}$ for every $m \leq n$. Then, for every i, we have $R_i^{*,n} \subseteq R_i^n$ and

$$\begin{aligned}
\mathrm{SB}_i\left(\prod_{i\in I} R_{-i}^{*,n}\right) &\overset{(\mathrm{IH},\mathrm{def.SB}_i)}{=} \bigcap_{h:S_{-i}(h)\cap \operatorname{proj}_{S_{-i}} \prod_{i\in I} R_i^n \neq \emptyset} \mathrm{B}_{i,h}\left(\prod_{i\in I} R_{-i}^{*,n}\right) \\
&\overset{(\mathrm{mon.B}_{i,h})}{\subseteq} \bigcap_{h:S_{-i}(h)\cap \operatorname{proj}_{S_{-i}} \prod_{i\in I} R_i^n \neq \emptyset} \mathrm{B}_{i,h}\left(\prod_{i\in I} R_{-i}^{n}\right) \overset{(\mathrm{def.SB}_i)}{=} \mathrm{SB}_i\left(\prod_{i\in I} R_{-i}^{n}\right),
\end{aligned}$$

where the equalities follow from the inductive hypothesis and the definition of SB_i (taking into account that, for every event $E_{-i} \subseteq S_{-i} \times T_{-i}$, we have $(S_{-i}(h) \times T_{-i}) \cap E_{-i} \neq \emptyset$ if and only if $S_{-i}(h) \cap \operatorname{proj}_{S_{-i}} E_{-i} \neq \emptyset$), and the inclusion follows from the monotonicity of each $\mathrm{B}_{i,h}$ operator. Since $R_i^{*,n+1} = R_i^{*,n} \cap \mathrm{SB}_i\left(R_{-i}^{*,n}\right)$ and $R_i^{n+1} = R_i^n \cap \mathrm{SB}_i\left(R_{-i}^n\right)$, it follows that $R_i^{*,n+1} \subseteq R_i^{n+1}$. Thus, $\prod_{i\in I} R_i^{*,n+1} \subseteq \prod_{i\in I} R_i^{n+1}$, which implies $\operatorname{proj}_S \prod_{i\in I} R_i^{*,n+1} \subseteq \operatorname{proj}_S \prod_{i\in I} R_i^{n+1}$.

To prove the converse, fix any personal state $(s_i, t_i) \in R_i^{n+1}$. Let $\beta_{i,-i}^1(t_i) := \left(\operatorname{marg}_{S_{-i}} \beta_{i,h}(t_i)\right)_{h\in H}$ denote the marginal first-order belief of t_i about coplayers, which is a CPS because rationality requires independence. By consistency, $\beta_{i,i}^1(t_i) = \delta_{s_i}^* := \left(\delta_{s_i^h}\right)_{h\in H}$ is the CPS concentrated on s_i^h for each $h \in H$. The first-order CPS of t_i, viz. $\beta_i^1(t_i)$, is the "product" of $\beta_{i,i}^1(t_i)$ and $\beta_{i,-i}^1(t_i)$. By Remark 10,

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

By the inductive hypothesis (IH), $\text{proj}_{S_{-i}} R_{-i}^m = \text{proj}_{S_{-i}} R_{-i}^{*,m}$ for every $m \in \{1, ..., n\}$. Thus, for each $m \in \{1, ..., n\}$ and every $h \in H$, $S_{-i}(h) \cap \text{proj}_{S_{-i}} R_{-i}^{*,m} \neq \emptyset$ implies $\beta_{i,-i}^1(t_i)\left(\text{proj}_{S_{-i}} R_{-i}^{*,m}\right) = 1$. By Lemma 3 in Battigalli and Tebaldi (2019) (cf. Lemma 7), there exists $\mu_{i,-i} := \left(\mu_{i,-i,h}\right)_{h \in H} \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}(S_{-i} \times T_{-i})$ such that

(1) $\beta_{i,-i}^1(t_i) = \left(\text{marg}_{S_{-i}} \mu_{i,-i,h}\right)_{h \in H}$, and

(2) for each $m \in \{1, ..., n\}$ and every $h \in H$, if $S_{-i}(h) \cap \text{proj}_{S_{-i}} R_{-i}^{*,m} \neq \emptyset$ then $\mu_{i,-i,h}\left(R_{-i}^{*,m}\right) = 1$.

Let $\mu_i := \left(\delta_{s_i^h} \times \mu_{i,-i,h}\right)_{h \in H} \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ the CPS obtained by taking the "product" of $\beta_{i,i}^1(t_i) = \delta_{s_i}^*$ and $\mu_{i,-i}$. By completeness of the type structure, β_i is onto; therefore, there is a type $t_i^* \in T_i$ such that $\beta_i(t_i^*) = \mu_i$. With this,

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

where the inclusion holds by construction and the equality holds by Remark 10. This shows that $\text{proj}_S \prod_{i \in I} R_i^{n+1} \subseteq \text{proj}_S \prod_{i \in I} R_i^{*,n+1}$. ■

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

Lemma 9 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Fix also $i \in I$ and $(s_i, t_i) \in S_i \times T_i$. Then $(s_i, t_i) \in C_i^*$ if and only if $(s_i, t_i) \in C_i$ and $\beta_i(t_i)$ fully believes $S_i \times C_{-i}^*$, where $C_{-i}^* := \prod_{j \neq i} C_j^*$.*

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

$$\begin{aligned} C^* &= \bigcap_{m \in \mathbb{N}_0} \text{B}^m(C) = \text{B}^0(C) \cap \left(\bigcap_{m \in \mathbb{N}} \text{B}^m(C)\right) = C \cap \left(\bigcap_{m \in \mathbb{N}_0} \text{B}\left(\text{B}^m(C)\right)\right) \\ &= C \cap \text{B}\left(\bigcap_{m \in \mathbb{N}_0} \left(\text{B}^m(C)\right)\right) = C \cap \text{B}(C^*). \end{aligned}$$

So the statement immediately follows. ■

Lemma 10 *Fix a finite game Γ and a complete Γ-based type structure $\mathcal{T}$. Then, for all $i \in I$ and $h \in H$,*

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

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

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

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

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

Since $C^* := \cap_{n \in \mathbb{N}_0} \mathrm{B}^n(C)$, this implies the thesis.

Basis step. Fix $i \in I$ and $\bar{h} \in H$. Pick any $s_i \in S_i\left(\bar{h}\right)$, 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

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

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 t_i satisfies independence and, for all $h \in H$,

$$\sigma_{t_i,i}(s_i(h) | h) \geq \beta_{i,i}(t_i)\left(s_i^h | S_i(h)\right) = \delta^*_{s_i}\left(s_i^h | S_i(h)\right) = 1,$$

where the inequality holds because $s_i^h \in S_i(h, s_i(h))$. Hence $\text{proj}_{S_i \times T_i} \mathrm{B}^0(C) \cap \left(S_i\left(\bar{h}\right) \times T_i\right) = C_i \cap \left(S_i\left(\bar{h}\right) \times T_i\right) \neq \emptyset$. As i and $\bar{h}$ are arbitrary, the proof of the basis step is complete.

Inductive step. Assume that the result is true for $n \geq 0$. We show that it is also true for $n+1$.

Fix $i \in I$ and $\bar{h} \in H$. Pick any $s_i \in S_i\left(\bar{h}\right)$. By the inductive hypothesis, the (closed) set $\text{proj}_{S_{-i} \times T_{-i}} \mathrm{B}^n(C)$ is nonempty, and for all $h \in H$,

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

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 $\text{proj}_{S_{-i} \times T_{-i}} \mathrm{B}^n(C)$. With this, consider the CPS $\mu_i \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h \in H$,

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

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 $(s_i, t_i) \in \mathrm{B}_i\left(\text{proj}_{S_{-i} \times T_{-i}} \mathrm{B}^n(C)\right)$, because $\beta_{i,-i}(t_i) = \mu_{i,-i}$. It follows that $(s_i, t_i) \in \text{proj}_{S_i \times T_i} \mathrm{B}^{n+1}(C) \cap \left(S_i\left(\bar{h}\right) \times T_i\right)$. Since i and $\bar{h}$ are arbitrary, the conclusion follows. ■

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

Proof of Theorem 4. *Part (i)*: First note that, by Lemma 10, $C_i^* \neq \emptyset$ for each $i \in I$. Moreover

$$\forall i \in I, S_i = \text{proj}_{S_i}(C_i^*). \tag{7.3}$$

The inclusion $\text{proj}_{S_i}(C_i^*) \subseteq S_i$ is obvious. Conversely, pick any $s_i \in S_i$. By Lemma 10, 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

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

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 t_{s_i} satisfies independence and, for all $h \in H$,

$$\sigma_{t_{s_i},i}(s_i(h)|h) \geq \beta_{i,i}(t_{s_i})\left(s_i^h|S_i(h)\right) = \delta_{s_i}^*\left(s_i^h|S_i(h)\right) = 1,$$

where the inequality holds because $s_i^h \in S_i(h, s_i(h))$. Since $\beta_{i,-i}(t_i) := \mu_{s_i,-i}$ fully believes C_{-i}^*, Lemma 9 yields $(s_i, t_{s_i}) \in C_i^*$. Therefore $s_i \in \text{proj}_{S_i}(C_i^*)$.

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

Conversely, pick any $s_i \in S_i^1$. By definition, there exists $\nu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\nu_i)$. Part (ii) of Lemma 7 yields the existence of $\mu_{i,-i} \in \Delta^{\mathcal{S}_{-i}\times T_{-i}}(S_{-i} \times T_{-i})$ such that $\mu_{i,-i}$ fully believes C_{-i}^* and $\text{marg}_{S_{-i}}\mu_{i,-i} = \nu_i$. Consider the CPS $\mu_i \in \Delta^{\mathcal{S}\times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h \in H$,

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

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

$$\sigma_{t_i,i}(s_i(h)|h) \geq \beta_{i,i}(t_i)\left(s_i^h|S_i(h)\right) = \delta_{s_i}^*\left(s_i^h|S_i(h)\right) = 1$$

for all $h \in H$. Since $\mu_{i,-i}$ fully believes C_{-i}^*, Lemma 9 yields $(s_i, t_i) \in C_i^*$. Moreover, type t_i plans optimally (see Remark 3) because, for all $h \in H$,

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

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 $m \leq n+1$.

Pick any $s_i \in \text{proj}_{S_i} R_i^{*,n+1}$, so that $(s_i, t_i) \in R_i^{*,n+1}$ for some $t_i \in T_i$. Then, by Remark 10, $(s_i, t_i) \in R_i^{*,1} \cap \left(\cap_{m \leq n} \text{SB}_i\left(R_{-i}^{*,m}\right)\right)$. 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 $s_i \in \rho_i(\nu_i)$. Part (i) of Lemma 7 entails that ν_i strongly believes $\left(\text{proj}_{S_{-i}} R_{-i}^{*,m}\right)_{m=1}^{n}$, hence, by the inductive hypothesis, ν_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, pick any $s_i \in S_i^{n+1}$. By definition, there exists $\nu_i \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\nu_i)$ and ν_i strongly believes $\left(S_{-i}^{m}\right)_{m=1}^{n}$. By the inductive hypothesis, ν_i strongly believes $\left(\text{proj}_{S_{-i}} R_{-i}^{*,m}\right)_{m=1}^{n}$. Moreover, ν_i fully believes S_{-i} by definition, and so, by (7.3), ν_i fully believes $\text{proj}_{S_{-i}} C_{-i}^{*}$. Hence part (ii) of Lemma 7 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) $\text{marg}_{S_{-i}} \mu_{i,-i} = \nu_i$.

Consider the CPS $\mu_i \in \Delta^{\mathcal{S} \times T_{-i}}(S \times T_{-i})$ defined as follows: for all $h \in H$,

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

Since β_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 10, this is equivalent to showing 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 $\left(\prod_{i \in I} R_i^{*,n}\right)_{n \in \mathbb{N}}$ 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. ■

Proof of Theorem 3. Part (i) follows from Lemma 8 and Theorem 4. Part (ii) follows from the same argument as in the proof of Theorem 4 (ii). ■

Appendix B: Supplementary material

This appendix provides a formal analysis of some results mentioned in the main text that are not stated there as formal theorems. First, in Appendix B.1 we show how backwards rationalizability can be given a characterization in terms of the so-called "backwards procedure" (Penta 2015). Second, Appendix B.2 provides a formal epistemic analysis of initial rationalizability in terms of the epistemic notions as (informally) defined in the main text. Finally, Appendix B.3 illustrates a possible way to extend the epistemic analysis of backwards rationalizability to the general class of finite games with perfect recall.

Appendix B.1: 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 formally introduce the "backwards procedure" for the games considered in this paper; then we show that the solution concept of backwards rationalizability—as per Definition 6—yields a subset of the profiles surviving the backwards procedure; we show that the equivalence between the two concepts obtains if the notion of CPS is replaced by the weaker notion of forward CPS (see Section 7). We have to introduce additional notation and definitions. To ease language, in this appendix we slightly modify our terminology (cf. Section 7). 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 Γ. 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

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

For each $i \in I$ and $h \in H$, let $\pi_i^h : S_i \to S_i^{\succeq h}$ be the 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 π_j^h $(j \neq i)$, that is, $\pi_{-i}^h(s_{-i}) = \left(\pi_j^h(s_j)\right)_{j \neq i}$. The map $\pi^h : S \to S^{\succeq h}$ is defined in the usual way: $\pi^h(s) = \left(\pi_j^h(s_j)\right)_{j \in I}$ for each $s = (s_j)_{j \in I} \in S$.

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 $h \in H$, define the path function $\zeta(\cdot|h) : S^{\succeq h} \to Z$. Note: if $s \in S(h)$, then $\zeta\left(\pi^h(s)|h\right) = \zeta(s)$, i.e., $\zeta(s|h)$ is the terminal history induced by profile s from history h. With this,

$$U_i(\cdot|h) := u_i \circ \zeta(\cdot|h) : S^{\succeq h} \to \mathbb{R}$$

is the utility of player i as a function of profile $s^{\succeq h} = \left(s_i^{\succeq h}\right)_{i \in I}$. 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 $BR_i(\nu_i)$.

Next, 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 $k \in \{1, ..., K\}$,

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

is the set of all histories of height k. Next, fix some $k > 1$. For each $h \in H^k$,

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

is the set of all histories of height $k-1$ that strictly follow h.

We can formally introduce the "backwards procedure," which starts by considering first all histories of height 1, and then proceeds 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*

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

$(k>1)$ *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*

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

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 first main result of this section is the following:

Proposition 1 *Fix a finite game with observable actions* Γ*. Then*

$$\hat{S}^{\infty} \subseteq P^{K,\infty}(\varnothing).$$

To prove the result, we first record an ancillary result that will be used also in the proof of Proposition 2 below.

Lemma 11 *Fix* $i \in I$, $\bar{h} \in H$ *and a nonempty set* $Q_i \subseteq S_i$*. Then, for every* $h \in H(\bar{h})$,

$$\pi_i^h\left(\chi_i^{\bar{h}}(Q_i)\right) \subseteq \pi_i^h(Q_i)$$

and

$$\pi_i^h\left(\chi_i^h(Q_i)\right) = \pi_i^h(Q_i).$$

Proof. Fix some $h \in H(\bar{h})$. To show the first inclusion, pick any $s_i^{\succeq h} \in \pi_i^h\left(\chi_i^{\bar{h}}(Q_i)\right)$. We show the existence of $\bar{s}_i \in Q_i$ such that $(\bar{s}_i|h) = s_i^{\succeq h}$. By definition, there exists $s_i \in \chi_i^{\bar{h}}(Q_i)$ such that $(s_i|h) = s_i^{\succeq h}$. Moreover, since $s_i \in \chi_i^{\bar{h}}(Q_i)$, there exists $\bar{s}_i \in Q_i$ such that $s_i(h') = \bar{s}_i(h')$ for all $h' \in H\left(\bar{h}\right)$. Since $h \in H(\bar{h})$, we obtain $(\bar{s}_i|h) = s_i^{\succeq h}$, as required.

We now prove that $\pi_i^h\left(\chi_i^h(Q_i)\right) = \pi_i^h(Q_i)$. Note that the inclusion $\pi_i^h\left(\chi_i^h(Q_i)\right) \subseteq \pi_i^h(Q_i)$ follows from the first one (with $h = \bar{h}$). For the converse, fix some $s_i^{\succeq h} \in \pi_i^h(Q_i)$. 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 $\hat{s}_i \in Q_i$ such that $(\hat{s}_i|h) = s_i^{\succeq h}$. Next, pick any $s_i^* \in S_i(h)$, and define $s_i \in S_i$ as follows: for all $h' \in H$,

$$s_i(h') := \begin{cases} \hat{s}_i(h'), & \text{if } h' \in H(h), \\ s_i^*(h'), & \text{otherwise.} \end{cases}$$

We have $s_i \in S_i(h)$, and $s_i(h') = \hat{s}_i(h')$ for all $h' \in H(h)$; it follows that $s_i \in \chi_i^h(Q_i)$. Therefore, $(s_i|h) = (\hat{s}_i|h) = s_i^{\succeq h}$. ■

Proof of Proposition 1. 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 $n \geq 0$. We show that it is true for $n + 1$. 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 $h' \in H$. Lemma 11 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 inductive 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 $s_{-i}^{\succeq h} \in S_{-i}^{\succeq h}$,

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

That is, ν_i is the image measure of $\mu_i(\cdot|S_{-i}(h))$ 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 inductive 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 $n \geq 0$. We show that it is true for $n+1$. The argument proceeds in the same way as in step $k = 1$. Pick any $s_i \in \hat{S}_i^{\infty}$. Then, by definition, 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 Lemma 11 and the inductive 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 $\mu_i(\cdot|S_{-i}(h))$ 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 the equivalence between the backwards procedure and backwards rationalizability holds provided that the notion of CPS is replaced by *forward* CPS (cf. Section 7).

Definition 9 *A **forward CPS** is an array of probability measures $\mu_i = (\mu_i(\cdot|h))_{h \in H} \in (\Delta(S_{-i}))^H$ such that:*

(i) $\mu_i(S_{-i}(h)|h) = 1$ for all $h \in H$; and

(ii) for all $h, h' \in H$ such that $h \prec h'$, and for all $E_{-i} \subseteq S_{-i}(h')$,

$$\mu_i(E_{-i}|h) = \mu_i(E_{-i}|h')\,\mu_i(S_{-i}(h')|h).$$

Let $\left(\widetilde{S}^n\right)_{n\in\mathbb{N}_0}$ be the solution procedure obtained by replacing, in the definition of backwards rationalizability (Definition 6), CPSs with forward CPSs. The second main result of this section is the following:

Proposition 2 *Fix a finite game with observable actions Γ. Then*

$$\widetilde{S}^\infty = P^{K,\infty}(\varnothing).$$

Proof. The proof showing that $\widetilde{S}^\infty \subseteq P^{K,\infty}(\varnothing)$ is analogous to the one in Proposition 1. Thus, we prove that $P^{K,\infty}(\varnothing) \subseteq \widetilde{S}^\infty$. We do this by showing that $P^{K,\infty}(\varnothing) \subseteq \widetilde{S}^n$ for all $n \in \mathbb{N}_0$. For $n = 0$, the result is immediate because $\widetilde{S}^0 = S$. Thus, suppose that the result is true for $n \geq 0$. We prove the result for $n+1$.

We first record a consequence of the inductive hypothesis.

Claim 3 *For every $i \in I$, $k \in \{1, ..., K\}$ and $h \in H^k$,*

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

Proof of Claim 3. The first equality follows from the definition of the backwards procedure, while the set inclusion follows from the inductive hypothesis. Lemma 11 yields the last equality. □

We make use of this result to construct, for each $h \in H$, a profile of maps $\left(\varphi_i^h : S_i^{\succ h} \to S_i\right)_{i\in I}$ 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 3 yields, for each $s_i^{\succ h} \in P_i^{k,\infty}(h)$, the existence of $s_i \in \chi_i^h\left(\widetilde{S}_i^n\right)$ such that $\pi_i^h(s_i) = s_i^{\succ h}$. Hence, for every $s_i^{\succ h} \in P_i^{k,\infty}(h)$, we choose and fix some $s_i \in \chi_i^h\left(\widetilde{S}_i^n\right)$ such that $\pi_i^h(s_i) = s_i^{\succ h}$; we also choose an arbitrary $s_i^0 \in S_i$, and we define the map $\varphi_i^h : S_i^{\succ h} \to S_i$ as follows:

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

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

$$P_i^{k,\infty}(h) \subseteq \left(\varphi_i^h\right)^{-1}\left(\chi_i^h\left(\widetilde{S}_i^n\right)\right). \tag{7.4}$$

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 φ_{j}^{h} $(j \neq i)$, that is, $\varphi_{-i}^{h}\left(s_{-i}^{\succeq h}\right) := \left(\varphi_{j}^{h}\left(s_{j}^{\succeq h}\right)\right)_{j \neq i}$ for all $s_{-i}^{\succeq h} = \left(s_{j}^{\succeq h}\right)_{j \neq i} \in S_{-i}^{\succeq h}$.

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 forward CPS $\mu_i \in (\Delta(S_{-i}))^{H}$ such that $s_i \in \rho_i(\mu_i)$ and $\mu_i\left(\chi_{-i}^{h}\left(\widetilde{S}_{-i}^{n}\right)|h\right) = 1$ 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 forward CPS $\mu_i \in (\Delta(S_{-i}))^{H}$ 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

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

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

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

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

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

For all other histories, we proceed as above, in order to obtain an array of probability measures $\mu_i = (\mu_i(\cdot|h))_{h \in H}$ such that the chain rule holds for all $h, h' \in H$ such that $h \prec h'$; hence μ_i is a forward CPS.

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

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

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

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

where the first equality is by definition, the inequality follows from Lemma 4, 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 (7.4) that

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

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.

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

Remark 11 *The proof of Proposition 2 is analogous to the proof of Proposition 6 in Penta (2015). It should be noted that Penta adopts CPSs as form of belief system (Penta 2015, Appendix A.2). Yet, his proof of Proposition 6 yields a forward CPS as output (see Penta 2015, pp. 306-307).*

Appendix B.2: Epistemic analysis of initial rationalizability

The following notion of initial, or weak rationalizability 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.

Definition 10 *Consider the following procedure.*

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

(Step $n > 0$**)** *For every $i \in I$ and every $s_i \in S_i$, let $s_i \in W_i^n$ 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. *$\mu_i\left(W_{-i}^{n-1} | S_{-i}\right) = 1$.*
   Also, let $W_{-i}^n := \prod_{j \neq i}^n W_j^n$ and $W^n := \prod_{i \in I} W_i^n$.

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

One can show by standard arguments that initial rationalizability is a nonempty solution concept:

Remark 12 *$W^\infty \neq \emptyset$.*

Fix a game Γ and an associated Γ-based type structure $\mathcal{T}$. For each $i \in I$, let $R_{i,\varnothing}^1 := R_i$, and, for each $n \in \mathbb{N}$, define $R_{i,\varnothing}^{n+1}$ recursively by

$$R_{i,\varnothing}^{n+1} := R_{i,\varnothing}^n \cap \mathrm{B}_{i,\varnothing}\left(R_{-i,\varnothing}^n\right),$$

where $R_{-i,\varnothing}^n := \prod_{j \neq i} R_{j,\varnothing}^n$. The set of states consistent with rationality and common initial belief in rationality (RCIBR) is therefore defined as

$$R_\varnothing^\infty := \prod_{i \in I} R_{i,\varnothing}^\infty,$$

where $R_{i,\varnothing}^\infty := \bigcap_{n \in \mathbb{N}} R_{i,\varnothing}^n$ for every $i \in I$.

For each $i \in I$, let $\hat{R}_{i,\varnothing}^1 := C_i^* \cap OP_i$, and, for each $n \in \mathbb{N}$, define $\hat{R}_{i,\varnothing}^{n+1}$ recursively by

$$\hat{R}_{i,\varnothing}^{n+1} := \hat{R}_{i,\varnothing}^n \cap \mathrm{B}_{i,\varnothing}\left(\hat{R}_{-i,\varnothing}^n\right),$$

where $\hat{R}^n_{-i,\varnothing} := \prod_{j\neq i} \hat{R}^n_{j,\varnothing}$. For each $i \in I$, let $\hat{R}^\infty_{i,\varnothing} := \bigcap_{n\in\mathbb{N}} \hat{R}^n_{i,\varnothing}$. Then

$$\hat{R}^\infty_{\varnothing} := \prod_{i\in I} \hat{R}^\infty_{i,\varnothing}$$

is the set of states in which there is optimal planning, transparency of consistency, and common initial belief in optimal planning.

Remark 13 *Event $\hat{R}^\infty_{\varnothing}$ is the set of states in which*

(a) there is optimal planning and transparency of consistency,

(b) there is common initial belief in (a).

To see this, note that event C^* is self-evident, i.e., $C^* \subseteq \mathrm{B}(C^*)$, and full belief in C^* implies initial belief in C^*.

Theorem 5 *Fix a finite game Γ and a Γ-based* complete *type structure $\mathcal{T}$. Then,*

(i) for every $n \in \mathbb{N}$, $\mathrm{proj}_S \prod_{i\in I} \hat{R}^n_{i,\varnothing} = \mathrm{proj}_S \prod_{i\in I} R^n_{i,\varnothing} = W^n$;

(ii) $\mathrm{proj}_S \hat{R}^\infty_{\varnothing} = \mathrm{proj}_S R^\infty_{\varnothing} = W^\infty$.

The proof of this result is a simplified version of the proof of Theorem 4 and so it is omitted. We instead provide an alternative epistemic justification of initial rationalizability which is closer to the one provided for backwards rationalizability.

In what follows, fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. We say that each player i **initially believes in** the **consistency** of the other players if i believes $C_{-i} := \prod_{j\neq i} C_j$ at the beginning of the game. The corresponding events are

$$\begin{aligned} IBC_i &:= \mathrm{B}_{i,\varnothing}(C_{-i}), \\ IBC &:= \prod_{i\in I} IBC_i. \end{aligned}$$

Note that $BCC \subseteq IBC$, that is, a player who believes in continuation consistency also initially believes in consistency of the co-players. Since each C_{-i} is a product of closed sets, hence itself closed, IBC_i is closed as well.

With this, define recursively the following epistemic events:

- $\widehat{OP}^1_i := OP_i \cap IBC_i$,
- $\widehat{OP}^{m+1}_i := \widehat{OP}^m_i \cap \mathrm{B}_i\left(\widehat{OP}^m_{-i}\right)$, where $\widehat{OP}^m_{-i} := \prod_{j\neq i} \widehat{OP}^m_j$.

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

Finally, let $\widehat{OP}^\infty := \cap_{m \in \mathbb{N}} \widehat{OP}^m$. Then

$$\widehat{OP}^\infty \cap C$$

is the set of states in which there is consistency and transparency of optimal planning and of initial belief in consistency.

Theorem 6 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Then,*

(i) for every $n \in \mathbb{N}$, $\mathrm{proj}_S\left(\widehat{OP}^n \cap C\right) \subseteq W^n$*;*

(ii) $\mathrm{proj}_S\left(\widehat{OP}^\infty \cap C\right) \subseteq W^\infty$*.*

Furthermore, if structure $\mathcal{T}$ is complete *these weak inclusions hold as equalities.*

The proof of Theorem 6 relies on Lemma 12 below. To ease the statement and proof, let $\widehat{OP}_i^0 := S_i \times T_i$ for each $i \in I$. The sets $\widehat{OP}^0$ and $\widehat{OP}_{-i}^0$ are defined in the obvious way: $\widehat{OP}^0 := \prod_{i \in I} \widehat{OP}_i^0$ and $\widehat{OP}_{-i}^0 := \prod_{j \neq i} \widehat{OP}_j^0$.

Lemma 12 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. The following statements hold:*

(i) for all $n \in \mathbb{N}_0$,

$$\mathrm{proj}_S\left(\widehat{OP}^n \cap C\right) \subseteq W^n;$$

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

$$\mathrm{proj}_S\left(\widehat{OP}^n \cap C\right) = W^n.$$

Proof. Part (i): We prove the following claim:

$$\forall i \in I, \forall n \in \mathbb{N}_0, \mathrm{proj}_{S_i}\left(\widehat{OP}_i^n \cap C\right) \subseteq W_i^n.$$

The proof is by induction on $n \in \mathbb{N}_0$.

Basis step. Note that, for every $i \in I$,

$$\begin{aligned} \mathrm{proj}_{S_i}\left(\widehat{OP}_i^0 \cap C_i\right) &= \mathrm{proj}_{S_i}\left(C_i\right) \\ &\subseteq S_i \\ &= W_i^0, \end{aligned}$$

so the result follows immediately.

Inductive step. Assume that the result is true for $n \geq 0$. We show that it is also true for $n+1$.

Fix $i \in I$. Pick any $s_i \in \mathrm{proj}_{S_i}\left(\widehat{OP}_i^{n+1} \cap C_i\right)$, so that $(s_i, t_i) \in \widehat{OP}_i^{n+1} \cap C_i$ for some $t_i \in T_i$. Since $\widehat{OP}_i^{n+1} \subseteq \widehat{OP}_i^n$, it follows that $(s_i, t_i) \in \widehat{OP}_i^n \cap C_i$, and so, by the inductive hypothesis, $s_i \in W_i^n$. Hence $s_i \in \rho_i\left(\nu_i\right)$, where ν_i denotes the marginal of $\beta_i\left(t_i\right)$ on $\left(S_{-i}, \mathcal{S}_{-i}\right)$. So, in order to show that $s_i \in W_i^{n+1}$, we have to show that $\nu_i\left(W_{-i}^n | S_{-i}\right) = 1$.

To this end, first note that $(s_i, t_i) \in \widehat{OP}_i^{n+1}$ implies $(s_i, t_i) \in \mathrm{B}_i\left(\widehat{OP}_{-i}^n\right) := \cap_{h' \in H} \mathrm{B}_{i,h'}\left(\widehat{OP}_{-i}^n\right)$. Note also that $(s_i, t_i) \in IBC_i := \mathrm{B}_{i,\varnothing}\left(C_{-i}\right)$; hence, by the conjunction property of the operator $\mathrm{B}_{i,\varnothing}\left(\cdot\right)$, it follows that $(s_i, t_i) \in \mathrm{B}_{i,\varnothing}\left(\widehat{OP}_{-i}^n \cap C_{-i}\right)$. Using this fact, we get that

$$\begin{aligned} \nu_i\left(W_{-i}^n | S_{-i}\right) &\geq \nu_i\left(\prod_{j \neq i} \mathrm{proj}_{S_j}\left(\widehat{OP}_j^n \cap C_j\right) | S_{-i}\right) \\ &= \mathrm{marg}_{S_{-i}} \beta_{i,\varnothing}\left(t_i\right)\left(\prod_{j \neq i} \mathrm{proj}_{S_j}\left(\widehat{OP}_j^n \cap C_j\right)\right) \\ &= \beta_{i,\varnothing}\left(t_i\right)\left(S_i \times \prod_{j \neq i}\left(\mathrm{proj}_{S_j}\left(\widehat{OP}_j^n \cap C_j\right) \times T_j\right)\right) \\ &\geq \beta_{i,\varnothing}\left(t_i\right)\left(S_i \times \prod_{j \neq i}\left(\widehat{OP}_j^n \cap C_j\right)\right) = 1, \end{aligned}$$

where the first inequality follows from the inductive hypothesis, the first and second equalities follow by definition, and the second inequality is immediate. This shows that ν_i satisfies the required properties. Since $i \in I$ is arbitrary, the conclusion follows.

Part (ii): Let $\mathcal{T}$ be complete. We prove the following claim: for every $i \in I$ and $n \in \mathbb{N}_0$,

$$W_i^n = \mathrm{proj}_{S_i}\left(\widehat{OP}_i^n \cap C_i\right).$$

The proof is by induction on $n \in \mathbb{N}_0$.

Basis step. Note that, for every $i \in I$,

$$\text{proj}_{S_i}\left(\widehat{OP}_i^0 \cap C_i\right) = \text{proj}_{S_i} C_i = S_i = W_i^0,$$

where the first equality holds because $\widehat{OP}_i^0 := S_i \times T_i$, the second equality follows from (7.1) in Appendix A, and the last equality holds by definition.

Inductive step. Assume that the result is true for $n \geq 0$. That is, the inductive hypothesis is

$$\forall i \in I, W_i^n = \text{proj}_{S_i}\left(\widehat{OP}_i^n \cap C_i\right). \tag{7.5}$$

To show that the result is also true for $n+1$, we first establish—as a consequence of the inductive hypothesis—the existence of a profile of maps $(\varphi_i)_{i \in I}$ satisfying some desirable properties.

By (7.5), it follows that, for each $i \in I$, if $s_i \in W_i^n$ then there exists $t_{s_i} \in T_i$ such that $(s_i, t_{s_i}) \in \widehat{OP}_i^n \cap C_i$. Thus, for each $i \in I$ and $s_i \in W_i^n$, we choose and fix some t_{s_i} satisfying this property. We also fix an arbitrary type $\bar{t}_i \in T_i$ and we define the map $\psi_i : S_i \to S_i \times T_i$ as

$$\psi_i(s_i) := \begin{cases} (s_i, t_{s_i}), & \text{if } s_i \in W_i^n, \\ (s_i, \bar{t}_i), & \text{if } s_i \in S_i \backslash W_i^n. \end{cases}$$

So, the profile of maps $(\psi_i)_{i \in I}$ satisfies the following property:

$$\forall i \in I, \psi_i(W_i^n) \subseteq \widehat{OP}_i^n \cap C_i. \tag{7.6}$$

Moreover, Remark 12 and (7.5) imply that $\widehat{OP}_i^n \neq \emptyset$ for every $i \in I$. Since the definition of the sets $\widehat{OP}_i^n$ $(i \in I)$ puts restrictions only on the type sets, we have that

$$\forall i \in I, S_i = \text{proj}_{S_i} \widehat{OP}_i^n.$$

This implies that, for each $i \in I$ and each $s_i \in S_i$, there exists $t_{s_i} \in T_i$ such that $(s_i, t_{s_i}) \in \widehat{OP}_i^n$. Hence, for each $i \in I$ and $s_i \in S_i$, we choose and fix some t_{s_i} satisfying this property. For each $i \in I$, we let $\gamma_i : S_i \to S_i \times T_i$ denote the map that associates each $s_i \in S_i$ with the chosen type t_{s_i}. So, the profile of maps $(\gamma_i)_{i \in I}$ satisfies the following property:

$$\forall i \in I, \gamma_i(S_i) \subseteq \widehat{OP}_i^n. \tag{7.7}$$

We can now define, for each $i \in I$, the map $\varphi_i : S_i \to S_i \times T_i$ as follows:

$$\varphi_i(s_i) := \begin{cases} \psi_i(s_i), & \text{if } s_i \in W_i^n, \\ \gamma_i(s_i), & \text{if } s_i \in S_i \backslash W_i^n. \end{cases}$$

It follows from (7.6) that $\varphi_i(W_i^n) \subseteq \widehat{OP}_i^n \cap C_i$ for every $i \in I$. Furthermore, (7.6) and (7.7) entail that $\varphi_i(S_i) \subseteq \widehat{OP}_i^n$ for every $i \in I$. Hence, we record, for future reference, the following properties of the profile of maps $(\varphi_i)_{i \in I}$:

$$\forall i \in I, W_i^n \subseteq (\varphi_i)^{-1}\left(\widehat{OP}_i^n \cap C_i\right), \tag{7.8}$$

and

$$\forall i \in I, S_i = (\varphi_i)^{-1}\left(\widehat{OP}_i^n\right). \tag{7.9}$$

Having done these preparations, we can now provide the proof of the inductive step.

Fix $i \in I$. Part (i) gives that $\mathrm{proj}_{S_i}\left(OP_i^{n+1} \cap C_i\right) \subseteq W_i^{n+1}$. For the converse, pick any $s_i \in W_i^{n+1}$. By definition, there exists $\nu_{s_i} \in \Delta^{\mathcal{S}_{-i}}(S_{-i})$ such that $s_i \in \rho_i(\nu_{s_i})$ and $\nu_{s_i}\left(W_{-i}^n | S_{-i}\right) = 1$. Consider the CPS $\mu_{s_i,-i} \in \Delta^{\mathcal{S}_{-i} \times T_{-i}}(S_{-i} \times T_{-i})$ defined as follows: for all events $E_{-i} \subseteq S_{-i} \times T_{-i}$ and $h \in H$,

$$\mu_{s_i,-i}(E_{-i} | S_{-i}(h) \times T_{-i}) := \nu_{s_i}\left(\left(\varphi_{-i}\right)^{-1}(E_{-i}) | S_{-i}(h)\right),$$

where $\varphi_{-i} := \left(\varphi_j\right)_{j \neq i}$. 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 ν_{s_i}. Furthermore, CPS $\mu_{s_i,-i}$ satisfies the following properties:

$$\forall h \in H, \mu_{s_i,-i}\left(\widehat{OP}_{-i}^n | S_{-i}(h) \times T_{-i}\right) = 1, \tag{7.10}$$

and

$$\mu_{s_i,-i}(C_{-i} | S_{-i} \times T_{-i}) = 1. \tag{7.11}$$

To see why (7.10) holds, note that, for all $h \in H$,

$$\begin{aligned} \mu_{s_i,-i}\left(\widehat{OP}_{-i}^n | S_{-i}(h) \times T_{-i}\right) &= \nu_{s_i}\left(\left(\varphi_{-i}\right)^{-1}\left(\widehat{OP}_{-i}^n\right) | S_{-i}(h)\right) \\ &= \nu_{s_i}(S_{-i} | S_{-i}(h)) \\ &= 1, \end{aligned}$$

where the first equality holds by definition, and the second equality follows from (7.9). Moreover, (7.11) holds because

$$\begin{aligned}\mu_{s_i,-i}\left(C_{-i}|S_{-i}\times T_{-i}\right) &= \nu_{s_i}\left(\left(\varphi_{-i}\right)^{-1}\left(C_{-i}\right)|S_{-i}\right)\\ &\geq \nu_{s_i}\left(\left(\varphi_{-i}\right)^{-1}\left(\widehat{OP}_{-i}^{n}\cap C_{-i}\right)|S_{-i}\right)\\ &\geq \nu_{s_i}\left(W_{-i}^{n}|S_{-i}\right)\\ &= 1,\end{aligned}$$

where the first equality holds by definition, the first inequality is obvious, and the second one follows from (7.8).

Consider the CPS $\mu_{s_i}\in\Delta^{\mathcal{S}\times T_{-i}}\left(S\times T_{-i}\right)$ defined as follows: for all $h\in H$,

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

By completeness, there exists $t_i\in T_i$ such that $\beta_i\left(t_i\right)=\mu_{s_i}$.

We now show that

$$\left(s_i,t_i\right)\in\widehat{OP}_i^{n+1}\cap C_i=OP_i\cap IBC_i\cap\left(\bigcap_{l=0}^{n}\mathrm{B}_i\left(\widehat{OP}_{-i}^{l}\right)\right)\cap C_i.$$

The proof showing that $\left(s_i,t_i\right)\in OP_i\cap C_i$ is the same as that in Lemma 6 of Appendix A. Next, we check that $\left(s_i,t_i\right)\in IBC_i:=\mathrm{B}_{i,\varnothing}\left(C_{-i}\right)$:

$$\begin{aligned}\beta_{i,\varnothing}\left(t_i\right)\left(S_i\times C_{-i}\right) &= \beta_i\left(t_i\right)\left(S_i\times C_{-i}|S\times T_{-i}\right)\\ &= \delta_{s_i}^{*}\left(S_i|S_i\right)\times\mu_{s_i,-i}\left(C_{-i}|S_{-i}\times T_{-i}\right)\\ &= 1,\end{aligned}$$

where the third equality follows from (7.11) and from the definition of $\delta_{s_i}^{*}$. It remains to show that $\left(s_i,t_i\right)\in\mathrm{B}_i\left(\widehat{OP}_{-i}^{n}\right):=\cap_{h\in H}\mathrm{B}_{i,h}\left(\widehat{OP}_{-i}^{n}\right)$; since the sequence $\left(\widehat{OP}_{-i}^{l}\right)_{l=0,1,\ldots,n}$ is decreasing, monotonicity of the operator $\mathrm{B}_i\left(\cdot\right)$ implies $\left(s_i,t_i\right)\in\cap_{l=0}^{n}\mathrm{B}_i\left(\widehat{OP}_{-i}^{l}\right)$. By (7.10), we obtain, for all $h\in H$,

$$\begin{aligned}\beta_{i,h}\left(t_i\right)\left(S_i\times\widehat{OP}_{-i}^{n}\right) &= \beta_i\left(t_i\right)\left(S_i\times\widehat{OP}_{-i}^{n}|S\left(h\right)\times T_{-i}\right)\\ &= \delta_{s_i}^{*}\left(S_i|S_i\left(h'\right)\right)\times\mu_{s_i,-i}\left(\widehat{OP}_{-i}^{n}|S_{-i}\left(h\right)\times T_{-i}\right)\\ &= 1.\end{aligned}$$

Therefore $(s_i, t_i) \in \widehat{OP}_i^{n+1} \cap C_i$, and so $s_i \in \text{proj}_{S_i}\left(\widehat{OP}_i^{n+1} \cap C_i\right)$. Since $i \in I$ is arbitrary, the proof of the inductive step is complete. ■

We can now provide the proof of Theorem 6.

Proof of Theorem 6. Parts (i)-(ii) of the main statement of the theorem immediately follow from Lemma 12 (i). The statement about complete type structures immediately follows from Lemma 12 (ii) as long as we consider finitely many steps, i.e., $n \in \mathbb{N}$. To see that it holds also in the limit, first note that $\text{proj}_S\left(\widehat{OP}^n \cap C\right) = W^n \neq \emptyset$ for every $n \in \mathbb{N}$, hence $\widehat{OP}^n \cap C \neq \emptyset$ for every $n \in \mathbb{N}$. Then

$$\text{proj}_S\left(\widehat{OP}^\infty \cap C\right) = \text{proj}_S \bigcap_{n=1}^{\infty}\left(\widehat{OP}^n \cap C\right) = \bigcap_{n=1}^{\infty} \text{proj}_S\left(\widehat{OP}^n \cap C\right) = \bigcap_{n=1}^{\infty} W^n,$$

where the first equality holds by definition, the second follows from Lemma 2 in Appendix A (as $\left(\widehat{OP}^n \cap C\right)_{n=1}^{\infty}$ is a decreasing sequence of closed and nonempty sets), and the third follows from Lemma 12. ■

Appendix B.3: Dynamic games with perfect recall

B.3.1 Games with perfect recall

A **finite game with perfect recall** is a structure

$$\left\langle I, \bar{X}, \iota, (A_i, u_i, H_i)_{i \in I} \right\rangle$$

given by the following elements:

- I is a nonempty finite set of **players**.
- For each $i \in I$, A_i is a nonempty finite set of potentially feasible **actions**. For each nonempty subset $J \subseteq I$, we let $A_J := \prod_{i \in J} A_i$ denote the set of action profiles for players in J.
- $\bar{X}$ is a finite set of finite sequences of action profiles, that is,

$$\bar{X} \subseteq \left(\bigcup_{\emptyset \neq J \subseteq I} A_J\right)^{<\mathbb{N}_0}.$$

$\bar{X}$ is closed with respect to the prefix-of relation: for every $y \in \bar{X}$, and every $x \prec y$, $x \in \bar{X}$; thus, $\varnothing \in \bar{X}$. Elements of $\bar{X}$ are called **histories**. A history is a sequence of action profiles $\bar{x} := (a^1, ..., a^n)$ such that, for every $m = 1, ..., n$, $a^m = (a_i^m)_{i \in J} \in A_J$ for some nonempty $J \subseteq I$.

- $\iota : \bar{X} \to 2^I$ is the **alert-player correspondence**, and it is such that

$$\left(a^1, ..., a^{n-1}, a^n\right) \in \bar{X}$$

only if $\iota\left(a^1, ..., a^{n-1}\right) \neq \emptyset$ and $a^n \in A_{\iota(a^1,...,a^{n-1})}$. Furthermore, the set of feasible action profiles at any $x \in \bar{X}$, viz. $A_{\iota(x)}(x) := \left\{a_{\iota(x)} : \left(x, a_{\iota(x)}\right) \in \bar{X}\right\}$, is a Cartesian product:

$$A_{\iota(x)}(x) = \prod_{i \in \iota(x)} A_i(x),$$

where $A_i(x) := \text{proj}_{A_i} A_{\iota(x)}(x)$ is the set of feasible actions of $i \in \iota(x)$ at x. We say that i is **alert** at x if $i \in \iota(x)$. The reason is that we allow the set of feasible actions $A_i(x)$ to be a singleton; in this case, i is alert but not active at x. If $|A_i(x)| \geq 2$, then i is **active** at x. We assume that *all players are alert at the empty history* (the beginning of the game): $\iota(\varnothing) = I$. We let $Z := \left\{\bar{x} \in \bar{X} : \iota(\bar{x}) = \emptyset\right\}$ denote the set of **terminal histories**, and $X := \bar{X} \backslash Z$ is the set of **nonterminal histories**.

- For each $i \in I$, $u_i : Z \to \mathbb{R}$ is the **payoff function** of player i.

- For each $i \in I$, H_i is i's **information partition** of the set

$$X_i := \{x \in X : i \in \iota(x)\}$$

of histories where i is alert; H_i is such that the feasibility correspondence $x \mapsto A_i(x)$ is H_i-measurable: for all $h_i \in H_i$ and $x, y \in h_i$, $A_i(x) = A_i(y)$. With this, we let $A_i(h_i)$ denote the set of feasible actions of i given information set h_i.

We assume that each H_i satisfies the *perfect recall* property: For every $x \in X_i$, let $\exp_i(x)$ denote i's experience along history x; that is, $\exp_i(x)$ is the ordered list of all information sets $h_i \in H_i$ encountered along the history x, and the actions i played there. Perfect recall requires that, for all $h_i \in H_i$ and $x, y \in X_i$, if $x, y \in h_i$, then $\exp_i(x) = \exp_i(y)$.

Note, it has to be the case that $\{\varnothing\} \in H_i$ for each i. Indeed, by assumption each i is alert at $\varnothing$; hence, $\varnothing \in h_i$ for some $h_i \in H_i$; with this, perfect recall implies

$h_i = \{\varnothing\}$. Our *interpretation* (which justifies our notation) is that h_i is determined by a *personal history*, that is, a sequence of actions played by i and messages about previous play received by i. A player who receives a message about previous play is alert. The commonly known rules of the game determine how messages are generated by previous play. Thus, a player who perfectly recalls (when alert) the sequence of actions he played and messages he received can infer what sequences of action profiles x are consistent with such personal history; his information set h_i comprises such histories. It follows that H_i has to satisfy the perfect recall property.

We let $H := \cup_{i \in I} H_i$, with typical element $h \in H$. We endow H with the strict precedence relation $\prec$ inherited from tree X. Thus, given $h, h' \in H$, we say that h' **strictly follows** h, and we write $h \prec h'$ or $h' \succ h$, if for every $y \in h'$ there exists $x \in h$ such that $x \prec y$. We say that h and h' are **simultaneous** if there is a history $x \in X$ such that $x \in h \cap h'$. We say that h' **weakly follows** h, and we write $h \preceq h'$ or $h' \succeq h$, if either $h \prec h'$ or h and h' are simultaneous. Note that $\preceq$ is *not* antisymmetric on H: that is, even if $h \preceq h'$ and $h' \preceq h$, h and h' may not be simultaneous—see the game in Figure B.1 below. However, by perfect recall, the restriction of $\preceq$ on each H_i is antisymmetric: for every $h_i, h_i' \in H_i$, if $h_i \preceq h_i'$ and $h_i' \preceq h_i$, then $h_i = h_i'$.

Structure $\left\langle I, \bar{X}, \iota, (A_i, u_i, H_i)_{i \in I} \right\rangle$ is a **multistage** game if H has the following property: for every $h \in H$ and $x, y \in h$, x and y must have the same length. This means that the players always know how many action profiles have already been chosen. If each $h \in H$ is a singleton, then $\left\langle I, \bar{X}, \iota, (A_i, H_i)_{i \in I} \right\rangle$ is a game structure with **observable actions**.[43]

B.3.2 External states

For each $i \in I$, let $S_i := \prod_{h_i \in H_i} A_i(h_i)$ 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 $S_{-i} := \prod_{j \in I \setminus \{i\}} S_j$.

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. With this, we find it convenient to define the set of external states reaching a nonterminal history. For each $x \in X$, let $S(x)$ denote the set of external states inducing x:

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

[43] The definition of game with observable actions in the main text is a special case, as it is assumed that players are always alert.

The projection

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

is the set of external states of i that allow x. Similarly, the projection

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

is the set of profiles of external states of players other than i that allow x. Note that, for every $x \in X$,

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

Similarly, for each $h \in H$, let $S(h)$ denote the set of external states inducing h:

$$S(h) := \bigcup_{x \in h} S(x) = \{s \in S : \exists x \in h, x \prec \zeta(s)\},$$

while the projections

$$\begin{aligned} S_i(h) &: = \bigcup_{x \in h} S_i(x), \\ S_{-i}(h) &: = \bigcup_{x \in h} S_{-i}(x), \end{aligned}$$

are, respectively, the sets of profiles of external states of i and of his co-players that allow h.[44]

Perfect recall implies the following factorization: for each player i and each information set $h_i \in H_i$,

$$S(h_i) = S_i(h_i) \times S_{-i}(h_i).$$

That is, the information about behavior encoded in $h_i \in H_i$ can be decomposed into information about own behavior and information about the co-players' behavior, because i remembers what he knew and did at earlier histories. Furthermore, perfect recall also implies the following property: for all $i \in I$, $h_i \in H_i$ and $x, y \in h_i$,

$$S_i(x) = S_i(y).$$

[44]The following equalities are immediate by inspection of definitions:

$$\begin{aligned} S_i(h) &= \{s_i \in S_i : \exists s_{-i} \in S_{-i}, (s_i, s_{-i}) \in S(h)\}, \\ S_{-i}(h) &= \{s_{-i} \in S_{-i} : \exists s_i \in S_i, (s_i, s_{-i}) \in S(h)\}. \end{aligned}$$

As in the main text,

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

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

The following example illustrates a game with imperfect monitoring of past actions and without a multistage structure. Consider the game Γ represented in Figure B.1.

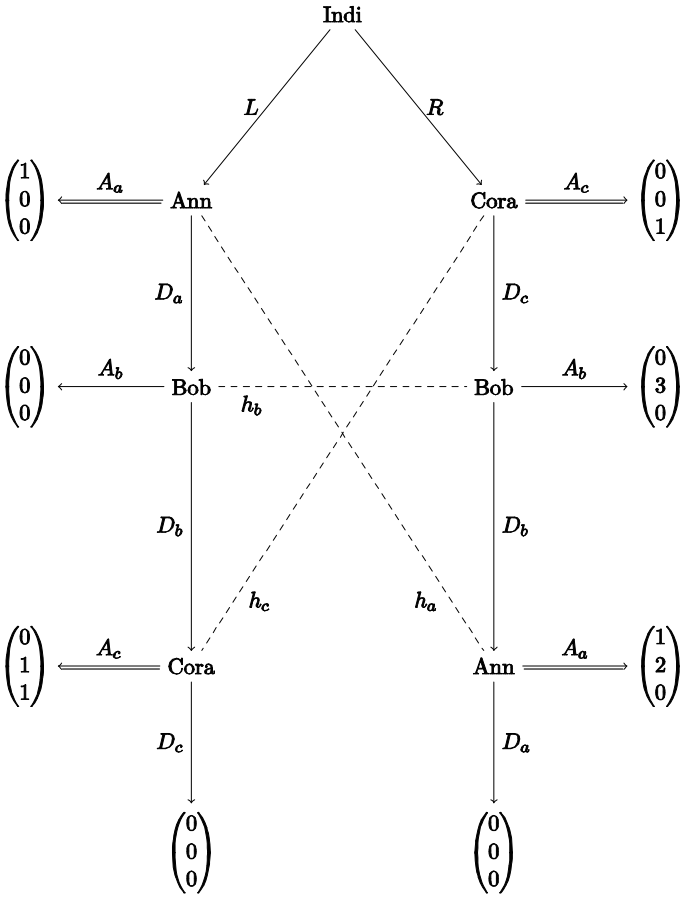


Figure B.1: A game with unobservable actions

At the beginning of the game, Indi chooses *L*eft (L) or *R*ight (R). If she chooses L (resp. R), then Ann (resp. Cora) is called to make a choice. Both Ann and Cora can choose between two possible actions, *A*cross and *D*own—such actions are denoted by A_a and D_a for Ann, and by A_c and D_c for Cora.

However, Indi's action is *not* observable. For instance, if Ann is called to choose,

then she does not know the sequence of action profiles: the first possibility is that Indi chose L, so that Ann is the second player to move in the game. The other possibility is that Indi chose R: in this case, if Cora chose D_c, then Bob can choose between *A*cross (A_b) and *D*own (D_b); and if Bob chose D_b, then Ann is the player that is called to move. Analogous considerations hold for Cora.

The information partitions of Ann and Cora are, respectively, $H_a = \{\{\varnothing\}, h_a\}$ and $H_c = \{\{\varnothing\}, h_c\}$, where

$$h_a = \{(L), (R, D_c, D_b)\} \quad \text{and} \quad h_c = \{(R), (L, D_a, D_b)\}.$$

The information partition of Bob is $H_b = \{\{\varnothing\}, h_b\}$, where

$$h_b = \{(L, D_a), (R, D_c)\};$$

in words, if he is called to move, Bob does not know if the second player was Ann or Cora. Note that $h_i \preceq h_j$ for every $i, j \in \{a, b, c\}$ such that $i \neq j$, but they are not simultaneous.

Indi's payoffs do not appear in Figure B.1 (she is *indi*fferent). The numbers in the column vectors associated with terminal histories are the payoffs of Ann, Bob and Cora; for instance, at history (R, D_c, D_b, A_a), Ann gets 1, Bob gets 2, and Cora gets 0. Note that A_a and A_c are dominant actions; see Figure B.1, where underlined arcs of Ann and Cora represent their optimal planned actions. Hence, if Bob is called to choose, then he is certain that either Ann or Cora deviated from the optimal plan.

B.3.2 Conditional beliefs and type structures

First-order beliefs of player i are CPSs on $(S, \mathcal{S}_i)$, where $\mathcal{S}_i$ is the collection of conditioning events about behavior corresponding to information sets:

$$\mathcal{S}_i := \{F \subseteq S : \exists h_i \in H_i, F = S(h_i)\}.$$

For any $i \in I$, let T_{-i} denote the set of possible "types" of the other players; then the conditioning event for i corresponding to $h_i \in H_i$ is $S(h_i) \times T_{-i}$.

With this, a Γ-based **type structure** is a tuple

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

such that, for every $i \in I$, the type set T_i is a compact metrizable space, and the belief map $\beta_i : T_i \to \Delta^{\mathcal{S}_i \times T_{-i}}(S \times T_{-i})$ is continuous. As in the main text, a **personal state** of player i is a pair $(s_i, t_i) \in S_i \times T_i$, and a **state of the world** is a profile $(s_i, t_i)_{i \in I} \in \prod_{i \in I}(S_i \times T_i)$. Moreover, we will write $\beta_{i,h_i}(t_i)$ to denote the beliefs of type t_i conditional on information set $h_i \in H_i$, that is,

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

B.3.3 Independence, optimal planning and rationality

The notion of independence given in the main text naturally extends to the current framework. To see this, we first introduce some notation. Let $\mathcal{S}_{i,-i}$ denote the collection of conditioning events for player i about the behavior of the co-players; similarly, let $\mathcal{S}_{i,i}$ denote the collection of conditioning events about i's behavior. Formally,

$$\begin{aligned} \mathcal{S}_{i,-i} &: = \left\{F \subseteq S_{-i} : \exists h_i \in H_i, F = S_{-i}\left(h_i\right)\right\}, \\ \mathcal{S}_{i,i} &: = \left\{F \subseteq S_i : \exists h_i \in H_i, F = S_i\left(h_i\right)\right\}. \end{aligned}$$

A type t_i in a Γ-based type structure $\mathcal{T}$ satisfies independence if there are two CPSs $\beta_{i,i}\left(t_i\right) \in \Delta^{\mathcal{S}_{i,i}}\left(S_i\right)$ and $\beta_{i,-i}\left(t_i\right) \in \Delta^{\mathcal{S}_{i,-i} \times T_{-i}}\left(S_{-i} \times T_{-i}\right)$ such that

$$\forall h_i \in H_i, \beta_i\left(t_i\right)\left(\cdot | S\left(h_i\right) \times T_{-i}\right) = \beta_{i,i}\left(t_i\right)\left(\cdot | S_i\left(h_i\right)\right) \times \beta_{i,-i}\left(t_i\right)\left(\cdot | S_{-i}\left(h_i\right) \times T_{-i}\right).$$

We do not restate definitions and concepts introduced in Sections 4.1-4.3 for this environment, since the only required changes are easy to identify; essentially, histories need to be replaced by information sets. We describe in detail only those few cases where extra care is required in the notation.

For each $x \in X$ and each $i \in I$, we let

$$H_i^{\succeq}\left(x\right) := \left\{h_i \in H_i : \exists y \in h_i, x \preceq y\right\}$$

denote the set of information sets of player i that weakly follow history x.

Fix a Γ-based type structure $\mathcal{T}$. We say that player i is **consistent from** history x at personal state $\left(s_i, t_i\right)$ of a Γ-based type structure $\mathcal{T}$ if t_i satisfies independence and $\sigma_{t_i,i}\left(s_i\left(h_i\right) | h_i\right) = 1$ for all $h_i \in H_i^{\succeq}\left(x\right)$; player i is **consistent** at $\left(s_i, t_i\right)$ if he is consistent from the empty history $\varnothing$; player i is **rational** at $\left(s_i, t_i\right)$ if he is consistent at $\left(s_i, t_i\right)$ and type t_i plans optimally.

The sets of personal states where i is consistent from history x, consistent, and rational are respectively denoted by

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

Note: if $H_i^{\succeq}\left(x\right) = \emptyset$, then $C_i^{\succeq x} = S_i \times T_i^*$, where T_i^* denotes the set of i's types that satisfy independence; this occurs when player i is not alert at x and at every history $y \succeq x$. As in the main text, it can be shown that these sets are events, because they are closed.

B.3.4 Backwards rationalizability

We introduce the solution concept of backwards rationalizability for the class of games with perfect recall. For every $x \in X$, let $\chi^x : \mathcal{Q} \to \mathcal{Q}$ be the operator defined as follows: for all $i \in I$ and $Q \in \mathcal{Q}$, if $H_i^{\succ}(x) = \emptyset$, then

$$\chi_i^x(Q_i) := S_i(x);$$

if $H_i^{\succ}(x) \neq \emptyset$, then

$$\chi_i^x(Q_i) := \left\{ s_i \in S_i(x) : \exists \bar{s}_i \in Q_i, \forall h_i \in H_i^{\succ}(x), s_i(h_i) = \bar{s}_i(h_i) \right\}.$$

With this, let

$$\begin{aligned} \chi_{-i}^x(Q_{-i}) &:= \prod_{j \neq i} \chi_j^x(Q_j), \\ \chi^x(Q) &:= \prod_{i \in I} \chi_i^x(Q_i). \end{aligned}$$

In words, if $H_i^{\succ}(x) \neq \emptyset$, then each $\chi_i^x(Q_i)$ is the set of all $s_i \in S_i(x)$ whose projection onto $H_i^{\succ}(x)$ coincides with the projection onto $H_i^{\succ}(x)$ of some $\bar{s}_i \in Q_i$. Note that $\chi^{\varnothing}(Q) = Q$ because $H_i^{\succ}(\varnothing) = H_i \neq \emptyset$ for every $i \in I$, and $\chi^x(S) = S(x)$ for all $x \in X$.

Definition 11 *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 $\mu_i \in \Delta^{\mathcal{S}_{i,-i}}(S_{-i})$ such that*

1. *$s_i \in \rho_i(\mu_i)$;*
2. *for every $h_i \in H_i$ and every $x \in h_i$,*

$$\mu_i\left(S_{-i}(x) \,|S_{-i}(h_i)\right) > 0 \Rightarrow \frac{\mu_i\left(\chi_{-i}^x\left(\hat{S}_{-i}^{n-1}\right) |S_{-i}(h_i)\right)}{\mu_i\left(S_{-i}(x) \,|S_{-i}(h_i)\right)} = 1.$$

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****.*

Definition 11 is a generalization of the definition of backwards rationalizability given in the main text (Definition 6). The difference relies on Condition 2 of the recursive step. To elaborate, consider an arbitrary $h_i \in H_i$. Every μ_i justifying $s_i \in \hat{S}_i^n$ must satisfy the following requirement: if a history $x \in h_i$ is considered possible under μ_i (i.e., if $\mu_i\left(S_{-i}\left(x\right)|S_{-i}\left(h_i\right)\right) > 0$), then, conditional on x, CPS μ_i assigns probability 1 to the continuations of behaviors from $\hat{S}_{-i}^{n-1}$. To see this, note that

$$\frac{\mu_i\left(\chi_{-i}^x\left(\hat{S}_{-i}^{n-1}\right)|S_{-i}\left(h_i\right)\right)}{\mu_i\left(S_{-i}\left(x\right)|S_{-i}\left(h_i\right)\right)} = \frac{\mu_i\left(\chi_{-i}^x\left(\hat{S}_{-i}^{n-1}\right) \cap S_{-i}\left(x\right)|S_{-i}\left(h_i\right)\right)}{\mu_i\left(S_{-i}\left(x\right)|S_{-i}\left(h_i\right)\right)} = 1,$$

since, by definition, $\chi_{-i}^x\left(\hat{S}_{-i}^{n-1}\right)$ is a subset of $S_{-i}\left(x\right)$. In other words, only the histories/nodes in an information set with strictly positive conditional probability matter, and we can determine a well defined belief conditional on each one of those nodes. This belief assigns probability 1 to the continuation from $\hat{S}_{-i}^{n-1}$ conditional on such nodes.

If Γ is a multistage game with observable actions and the players are always alert, then each information set h_i is a singleton, and H_i is identified with X; in this case, Definition 11 coincides with Definition 6.

To illustrate the procedure, refer back to the game in Figure B.1. The set of backwards rationalizable profiles is

$$\hat{S}^\infty = \{L, R\} \times \{A_a\} \times \{A_b, D_b\} \times \{A_c\}.$$

Indi is indifferent, while A_a and A_c are dominant actions for Ann and Cora, respectively. Hence, if Bob is called to move, then he is certain that either Ann or Cora deviated from the optimal plan. To see why both A_b and D_b are backwards rationalizable, let us formally write the conditioning event for Bob as follows:

$$\begin{aligned} S_{-b}\left(h_b\right) &= \cup_{x \in h_b} S_{-b}\left(x\right) \\ &= S_{-b}\left(L, D_a\right) \cup S_{-b}\left(R, D_c\right) \\ &= \left(\{L\} \times \{D_a\} \times \{A_c, D_c\}\right) \cup \left(\{R\} \times \{A_a, D_a\} \times \{D_c\}\right). \end{aligned}$$

Consider a CPS μ_b such that

$$\begin{aligned} \mu_b\left(S_{-b}\left(L, D_a\right)|S_{-b}\left(h_b\right)\right) &= \mu_b\left(\{(L, D_a, A_c)\}|S_{-b}\left(h_b\right)\right) > 0, \\ \mu_b\left(S_{-b}\left(R, D_c\right)|S_{-b}\left(h_b\right)\right) &= \mu_b\left(\{(R, A_a, D_c)\}|S_{-b}\left(h_b\right)\right) > 0. \end{aligned}$$

Then μ_b satisfies part 2 of the recursive step in Definition 11, because $\chi_{-b}^{(L,D_a)}\left(\hat{S}_{-b}^\infty\right) = \{(L, D_a, A_c)\}$ and $\chi_{-b}^{(R,D_c)}\left(\hat{S}_{-b}^\infty\right) = \{(R, A_a, D_c)\}$.

Let $\bar{\mu} := \mu_b\left(S_{-b}\left(L, D_a\right) | S_{-b}\left(h_b\right)\right)$. Then,

$$\begin{aligned}\mathbb{E}_{\mu_b}\left[U_b\left(A_b, \cdot\right) | h_b\right] &= 3\left(1-\bar{\mu}\right), \\ \mathbb{E}_{\mu_b}\left[U_b\left(D_b, \cdot\right) | h_b\right] &= \bar{\mu} + 2\left(1-\bar{\mu}\right).\end{aligned}$$

If $\bar{\mu} < 1/2$, then A_b is optimal under μ_b; if $\bar{\mu} \geq 1/2$, then D_b is optimal under μ_b. Therefore, both A_b and D_b are backwards rationalizable.

Consider a small modification of the game in Figure B.1. That is, the only change is that, at history (R, D_c, D_b, A_a), the payoff for Bob is 4. In this case, A_b is not backwards rationalizable. Indeed, every CPS μ_b such that $\bar{\mu} > 0$ yields $\mathbb{E}_{\mu_b}\left[U_b\left(D_b, \cdot\right) | h_b\right] = \bar{\mu} + 3\left(1-\bar{\mu}\right) > 3\left(1-\bar{\mu}\right) = \mathbb{E}_{\mu_b}\left[U_b\left(A_b, \cdot\right) | h_b\right]$, and A_b is not optimal under a CPS μ_b such that $\bar{\mu} = 0$.

We now show that backwards rationalizability—as per Definition 11—is always consistent with epistemic assumptions that generalize those stated in the main text. Fix a Γ-based type structure $\mathcal{T}$. We say that player i **believes** at state (s_i, t_i) **in** the **continuation consistency** of the other players if, for every $h_i \in H_i$ and every $x \in h_i$,

$$\beta_{i,h_i}\left(S\left(x\right) \times T_{-i}\right) > 0 \Rightarrow \frac{\beta_{i,h_i}\left(t_i\right)\left(\left(S_i \times C_{-i}^{\succeq x}\right) \cap \left(S\left(x\right) \times T_{-i}\right)\right)}{\beta_{i,h_i}\left(S\left(x\right) \times T_{-i}\right)} = 1,$$

where $C_{-i}^{\succeq x} := \prod_{j \neq i} C_j^{\succeq x}$. In words, player i believes in the co-players' consistency starting from every history/node x he deems possible conditional on the information set containing x. If Γ is a multistage game with observable actions and the players are always alert, then each information set h_i is a singleton, and H_i is identified with X; in this case, the above definition of belief in continuation consistency coincides with the one given in the main text.

We let BCC_i denote the set of personal external states of player i where belief in continuation consistency holds, and

$$BCC := \prod_{i \in I} BCC_i.$$

Remark 14 *BCC is an event of $S \times T$.*

Proof. Fix $h_i \in H_i$ and $x \in h_i$ arbitrarily. We first show that the following sets are events:

$$\begin{aligned}E_x &:= \left\{\mu_i \in \Delta^{\mathcal{S}_i}\left(S \times T_{-i}\right) : \mu_i\left(S\left(x\right) \times T_{-i} | S\left(h_i\right) \times T_{-i}\right) > 0\right\}, \\ F_x &:= \left\{\mu_i \in \Delta^{\mathcal{S}_i}\left(S \times T_{-i}\right) : \begin{array}{l}\mu_i\left(\left(S_i \times C_{-i}^{\succeq x}\right) \cap \left(S\left(x\right) \times T_{-i}\right) | S\left(h_i\right) \times T_{-i}\right) \\ = \mu_i\left(S\left(x\right) \times T_{-i} | S\left(h_i\right) \times T_{-i}\right)\end{array}\right\}.\end{aligned}$$

Recall that if O is an open subset of a metrizable space X, then

$$\{\mu \in \Delta(X) : \mu(O) > 0\}$$

is an open subset of $\Delta(X)$ (see Aliprantis and Border 2006, Corollary 15.6). This implies that, if O is an open subset of X, then

$$\{\mu \in \Delta(X) : \mu(O) = 0\}$$

is a closed subset of $\Delta(X)$. Since $S(x) \times T_{-i}$ is (cl)open in $S \times T_{-i}$,

$$\left\{\mu_i \in (\Delta(S \times T_{-i}))^{S_i} : \mu_i(S(x) \times T_{-i} | S(h_i) \times T_{-i}) > 0\right\}$$

is open, hence E_x is open in (the relative topology of) $\Delta^{S_i}(S \times T_{-i})$. Next, define the following sets:

$$\begin{aligned} M_x &: = \left(S_i \times C_{-i}^{\succeq x}\right) \cap (S(x) \times T_{-i}), \\ N_x &: = S(x) \times T_{-i}. \end{aligned}$$

Clearly $M_x \subseteq N_x$, and M_x is a closed subset of $S \times T_{-i}$, because $C_{-i}^{\succeq x}$ is closed in $S_{-i} \times T_{-i}$. As N_x is (cl)open in $S \times T_{-i}$, we conclude that $N_x \backslash M_x$ is open in $S \times T_{-i}$. Note that F_x can be written as

$$F_x = \left\{\mu_i \in \Delta^{S_i}(S \times T_{-i}) : \mu_i(N_x \backslash M_x | S(h_i) \times T_{-i}) = 0\right\}.$$

Since

$$\left\{\mu_i \in (\Delta(S \times T_{-i}))^{S_i} : \mu_i(N_x \backslash M_x | S(h_i) \times T_{-i}) = 0\right\}$$

is closed, it follows that F_x is closed in (the relative topology of) $\Delta^{S_i}(S \times T_{-i})$. Hence $G_x = E_x \cap F_x$ is Borel in $\Delta^{S_i}(S \times T_{-i})$.

With this, we can now claim that $\cap_{x \in h_i} G_x$ is a Borel subset of $\Delta^{S_i}(S \times T_{-i})$; continuity of the belief map implies that the set

$$\begin{aligned} BCC_i &= \bigcap_{h_i \in H_i} \left(S_i \times \left(\left(\beta_{i,h_i}\right)^{-1} \left(\bigcap_{x \in h_i} G_x\right)\right)\right) \\ &= S_i \times \bigcap_{h_i \in H_i} \left\{ t_i \in T_i : \begin{array}{l} \forall x \in h_i, \\ \beta_{i,h_i}(S(x) \times T_{-i}) > 0 \Rightarrow \frac{\beta_{i,h_i}(t_i)\left(\left(S_i \times C_{-i}^{\succeq x}\right) \cap (S(x) \times T_{-i})\right)}{\beta_{i,h_i}(S(x) \times T_{-i})} = 1 \end{array} \right\} \end{aligned}$$

is Borel in $S_i \times T_i$. Thus, BCC is a Borel subset of $S \times T$. ■

We can therefore define recursively the following epistemic events. First, let $OP_i^0 := S_i \times T_i$ for each $i \in I$. The sets OP^0 and OP_{-i}^0 are defined in the obvious way: $OP^0 := \prod_{i \in I} OP_i^0$ and $OP_{-i}^0 := \prod_{j \neq i} OP_j^0$. Then:

- $OP_i^1 := OP_i \cap BCC_i$,
- $OP_i^{m+1} := OP_i^m \cap \mathrm{B}_i\left(OP_{-i}^m\right)$, 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, $OP^m := \prod_{i\in I} OP_i^m$. Finally, we let $OP^\infty := \cap_{m\in\mathbb{N}} OP^m$.

We have a partial analogue of Theorem 2.

Theorem 7 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Then,*
(i) for every $n \in \mathbb{N}$, $\mathrm{proj}_S(OP^n \cap C) \subseteq \hat{S}^n$;
(ii) $\mathrm{proj}_S(OP^\infty \cap C) \subseteq \hat{S}^\infty$.

To prove Theorem 7, we need to adapt the results in Appendix A. For the reader's convenience, we provide a self-contained proof of Theorem 7, with all the required modifications of the results in Appendix A.

Fix a finite game Γ. For a given $x \in X$, let

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

where $s_i^{h_i}$ denotes the minimal modification of s_i that allows h_i. Note that $\rho_i^{\succeq \varnothing}(\mu_i) = \rho_i(\mu_i)$.

Remark 15 *Fix $x \in X$ and a CPS μ_i on $(S_{-i}, \mathcal{S}_{i,-i})$.*
(i) If $s_i \in \rho_i(\mu_i)$, then $s_i \in \rho_i^{\succeq x}(\mu_i)$.
(ii) If $s_i \in \rho_i^{\succeq x}(\mu_i)$, then there exists $\bar{s}_i \in S_i$ such that $\bar{s}_i \in \rho_i(\mu_i)$ and $s_i(h_i) = \bar{s}_i(h_i)$ for all $h_i \in H_i^{\succeq}(x)$.

Lemma 13 *For every $i \in I$, $x \in X$ and $n \in \mathbb{N}$,*

$$\chi_i^x\left(\hat{S}_i^n\right) = \left\{ \begin{array}{ll} s_i \in S_i(x): & \exists \mu_i \in \Delta^{\mathcal{S}_{i,-i}}(S_{-i}), \\ & 1)\ s_i \in \rho_i^{\succeq x}(\mu_i), \\ & 2)\ \forall h_i \in H_i, \forall y \in h_i, \\ & \mu_i(S_{-i}(y)\,|S_{-i}(h_i)) > 0 \Rightarrow \frac{\mu_i\left(\chi_{-i}^y\left(\hat{S}_{-i}^{n-1}\right)|S_{-i}(h_i)\right)}{\mu_i(S_{-i}(y)|S_{-i}(h_i))} = 1 \end{array} \right\}.$$

Proof. Pick any $s_i \in \chi_i^x\left(\hat{S}_i^n\right)$. Then, by definition, there exists $\bar{s}_i \in \hat{S}_i^n$ such that $s_i(h_i) = \bar{s}_i(h_i)$ for all $h_i \in H_i^{\succeq}(x)$. Hence $\bar{s}_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta^{\mathcal{S}_{i,-i}}(S_{-i})$ satisfying condition 2 in Definition 11. Remark 15.(i) entails that $\bar{s}_i \in \rho_i^{\succeq x}(\mu_i)$, and since $\bar{s}_i$ coincides with s_i at all $h_i \in H_i^{\succeq}(x)$, we have $s_i \in \rho_i^{\succeq x}(\mu_i)$.

For the other direction, pick any $s_i \in S_i(x)$ such that $s_i \in \rho_i^{\succeq x}(\mu_i)$ for some $\mu_i \in \Delta^{\mathcal{S}_{i,-i}}(S_{-i})$ satisfying condition 2 in Definition 11. By Remark 15.(ii), there exists $\bar{s}_i \in S_i$ such that $\bar{s}_i \in \rho_i(\mu_i)$ and $s_i(h_i) = \bar{s}_i(h_i)$ for all $h_i \in H_i^{\succeq}(x)$. By Definition 11, $\bar{s}_i \in \hat{S}_i^n$. Hence $s_i \in \chi_i^x\left(\hat{S}_i^n\right)$. ■

The following lemma is the analogue of Lemma 5.

Lemma 14 *Fix a finite game Γ and a Γ-based type structure $\mathcal{T}$. Then, for all $n \in \mathbb{N}_0$ and $x \in X$,*

$$\chi^x\left(\text{proj}_S\left(OP^n \cap C\right)\right) \subseteq \chi^x\left(\hat{S}^n\right).$$

Proof. We first prove the following auxiliary result:

Claim 4 *Fix $n \in \mathbb{N}_0$ and $x \in X$. Then*

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

Proof of Claim 4. First note that $C_i \subseteq C_i^{\succeq x}$ and $\chi_i^x\left(\text{proj}_{S_i}\left(OP_i^n \cap C_i\right)\right) \subseteq S_i(x)$ for each $i \in I$. Consequently, if $OP_i^n \cap C_i$ or $OP_i^n \cap C_i^{\succeq x}$ are empty, then the result is immediate. So in what follows we will assume that $OP_i^n \cap C_i$ is nonempty. Pick any $s_i \in \chi_i^x\left(\text{proj}_{S_i}\left(OP_i^n \cap C_i\right)\right)$. Then $s_i \in S_i(x)$, and so we only need to show the existence of $t_i \in T_i$ such that $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq x}$; this will imply $s_i \in \text{proj}_{S_i}\left(OP_i^n \cap C_i^{\succeq x}\right) \cap S_i(x)$, as required. By definition of $\chi_i^x(\cdot)$, there exists $\bar{s}_i \in \text{proj}_{S_i}\left(OP_i^n \cap C_i\right)$ such that $s_i(h_i) = \bar{s}_i(h_i)$ for all $h_i \in H_i^{\succeq}(x)$. 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 $(\bar{s}_i, t_i)$ entail that $\bar{s}_i \in \rho_i(\nu_i)$, where $\nu_i := \text{marg}_{S_{-i}}\beta_i(t_i)$. Remark 15.(i) implies that $\bar{s}_i \in \rho_i^{\succeq x}(\nu_i)$, and since $\bar{s}_i(h_i) = s_i(h_i)$ for every $h_i \in H_i^{\succeq}(x)$, we obtain $s_i \in \rho_i^{\succeq x}(\nu_i)$. Therefore $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq x}$. □

We now prove the following claim:

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

With this, the result follows from Claim 4. The proof is by induction on $n \in \mathbb{N}_0$.

Basis step. Note that, for every $i \in I$ and $x \in X$,

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

so the result follows immediately.

Inductive step. Assume that the result is true for $n \geq 0$. We show that it is also true for $n+1$.

Fix $i \in I$ and $x \in X$ arbitrarily. Pick any $s_i \in \text{proj}_{S_i}\left(OP_i^{n+1} \cap C_i^{\succeq x}\right) \cap S_i(x)$, so that $(s_i, t_i) \in OP_i^{n+1} \cap C_i^{\succeq x}$ for some $t_i \in T_i$. Since $OP_i^{n+1} \subseteq OP_i^n$, it follows that $(s_i, t_i) \in OP_i^n \cap C_i^{\succeq x}$, and so, by the inductive hypothesis, $s_i \in \chi_i^x\left(\hat{S}_i^n\right)$. By Remark 15.(i), $s_i \in \rho_i^{\succeq x}(\nu_i)$, where $\nu_i := \text{marg}_{S_{-i}}\beta_i(t_i)$. So, in order to show that $s_i \in \chi^x\left(\hat{S}_i^{n+1}\right)$, it is enough to show (by Lemma 13) that

$$\frac{\nu_i\left(\chi_{-i}^y\left(\hat{S}_{-i}^n\right)|S_{-i}(h_i)\right)}{\nu_i\left(S_{-i}(y)\,|S_{-i}(h_i)\right)} = 1$$

for every $h_i \in H_i$ and for every $y \in h_i$ such that $\nu_i\left(S_{-i}(y)\,|S_{-i}(h_i)\right) > 0$.

To this end, first note that $(s_i, t_i) \in OP_i^{n+1}$ implies $(s_i, t_i) \in \text{B}_i\left(OP_{-i}^n\right) := \cap_{h_i \in H_i}\text{B}_{i,h_i}\left(OP_{-i}^n\right)$. Note also that $(s_i, t_i) \in BCC_i$, hence

$$\frac{\beta_{i,h_i}(t_i)\left(\left(S_i \times \left(OP_{-i}^n \cap C_{-i}^{\succeq y}\right)\right) \cap \left(S(y) \times T_{-i}\right)\right)}{\beta_{i,h_i}(t_i)\left(S(y) \times T_{-i}\right)} = 1$$

for every $h_i \in H_i$ and for every $y \in h_i$ such that $\beta_{i,h_i}(t_i)\left(S(y) \times T_{-i}\right) > 0$. Using this fact, for every $h_i \in H_i$ and for every $y \in h_i$ such that $\beta_{i,h_i}(t_i)\left(S(y) \times T_{-i}\right) > 0$,

we have

$$\begin{aligned}
& \frac{\nu_i\left(\chi_{-i}^{y}\left(\hat{S}_{-i}^{n}\right)|S_{-i}(h_i)\right)}{\nu_i\left(S_{-i}(y)|S_{-i}(h_i)\right)} \\
\geq\ & \frac{\nu_i\left(\prod_{j\neq i}\left(\mathrm{proj}_{S_j}\left(OP_j^n\cap C_j^{\succeq y}\right)\cap S_j(y)\right)|S_{-i}(h_i)\right)}{\nu_i\left(S_{-i}(y)|S_{-i}(h_i)\right)} \\
=\ & \frac{\mathrm{marg}_{S_{-i}}\beta_{i,h_i}(t_i)\left(\prod_{j\neq i}\left(\mathrm{proj}_{S_j}\left(OP_j^n\cap C_j^{\succeq y}\right)\cap S_j(y)\right)\right)}{\mathrm{marg}_{S_{-i}}\beta_{i,h_i}(t_i)\left(\mathrm{proj}_{S_{-i}}\left(S(y)\times T_{-i}\right)\right)} \\
=\ & \frac{\beta_{i,h_i}(t_i)\left(S_i\times\left(\prod_{j\neq i}\left(\mathrm{proj}_{S_j}\left(OP_j^n\cap C_j^{\succeq y}\right)\times T_j\right)\cap\prod_{j\neq i}\left(S_j(y)\times T_j\right)\right)\right)}{\beta_{i,h_i}(t_i)\left(S(y)\times T_{-i}\right)} \\
\geq\ & \frac{\beta_{i,h_i}(t_i)\left(S_i\times\left(\prod_{j\neq i}\left(OP_j^n\cap C_j^{\succeq y}\right)\cap\prod_{j\neq i}\left(S_j(y)\times T_j\right)\right)\right)}{\beta_{i,h_i}(t_i)\left(S(y)\times T_{-i}\right)} \\
\geq\ & \frac{\beta_{i,h_i}(t_i)\left(\left(S_i\times\left(OP_{-i}^n\cap C_{-i}^{\succeq y}\right)\right)\cap\left(S(y)\times T_{-i}\right)\right)}{\beta_{i,h_i}(t_i)\left(S(y)\times T_{-i}\right)} \\
=\ & 1,
\end{aligned}$$

where the first inequality follows from the inductive hypothesis, the first and second equalities follow by definition, the second inequality is immediate, and the third inequality holds because $S(y)\subseteq S_i\times S_{-i}(y)$. This shows that ν_i satisfies the required properties. Since $i\in I$ and $x\in X$ are arbitrary, the conclusion follows. ■

Proof of Theorem 7. Immediate from Lemma 14 (with $x=\varnothing$). ■

References

[1] Aliprantis, C., and K. Border (2006): *Infinite Dimentional Analysis.* Berlin: Springer-Verlag.

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

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

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

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

[6] Bach, C.W., and C. Heilmann (2011): "Agent Connectedness and Backward Induction," *International Game Theory Review,* 13, 195-208.

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

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

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

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

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

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

[13] Battigalli, P., and M. Dufwenberg (2020): "Belief-Dependent Motivations and Psychological Game Theory," *Journal of Economic Literature*, forthcoming.

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

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

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

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

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

[19] Battigalli, P., and P. Tebaldi (2019): "Interactive Epistemology in Simple Dynamic Games with a Continuum of Strategies," *Economic Theory*, 68, 737-763.

[20] Battigalli, P., R. Corrao, and F. Sanna (2020): "Epistemic Game Theory Without Type Structures. An Application to Psychological Games," *Games and Economic Behavior,* 120, 28-57.

[21] 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.

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

[23] Battigalli, P., E. Catonini, G. Lanzani, and M. Marinacci (2019b): "Ambiguity Attitudes and Self-Confirming Equilibrium in Sequential Games," forthcoming, *Games and Economic Behavior.*

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

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

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

[27] Bonanno, G. (2014): "A Doxastic Behavioral Characterization of Generalized Backward Induction," *Games and Economic Behavior,* 88, 221–241.

[28] Chen, J., and S. Micali (2012): "The Order Independence of Iterated Dominance in Extensive Games, with Connections to Mechanism Design and Backward Induction," Technical Report 2012-023, Computer Science and Artificial Intelligence Laboratory, MIT.

[29] 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.

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

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

[32] 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.

[33] 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.

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

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

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

[37] 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.

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

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

[40] 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.

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

[42] Renyi, A. (1955): "On a New Axiomatic Theory of Probability," *Acta Mathematica Academiae Scientiarum Hungaricae,* 6, 285-335.

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

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

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

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

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

[48] Siniscalchi, M. (2020): "Structural Rationality in Dynamic Games," mimeo, Northwestern University.