Monotonicity and Robust Implementation Under Forward-Induction Reasoning*

Pierpaolo Battigalli
Bocconi University and IGIER, pierpaolo.battigalli@unibocconi.it

Emiliano Catonini
NYU Shanghai, emiliano.catonini@nyu.edu

February 22, 2024

Abstract

We prove that, in sequential games with payoff uncertainty, strong rationalizability characterizes the path-predictions of forward-induction reasoning across all possible restrictions to players' initial hierarchies of beliefs on the exogenous uncertainty. With this, we can show that the implementation of social choice functions through sequential mechanisms under common strong belief in rationality—which considerably expands the realm of implementable functions compared with simultaneous-move mechanisms (Mueller, *J. Econ. Theory* 2016)—is robust in the sense of Bergemann and Morris (*Theor. Econ.* 2009).

1 Introduction

We prove a monotonicity result for a strong version of rationalizability in sequential games with incomplete information that captures forward-induction reasoning. To illustrate the importance of this result, we build on work by Bergemann & Morris (2009) and Mueller

*We thank Carlo Andreatta, Nicodemo De Vito, Shuige Liu, Nicolas Sourisseau, and all the attendants of the presentation of the paper at SAET 2023, the University of Michigan, NYU, and Bocconi University.

(2016) to merge the forward-induction analysis of sequential games with the theory of robust full implementation.[1]

In a nutshell, Bergemann & Morris (2009) show that robust (virtual) full implementation of social choice functions with *static* mechanisms—which amounts to rationalizable implementation—is severely limited when agents' valuations of outcomes exhibit a mild degree of interdependence. Mueller (2016) instead proves that using *sequential* mechanisms and assuming that agents reason by *forward induction*—as captured by the strong rationalizability solution concept (Battigalli 1999, 2003)—yields a very significant expansion of the implementable scf's. Yet, Mueller did not prove that such implementation is robust to considering contextual restrictions on agents' exogenous interactive beliefs about each other's types. This difficulty is related to the *non-monotonicity of* the *strong belief* concept used to represent forward-induction reasoning (Battigalli & Siniscalchi 2002 and Battigalli & Friedenberg 2012).

We obtain the robustness of implementation w.r.t. strong rationalizability from a **monotonicity result** concerning the solution concept: *the set of* state-dependent *strongly rationalizable paths of play* (but not the set of strategy profiles) *is monotone w.r.t. contextual restrictions on exogenous interactive beliefs*. In the rest of this Introduction we provide more background and details.

1.1 Robust implementation and rationalizability, static mechanisms

To set the stage, we first remind the reader of the conceptual connection between robust implementation and rationalizability, which is not a typical textbook topic, focusing first on static mechanisms. Consider an **economic environment** $\mathcal{E}$ with asymmetric information. There is a set I of agents and a set Y of economic outcomes (possibly, lotteries), a subset of some Euclidean space. The (expected) value to player i of outcome y is $v_i(\theta, y)$, where $\theta = (\theta_i)_{i\in I} \in \Theta = \times_{i\in I}\Theta_i$ is a state of nature and θ_i is i's private information about θ, or i's "payoff type."

Agents hold interactive hierarchical beliefs about each other's payoff types, which can be represented by means of a **type structure** $\mathcal{T}$ *à la* Harsanyi (1967-68). In words, $\mathcal{T}$

[1] On robust implementation see the survey by Bergemann & Morris (2012) and the relevant references therein.

captures what belief hierarchies are commonly believed possible, given some exogenous contextual restrictions on beliefs. Without contextual restrictions, $\mathcal{T}$ is the universal type structure containing all the collectively coherent belief hierarchies (e.g., Mertens & Zamir 1985, Brandenburger & Dekel 1993).

A planner (she) can commit to make the agents interact according to a mechanism $\mathcal{M}$, that is, some commonly known set of rules that yield a set Z of possible paths of play coupled with an outcome function $g : Z \to Y$. In static mechanisms, $Z = A$ is just the set of possible action profiles; in the subset of direct mechanisms, Z is isomorphic to Θ. The triple $\Gamma^b = (\mathcal{M}, \mathcal{E}, \mathcal{T})$ describes a situation of strategic interaction called "**Bayesian game**." In the *traditional full implementation problem*, it assumed that the *planner knows* both $\mathcal{E}$ *and* $\mathcal{T}$; with this, she wants to implement a map f (social choice function, scf) associating each state θ with a desirable outcome $y = f(\theta) \in Y$ by letting agents strategically interact according to an "appropriate" solution concept (e.g., Bayesian equilibrium, or rationalizability).[2] The solution concept yields, for each state $\theta \in \Theta$, a set $\mathbf{Z}^{\Gamma^b}(\theta)$ of possible paths of play. A mechanism $\mathcal{M}$ **fully implements** scf f if, for each state of nature θ, the image set of possible outcomes $g\left(\mathbf{Z}^{\Gamma^b}(\theta)\right)$ contains only the desired outcome $y = f(\theta)$, that is, $g\left(\mathbf{Z}^{\Gamma^b}(\theta)\right) = \{f(\theta)\}$ for all θ.[3] However, the planner often ignores the contextual features captured by type structure $\mathcal{T}$. If she deems all type structures possible, in compliance with "Wilson's doctrine," a natural notion of **robust full implementation** requires that $\mathcal{M}$ fully implements the scf f for all Bayesian games Γ^b based on $(\mathcal{M}, \mathcal{E})$, that is, across all type structures $\mathcal{T}$ (see Wilson 1987 and Bergemann & Morris 2009, 2012).[4] Since this paper is only concerned with different forms of full implementation, from now on in this introduction we will omit the adjective "full."

Robust implementation is conceptually related to **rationalizability**, that is, the solution concept characterizing the behavioral implications of *Rationality and Common Belief*

[2] We limit our attention to social choice functions. Similar considerations apply to social choice correspondences.

[3] Partial implementation relies on equilibrium analysis and requires instead that $g(\mathbf{z}(\cdot)) = f(\cdot)$ for at least one equilibrium selection $\mathbf{z}(\cdot)$ from equilibrium correspondence $\mathbf{Z}^{\Gamma^b}(\cdot)$.

[4] Quoting Wilson (1987): "I foresee progress of game theory as depending on successive reductions in the base of common knowledge required to conduct useful analyses of practical problems. Only by repeated weakening of common knowledge assumptions will the theory approximate reality."

in Rationality (RCBR).[5] On the one hand, not relying on the assumption that players' endogenous beliefs about each other's behavior serendipitously coordinate on a Bayesian equilibrium is in itself a form of robustness in the spirit of Wilson's doctrine. On the other hand, it has been observed that *the state-dependent outcomes consistent with Bayesian equilibrium across all type structures are precisely those allowed by a version of rationalizability for games with payoff uncertainty*—aka "belief-free rationalizability" —that applies to structure $(\mathcal{M}, \mathcal{E})$, i.e., to a description of the game that does not specify interactive beliefs about payoff types.[6] In particular, *restricting attention to static* (e.g., direct) *mechanisms*, robust Bayesian-equilibrium implementation is equivalent to implementation w.r.t. rationalizability for games with payoff uncertainty. Maintaining the viewpoint that rationalizable implementation is in itself a form of robustness, it is also worth noting that *robust* implementation w.r.t. rationalizability for Bayesian games is equivalent to implementation w.r.t. rationalizability for games with payoff uncertainty. The intuition for this result is relatively straightforward: (probability-1) belief is a **monotone** operator, that is, believing a weak proposition (large event) is easier than believing a logically stronger proposition (smaller event included in the former one). It follows that common belief in rationality and in contextual restrictions on exogenous interactive beliefs (which yields rationalizability in Bayesian games, see Battigalli *et al.* 2011) implies mere common belief in rationality. Since "no restriction" is a particular kind of contextual restriction (represented by the universal type structure), the robustness result follows. With this, we refer to robust implementation (with static mechanisms) also as "implementation under RCBR."

Finally, we are going to consider a weaker form of "virtual implementation," or **v-implementation**, that only requires to approximate the desired outcome $f(\theta)$ with an arbitrary degree of precision (see Abreu & Matsushima 1992 and Bergemann & Morris 2009). Clearly, robust v-implementation is easier to achieve than robust implementa-

[5] See, e.g., Battigalli & Siniscalchi (1999, 2002) and the relevant references therein. Note that here "**rationality**" means only expected utility maximization given whatever *subjective* beliefs a player holds about co-players' behavior and exogenous uncertainty. Every other restriction on behavior is the result of additional assumptions on interactive beliefs.

[6] See Battigalli (1999), Battigalli & Siniscalchi (2003), and the relevant references therein. Technically, rationalizability for games with payoff uncertainty is slightly different from what Bergemann & Morris (GEB, 2017) eventually called "belief-free rationalizability." We use the term with its original and most natural meaning.

tion. At this stage of the discussion it is not important to explain the details of v-implementation. Bergemann & Morris (2009) show that v-implementation with static mechanisms under RCBR is hard when valuations are highly, or even just mildly dependent on the types of others. Consider the following example. A single good must be allocated to one of many agents through a static mechanism with monetary transfers. Each agent/player i values the good

$$v_i(\theta_i, \theta_{-i}) = \theta_i + \gamma \sum_{j \neq i} \theta_j \quad (\gamma \geq 0),$$

where θ_i is private information of i and belongs to a finite set of payoff types Θ_i that satisfies $\{0,1\} \subseteq \Theta_i \subseteq [0,1]$. As i's valuation also depends on θ_{-i}, players have *interdependent* valuations for the good. The degree of interdependence is increasing in γ. It turns out that, for $\gamma > \frac{1}{|I|-1}$, only constant social choice functions can be v-implemented under RCBR with static mechanisms. This is problematic because, in the extant literature, only the latter form of implementation is known to be robust.

1.2 Robust implementation and rationalizability, sequential mechanisms

Using sequential mechanisms gives more flexibility and could significantly enlarge the set of robustly implementable scf's. Yet, the picture becomes more complex (and interesting) if we allow for sequential mechanisms, because there are different versions of rationalizability for sequential games characterizing the behavioral implications of different specifications of "common belief in rationality".[7] The weakest one, aka "weak rationalizability" or "initial rationalizability," relies on the assumption of *Rationality and Common Initial Belief in Rationality* (RCIBR, see Battigalli 2003 and Battigalli & Siniscalchi 1999). Therefore, we refer to (robust v-) implementation w.r.t. this version of rationalizability as "implementation under RCIBR."

Since initial (probability-1) belief is monotone, the aforementioned results for static mechanisms extend to sequential mechanisms (a weak version of perfect Bayesian equilib-

[7] Where "rationality" is now meant in the *sequential* sense of subjective expected utility maximization *conditional on* observations about previous moves.

rium) and implementation under RCIBR. However, since weak rationalizability typically allows for a large set of outcomes, it is unlikely that relevant scf's can be implemented under RCIBR. In particular, allowing for sequential mechanisms in the previous example, one can show that, for $\gamma > \frac{1}{|I|-1}$, only constant scf's can be robustly implemented under RCIBR.[8]

A stronger and more interesting version of rationalizability for sequential games captures a form of forward-induction (FI) reasoning, as it characterizes the behavioral implications of *Rationality and Common Strong Belief in Rationality* (RCSBR). The basic assumption—called **strong belief in rationality**—is that players, even if surprised by the co-players' behavior, hold on to the assumption that co-players are rational as long as their observations do not contradict co-players' rationality. Similar assumptions apply to higher levels of strategic sophistication, such as "co-players are rational and strongly believe in the rationality of others" (see Battigalli & Siniscalchi 2002). The simplest version of rationalizability capturing RCSBR in incomplete-information environments is **strong rationalizability** for games with payoff uncertainty, a kind of "belief-free strong rationalizability." Therefore, we refer to implementation w.r.t. strong rationalizability as "implementation under RCSBR."

Clearly, strong rationalizability refines weak/initial rationalizability. Thus, allowing for sequential mechanisms, v-implementation under RCSBR might considerably expand the set of v-implementable scf's. Indeed, considering a discretized environment, Mueller (2016) shows precisely this. For example, in the aforementioned implementation problem efficient allocations can be v-implemented under RCSBR for almost all parameter values $\gamma \geq 0$.

But is v-implementation under RCSBR *robust*? In other words, suppose agents' interactive exogenous beliefs about each other's payoff types satisfy some contextual restrictions represented by a (non-universal) Harsanyi type structure $\mathcal{T}$. Then, their behavior should satisfy *strong rationalizability for the Bayesian game* $\Gamma^b = (\mathcal{M}, \mathcal{E}, \mathcal{T})$. Robustness would require that the given scf f is v-implementable w.r.t. strong rationalizability in Bayesian games across all type structures $\mathcal{T}$. Unfortunately, we cannot replicate the aforementioned monotonicity argument to show that v-implementability w.r.t. "belief-free strong rationalizability" is necessarily robust in this sense, because *strong belief is*

[8] See Mueller 2016 and 2020.

not monotone: Indeed, while at the beginning of the game it is easier to believe a weak proposition such as "my co-players are rational" than a stronger one such as "my co-players are rational and their exogenous beliefs satisfy the contextual restrictions," there typically are more observations consistent with the weaker proposition, and therefore more instances in which strong belief requires to assign probability 1 to this proposition, making it more difficult to strongly believe it. With this, when contextual considerations (e.g., social norms) also shape endogenous beliefs about behavior, it is easy to show that the set of *paths of play* is non-monotone w.r.t. such contextual restrictions (see Battigalli & Friedenberg 2012).

Due to the non-monotonicity of strong belief, the extant literature does not show that v-implementation under RCSBR is robust to considering contextual restrictions on agents' *exogenous* interactive beliefs. Yet, existing examples and results concerning the (non)monotonicity of strongly rationalizable paths of play only refer to restrictions on interactive beliefs about behavior. We prove that this is not by chance, or lack of trying to find counterexamples. Indeed, our main game-theoretic result is that, although the strongly rationalizable set of (state-dependent) strategy profiles can be highly non-monotonic w.r.t. restrictions on exogenous interactive beliefs about payoff types, *the set of* state-dependent *strongly rationalizable paths of play is* (always nonempty and) *monotone w.r.t. such restrictions.* With this, we can also prove that *v-implementation under RCSBR is robust.* Fix an scf $f : \Theta \to Y$. Let $\Gamma = (\mathcal{M}, \mathcal{E})$ denote the game with payoff uncertainty (or "belief-free" game) induced by mechanism $\mathcal{M}$ with outcome function $g : Z \to Y$ in environment $\mathcal{E}$ and let $\theta \mapsto \mathbf{Z}^{\Gamma}(\theta)$ denote the strongly rationalizable-paths correspondence. Suppose that, for all states θ, $g\left(\mathbf{Z}^{\Gamma}(\theta)\right) \approx \{f(\theta)\}$ to an arbitrary degree of precision. Our theorem implies that, for all Bayesian games $\Gamma^b = (\mathcal{M}, \mathcal{E}, \mathcal{T})$ obtained by appending an Harsanyi type structure $\mathcal{T}$ to $\Gamma = (\mathcal{M}, \mathcal{E})$, $\emptyset \neq \mathbf{Z}^{\Gamma^b}(\theta) \subseteq \mathbf{Z}^{\Gamma}(\theta)$. Therefore, $g\left(\mathbf{Z}^{\Gamma^b}(\theta)\right) \approx \{f(\theta)\}$ for all such games Γ^b and states θ to an arbitrary degree of precision.

The paper is organized as follows. In Section 2, we provide the game-theoretic framework for the analysis. In Section 3 we state and prove our main game-theoretic result. In Section 4, we use the result for the analysis of Bayesian games. In Section 5, we apply our results to the robust implementation question. The Appendix collects the proofs of the key claims and lemmas that are omitted from the main body of the paper.

2 Fundamentals[9]

2.1 Multistage games with payoff uncertainty

We consider the following *finite* multistage game with *observable actions* and *payoff uncertainty*.[10] There is a set of players I and each $i \in I$ has a set of potentially available actions A_i. Let $A = \times_{i \in I} A_i$ denote the set of action profiles and $A^{<\mathbb{N}_0}$ the set of finite sequences of such profiles (including the empty sequence $\varnothing$). A subset of $A^{<\mathbb{N}_0}$ is a **tree** with root $\varnothing$ (the empty sequence) if it is closed under the "prefix-of" precedence relation $\preceq$ (note that $\varnothing$ is a prefix of every sequence). The rules of the game yield a *tree* $\bar{H} \subseteq A^{<\mathbb{N}_0}$ of possible sequences, called **histories**, and a feasibility correspondence $h \mapsto \mathcal{A}(h) = \left\{a \in A : (h,a) \in \bar{H}\right\}$ such that (1) $\mathcal{A}(h) = \times_{i \in I} \mathcal{A}_i(h)$ and (2) $\mathcal{A}(h) = \emptyset$ implies $\mathcal{A}_i(h) = \emptyset$ for every $i \in I$. The set of terminal histories—or possible paths of play—is $Z = \left\{z \in \bar{H} : \mathcal{A}(h) = \emptyset\right\}$, and the set of nonterminal histories is $H = \bar{H} \backslash Z$. *Nonterminal histories are publicly observed as soon as they realize.*

Each player i is *privately informed* of the true value of a payoff-relevant parameter θ_i, called the **payoff-type** of i, whereas the set Θ_i of possible values of θ_i is common knowledge. The parameterized payoff function of player i is

$$u_i : \Theta \times Z \to \mathbb{R},$$

where $\Theta = \times_{i \in I} \Theta_i$ is the set of all possible type profiles. Payoff uncertainty is represented by the dependence of u_i on θ. When convenient, we write $u_{i,\theta} : Z \to \mathbb{R}$ for the section of u_i at θ. Thus, a multistage game with payoff uncertainty and observable actions is given by

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

where all the featured sets are finite.

[9] The formalism is based on the (still incomplete) draft of textbook *Game Theory: Analysis of Strategic Thinking* by Battigalli, Catonini, and De Vito. Chapter 15 of the book (on the analysis of *finite multistage games with observable actions and incomplete information*) analyzes notions of rationalizability for such games, including strong directed rationalizability.

[10] We assume observable actions for simplicity of exposition: the analysis can be easily adapted to any finite game played by agents with perfect recall. The necessary modifications of the proofs are available upon request. We conjecture that the analysis also extends to games with infinite horizon.

We interpret function u_i as the composition of a parameterized **utility function** $v_i : \Theta \times Y \to \mathbb{R}$, where Y is the relevant space of outcomes, and an **outcome function** $g : Z \to Y$ specified by the rules of the game: $u_i(\theta, z) = v_i(\theta, g(z))$.

From these primitives, we can derive a set of **strategies** $S_i = \times_{h \in H} \mathcal{A}_i(h)$ for each player i. Let $S = \times_{i \in I} S_i$ and $S_{-i} = \times_{j \neq i} S_j$. Note, we take an interim perspective: the game starts with some exogenously given state of nature θ (e.g., representing players' traits), imperfectly and asymmetrically known by the players. Thus, strategies only describe how behavior depends on previous moves. Let $\zeta : S \to Z$ denote the **path function** that associates each strategy profile $s = (s_i)_{i \in I} \in S$ with the induced path $z = \zeta(s)$.[11] To ease notation, it is convenient to extend the path function to domain $\Theta \times S$ in the obvious way

$$(\theta, s) \mapsto \bar{\zeta}(\theta, s) = (\theta, \zeta(s))$$

and to define the (parameterized) strategic-form payoff function of player i as

$$U_i = u_i \circ \bar{\zeta} : \Theta \times S \to \mathbb{R}.$$

Finally, for each $h \in \bar{H}$,

$$S(h) = S_i(h) \times S_{-i}(h) := \{s \in S : h \preceq \zeta(s)\}$$

denotes the set of all strategy profiles inducing h.[12]

[11] Define recursively whether a history is induced by a given strategy profile s: the empty history $\varnothing$ is trivially induced by every $s \in S$. A history (h, a) is induced by s if h is induced by s and $a = (s_i(h))_{i \in I}$. With this, for every $s \in S$, $\zeta(s)$ is the terminal history induced by s.

[12] Actually, $S(h) = \times_{j \in I} S_j(h)$ for every $h \in \bar{H}$, but the relevant factorization is $S(h) = S_i(h) \times S_{-i}(h)$.

The primitive and derived elements are summarized by the following table:

Symbol	**Terminology**
$i \in I$	players
$a_i \in A_i$	actions of i
$a \in A = \times_{i \in I} A_i$	action profiles
$h \in \bar{H} \subseteq A^{<\mathbb{N}_0}$	histories ($\bar{H}$ is a tree)
$\mathcal{A}_i(h)$ ($\mathcal{A}(h) = \times_{i \in I} \mathcal{A}_i(h)$)	feasible actions (action profiles) given h
$z \in Z$	terminal histories, or paths of play
$H = \bar{H} \backslash Z$	nonterminal histories
$\theta_i \in \Theta_i$	payoff-types of i
$\theta \in \Theta = \times_{i \in I} \Theta_i$	states of nature
$u_i : \Theta \times Z \to \mathbb{R}$	(parameterized) payoff function of i
$s_i \in S_i = \times_{h \in H} \mathcal{A}_i(h)$	strategies of i
$s \in S = \times_{i \in I} S_i$	strategy profiles
$s \in S(h)$	strategy profiles inducing h
$\zeta : S \to Z$	path function
$\bar{\zeta} : \Theta \times S \to Z$	extended path function
$U_i = u_i \circ \bar{\zeta} : \Theta \times S \to \mathbb{R}$	(param.) strategic-form payoff function of i

2.2 Beliefs

We model the beliefs of each player i as the play unfolds by means of a **conditional probability system** (Renyi, 1955)

$$\mu^i = \left(\mu^i\left(\cdot | \Theta_{-i} \times S_{-i}(h)\right)\right)_{h \in H} \in \Delta^H\left(\Theta_{-i} \times S_{-i}\right),$$

abbreviated in $\mu^i = \left(\mu^i(\cdot|h)\right)_{h \in H}$. With this, $\Delta^H\left(\Theta_{-i} \times S_{-i}\right)$ is the subset of arrays of beliefs $\mu^i \in \left(\Delta\left(\Theta_{-i} \times S_{-i}\right)\right)^H$ such that, for every $h \in H$, $\mu^i\left(\Theta \times S_{-i}(h) | h\right) = 1$ and the *chain rule* holds, that is, for all $h, h' \in H$ and $E \subseteq \Theta_{-i} \times S_{-i}(h')$,

$$S_{-i}(h') \subseteq S_{-i}(h) \Longrightarrow \mu^i(E|h) = \mu^i(E|h')\, \mu^i\left(\Theta \times S_{-i}(h') | h\right).$$

Note that $h \preceq h'$ implies $S_{-i}(h') \subseteq S_{-i}(h)$, but the converse is not true because histories also represent behavior of player i (see, e.g., Battigalli *et al.* 2023).

We will consider type-dependent **restrictions on** players' **exogenous beliefs** (i.e., initial beliefs about the types of others), represented by subsets of probability measures: for all $i \in I$ and $\theta_i \in \Theta_i$,

$$\bar{\Delta}_{i,\theta_i} \subseteq \Delta\left(\Theta_{-i}\right).$$

With this, we introduce profiles $\Delta = \left(\Delta_{i,\theta_i}\right)_{i\in I,\theta_i\in\Theta_i}$ of type-dependent subsets of CPSs such that, for all $i \in I$ and $\theta_i \in \Theta_i$,[13]

$$\Delta_{i,\theta_i} = \left\{\mu^i \in \Delta^H\left(\Theta_{-i} \times S_{-i}\right) : \mathrm{marg}_{\Theta_{-i}}\mu^i\left(\cdot|\varnothing\right) \in \bar{\Delta}_{i,\theta_i}\right\}.$$

The proof of the main theorem will require to construct CPSs with certain features. It turns out that it is simpler to construct a "forward-consistent belief system" (Battigalli *et al.* 2023) with such features and then claim the existence of a CPS that preserves them. A **forward-consistent belief system** is an array of beliefs $\hat{\mu}^i = \left(\hat{\mu}^i(\cdot|h)\right)_{h\in H} \in \left(\Delta\left(\Theta_{-i} \times S_{-i}\right)\right)^H$ such that, for every $h \in H$, $\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h)|h) = 1$ and the *forward chain rule* holds: for all $h, h' \in H$ and $E \subseteq \Theta_{-i} \times S_{-i}(h')$,

$$h \preceq h' \Longrightarrow \hat{\mu}^i(E|h) = \hat{\mu}^i(E|h')\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h).$$

The forward chain rule is weaker than the chain rule, because, as noticed above, $S_{-i}(h') \subseteq S_{-i}(h)$ does not imply $h \preceq h'$.

For each $E_{-i} \subseteq \Theta_{-i} \times S_{-i}$, we say that a CPS, or—more generally, a forward-consistent belief system—μ^i **strongly believes** E_{-i} (Battigalli & Siniscalchi 2002) if μ^i assigns probability 1 to E_{-i} as long as E_{-i} is not contradicted by observation:

$$\forall h \in H, \quad E_{-i} \cap \left(\Theta_{-i} \times S_{-i}\left(h\right)\right) \neq \emptyset \Rightarrow \mu^i\left(E_{-i}|h\right) = 1.$$

Let $\Delta^H_{\mathrm{sb}}\left(E_{-i}\right)$ denote the set of CPSs μ^i that strongly believe E_{-i}.

For the transformation of belief systems into CPSs, we rely on the following result.

[13] Such resrictions are called "regular" in Battigalli (2003).

Let

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

denote the set of non-terminal histories that can occur if s_i is played.

Lemma 1 (Battigalli, Catonini and Manili, 2023) *Fix a strategy s_i and a forward-consistent belief system $\hat{\mu}^i$ that strongly believes $E^1_{-i}, \ldots, E^{n-1}_{-i}$, where $E^{n-1}_{-i} \subseteq \ldots \subseteq E^1_{-i}$. Then, there is a CPS $\widetilde{\mu}^i$ that strongly believes $E^1_{-i}, \ldots, E^{n-1}_{-i}$ such that $\widetilde{\mu}^i(\cdot|h) = \hat{\mu}^i(\cdot|h)$ for all $h \in H(s_i)$.*

2.3 Sequential optimality

We represent the behavior of a rational player i of type θ_i by means of a (weak) **sequential best reply** correspondence $\mu^i \mapsto r_{i,\theta_i}(\mu^i)$ defined as

$$r_{i,\theta_i}\left(\mu^i\right) = \left\{\bar{s}_i : \forall h \in H_i(\bar{s}_i), \bar{s}_i \in \arg\max_{s_i \in S_i(h)} \mathbb{E}_{\mu^i(\cdot|h)}\left(U_i(\theta_i, s_i, \cdot)\right)\right\}.$$

By Lemma 1 we can take μ^i to be a forward-consistent belief system; by known dynamic programming arguments $r_{i,\theta_i}(\mu^i) \neq \emptyset$ for all θ_i and forward-consistent belief systems μ^i.[14]

Fix a CPS $\mu^i \in \Delta^H(\Theta_{-i} \times S_{-i})$ and a type θ_i. For each strategy $\bar{s}_i$ and history $h \in H_i(\bar{s}_i)$, we say that $\bar{s}_i$ is a **continuation best reply** to $\mu^i(\cdot|h) \in \Delta(\Theta_{-i} \times S_{-i}(h))$ for θ_i if, for every $s_i \in S_i(h)$,

$$\mathbb{E}_{\mu^i(\cdot|h)}\left(U_i(\theta_i, \bar{s}_i, \cdot)\right) \geq \mathbb{E}_{\mu^i(\cdot|h)}\left(U_i(\theta_i, s_i, \cdot)\right).$$

Thus, $\bar{s}_i$ is a (weak) sequential best reply to μ^i for θ_i if $\bar{s}_i$ is a continuation best reply to $\mu^i(\cdot|h)$ for θ_i at every $h \in H_i(\bar{s}_i)$.[15] In the proof of the main theorem, we use the following dynamic programming result.[16]

[14] See Battigalli *et al.* (2023) and the relevant references therein, where this *weak* notion of sequential best reply (which applies to reduced strategies as well as strategies) is extensively discussed and motivated.

[15] Strictly speaking, "continuation" should refer to the substrategy on the subtree with root h.

[16] We conjecture that, using standard truncation arguments, the result can be extended to infinite-horizon games that satisfy continuity at infinity.

Lemma 2 *Fix a CPS μ^i, a type θ_i, and a strategy s_i. If, for every $h \in H_i(s_i)$, there exists a continuation best reply $s_i' \in S_i(h)$ to $\mu^i(\cdot|h)$ for θ_i such that $s_i'(h) = s_i(h)$, then s_i is a sequential best reply to μ^i for θ_i, that is, $s_i \in r_{i,\theta_i}(\mu^i)$.*

2.4 Strong (Directed) Rationalizability

We assume that players are *rational* and that the restrictions on exogenous beliefs are **transparent**, that is, the belief restrictions hold and there is common belief of this fact conditional on every nonterminal history. Moreover, we assume that players strongly believe that:

- the co-players are rational and the restrictions are transparent;
- the co-players are rational, the restrictions are transparent, and the co-players strongly believe that everyone else is rational and that the restrictions are transparent;
- and so on.

In brief, we assume *rationality, transparency of the belief restrictions, and common strong belief thereof.*

The previous hypotheses can be made formal in the language of epistemic game theory. As shown by Battigalli & Prestipino (2013), the behavioral implications of these epistemic hypotheses are characterized by **Strong Directed Rationalizability** (Battigalli 2003, Battigalli & Siniscalchi 2003).[17]

For each $i \in I$, let $C_{i,\text{sb}}^{\Delta,0} = \Theta_i \times S_i$. Then, for each $n > 0$, define the set of **strongly Δ-n-rationalizable** type-strategy pairs of i as

$$C_{i,\text{sb}}^{\Delta,n} = \left\{(\theta_i, s_i) : \exists \mu^i \in \cap_{m=0}^{n-1} \Delta_{\text{sb}}^H(C_{-i,\text{sb}}^{\Delta,m}) \cap \Delta_{i,\theta_i}, s_i \in r_{i,\theta_i}(\mu^i)\right\}.$$

[17] These articles use the term "(strong) Δ-rationalizability." We use "(strong) directed rationalizability" to refer to the correspondence that associates each profile of belief restrictions Δ with the corresponding strongly rationalizable behavior, so that Δ "directs" the resulting behavior.

With this, the set of strongly Δ-n-rationalizable strategies for θ_i is the section at θ_i of $C_{i,\text{sb}}^{\Delta,n}$

$$S_i^{\Delta,n}(\theta_i) = \left(C_{i,\text{sb}}^{\Delta,n}\right)_{\theta_i} = \left\{s_i : (\theta_i, s_i) \in C_{i,\text{sb}}^{\Delta,n}\right\},$$

and the set of strongly Δ-n-rationalizable strategy profiles at θ is

$$S^{\Delta,n}(\theta) = \times_{i\in I} S_i^{\Delta,n}(\theta_i).$$

Finally, let

$$C_{i,\text{sb}}^{\Delta,\infty} = \cap_{n>0} C_{i,\text{sb}}^{\Delta,n}$$

denote the set of strongly Δ-rationalizable type-strategy pairs of i, and let

$$\begin{aligned} S_i^{\Delta,\infty}(\theta_i) &= \left(C_{i,\text{sb}}^{\Delta,\infty}\right)_{\theta_i}, \\ S^{\Delta,\infty}(\theta) &= \times_{i\in I} S_i^{\Delta,\infty}(\theta_i). \end{aligned}$$

Recalling that the sequential best reply correspondence is non-empty valued and noting that mere restrictions on exogenous beliefs cannot contradict the restrictions on beliefs about type-dependent behavior implied strategic reasoning, one can prove by induction the following result:

Lemma 3 *(cf. Battigalli 2003) Since* Δ *represents* restrictions on exogenous beliefs*, for each* $\theta \in \Theta$*, the set of strongly* Δ*-rationalizable strategy profiles is non-empty:* $S^{\Delta,\infty}(\theta) \neq \emptyset$.

When there are no actual belief restrictions, i.e. when each Δ_{i,θ_i} is the set $\Delta^H(\Theta_{-i} \times S_{-i})$ of all CPSs of i, Strong Directed Rationalizability boils down to **Strong Rationalizability** (Pearce 1982, Battigalli 1997), which characterizes the behavioral implications of *Rationality and Common Strong Belief in Rationality* (Battigalli & Siniscalchi, 2002).[18] We omit the superscript Δ to denote strong rationalizability: $C_{i,\text{sb}}^{\infty}$ ($C_{i,\text{sb}}^{n}$) is the set of strongly (n-)rationalizable pairs of i, $S_i^{\infty}(\theta_i)$ ($S_i^{n}(\theta_i)$) is the set of strongly

[18] Strong rationalizability was once called "extensive-form rationalizability." Following Battigalli (2003) and the more recent literature, *we eschew this terminology*, because—as mentioned in the Introduction—there are multiple versions of the rationalizability idea for sequential games represented in extensive form, based on different versions of "common belief in rationality."

(n-)rationalizable strategies of i, and $(S^{\infty}(\theta))_{\theta\in\Theta}$ ($(S^{n}(\theta))_{\theta\in\Theta}$) is the set of profiles of strongly (n-)rationalizable strategies at θ.

A path (terminal history) $z \in Z$ is strongly Δ-rationalizable if there exists some strongly Δ-rationalizable profile (θ, s) such that $\zeta(s) = z$. Thus, the set of **strongly Δ-rationalizable paths** is $\bar{\zeta}\left(C_{\text{sb}}^{\Delta,\infty}\right)$, and the set of strongly Δ-rationalizable paths at state of nature θ is the section $\bar{\zeta}\left(C_{\text{sb}}^{\Delta,\infty}\right)_{\theta} = \zeta\left(S^{\Delta,\infty}(\theta)\right)$.

3 Main Theorem

We show that, *when we consider only restrictions on exogenous beliefs*, *the set of strongly Δ-rationalizable paths is monotone in* Δ, despite the non-monotonicity of strong belief.

Because it suffices for our application to implementation theory, here we just focus on the comparison between some profile Δ of subsets of CPSs that only restrict exogenous beliefs, and the case of no restrictions ($\Delta_{i,\theta_i} = \Delta^{H}(\Theta_{-i} \times S_{-i})$ for all i and θ_i, that is, undirected strong rationalizability). Thus, we prove that for any fixed profile of restrictions on exogenous beliefs Δ the set of strongly Δ-rationalizable paths is contained in the set of strongly rationalizable paths. It will be clear that the proof can be easily adapted to obtain the more general path-monotonicity claim.

Theorem 1 *Fix a profile $\Delta = (\Delta_{i,\theta_i})_{i,\in I,\theta_i\in\Theta_i}$ of restrictions on* exogenous *beliefs. Then, for all steps $n > 0$ and states $\theta \in \Theta$, $\emptyset \neq \zeta\left(S^{\Delta,n}(\theta)\right) \subseteq \zeta\left(S^{n}(\theta)\right)$, that is, for each $(\theta, s) \in C_{\text{sb}}^{\Delta,\infty} \neq \emptyset$, there exists $s' \in S$ such that $(\theta, s') \in C_{\text{sb}}^{\infty}$ and $\zeta(s) = \zeta(s')$.*

The assumption that the belief restriction only apply to exogenous beliefs is tight. In the literature, there are many examples of strong directed rationalizability with restrictions on the initial beliefs about the opponent's *strategy* yielding non-strongly-rationalizable outcomes (see, e.g., Battigalli & Friedenberg 2012 and Catonini 2019). In the supplemental appendix, we provide an analogous example of restrictions on *non-initial* beliefs about the opponent's type.

Before the proof of the theorem, we provide an example of path-monotonicity that highlights the main difficulty to tackle.

Example 1 Consider a signaling game between players 1 and 2 with $\Theta_1 = \{x, y, z\}$, $S_1 = A_1 = \{\ell, r\}$, $\mathcal{A}_2(\ell) = \{a, b\}$, $\mathcal{A}_2(r) = \{\mathrm{c}, \mathrm{d}, \mathrm{e}\}$, and the following payoffs:

Payoffs of *1* and 2:

after ℓ	a	b
$\theta_1 = x$	*3* 1	*1* 0
$\theta_1 = y$	*1* 0	*1* 1
$\theta_1 = z$	*3* 1	*1* 0

after r	c	d	e
$\theta_1 = x$	*0* 0	*0* 0	*0* 1
$\theta_1 = y$	*0* 0	*0* 1	*3* 0
$\theta_1 = z$	*0* 1	*2* 0	*2* 0

We start with *Strong Rationalizability*, showing all the eliminations for every step.

1. We can only eliminate action r for type x, as it is dominated by action ℓ. Thus, $S_1^1(x) = \{\ell\}$. [Since no strategy of player 2 is eliminated in the first step, it follows that in even (odd) steps only eliminations for player 2 (player 1) are possible.]

2. (*Player 2*) Every $\mu^2 \in \Delta^H_{\mathrm{sb}}\left(C^1_{1,\mathrm{sb}}\right)$ assigns probability 0 to type x upon observing action r (an instance of forward-induction reasoning). With this, action e is never a best reply. Hence, $C^2_{2,\mathrm{sb}} = \{\mathrm{a.c}, \mathrm{b.c}, \mathrm{a.d}, \mathrm{b.d}\}$.

3. (*Player 1*) For type y, action r is not a best reply to any belief over $C^2_{2,\mathrm{sb}}$. Thus, $S_1^3(y) = \{\ell\}$.

4. (*Player 2*) Every $\mu^2 \in \Delta^H_{\mathrm{sb}}\left(C^3_{1,\mathrm{sb}}\right)$ assigns probability 1 to type z given action r. Thus, $C^4_{2,\mathrm{sb}} = \{\mathrm{a.c}, \mathrm{b.c}\}$.

5. (*Player 1*) Given this, type z expects to obtain 0 from r and at least 1 from ℓ. Thus, $S_1^5(y) = \{\ell\}$.

No remaining strategy of player 2 can be eliminated. So we have

$$\begin{aligned} C^\infty_{1,\mathrm{sb}} &= \Theta_1 \times \{\ell\}, \\ C^\infty_{2,\mathrm{sb}} &= \{\mathrm{a.c}, \mathrm{b.c}\}. \end{aligned}$$

Thus, the *strongly rationalizable paths* are (ℓ, a) and (ℓ, b) for each of player 1.

Now consider the following restrictions to the exogenous beliefs of player 2 (only): let Δ_2 collect all the CPSs μ^2 that initially assign probability 1 to type z, i.e., $\mu^2(\{z\} \times S_1 | \varnothing) = 1$.

Strong Δ-Rationalizability is given by the following steps:

Δ,1. As above, action r is eliminated for type x, $S_1^{\Delta,1}(x) = S_1^1(x) = \{\ell\}$, but now some strategies of player 2 are eliminated. By the chain rule, every $\mu^2 \in \Delta_2$ assigns probability 1 to z given ℓ, if $\mu^2(\{z, \ell\} | \varnothing) > 0$, and/or given r, if $\mu^2(\{z, r\} | \varnothing) > 0$. Thus, player 2

best replies with a after ℓ and/or with c after r: $C_{2,\text{sb}}^{\Delta,1} = \{\text{a.c}, \text{b.c}, \text{a.d}, \text{a.e}\}$.
Δ,**2.** As in strong rationalizability, action e is never a best reply given r; hence, strategy $a.e$ of player 2 is eliminated: $C_{2,\text{sb}}^{\Delta,2} = \{\text{a.c}, \text{b.c}, \text{a.d}\}$. Moreover, for type z, r is dominated by ℓ w.r.t. strategies in $C_{2,\text{sb}}^{\Delta,1}$; thus, $S_1^{\Delta,2}(z) = \{\ell\}$.
Δ,**3.** For type y, r is dominated by ℓ over $C_{2,\text{sb}}^{\Delta,2}$; thus, $S_1^{\Delta,3}(z) = \{\ell\}$. Moreover, every $\mu^2 \in \Delta_{\text{sb}}^H\left(C_{1,\text{sb}}^{\Delta,2}\right)$ assigns probability 1 to type y given action r; thus, So, $C_{2,\text{sb}}^{\Delta,3} = \{\text{a.d}\}$.
We pinned down one strategy for (each type of) each player:

$$\begin{aligned} C_{1,\text{sb}}^{\Delta,\infty} &= \Theta_1 \times \{\ell\}, \\ C_{2,\text{sb}}^{\Delta,\infty} &= \{\text{a.d}\}. \end{aligned}$$

The *strongly* Δ*-rationalizable path* is (ℓ, a) for each type of player 1. In compliance with Theorem 1, this is one of the two strongly rationalizable paths. Note, however, that the strongly Δ-rationalizable reaction of player 2 to r is d, whereas the strongly rationalizable one was c. ▲

Given that the two elimination procedures may induce completely disjoint off-path behaviors, proving path-monotonicity is hard. It is even hard to grasp how path-monotonicity can hold in absence of any discipline on the rationalizable off-path behaviors. Before providing the proof of Theorem 1, we propose a different, intuitive argument of "why path-monotonicity cannot fail", which is a kind of proof by contradiction.

The set of strongly Δ-rationalizable paths has a sort of "best reply property", so that, if a player does not entertain deviations outside that set of paths, every move can be justified under the belief that the others will not leave those paths either. Now suppose for a moment that, at some step of strong rationalizability, for the first time, a player leaves one of the strongly Δ-rationalizable paths—or at least does so whenever she believes that the others will stay. Given what we said before, this player must be considering a deviation outside of the strongly Δ-rationalizable paths. Moreover, since this is the first step where this phenomenon occurs, at the previous step everyone could find a reason to stay on the strongly Δ-rationalizable paths while believing that the others were staying as well. This means that all the opponents of our player may be surprised by her deviation. Surprise implies belief revision, so that every belief about the continuation play is possible (i.e., consistent with the chain rule). But this means that

our player finds a reason to deviate under all possible conjectures about the opponents' reactions. These two considerations combined imply that the set of possible continuation paths after the deviation must have, again, a sort of best reply property. In light of this, and of the absence of restrictions to off-path beliefs, some of these continuation strategies should have survived also strong Δ-rationalizability. (A recursive argument is required to rule out deviations from those continuation paths.) But this contradicts the fact that strong Δ-rationalizability ruled out such a deviation for our player.

3.1 Proof of Theorem 1

Non-emptiness follows from Lemma 3. Here we only focus on the path-inclusion. Comparing directly strong rationalizability and strong Δ-rationalizability is difficult. As we have just seen through the example, the two procedures may substantially depart in terms of strategies, and this also makes it hard to compare them in terms of outcomes. To overcome this difficulty, we construct a sequence of elimination procedures that gradually transform strong Δ-rationalizability into strong rationalizability,[19] and we prove step-by-step path-inclusion between each pair of consecutive, "similar" procedures.

Let K be the number of steps that it takes for strong rationalizability to converge: $C_{\text{sb}}^{K-1} \subset C_{\text{sb}}^{K} = C_{\text{sb}}^{\infty}$ ($\subset$ denotes *strict* inclusion). Note that K is well defined because the game is finite.[20] For each $k = 0, ..., K$, we introduce Procedure k, which performs the *first* k steps of elimination *without* belief *restrictions* and the *following* steps *with* the belief *restrictions*. Thus, Procedure 0 coincides with strong Δ-rationalizability, while the first K steps of Procedure K coincide with strong rationalizability. Hence, the step-K path-inclusions between Procedure 0 and Procedure 1, Procedure 1 and Procedure 2, and so on up to Procedure K prove the theorem.

Now we define formally such elimination procedures, denoted by $((\mathrm{X}_k^n)_{n=0}^{\infty})_{k=0}^{K}$. If everything is strongly rationalizable, there is nothing to prove, so suppose that strong rationalizability deletes some pair (θ_i, s_i) for at least one player i, so that $K > 0$.

[19] The same idea appears in Perea (2018), who compares different orders of elimination of strong rationalizability in games with perfect information through a sequence of pairwise similar elimination orders. Other intuitions used in the proof appear in Catonini (2020).

[20] A generalization of our argument for (classes of) infinite games is available upon request.

As anticipated, for $k = 0$, we have strong Δ-rationalizability:

$$(\mathrm{X}_0^n)_{n=0}^{\infty} = \left(C_{\mathrm{sb}}^{\Delta,n}\right)_{n=0}^{\infty}.$$

For each $k = 1, ..., K$, define $((\mathrm{X}_{k,i}^n)_{i\in I})_{n=0}^{\infty}$ as follows. Let $\mathrm{X}_k^0 = \Theta \times S$.

For all $1 \leq n \leq k$ and $i \in I$,

$$\mathrm{X}_{k,i}^n = \left\{(\theta_i, s_i) \in \Theta_i \times S_i : \exists \mu^i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k,-i}^m), s_i \in r_{i,\theta_i}(\mu^i)\right\}. \tag{1}$$

Thus, for $k > 0$, steps $n = 1, ..., k$ of the k-procedure coincide with strong rationalizability: $\mathrm{X}_k^n = C_{\mathrm{sb}}^n$ for $n \leq k$.

For all $n > k$ and $i \in I$, let

$$\mathrm{X}_{k,i}^n = \left\{(\theta_i, s_i) \in \Theta_i \times S_i : \exists \mu^i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k,-i}^m) \cap \Delta_{i,\theta_i}, s_i \in r_{i,\theta_i}(\mu^i)\right\}. \tag{2}$$

Thus, the k-procedure deviates from strong rationalizability from step $n = k+1$ onwards, because it starts imposing the Δ-restrictions on justifying beliefs only from step $k + 1$.

It follows that, as anticipated, $(\mathrm{X}_K^n)_{n=0}^{\infty}$ is an elimination procedure which coincides with strong rationalizability $(C_{\mathrm{sb}}^n)_{n=0}^{\infty}$ for the first K steps, so obtaining the strongly rationalizable profiles, but then proceeds to (possibly) delete more profiles by adding the Δ-restrictions. More generally, no procedure needs to converge by step K (although some may also converge at an earlier step), but for our purpose we can focus on the first K steps of all procedures.

We are going to prove that, for each step of elimination n, *the set of θ-dependent paths that are consistent with step n weakly expands as k increases*, which implies the thesis. To do so, we proceed in this order: first we fix $k \in \{1, ..., K\}$ and consider Procedure $k-1$ and Procedure k; then, we prove the path-inclusion between the two procedures at every step of elimination n by induction on n.

First we provide an intuition of how we exploit the similarity between the two procedures and how the assumption of exogenous restrictions makes their comparison possible. From this intuition, we will derive the two-fold inductive hypothesis for the formal proof. To simplify notation, we drop the indexes $k-1$ and k of the two procedures and we call them "P" and "Q": $((\mathrm{P}_i^n)_{i\in I})_{n=0}^{\infty} = ((\mathrm{X}_{k-1,i}^n)_{i\in I})_{n=0}^{\infty}$ and $((\mathrm{Q}_i^n)_{i\in I})_{n=0}^{\infty} = ((\mathrm{X}_{k,i}^n)_{i\in I})_{n=0}^{\infty}$.

We are also going to apply the notation "$|_{\widehat{H}}$" to (profiles of) strategies or type-strategy pairs in order to restrict the domain of strategies to a subset of histories $\widehat{H}$. Furthermore, for any subset $\mathrm{X} \subseteq \Theta \times S$, we let

$$\begin{aligned} H(\mathrm{X}) &= \left\{h \in H : \exists\, (\theta, s) \in \mathrm{X}, h \prec \bar{\zeta}\,(\theta, s)\right\} \\ &= \left\{h \in H : \exists \theta \in \Theta, \exists s \in \mathrm{X}_\theta, h \prec \zeta\,(s)\right\} \end{aligned}$$

denote the set of non-terminal histories that realize for some $(\theta, s) \in \mathrm{X}$. With this, for any $\mathrm{X}_{-i} \subseteq \Theta_{-i} \times S_{-i}$, we also let

$$H(\mathrm{X}_{-i}) = H(\Theta_i \times S_i \times \mathrm{X}_{-i})$$

denote the set of non-terminal histories that realize for some $(\theta_{-i}, s_{-i}) \in \mathrm{X}_{-i}$ and $(\theta_i, s_i) \in \Theta_i \times S_i$.

P and Q coincide with Strong Rationalizability for steps $n \in \{1, ..., k-1\}$ and depart at step $n = k$.

At step $n = k$, P adopts the belief restrictions and Q does not, so:

$$\mathrm{P}^n \subseteq \mathrm{Q}^n \text{ for } n = k.$$

At step $n + 1 = k + 1$ both P and Q adopt the restrictions, but P imposes strong belief in smaller strategy sets and therefore, *along the paths consistent with these sets*, it remains more restrictive:

$$\mathrm{P}^{n+1}|_{H(\mathrm{P}^n)} \subseteq \mathrm{Q}^{n+1}|_{H(\mathrm{P}^n)} \text{ for } n = k. \tag{3}$$

At step $n + 2 = k + 2$, things get complicated.

First: Is this step of procedure P still still more restrictive than Q regarding beliefs about the co-players' types and moves at the histories in $H(\mathrm{P}^n)$, as equation (3) seems to suggest?

The answer is yes, but only thanks to the assumption of restrictions on the exogenous beliefs. Restrictions on the beliefs about the endogenous/strategic uncertainty could allow player i to believe in some $(\theta_{-i}, s_{-i}) \in \mathrm{P}_{-i}^{n+1}$ but not in its counterpart $\left(\theta_{-i}, s'_{-i}\right) \in$

Q_{-i}^{n+1} with $s_{-i}|_{H(\mathrm{P}^n)} = s'_{-i}|_{H(\mathrm{P}^n)}$. The role of restricting only the *initial* beliefs is more subtle. Strong belief in P_{-i}^{n+1} and in Q_{-i}^{n+1} may induce different beliefs about θ_{-i} at some history $h' \in \left(H(\mathrm{P}_{-i}^{n+1}) \cap H(\mathrm{Q}_{-i}^{n+1})\right) \setminus H(\mathrm{P}^n)$. If there were restrictions on such beliefs at h', it could well be that some of the beliefs derived from Q_{-i}^{n+1} were incompatible with the restrictions. Via the chain rule, this could also rule out some beliefs at some $h \in H(\mathrm{P}^n)$ such that $h \prec h'$.

But this not the end of the story. Strong belief in Q_{-i}^{n+1} may be more restrictive than in P_{-i}^{n+1} regarding behavior outside of $H(\mathrm{P}^n)$, even at histories that are consistent with both P_{-i}^{n+1} and Q_{-i}^{n+1}. This is because the inclusion of equation (3) is restricted to $H(\mathrm{P}^n)$. Thus, strong belief in Q_{-i}^{n+1} may rule out some belief about the reactions of the co-players to a deviation of i from $H(\mathrm{P}^n)$ which is instead allowed by strong belief in P_{-i}^{n+1}. With this, there could be a deviation from one of the paths consistent with P^{n+2} which player i expects to lead out of $H(\mathrm{P}^n)$ and be always profitable under strong belief in Q_{-i}^{n+1}. This is what makes it hard to prove that

$$\mathrm{Q}^{n+2}|_{H(\mathrm{P}^{n+1})} \supseteq \mathrm{P}^{n+2}|_{H(\mathrm{P}^{n+1})} \text{ for } n = k. \tag{4}$$

What guarantees that such a deviation does not exist? We are going to argue that $H(\mathrm{P}^n) \supseteq H(\mathrm{Q}^{n+1})$, so that no strategy in $\mathrm{Q}_i^{n+2} \subseteq \mathrm{Q}_i^{n+1}$ (i.e., no strategy that player i could ever find profitable at step $n+2$ of procedure Q) leads out of $H(\mathrm{P}^n)$ (actually, of $H(\mathrm{P}^{n+1}) \subseteq H(\mathrm{P}^n)$) under strong belief that the co-players follow strategies in Q_{-i}^{n+1}.

Here is where the similarity between the two procedures comes into play: $H(\mathrm{P}^n) \supseteq H(\mathrm{Q}^{n+1})$ is a reverse inclusion compared to the path-inclusion we want to prove, but with procedure Q one step ahead of procedure P. Thus, to see why the inclusion holds, we must flip the roles of the two procedures and start from the trivial observation that, since $\mathrm{Q}^n \subseteq \mathrm{Q}^{n-1} = \mathrm{P}^{n-1}$ for each $n \leq k$,

$$\mathrm{Q}^n \subseteq \mathrm{P}^{n-1} \text{ for } n = k.$$

Next, we consider step $n+1$ of Q and step n of P. Both steps use the belief restrictions, as Q introduces the restrictions only one step later than P. Thanks to this similarity, we

can argue as above (cf. equation (3)) to obtain

$$\mathrm{Q}^{n+1}|_{H(\mathrm{Q}^n)} \subseteq \mathrm{P}^n|_{H(\mathrm{Q}^n)} \text{ for } n = k. \tag{5}$$

Thus, as $H(\mathrm{Q}^n) \supseteq H(\mathrm{Q}^{n+1})$, we have

$$H(\mathrm{P}^n) \supseteq H(\mathrm{Q}^{n+1}) \text{ for } n = k.$$

Proving (5) was easy because we could rely on the inclusion $\mathrm{Q}^n \subseteq \mathrm{P}^{n-1}$, which is in terms of (full) strategies. For $n > k$, we run into the same complications we had for (4), but we can solve them exactly in the same way by first showing that $H(\mathrm{Q}^{n-1}) \supseteq H(\mathrm{P}^{n-1})$. The direction of this inclusion is the one of the path-inclusion we want to prove, but it involves step $n-1$ of both procedures, so we can take it as induction hypothesis. With this, we can then take as induction hypothesis also the reverse inclusion $H(\mathrm{P}^{n-1}) \supseteq H(\mathrm{Q}^n)$.

For the formal proof, our induction hypothesis is a bit stronger than a double inclusion between the histories (and also the paths) induced by the two procedures. At those histories, we will need to mimic the beliefs player i can have at some step of a procedure for the corresponding step of the other procedure. For this, we need the two procedures to correspond in terms of restricted strategies, in the following way:

IH1(n) for every $i \in I$ and $(\theta_i, s_i) \in \mathrm{X}^n_{k-1,i}$, there is $\hat{s}_i^{(\theta_i,s_i)} \in S_i$ such that $(\theta_i, \hat{s}_i^{(\theta_i,s_i)}) \in \mathrm{X}^n_{k,i}$ and $\hat{s}_i^{(\theta_i,s_i)}(h) = s_i(h)$ for all $h \in H(\mathrm{X}^{n-1}_{k-1})$ (thus, step n of the $(k-1)$-procedure path-refines step n of the k-procedure);

IH2(n) for every $i \in I$ and $(\theta_i, s_i) \in \mathrm{X}^n_{k,i}$, there is $\widetilde{s}_i^{(\theta_i,s_i)} \in S_i$ such that $(\theta_i, \widetilde{s}_i^{(\theta_i,s_i)}) \in \mathrm{X}^{n-1}_{k-1,i}$ and $\widetilde{s}_i^{(\theta_i,s_i)}(h) = s_i(h)$ for all $h \in H(\mathrm{X}^{n-1}_{k})$ (thus, step n of the k-procedure path-refines step $n-1$ of the $(k-1)$-procedure);

For $n = K$, IH1 implies that, for each $(\theta, s) \in \mathrm{X}^K_{k-1}$, there exists $s' \in S$ such that $(\theta, s') \in \mathrm{X}^K_k$ and $\zeta(s) = \zeta(s')$. Since k is arbitrary in $\{1, ..., K\}$, this implies that for each $(\theta, s) \in \mathrm{X}^K_0 \subseteq C^{\Delta,\infty}_{\mathrm{sb}}$, there exists $s' \in S$ such that $(\theta, s') \in \mathrm{X}^K_K = C^{\infty}_{\mathrm{sb}}$ and $\zeta(s) = \zeta(s')$, that is, strong Δ-rationalizability path-refines strong rationalizability.

The rest of this section is dedicated to proving IH1 and IH2, following the strategy we outlined above. The formal proofs of Claims 1-4 are deferred to the Appendix.

Basis steps

IH2($n = k$) comes from the observation that, by inspection of (1), $\mathrm{X}_k^k \subseteq \mathrm{X}_k^{k-1} = C_{\mathrm{sb}}^{k-1} = \mathrm{X}_{k-1}^{k-1}$; IH1($n = k$) comes from (for all $i \in I$)

$$\begin{aligned} \mathrm{X}_{k-1,i}^k &= \left\{(\theta_i, s_i) \in \Theta_i \times S_i : \exists \mu^i \in \cap_{m=0}^{k-1} \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k-1,-i}^m) \cap \Delta_{i,\theta_i}, s_i \in r_{i,\theta_i}(\mu^i)\right\} \\ &\subseteq \left\{(\theta_i, s_i) \in \Theta_i \times S_i : \exists \mu^i \in \cap_{m=0}^{k-1} \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k,-i}^m), s_i \in r_{i,\theta_i}(\mu^i)\right\} = \mathrm{X}_{k,i}^k, \end{aligned}$$

where the first equality holds by (2), the last equality holds by (1), and the inclusion follows from the fact that, by (1), $\mathrm{X}_{k-1,-i}^m = C_{\mathrm{sb},-i}^m = \mathrm{X}_{k,-i}^m$ for all $m = 0, ..., k-1$.

Inductive steps

The proofs of two inductive steps, IH1($n+1$) and IH2($n+1$), are essentially identical, because both procedures $\left(\mathrm{X}_{k-1}^n\right)_{n=0}^{\infty}$ and $(\mathrm{X}_k^n)_{n=0}^{\infty}$ are defined by (2) at each step $n > k$. We start from the proof for IH2($n+1$), which uses IH2(n) and IH1(n). We relegate the proof of IH1($n+1$), which uses IH1(n) and IH2($n+1$) (which we prove first), to the appendix.

Inductive step IH2

Suppose IH1(n)-IH2(n) hold. We must show that IH2($n+1$) holds. Fix $i \in I$ and $(\theta_i, s_i) \in \mathrm{X}_{k,i}^{n+1}$. We are going to show the existence of a CPS $\widetilde{\mu}^{(\theta_i,s_i)} = \widetilde{\mu}^i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k-1,-i}^m) \cap \Delta_{i,\theta_i}$ and of a strategy $\widetilde{s}_i^{(\theta_i,s_i)} = \widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}^i) \subseteq \mathrm{X}_{k-1,i}^n$ such that $\widetilde{s}_i(h) = s_i(h)$ for all $h \in H(\mathrm{X}_k^n)$ (both the CPS and the strategy depend on the fixed pair (θ_i, s_i), thus, $\widetilde{s}_i = \widetilde{s}_i^{(\theta_i,s_i)}$; to ease notation, in what follows we do not make the dependence explicit). Since the choice of $i \in I$ and $(\theta_i, s_i) \in \mathrm{X}_{k,i}^{n+1}$ is arbitrary, this will prove IH2($n+1$).

By definition of $\mathrm{X}_{k,i}^{n+1}$ (see eq. (2)), there is some $\mu^i \in \cap_{m=0}^n \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k,-i}^m) \cap \Delta_{i,\theta_i}$ such that $s_i \in r_{i,\theta_i}(\mu^i)$.

Using IH2(n), we can construct a CPS $\widetilde{\mu}^i$ for step n of procedure $k-1$ that mimics μ_i along the paths $\bar{\zeta}(\mathrm{X}_k^n)$ that are consistent with step n of procedure k.

Claim 1 *There exists $\widetilde{\mu}^i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k-1,-i}^m) \cap \Delta_{i,\theta_i}$ such that, for every $h \in H\left(\mathrm{X}_k^n\right) \cap H_i(s_i)$,*

$$\forall (\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_k^n), \quad \widetilde{\mu}^i(\{\theta_{-i}\} \times S_{-i}(z)|h) = \mu^i(\{\theta_{-i}\} \times S_{-i}(z)|h). \tag{6}$$

IH2(n) also implies that the histories along those paths, $H\left(\mathrm{X}_k^n\right)$, are also consistent with step $n-1$ of procedure $k-1$.

Claim 2 *$H(\mathrm{X}_k^n) \subseteq H(\mathrm{X}_{k-1}^{n-1})$.*

In what follows, we will also use the following implication of standard dynamic programming arguments.

Claim 3 *Fix a subset of histories $\widetilde{H}$ such that, for every $h \in \widetilde{H}$, s_i is a continuation best reply to $\widetilde{\mu}^i(\cdot|h)$ for θ_i. There exists $\widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}^i)$ such that $\widetilde{s}_i(h) = s_i(h)$ for every $h \in \widetilde{H}$.*

Claim 2 allows to apply IH1(n) and say that every sequential best reply $\widetilde{s}_i$ to $\widetilde{\mu}^i$, which survives step n of procedure $k-1$, has a counterpart $\widetilde{s}_i'$ that survives step n or procedure k and mimics $\widetilde{s}_i$ at each $h \in H\left(\mathrm{X}_k^n\right) \cap H(\widetilde{s}_i')$. So, since obviously $\widetilde{s}_i'$ does not leave the paths $\bar{\zeta}(\mathrm{X}_k^n)$, neither $\widetilde{s}_i$ does, and hence it yields the same expected payoff under $\widetilde{\mu}^i(\cdot|h)$ and $\mu^i(\cdot|h)$, just like s_i, *if* $h \in H(s_i)$. But then the fact that s_i is a continuation best reply to μ^i at h, implies that it is also a continuation best reply to $\widetilde{\mu}^i$ at h.[21]

Claim 4 *For each $h \in H\left(\mathrm{X}_k^n\right) \cap H_i(s_i)$, strategy s_i is a continuation best reply to $\widetilde{\mu}^i(\cdot|h)$ for θ_i.*

So, by Claim 3, starting from the initial history and moving downwards, we can construct a sequential best reply to $\widetilde{\mu}^i$ that prescribes the same moves as s_i at all $h \in H\left(\mathrm{X}_k^n\right) \cap H(s_i)$. Now fix $\widetilde{\mu}^i$ as per Claim 1. From equation (2) it follows that $\{\theta_i\} \times r_{i,\theta_i}(\widetilde{\mu}^i) \subseteq \mathrm{X}_{k-1,i}^n$. To conclude the proof, we show the existence of $\widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}^i)$ such that $\widetilde{s}_i(h) = s_i(h)$ for all $h \in H(\mathrm{X}_k^n)$. By Claim 3 with $\widetilde{H} = H(\mathrm{X}_k^n) \cap H_i(s_i)$, this is a consequence of Claim 4. (For each $h \in H(\mathrm{X}_k^n) \backslash H_i(s_i)$, since $h \notin H_i(\widetilde{s}_i)$, we can always set $\widetilde{s}_i(h) = s_i(h)$ because we use the notion of sequential best reply which only refers to the histories that are consistent with the strategy.) ■

[21] This argument requires history h to be consistent with some sequential best reply to $\widetilde{\mu}^i$, and this can ensured with an inductive application of Claim 3.

4 Bayesian games

In the game with payoff uncertainty Γ, players' types θ parameterize the payoff functions to express incomplete and asymmetric information about them. Yet, the previous analysis does not prevent the parameters from containing payoff-irrelevant components; that is, the analysis remains valid if, for some player i and some types $\theta_i' \neq \theta_i''$, we have $u_j(\theta_i', \theta_{-i}, z) = u_j(\theta_i'', \theta_{-i}, z)$ for all $j \in I$, $\theta_{-i} \in \Theta_{-i}$, and $z \in Z$. However, we want to introduce such payoff-irrelevant components explicitly, in the following way. An **elaboration** of $\Gamma = \langle I, (\Theta_i, A_i, \mathcal{A}_i(\cdot), u_i)_{i \in I} \rangle$ is a structure

$$\Gamma^{\mathrm{e}} = \langle I, (T_i, A_i, \mathcal{A}_i(\cdot), u_i^{\mathrm{e}})_{i \in I} \rangle$$

such that, for every player $i \in I$, $T_i = \Theta_i \times E_i$, where E_i is a finite nonempty set, $u_i^{\mathrm{e}} : (\times_{j \in I} T_j) \times Z \to \mathbb{R}$, and

$$u_i^{\mathrm{e}}\left((\theta_k, e_k)_{k \in I}, z\right) = u_i\left((\theta_k)_{k \in I}, z\right)$$

for all $(\theta_k, e_k)_{k \in I} \in \times_{k \in I} T_k$ and $z \in Z$. In words, each type $t_i = (\theta_i, e_i)$ is made of the payoff-relevant component θ_i and of a payoff-irrelevant component e_i.

We are going to use the new types $(T_i)_{i \in I}$ as parts of a type structure *à la* Harsanyi (1967-68). Hence, we assign to each type t_i a probability measure $\beta_i(t_i)$ over the opponents' types T_{-i}, so that t_i is ultimately associated with a hierarchy of beliefs about the payoff-relevant parameters θ: the first-order belief is the marginal of $\beta_i(t_i)$ over Θ_{-i}; the second-order belief is the pushforward of $\beta_i(t_i)$ through the maps

$$(\theta_j, t_j)_{j \neq i} \in T_{-i} \mapsto \left(\theta_j, \mathrm{marg}_{\Theta_{-j}} \beta_j(t_j)\right)_{j \neq i} \in (\Theta_j \times \Delta(\Theta_{-j}))_{j \neq i};$$

and so forth. A **Bayesian elaboration** of $\Gamma = \langle I, (\Theta_i, A_i, \mathcal{A}_i(\cdot), u_i)_{i \in I} \rangle$ is obtained from adding the profile of belief maps $(\beta_i : T_i \to \Delta(T_{-i}))_{i \in I}$ to an elaboration:

$$\Gamma^{\mathrm{b}} = \left\langle I, \left(T_i, A_i, \mathcal{A}_i(\cdot), u_i^{\mathrm{b}}, \beta_i\right)_{i \in I} \right\rangle,$$

where $u_i^{\mathrm{b}} = u_i^{\mathrm{e}}$ for each $i \in I$. Note that an elaboration is essentially the same as the original game with payoff uncertainty when each set E_i is a singleton $\{\bar{e}_i\}$, so that Θ and

T are isomorphic (in an obvious sense). In this particular case, a Bayesian elaboration is also called "simple Bayesian game" and it adds to Γ a particular kind of profile of type-dependent restrictions on exogenous beliefs: recalling that we let $\bar{\Delta}_{i,\theta_i} \subseteq \Delta(\Theta_{-i})$ denote the restricted set of initial marginal beliefs of type θ_i of player i about co-players' types, we have that $\bar{\Delta}_{i,\theta_i} = \{\beta_i(\theta_i, \bar{e}_i)\}$ is a singleton for all i and θ_i.

Obviously, we can define **strong rationalizability for an elaboration** Γ^{e} as for Γ with each set Θ_i replaced by T_i: for each $i \in I$, $C^{\mathrm{e},0}_{i,\mathrm{sb}} = T_i \times S_i$, and for each $n \in \mathbb{N}$

$$C^{\mathrm{e},n}_{i,\mathrm{sb}} = \left\{(t_i, s_i) : \exists \mu^i \in \cap_{m=0}^{n-1} \Delta^H_{\mathrm{sb}}(C^{\mathrm{e},m}_{-i,\mathrm{sb}}), s_i \in r^{\mathrm{e}}_{i,t_i}(\mu^i)\right\},$$

where,

$$r^{\mathrm{e}}_{i,t_i}\left(\mu^i\right) = \left\{\bar{s}_i : \forall h \in H_i(\bar{s}_i), \bar{s}_i \in \arg\max_{s_i \in S_i(h)} \mathbb{E}_{\mu^i(\cdot|h)}\left(u^{\mathrm{e}}_i\left(t_i, \cdot, \zeta\left(s_i, \cdot\right)\right)\right)\right\}$$

for every CPS $\mu^i \in \Delta^H(T_{-i} \times S_{-i})$. Of course, by taking the sections of these sets at any given type, we obtain the strongly n-rationalizable strategies for that type:

$$S^{\mathrm{e},n}_i(t_i) = \left(C^{\mathrm{e},n}_{i,\mathrm{sb}}\right)_{t_i} := \left\{s_i : (t_i, s_i) \in C^{\mathrm{e},n}_{i,\mathrm{sb}}\right\}.$$

The following lemma formalizes the idea that the payoff-irrelevant component of types does not affect strong rationalizability.

Lemma 4 *Fix any elaboration* Γ^{e} *of* Γ. *For all* $i \in I$, $n \in \mathbb{N}_0$, *and* $(\theta_i, e_i) \in T_i$, $S^{\mathrm{e},n}_i(\theta_i, e_i) = S^n_i(\theta_i)$.

Now impose the belief system μ^i that justifies a pair (t_i, s_i) to be consistent with $\beta_i(t_i)$ *at the outset*. In this way, we define **strong rationalizability for a Bayesian elaboration** Γ^{b}: for each $i \in I$, $C^{\mathrm{b},0}_{i,\mathrm{sb}} = T_i \times S_i$, and for each $n \in \mathbb{N}$

$$C^{\mathrm{b},n}_{i,\mathrm{sb}} = \left\{(t_i, s_i) : \exists \mu^i \in \cap_{m=0}^{n-1} \Delta^H_{\mathrm{sb}}(C^{\mathrm{b},m}_{-i,\mathrm{sb}}), \mathrm{marg}_{T_{-i}}\mu^i(\cdot|\varnothing) = \beta_i(t_i), s_i \in r^{\mathrm{b}}_{i,t_i}(\mu^i)\right\},$$

where $r^{\mathrm{b}}_{i,t_i}(\mu^i) = r^{\mathrm{e}}_{i,t_i}(\mu^i)$ (defined above) for each $\mu^i \in \Delta^H(T_{-i} \times S_{-i})$, because $u^{\mathrm{b}}_i = u^{\mathrm{e}}_i$.

The set of strongly n-rationalizable strategies for type t_i in Γ^{b} is the section

$$S_i^{\mathrm{b},n}(t_i) = \left(C_{i,\mathrm{sb}}^{\mathrm{b},n}\right)_{t_i} := \left\{s_i : (t_i, s_i) \in C_{i,\mathrm{sb}}^{\mathrm{b},n}\right\}.$$

Strong rationalizability for a Bayesian elaboration is tightly related to strong directed rationalizability for the original game with payoff uncertainty. The equivalence is obvious for a simple Bayesian game, where each T_i is isomorphic to Θ_i (thus set $T_i = \Theta_i$), and for each θ_i, $\beta_i(\theta_i)$ can be taken as the unique initial belief allowed by $\bar{\Delta}_{i,\theta_i}$. Hence, a corollary of Theorem 1 is that *for every $\theta \in \Theta$, the (nonempty) set of strongly rationalizable paths of any (finite) simple Bayesian game based on a given (finite) multistage game with payoff uncertainty is included in the set of strongly rationalizable paths of the latter.* For a non-simple Bayesian elaboration Γ^{b} of Γ, one can perform an analogous exercise after defining an ancillary game with payoff uncertainty $\hat{\Gamma}$ with type sets $\hat{\Theta}_i = T_i$ in place of Θ_i. Strong rationalizability in Γ^{b} coincides with strong Δ-rationalizability in $\hat{\Gamma}$ with $\bar{\Delta}_{i,t_i} = \{\beta_i(t_i)\}$ for all $i \in I$ and $t_i \in \hat{\Theta}_i$; strong rationalizability in Γ coincides with strong rationalizability in $\hat{\Gamma}$ because $\hat{\Gamma}$ is an elaboration of Γ and thus Lemma 4 applies; the two things combined, via Theorem 1, yield the following result (the proof is omitted).

Theorem 2 *Fix any Bayesian elaboration Γ^{b} of Γ. Then, for every $n > 0$, for each $(\theta, e) \in T$, $\emptyset \neq \zeta\left(S^{\mathrm{b},n}(\theta, e)\right) \subseteq \zeta\left(S^n(\theta)\right)$, that is, for each $(\theta, e, s) \in C_{\mathrm{sb}}^{\mathrm{b},\infty} \neq \emptyset$, there exists $s' \in S$ such that $(\theta, s') \in C_{\mathrm{sb}}^{\infty}$ and $\zeta(s) = \zeta(s')$.*

5 Robust Implementation

We consider a classical mechanism design setting, which we formalize as follows. Fix a finite **economic environment**

$$\mathcal{E} = \left\langle I, Y, (\Theta_i, v_i)_{i \in I} \right\rangle,$$

where Y—a subset of a Euclidean space—is an outcome space and each $v_i : \Theta \times Y \to \mathbb{R}$ is a parameterized utility function. A special case of interest for the outcome space is a space of lotteries: $Y = \Delta(X)$, where X is a finite set of deterministic outcomes. In this case, $v_i(\theta, y)$ has to be interpreted as the vNM expected utility of lottery y given state of

nature θ. The economic environment collects the outcomes that the designer can assign to players and their preferences for such outcomes. A **multistage mechanism** (with observable actions) is a game form

$$\mathcal{M} = \left\langle I, \bar{H}, g \right\rangle,$$

where $g : Z \to Y$ is an outcome function defined on the set of terminal histories determined by the game tree $\bar{H}$. Thus, the mechanism specifies the rules of the game that determine the outcome. A pair $(\mathcal{E}, \mathcal{M})$ yields a game with payoff uncertainty

$$\Gamma(\mathcal{E}, \mathcal{M}) = \left\langle I, \bar{H}, \left(\Theta_i, (u_{i,\theta} = v_{i,\theta} \circ g)_{\theta \in \Theta}\right)_{i \in I} \right\rangle,$$

which contains both the rules of the game and the payoffs associated with the terminal histories: $u_{i,\theta}(z) = v_{i,\theta}(g(z))$ for all $\theta \in \Theta$ and $z \in Z$. Finally, we introduce a **social choice function**

$$f : \Theta \to Y,$$

representing the outcome the designer would want to realize as a function of players' types.

We are interested in the possibility of *implementing*, or at least *virtually implementing*, the social choice function; that is, we look for a mechanism where players of any types θ will always reach a terminal history z so that $g(z) = f(\theta)$, or at least $g(z) \approx f(\theta)$ in a sense to be made precise. Of course, the θ-dependent predicted path depends on the adopted solution concept. Following Mueller (2016), we adopt strong rationalizability and we focus on virtual implementation (v-implementation). Everything in the analysis is also valid for "exact" implementation.

Definition 1 *Social choice function f is **v-implementable under strong rationalizability** (in environment $\mathcal{E}$) if, for every $\varepsilon > 0$, there exists a multistage mechanism $\mathcal{M}$ such that, in game with payoff uncertainty $\Gamma(\mathcal{E}, \mathcal{M})$, for every $\theta \in \Theta$ and $s \in S^{\infty}(\theta) \neq \emptyset$, $\|g(\zeta(s)) - f(\theta)\| < \varepsilon$.*[22]

[22] In the definition, we require that $S^{\infty}(\theta) \neq \emptyset$ to avoid that the "for all ..." condition hold vacuously. In fact, we know from Lemma 3 that $S^{\infty}(\theta) \neq \emptyset$ for all $\theta \in \Theta$.

Bergemann & Morris (2009) introduce the notion of *robust implementation*, which requires the mechanism to implement the social choice function for any exogenous restrictions to players' collectively coherent hierarchies of beliefs about types, such as the existence of a common prior. As anticipated in the Introduction, in a static setting, one can show that implementation under rationalizability for static games with payoff uncertainty is robust, since—by monotonicity of probability-1 belief—the introduction of a Harsanyi type structure that restricts players' belief hierarchies can only reduce the set of their (interim correlated) rationalizable strategies.[23] As shown in Example 1, this is not true for strong rationalizability, due to the non-monotonicity of strong belief. For this reason, it was an open question whether Mueller's (2016) notion of implementation is robust in the sense of Bergemann & Morris (2009).

Definition 2 *Social choice function $f : \Theta \to Y$ is* ***robustly v-implementable under strong rationalizability*** *(in environment $\mathcal{E}$) if, for every $\varepsilon > 0$, there exists a multistage mechanism $\mathcal{M}$ such that, in every Bayesian elaboration $\Gamma^{\mathrm{b}}(\mathcal{E}, \mathcal{M})$ of the game with payoff uncertainty $\Gamma(\mathcal{E}, \mathcal{M})$, for all $t = (\theta, e) \in T$ and $s \in S^{\mathrm{b},\infty}(t) \neq \emptyset$, $\|g(\zeta(s)) - f(\theta)\| < \varepsilon$.*

In light of Theorem 2, we can give a positive answer to the open question.

Corollary 1 *Fix a finite economic environment $\mathcal{E}$ and a social choice function $f : \Theta \to Y$. If f is v-implementable under strong rationalizability, then f is also* robustly *v-implementable under strong rationalizability.*

Proof. Suppose that f is v-implementable under strong rationalizability and let $\mathcal{M}$ be a mechanism such that, in game with payoff uncertainty $\Gamma(\mathcal{E}, \mathcal{M})$, for all $\theta \in \Theta$ and $s \in S^{\infty}(\theta)$, $\|g(\zeta(s)) - f(\theta)\| < \varepsilon$. Take any Bayesian elaboration $\Gamma^{\mathrm{b}}(\mathcal{E}, \mathcal{M})$ of $\Gamma(\mathcal{E}, \mathcal{M})$. By Theorem 2, for all $(\theta, e) \in \Theta \times E = T$ and $s \in S^{\mathrm{b},\infty}(\theta, e)$, $\emptyset \neq \zeta\left(S^{\mathrm{b},\infty}(\theta, e)\right) \subseteq \zeta(S^{\infty}(\theta))$. It follows that, for all $t = (\theta, e) \in T$ and $s \in S^{\mathrm{b},\infty}(t) \neq \emptyset$, $\|g(\zeta(s)) - f(\theta)\| < \varepsilon$. ■

[23] Interim correlated rationalizbaility is the appropriate notion of rationalizability for Bayesian games. See Bergemann & Morris (2012), Battigalli *et al.* (2011), and the relevant references therein.

6 Appendix

This section contains the proofs omitted from the main body of the paper (with the exception of the detailed proof of inductive step IH1 in the proof of Theorem 1, which is contained in the Supplemental Appendix).

6.1 Proof of Lemma 2

We prove this result by contraposition. Suppose that $s_i \notin r_{i,\theta_i}(\mu^i)$. We need to show that there is some $\bar{h} \in H_i(s_i)$ such that, for every $s_i' \in S_i\left(\bar{h}\right)$, if $s_i'\left(\bar{h}\right) = s_i\left(\bar{h}\right)$, then s_i' is not a continuation best reply to $\mu^i(\cdot|\bar{h})$ for θ_i. Let $H_i^D(s_i, \mu^i)$ denote the set of histories $h \in H_i(s_i)$ such that s_i is not a continuation best reply to $\mu^i(\cdot|h)$. Since the game is finite, $H_i^D(s_i, \mu^i)$ has at least one maximal element $\bar{h}$, that is, $\bar{h} \in H_i^D(s_i, \mu^i)$ is not the strict prefix of any other $h \in H_i^D(s_i, \mu^i)$. Since $\bar{h} \in H_i^D(s_i, \mu^i)$, there is some $\bar{s}_i \in S_i\left(\bar{h}\right)$ such that

$$\mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, \bar{s}_i, \cdot)\right) > \mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i, \cdot)\right). \tag{7}$$

Pick any $s_i' \in S_i\left(\bar{h}\right)$ such that $s_i'\left(\bar{h}\right) = s_i\left(\bar{h}\right)$ (this includes $s_i' = s_i$). To take care of the possibility that $\left(\bar{h}, \left(s_i\left(\bar{h}\right), a_{-i}\right)\right) \in Z$ for some a_{-i} and to ease notation, for all z such that $\mu^i\left(\Theta_{-i} \times S_{-i}(z)\,|\bar{h}\right) > 0$ and all $(\theta_{-i}, s_{-i}) \in \Theta_{-i} \times S_{-i}(z)$, write

$$\mu^i\left(\theta_{-i}, s_{-i}|z\right) = \frac{\mu^i\left(\theta_{-i}, s_{-i}|\bar{h}\right)}{\mu^i\left(\Theta_{-i} \times S_{-i}(z)\,|\bar{h}\right)},$$

so that

$$\mathbb{E}_{\mu^i(\cdot|z)}\left(U_i(\theta_i, s_i', \cdot)\right) = \sum_{\theta_{-i} \in \Theta_{-i}} \mu^i\left(\{\theta_{-i}\} \times S_{-i}(z)\,|z\right) u_i\left(\theta_i, \theta_{-i}, z\right).$$

With this, $\mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i', \cdot)\right)$ can be decomposed as follows:

$$\begin{aligned} &\mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i', \cdot)\right) \\ = &\sum_{a_{-i}:\mu^i\left(\Theta_{-i} \times S_{-i}\left(\bar{h}, a_{-i}\right)|\bar{h}\right) > 0} \mu^i\left(\Theta_{-i} \times S_{-i}\left(\bar{h}, a_{-i}\right)|\bar{h}\right) \mathbb{E}_{\mu^i\left(\cdot|\left(\bar{h}, \left(s_i\left(\bar{h}\right), a_{-i}\right)\right)\right)}\left(U_i(\theta_i, s_i', \cdot)\right). \end{aligned}$$

By choice of $\bar{h}$, s_i is a continuation best reply to each $\mu^i(\cdot|h)$ with $h = \left(\bar{h}, \left(s_i\left(\bar{h}\right), a_{-i}\right)\right) \in H$. Thus,

$$\mathbb{E}_{\mu^i\left(\cdot|\left(\bar{h},\left(s_i\left(\bar{h}\right),a_{-i}\right)\right)\right)}\left(U_i(\theta_i, s_i, \cdot)\right) \geq \mathbb{E}_{\mu^i\left(\cdot|\left(\bar{h},\left(s_i\left(\bar{h}\right),a_{-i}\right)\right)\right)}\left(U_i(\theta_i, s_i', \cdot)\right)$$

for all a_{-i} with $\mu^i\left(\Theta_{-i} \times S_{-i}\left(\bar{h}, a_{-i}\right)|h\right) > 0$ (the other action profiles in $\mathcal{A}_{-i}\left(\bar{h}\right)$ do not affect expected payoff calculations). It follows that

$$\mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i, \cdot)\right) \geq \mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i', \cdot)\right). \tag{8}$$

Equations (7) and (8) combined yield

$$\mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, \bar{s}_i, \cdot)\right) > \mathbb{E}_{\mu^i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i', \cdot)\right),$$

so s_i' cannot be a continuation best reply to $\mu^i\left(\cdot|\bar{h}\right)$. ■

6.2 Omitted parts of the proof of Theorem 1

6.2.1 Proof of Claim 1

We are going to construct an array of beliefs $\hat{\mu}^i = \left(\hat{\mu}^i(\cdot|h)\right)_{h \in H_i}$ such that, for each $h \in H_i$:

F0. $\hat{\mu}^i\left(\Theta_{-i} \times S_{-i}(h)|h\right) = 1$;

F1. for all h' such that $h \prec h'$, for each $E \subseteq \Theta_{-i} \times S_{-i}(h')$,

$$\hat{\mu}^i\left(E|h'\right)\hat{\mu}^i\left(\Theta_{-i} \times S_{-i}(h')|h\right) = \hat{\mu}^i\left(E|h\right);$$

F2. for each $h \in H$ and $m = 0, ..., n-1$, if $h \in H(\mathrm{X}_{k-1,-i}^m)$, then $\hat{\mu}^i\left(\mathrm{X}_{k-1,-i}^m|h\right) = 1$;

F3. $\mathrm{marg}_{\Theta_{-i}}\hat{\mu}^i(\cdot|\varnothing) = \mathrm{marg}_{\Theta_{-i}}\mu^i(\cdot|\varnothing)$;

F4. for every $h \in H\left(\mathrm{X}_k^n\right) \cap H_i(s_i)$,

$$\forall(\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_k^n), \quad \hat{\mu}^i(\{\theta_{-i}\} \times S_{-i}(z)|h) = \mu^i(\{\theta_{-i}\} \times S_{-i}(z)|h). \tag{9}$$

By F0 and F1, $\hat{\mu}^i$ is a forward-consistent belief system. By F2, it strongly believes $\mathrm{X}^1_{k-1,-i}, ..., \mathrm{X}^{n-1}_{k-1,-i}$. Hence, by Lemma 1, there exists a CPS $\widetilde{\mu}^i \in \cap_{m=0}^{n-1}\Delta^H_{\mathrm{sb}}(\mathrm{X}^m_{k-1,-i})$ such that $\widetilde{\mu}^i(\cdot|h) = \hat{\mu}^i(\cdot|h)$ for all $h \in H(s_i)$. By $\widetilde{\mu}^i(\cdot|\varnothing) = \hat{\mu}^i(\cdot|\varnothing)$ and F3, we get $\widetilde{\mu}^i \in \Delta_{i,\theta_i}$. Finally, for every $h \in H\left(\mathrm{X}^n_k\right) \cap H_i(s_i)$, $\widetilde{\mu}^i(\cdot|h) = \hat{\mu}^i(\cdot|h)$ and F4 yield (6).

Now we start with the construction. By IH2(n), for every $(\theta_{-i}, s_{-i}) \in \mathrm{X}^n_{k,-i}$, there exists a profile $(\widetilde{s}_j^{(\theta_j,s_j)})_{j\neq i} \in S_{-i}$ such that $(\theta_j, \widetilde{s}_j^{(\theta_j,s_j)})_{j\neq i} \in \mathrm{X}^{n-1}_{k-1,-i}$ and, for each $j \neq i$, $\widetilde{s}_j^{(\theta_j,s_j)}(h) = s_j(h)$ for every $h \in H(\mathrm{X}^{n-1}_k)$. With this, define a map $\widetilde{\eta} : \Theta_{-i} \times S_{-i} \to \Theta_{-i} \times S_{-i}$ as follows:

$$\forall\left(\theta_{-i}, s_{-i}\right) \in \left(\Theta_{-i} \times S_{-i}\right), \quad \widetilde{\eta}\left(\theta_{-i}, s_{-i}\right) = \begin{cases} (\theta_j, \widetilde{s}_j^{(\theta_j,s_j)})_{j\neq i} & \text{if } (\theta_{-i}, s_{-i}) \in \mathrm{X}^n_{k,-i} \\ (\theta_{-i}, s_{-i}) & \text{otherwise} \end{cases}.$$

For each $h \in H\left(\mathrm{X}^n_k\right)$, define $\hat{\mu}^i\left(\cdot|h\right)$ as the $\widetilde{\eta}$-pushforward of $\mu^i\left(\cdot|h\right)$. For future reference, observe that

$$\hat{\mu}^i\left(\mathrm{X}^{n-1}_{k-1,-i}|h\right) = \mu^i\left(\widetilde{\eta}^{-1}(\mathrm{X}^{n-1}_{k-1,-i})|h\right) \geq \mu^i\left(\mathrm{X}^n_{k,-i}|h\right) = 1, \tag{10}$$

where the first equality is by construction, the inequality is by $\widetilde{\eta}(\mathrm{X}^n_{k,-i}) \subseteq \mathrm{X}^{n-1}_{k-1,-i}$, and the last equality is by strong belief in $\mathrm{X}^n_{k,-i}$. Now define

$$\widetilde{H} = \left\{h \in H \backslash H\left(\mathrm{X}^n_k\right) : \exists \bar{h} \in H\left(\mathrm{X}^n_k\right), \bar{h} \prec h, \hat{\mu}^i\left(\Theta_{-i} \times S_{-i}(h)|\bar{h}\right) > 0\right\}.$$

For each $h \in \widetilde{H}$, let $p^*(h)$ denote the longest $\bar{h} \prec h$ with $\bar{h} \in H\left(\mathrm{X}^n_k\right)$ such that $\hat{\mu}^i\left(\Theta_{-i} \times S_{-i}(h)|\bar{h}\right) > 0$, and derive $\hat{\mu}^i\left(\cdot|h\right)$ by conditioning $\hat{\mu}^i\left(\cdot|p^*(h)\right)$. To conclude the construction, fix $\bar{\mu}^i \in \cap_{m=0}^{n-1}\Delta^H_{\mathrm{sb}}(\mathrm{X}^m_{k-1,-i}) \cap \Delta_{i,\theta_i}$, and for each $h \in H \backslash \left(H\left(\mathrm{X}^n_k\right) \cup \widetilde{H}\right) =: \hat{H}$, let $\hat{\mu}^i\left(\cdot|h\right) = \bar{\mu}^i\left(\cdot|h\right)$.

First, we show that $\hat{\mu}^i$ satisfies F2. For each $h \in H\left(\mathrm{X}^n_k\right)$, equation (10) yields $\hat{\mu}^i\left(\mathrm{X}^{n-1}_{k-1,-i}|h\right) = 1$. For each $h \in \widetilde{H}$, equation (10) yields $\hat{\mu}^i\left(\mathrm{X}^{n-1}_{k-1,-i}|p^*(h)\right) = 1$, from which $\hat{\mu}^i\left(\mathrm{X}^{n-1}_{k-1,-i}|h\right) = 1$ follows by construction. For each $h \in \hat{H}$ and $m = 0, ..., n-1$, if $h \in H(\mathrm{X}^m_{k-1,-i})$, $\hat{\mu}^i\left(\mathrm{X}^m_{k-1,-i}|h\right) = 1$ follows from $\hat{\mu}^i\left(\cdot|h\right) = \bar{\mu}^i\left(\cdot|h\right)$ and $\bar{\mu}^i \in \Delta^H_{\mathrm{sb}}(\mathrm{X}^m_{k-1,-i})$.

Next, we show that, for every $h \in H\left(\mathrm{X}^n_k\right)$ and $(\theta_{-i}, h') \in \Theta_{-i} \times (H(\mathrm{X}^n_k) \cup \bar{\zeta}(\mathrm{X}^n_k))$,

$$\hat{\mu}^i(\{\theta_{-i}\} \times S_{-i}(h')|h) = \mu^i(\{\theta_{-i}\} \times S_{-i}(h')|h), \tag{11}$$

which yields:

- condition (9) when $h' \in \bar{\zeta}(\mathrm{X}_k^n)$, thus F4;

- F3 when h and h' coincide with the initial history;

- and, for future reference,

$$\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) = \mu^i(\Theta_{-i} \times S_{-i}(h')|h). \tag{12}$$

By construction, we have

$$\hat{\mu}^i(\{\theta_{-i}\} \times S_{-i}(h')|h) = \mu^i(\widetilde{\eta}^{-1}(\{\theta_{-i}\} \times S_{-i}(h'))|h).$$

We need to show that

$$\widetilde{\eta}^{-1}\left(\{\theta_{-i}\} \times S_{-i}(h')\right) = \{\theta_{-i}\} \times S_{-i}(h'). \tag{13}$$

Fix first $s_{-i} \in S_{-i}$ such that $(\theta_{-i}, s_{-i}) \in \widetilde{\eta}^{-1}\left(\{\theta_{-i}\} \times S_{-i}(h')\right)$. Then, there exists $s'_{-i} \in S_{-i}(h')$ such that $\widetilde{\eta}(\theta_{-i}, s_{-i}) = (\theta_{-i}, s'_{-i})$. By definition of $\widetilde{\eta}$, either $s'_{-i} = s_{-i}$, or $s_{-i}(\widetilde{h}) = s'_{-i}(\widetilde{h})$ for each $\widetilde{h} \in H(\mathrm{X}_k^{n-1})$, hence for each $\widetilde{h} \prec h'$, given that $h' \in H(\mathrm{X}_k^n) \cup \bar{\zeta}(\mathrm{X}_k^n)$. So, $s'_{-i} \in S_{-i}(h')$ implies $s_{-i} \in S_{-i}(h')$, i.e., $(\theta_{-i}, s_{-i}) \in \{\theta_{-i}\} \times S_{-i}(h')$. Now fix $s_{-i} \in S_{-i}(h')$. Let $(\theta_{-i}, s'_{-i}) = \widetilde{\eta}(\theta_{-i}, s_{-i})$. By definition of $\widetilde{\eta}$, either $s'_{-i} = s_{-i}$, or $s'_{-i}(\widetilde{h}) = s_{-i}(\widetilde{h})$ for each $\widetilde{h} \in H(\mathrm{X}_k^{n-1})$, hence for each $\widetilde{h} \prec h'$, given that $h' \in H(\mathrm{X}_k^n) \cup \bar{\zeta}(\mathrm{X}_k^n)$. So $s_{-i} \in S_{-i}(h')$ implies $s'_{-i} \in S_{-i}(h')$, which means $(\theta_{-i}, s_{-i}) \in \widetilde{\eta}^{-1}\left(\{\theta_{-i}\} \times S_{-i}(h')\right)$.

Finally, we show that $\hat{\mu}^i$ satisfies F0 and F1.

For each $h \in H\left(\mathrm{X}_k^n\right)$, since $\mu^i(\Theta_{-i} \times S_{-i}(h)|h) = 1$, equation 11 with $h' = h$ yields F0. For each $h \in \widetilde{H}$, F0 follows by conditioning. For each $h \in \hat{H}$, F0 holds by $\hat{\mu}^i\left(\cdot|h\right) = \bar{\mu}^i\left(\cdot|h\right)$.

For F1, fix $h, h' \in H$ such that $h \prec h'$. We want to show that

$$\forall E \subseteq \Theta_{-i} \times S_{-i}(h'),\ \ \hat{\mu}^i(E|h')\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) = \hat{\mu}^i(E|h). \tag{14}$$

This is true if $\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) = 0$, because then $\hat{\mu}^i(E|h) = 0$, so suppose that $\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) > 0$.

Case 1: $h \in \hat{H}$. Then $h' \in \hat{H}$ too. Hence, $\hat{\mu}^i(\cdot|h) = \bar{\mu}^i(\cdot|h)$ and $\hat{\mu}^i(\cdot|h') = \bar{\mu}^i(\cdot|h')$, so $\hat{\mu}^i$ inherits (14) from $\bar{\mu}^i$, which is a CPS.
Case 2: $h \in \widetilde{H}$. Then $\hat{\mu}^i(\cdot|h)$ is derived from $\hat{\mu}^i(\cdot|p^*(h))$ by conditioning. By $\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) > 0$, we have $\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|p^*(h)) > 0$, hence $h' \in \widetilde{H}$ too and $p^*(h) = p^*(h')$. Thus, $\hat{\mu}^i(\cdot|h')$ is derived from $\hat{\mu}^i(\cdot|p^*(h))$ too, and (14) follows.
Case 3: $h \in H(\mathrm{X}_k^n)$. If $h' \in H(\mathrm{X}_k^n)$, let $\bar{h} = h'$, otherwise, by $\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) > 0$, $h' \in \widetilde{H}$, and in this case let $\bar{h} = p^*(h')$. Thus, $\bar{h} \in H(\mathrm{X}_k^n)$. For each $E \subseteq \Theta_{-i} \times S_{-i}(\bar{h})$, by construction of $\hat{\mu}^i$ and equation (12), we get

$$\hat{\mu}^i(E|\bar{h})\hat{\mu}^i(\Theta_{-i} \times S_{-i}(\bar{h})|h) = \mu^i(\widetilde{\eta}^{-1}(E)|\bar{h})\mu^i(\Theta_{-i} \times S_{-i}(\bar{h})|h).$$

Equation (13) implies that $\widetilde{\eta}^{-1}(E) \subseteq \Theta_{-i} \times S_{-i}(\bar{h})$, so, since μ^i is a CPS, we have

$$\mu^i(\widetilde{\eta}^{-1}(E)|\bar{h})\mu^i(\Theta_{-i} \times S_{-i}(\bar{h})|h) = \mu^i(\widetilde{\eta}^{-1}(E)|h),$$

and $\mu^i(\widetilde{\eta}^{-1}(E)|h) = \hat{\mu}^i(E|h)$ by construction of $\hat{\mu}^i$. So,

$$\hat{\mu}^i(E|\bar{h})\hat{\mu}^i(\Theta_{-i} \times S_{-i}(\bar{h})|h) = \hat{\mu}^i(E|h). \tag{15}$$

If $\bar{h} = h'$, we are done. Otherwise, for each $E \subseteq \Theta_{-i} \times S_{-i}(h')$, we have

$$\begin{aligned}
\hat{\mu}^i(E|h')\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) &= \frac{\hat{\mu}^i(E|p^*(h'))}{\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|p^*(h'))}\hat{\mu}^i(\Theta_{-i} \times S_{-i}(h')|h) \\
&= \hat{\mu}^i(E|p^*(h'))\hat{\mu}^i(\Theta_{-i} \times S_{-i}(p^*(h'))|h) \\
&= \hat{\mu}^i(E|h),
\end{aligned}$$

where the first equality is by definition of $\hat{\mu}^i(E|h')$, the second equality follows from Case 2, and the third equality holds by equation (15) with $\bar{h} = p^*(h')$. □

6.2.2 Proof of Claim 2

Fix $\hat{s} \in \mathrm{Proj}_S \mathrm{X}_k^n$. By IH2($n$), there exists $\hat{s}' \in \mathrm{Proj}_S \mathrm{X}_{k-1}^{n-1}$ such that $\hat{s}'(\widetilde{h}) = \hat{s}(\widetilde{h})$ for every $\widetilde{h} \in H(\mathrm{X}_k^{n-1}) \supseteq H(\mathrm{X}_k^n)$. It follows that $\zeta(\hat{s}) = \zeta(\hat{s}') \in \bar{\zeta}(\mathrm{X}_{k-1}^{n-1})$. □

6.2.3 Proof of Claim 3

Construct $\widetilde{s}_i$ as follows. For each $h \in \widetilde{H}$, let $\widetilde{s}_i(h) = s_i(h)$. For each $h \in H \backslash \widetilde{H}$, let $\widetilde{s}_i(h) = s'_i(h)$ for some continuation best reply s'_i to $\widetilde{\mu}^i(\cdot|h)$ for θ_i. It follows from Lemma 2 that $\widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}^i)$. □

6.2.4 Proof of Claim 4

First note that $H\left(\mathrm{X}_k^n\right) \cap H_i(s_i)$ is closed with respect to prefixes (predecessors): for each $h \in H\left(\mathrm{X}_k^n\right) \cap H_i(s_i)$ each prefix $h' \prec h$ belongs to $H\left(\mathrm{X}_k^n\right) \cap H_i(s_i)$. So, suppose by way of induction that Claim 4 holds for every $h' \prec h$ — this is vacuously true if $h = \varnothing$. Then, setting $\widetilde{H} = \{h' \in H : h' \prec h\}$, Claim 3 guarantees the existence of some $\widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}^i)$ such that $\widetilde{s}_i(h') = s_i(h')$ for every $h' \prec h$, thus $\widetilde{s}_i \in S_i(h)$. First, we need to show that $\zeta(\widetilde{s}_i, \widetilde{s}_{-i}) \in \bar{\zeta}(\mathrm{X}_k^n)$ for every $(\theta_{-i}, \widetilde{s}_{-i}) \in \mathrm{Supp}\widetilde{\mu}^i(\cdot|h)$. So, fix $(\theta_{-i}, \widetilde{s}_{-i}) \in \mathrm{Supp}\widetilde{\mu}^i(\cdot|h)$. By Claim 1, $\{\theta_i\} \times r_{i,\theta_i}(\widetilde{\mu}^i) \subseteq \mathrm{X}_{k-1,i}^n$, and hence $\widetilde{s}_i \in \mathrm{Proj}_{S_i}\mathrm{X}_{k-1,i}^n$. So, by IH1($n$) there exists $\widetilde{s}'_i \in \mathrm{Proj}_{S_i}\mathrm{X}_{k,i}^n$ such that $\widetilde{s}'_i(h) = \widetilde{s}_i(h)$ for every $h \in H(\mathrm{X}_{k-1}^{n-1})$.[24] Fix $\left(\theta_{-i}, \widetilde{s}'_{-i}\right) \in \widetilde{\eta}^{-1}((\theta_{-i}, \widetilde{s}_{-i})) \subseteq \mathrm{X}_{k,-i}^n$ (map $\widetilde{\eta}$ is defined in the proof of Claim 1). Obviously, $\zeta(\widetilde{s}'_i, \widetilde{s}'_{-i}) \in \bar{\zeta}(\mathrm{X}_k^n)$. For every $\widetilde{h} \prec \zeta(\widetilde{s}'_i, \widetilde{s}'_{-i})$, we have $\widetilde{h} \in H(\mathrm{X}_k^n) \subseteq H(\mathrm{X}_k^{n-1})$, hence $\widetilde{s}_{-i}(\widetilde{h}) = \widetilde{s}'_{-i}(\widetilde{h})$ by construction of $\widetilde{\eta}$. Claim 2 gives $H(\mathrm{X}_k^n) \subseteq H(\mathrm{X}_{k-1}^{n-1})$, therefore $\widetilde{s}_i(\widetilde{h}) = \widetilde{s}'_i(\widetilde{h})$ as well. It follows that $\zeta(\widetilde{s}_i, \widetilde{s}_{-i}) = \zeta(\widetilde{s}'_i, \widetilde{s}'_{-i}) \in \bar{\zeta}(\mathrm{X}_k^n)$.

For each $(\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_k^n)$, the probability of (θ_{-i}, z) induced by $\widetilde{s}_i$ and $\widetilde{\mu}^i(\cdot|h)$ (resp., $\mu^i(\cdot|h)$) is 0, if $\widetilde{s}_i \notin S_i(z)$, or $\widetilde{\mu}^i(\{\theta_{-i}\} \times S_{-i}(z)|h)$ (resp., $\mu^i(\{\theta_{-i}\} \times S_{-i}(z)|h)$) otherwise. Then, by equation (6), $\widetilde{s}_i$ induces the same probability over each $(\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_k^n)$ under $\widetilde{\mu}^i(\cdot|h)$ and under $\mu^i(\cdot|h)$, hence the same distribution over $\Theta_{-i} \times Z$, because the probability induced by $\widetilde{s}_i$ and $\widetilde{\mu}^i(\cdot|h)$ over $\Theta_{-i} \times \left(Z \backslash \bar{\zeta}(\mathrm{X}_k^n)\right)$ is zero: as we have previously shown, for each $(\theta_{-i}, \widetilde{s}_{-i}) \in \mathrm{Supp}\widetilde{\mu}^i(\cdot|h)$, $\zeta(\widetilde{s}_i, \widetilde{s}_{-i}) \in \bar{\zeta}(\mathrm{X}_k^n)$. The same conclusion can be reached for s_i in the same way, after observing that for each $(\theta_{-i}, s_{-i}) \in \mathrm{Supp}\mu^i(\cdot|h)$, since $(\theta_i, s_i, \theta_{-i}, s_{-i}) \in \mathrm{X}_k^n$, we have $\zeta(s_i, s_{-i}) \in \bar{\zeta}(\mathrm{X}_k^n)$. So, call $\pi^{\widetilde{s}_i}$ and π^{s_i} the unique expected payoffs induced by, respectively, $(\theta_i, \widetilde{s}_i)$ and (θ_i, s_i) under both beliefs ($\mu^i(\cdot|h)$ and $\widetilde{\mu}^i(\cdot|h)$). Since $\widetilde{s}_i$ and s_i are continuation best replies for θ_i to, respectively, $\widetilde{\mu}^i(\cdot|h)$ and $\mu^i(\cdot|h)$, we have $\pi^{\widetilde{s}_i} \geq \pi^{s_i}$ and $\pi^{s_i} \geq \pi^{\widetilde{s}_i}$. Hence, $\pi^{s_i} = \pi^{\widetilde{s}_i}$. But then, also s_i is a continuation best reply for θ_i to $\widetilde{\mu}^i(\cdot|h)$. □

[24] This is the only passage where we use IH1(n) at full power, namely, where it is important (to then apply Claim 3) that IH1(n) involves all the histories in $H(\mathrm{X}_{k-1}^{n-1})$ and not just those in $H(\mathrm{X}_{k-1}^n)$.

6.3 Proof of Lemma 4

The statement is trivially true for $n = 0$. Suppose by way of induction that it is true for each $m \leq n$; fix $i \in I$ and $(\theta_i, e_i) \in T_i = \Theta_i \times E_i$ arbitrarily. Let $\bar{s}_i \in S_i^n(\theta_i)$. Then there is a CPS $\mu^i \in \cap_{m=0}^{n-1} \Delta_{\text{sb}}^H(C_{-i,\text{sb}}^m)$ such that $\bar{s}_i \in r_{i,\theta_i}(\mu^i)$. Define $\mu^{\text{e},i} \in \Delta(T_{-i} \times S_{-i})^H$ as follows: for all $h \in H$, $s_{-i} \in S_{-i}(h)$, $(\theta_{-i}, e_{-i}) \in T_{-i}$,

$$\mu^{\text{e},i}(\theta_{-i}, e_{-i}, s_{-i}|h) = \frac{1}{|E_{-i}|} \mu^i(\theta_{-i}, s_{-i}|h).$$

It can be checked that $\mu^{\text{e},i}$ is a CPS, that is, $\mu^{\text{e},i} \in \Delta^H(T_{-i} \times S_{-i})$. Furthermore, since $\mu^i(\cdot|h) = \text{marg}_{\Theta_{-i} \times S_{-i}(h)} \mu^{\text{e},i}(\cdot|h)$ for each $h \in H$, and the e_j-component of the type of each player $j \in I$ is payoff-irrelevant, $\bar{s}_i \in r_{i,(\theta_i,e_i)}(\mu^{\text{e},i})$. Finally, the aforementioned marginalization relationship between μ^i and $\mu^{\text{e},i}$ and the inductive hypothesis imply that $\mu^{\text{e},i} \in \cap_{m=0}^{n-1} \Delta_{\text{sb}}^H(C_{-i,\text{sb}}^{\text{e},m})$. Therefore, $\bar{s}_i \in S_i^{\text{e},n}(\theta_i, e_i)$. Conversely, suppose that $\bar{s}_i \in S_i^{\text{e},n}(\theta_i, e_i)$. Then there is a CPS $\mu^{\text{e},i} \in \cap_{m=0}^{n-1} \Delta_{\text{sb}}^H(C_{-i,\text{sb}}^{\text{e},m})$ such that $\bar{s}_i \in r_{i,(\theta_i,e_i)}(\mu^{\text{e},i})$. Define $\mu^i \in (\Theta_{-i} \times S_{-i})^H$ as $\mu^i(\cdot|h) = \text{marg}_{\Theta_{-i} \times S_{-i}(h)} \mu^{\text{e},i}(\cdot|h)$ for each $h \in H$. It can be checked that μ^i is a CPS, that is, $\mu^i \in \Delta^H(\Theta_{-i} \times S_{-i})$. Similarly to the previous argument, since the e_j-component of the type of each player $j \in I$ is payoff-irrelevant, $\bar{s}_i \in r_{i,\theta_i}(\mu^i)$. Furthermore, the marginalization relationship between μ^i and $\mu^{\text{e},i}$ and the inductive hypothesis imply that $\mu^i \in \cap_{m=0}^{n-1} \Delta_{\text{sb}}^H(C_{-i,\text{sb}}^m)$. ■

References

[1] Abreu, D., and H. Matsushima (1992): "Virtual Implementation in Iteratively Undominated Strategies: Complete Information," *Econometrica*, 60, 993-1008.

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

[3] Battigalli, P. (1999): "Rationalizability in Incomplete Information Games," EUI Working Paper ECO No. 99/17.

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

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

[6] Battigalli, P., and A. Prestipino (2013): "Transparent Restrictions on Beliefs and Forward Induction Reasoning in Games with Asymmetric Information," *The B.E. Journal of Theoretical Economics (Contributions)*, 13 (1), 1-53.

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

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

[9] Battigalli, P., and M. Siniscalchi (2003): "Rationalization and Incomplete Information," *Advances in Theoretical Economics*, 3 (1), Art. 3.

[10] Battigalli, P., Catonini, E., Manili, J. (2023): "Belief Change, Rationality, and Strategic Reasoning in Sequential Games," *Games and Economic Behavior*, 142, 527-551.

[11] Battigalli, P., Catonini, E., De Vito, N. (2023): *Game Theory: Analysis of Strategic Thinking*, manuscript.

[12] Battigalli, P., A. Di Tillio, E. Grillo and A. Penta (2011): "Interactive Epistemology and Solution Concepts for Games with Asymmetric Information," *The B.E. Journal of Theoretical Economics* (*Advances*), 11 (1), Article 6.

[13] Bergemann, D., and S. Morris (2009): "Robust Virtual Implementation," *Theoretical Economics*, 4, 45-88.

[14] Bergemann, D., and S. Morris (2012): "An Introduction to Robust Mechanism Design," *Foundations and Trends in Microeconomics*, 3, 169-230.

[15] Bergemann, D., and S. Morris (2017): "Belief-Free Rationalizability and Informational Robustness," *Games and Economic Behavior*, 104. 744–759.

[16] Brandenburger A. and E. Dekel (1993): "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory*, 59, 189-198.

[17] Catonini, E. (2019): "Rationalizability and epistemic priority orderings," *Games and Economic Behavior*, 114, 101-117.

[18] Catonini, E. (2020): "On Non-Monotonic Strategic Reasoning," *Games and Economic Behavior*, 120, 209-224.

[19] Harsanyi, J. (1967-68): "Games of Incomplete Information Played by Bayesian Players. Parts I, II, III," *Management Science*, 14, 159-182, 320-334, 486-502.

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

[21] Mueller, C. (2016): "Robust Virtual Implementation under Common Strong Belief in Rationality," *Journal of Economic Theory,* 162, 407–450.

[22] Mueller, C. (2020): "Robust Implementation in Weakly Perfect Bayesian Strategies," *Journal of Economic Theory,* 189, 105038.

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

[24] Wilson, R. (1987): "Game-Theoretic Analyses of Trading Processes," in (T. Bewley, Ed.) *Advances in Economic Theory, Fifth World Congress,* Vol. 1„ 33-70. New York. Cambridge University Press.

7 Supplemental Appendix

The first subsection gives a rigorous proof of the inductive step IH1 in the proof of Theorem 1. The second subsection contains an example where path monotonicity fails due to a kind of restriction on *endogenous* beliefs, i.e., beliefs about the co-player type conditional on the observed action of the co-player.

7.1 Proof of inductive step IH1 in the proof of Theorem 1

Suppose IH1(n)-IH2(n) hold. We proved that IH2($n+1$) holds as well. Thus, we have IH1(n)-IH2($n+1$). We must show that IH1($n+1$) holds, that is, step $n+1$ of the $(k-1)$ procedure path-refines step $n+1$ of the k-procedure. Fix $i \in I$ and $(\theta_i, s_i) \in \mathrm{X}_{k-1,i}^{n+1}$. Similarly to the proof of IH2($n+1$), we are going to show the existence of a CPS $\hat{\mu}^{(\theta_i,s_i)} = \hat{\mu}^i \in \cap_{m=0}^n \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k,-i}^m) \cap \Delta_{i,\theta_i}$ and of a strategy $\hat{s}_i^{(\theta_i,s_i)} = \hat{s}_i \in r_{i,\theta_i}(\hat{\mu}^i) \subseteq \mathrm{X}_{k,i}^{n+1}$ such that $\hat{s}_i(h) = s_i(h)$ for all $h \in H(\mathrm{X}_{k-1}^n)$.

By definition of $\mathrm{X}_{k-1,i}^{n+1}$ (see eq. (2)), there is some $\mu^i \in \cap_{m=0}^n \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k-1,-i}^m) \cap \Delta_{i,\theta_i}$ such that $s_i \in r_{i,\theta_i}(\mu^i)$.

Claim 1-bis. There exists $\hat{\mu}^i \in \cap_{m=0}^n \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k,-i}^m) \cap \Delta_{i,\theta_i}$ such that, for every $h \in H\left(\mathrm{X}_{k-1}^n\right) \cap H_i(s_i)$,

$$\forall(\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_{k-1}^n), \quad \hat{\mu}^i(\{\theta_{-i}\} \times S_{-i}(z)|h) = \mu^i(\{\theta_{-i}\} \times S_{-i}(z)|h). \tag{16}$$

The proof of Claim 1-bis is identical to the proof of Claim 1 in inductive step IH2, so we omit it.

Claim 2-bis: $H(\mathrm{X}_{k-1}^n) \subseteq H(\mathrm{X}_k^n)$.

Proof. Fix $\hat{s} \in \mathrm{Proj}_S \mathrm{X}_{k-1}^n$. By IH1($n$), there exists $\hat{s}' \in \mathrm{Proj}_S \mathrm{X}_k^n$ such that $\hat{s}'(\hat{h}) = \hat{s}(\hat{h})$ for every $\hat{h} \in H(\mathrm{X}_{k-1}^{n-1}) \supseteq H(\mathrm{X}_{k-1}^n)$. It follows that $\zeta(\hat{s}) = \zeta(\hat{s}') \in \bar{\zeta}(\mathrm{X}_k^n)$. □

Claim 3-bis: Fix a subset of histories $\hat{H}$ such that, for every $h \in \hat{H}$, s_i is a continuation best reply to $\hat{\mu}^i(\cdot|h)$ for θ_i. There exists $\hat{s}_i \in r_{i,\theta_i}(\hat{\mu}^i)$ such that $\hat{s}_i(h) = s_i(h)$ for every $h \in \hat{H}$.

Proof. Construct $\hat{s}_i$ as follows. For each $h \in \hat{H}$, let $\hat{s}_i(h) = s_i(h)$. For each $h \in H \backslash \hat{H}$, let $\hat{s}_i(h) = s'_i(h)$ for some continuation best reply s'_i to $\hat{\mu}^i(\cdot|h)$ for θ_i. It follows from Lemma 2 that $\hat{s}_i \in r_{i,\theta_i}(\hat{\mu}^i)$. □

Now fix $\hat{\mu}^i$ of Claim 1-bis. From equation (2), it follows that $\{\theta_i\} \times r_{i,\theta_i}(\hat{\mu}^i) \subseteq \mathrm{X}_{k,i}^{n+1}$. To conclude the proof, we show the existence of $\hat{s}_i \in r_{i,\theta_i}(\hat{\mu}^i)$ such that $\hat{s}_i(h) = s_i(h)$ for all $h \in H(\mathrm{X}_{k-1}^n)$. By Claim 3-bis with $\hat{H} = H(\mathrm{X}_{k-1}^n) \cap H_i(s_i)$, this is a consequence of the following result. (For each $h \in H(\mathrm{X}_{k-1}^n) \backslash H_i(s_i)$, since $h \notin H_i(\hat{s}_i)$, we can always set $\hat{s}_i(h) = s_i(h)$ because we use the notion of sequential best reply which only refers to the histories that are consistent with the strategy.)

Claim 4-bis: For each $h \in H\left(\mathrm{X}_{k-1}^n\right) \cap H_i(s_i)$, strategy s_i is a continuation best reply to $\hat{\mu}^i(\cdot|h)$ for θ_i.

Proof. First note that $H\left(\mathrm{X}_{k-1}^n\right) \cap H_i(s_i)$ is closed with respect to prefixes (predecessors): for each $h \in H\left(\mathrm{X}_{k-1}^n\right) \cap H_i(s_i)$ each prefix $h' \prec h$ belongs to $H\left(\mathrm{X}_{k-1}^n\right) \cap H_i(s_i)$. So, suppose by way of induction that Claim 4-bis holds for every $h' \prec h$ — this is vacuously true if $h = \varnothing$. Then, setting $\hat{H} = \{h' \in H : h' \prec h\}$, Claim 3-bis guarantees the existence of some $\hat{s}_i \in r_{i,\theta_i}(\hat{\mu}^i)$ such that $\hat{s}_i(h') = s_i(h')$ for every $h' \prec h$, thus $\hat{s}_i \in S_i(h)$. First, we need to show that $\zeta(\hat{s}_i, \hat{s}_{-i}) \in \bar{\zeta}(\mathrm{X}_{k-1}^n)$ for every $(\theta_{-i}, \hat{s}_{-i}) \in \mathrm{Supp}\hat{\mu}^i(\cdot|h)$. So, fix $(\theta_{-i}, \hat{s}_{-i}) \in \mathrm{Supp}\hat{\mu}^i(\cdot|h)$. By Claim 1-bis, $\hat{s}_i \in \mathrm{Proj}_{S_i}\mathrm{X}_{k,i}^{n+1}$, hence by IH2$(n+1)$, there exists $\hat{s}'_i \in \mathrm{Proj}_{S_i}\mathrm{X}_{k-1,i}^n$ such that $\hat{s}'_i(h) = \hat{s}_i(h)$ for every $h \in H(\mathrm{X}_k^n)$. Fix $\left(\theta_{-i}, \hat{s}'_{-i}\right) \in \hat{\eta}^{-1}((\theta_{-i}, \hat{s}_{-i})) \subseteq \mathrm{X}_{k-1,-i}^n$. Obviously, $\zeta(\hat{s}'_i, \hat{s}'_{-i}) \in \bar{\zeta}(\mathrm{X}_{k-1}^n)$. For every $\hat{h} \prec \zeta(\hat{s}'_i, \hat{s}'_{-i})$, we have $\hat{h} \in H(\mathrm{X}_{k-1}^n) \subseteq H(\mathrm{X}_{k-1}^{n-1})$, hence $\hat{s}_{-i}(\hat{h}) = \hat{s}'_{-i}(\hat{h})$ by construction of $\hat{\eta}$. Claim 2-bis gives $H(\mathrm{X}_{k-1}^n) \subseteq H(\mathrm{X}_k^n)$, therefore $\hat{s}_i(\hat{h}) = \hat{s}'_i(\hat{h})$ as well. It follows that $\zeta(\hat{s}_i, \hat{s}_{-i}) = \zeta(\hat{s}'_i, \hat{s}'_{-i}) \in \bar{\zeta}(\mathrm{X}_{k-1}^n)$.

For each $(\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_{k-1}^n)$, the probability of (θ_{-i}, z) induced by $\hat{s}_i$ and $\hat{\mu}^i(\cdot|h)$ (resp., $\mu^i(\cdot|h)$) is 0, if $\hat{s}_i \notin S_i(z)$, or $\hat{\mu}^i(\{\theta_{-i}\} \times S_{-i}(z)|h)$ (resp., $\mu^i(\{\theta_{-i}\} \times S_{-i}(z)|h)$) otherwise. Then, by equation (16), $\hat{s}_i$ induces the same probability over each $(\theta_{-i}, z) \in \Theta_{-i} \times \bar{\zeta}(\mathrm{X}_{k-1}^n)$ under $\hat{\mu}^i(\cdot|h)$ and under $\mu^i(\cdot|h)$, hence the same distribution over $\Theta_{-i} \times Z$, because the probability induced by $\hat{s}_i$ and $\hat{\mu}^i(\cdot|h)$ over $\Theta_{-i} \times \left(Z \backslash \bar{\zeta}(\mathrm{X}_{k-1}^n)\right)$ is zero: as we have previously shown, for each $(\theta_{-i}, \hat{s}_{-i}) \in \mathrm{Supp}\hat{\mu}^i(\cdot|h)$, $\zeta(\hat{s}_i, \hat{s}_{-i}) \in \bar{\zeta}(\mathrm{X}_{k-1}^n)$. The same conclusion can be reached for s_i in the same way, after observing that for each $(\theta_{-i}, s_{-i}) \in \mathrm{Supp}\mu^i(\cdot|h)$, since $(\theta_i, s_i, \theta_{-i}, s_{-i}) \in \mathrm{X}_{k-1}^n$, we have $\zeta(s_i, s_{-i}) \in \bar{\zeta}(\mathrm{X}_{k-1}^n)$. So,

call $\pi^{\hat{s}_i}$ and π^{s_i} the unique expected payoffs induced by, respectively, $(\theta_i, \hat{s}_i)$ and (θ_i, s_i) under both beliefs ($\mu^i(\cdot|h)$ and $\hat{\mu}^i(\cdot|h)$). Since $\hat{s}_i$ and s_i are continuation best replies for θ_i to, respectively, $\hat{\mu}^i(\cdot|h)$ and $\mu^i(\cdot|h)$, we have $\pi^{\hat{s}_i} \geq \pi^{s_i}$ and $\pi^{s_i} \geq \pi^{\hat{s}_i}$. Hence, $\pi^{s_i} = \pi^{\hat{s}_i}$. But then, also s_i is a continuation best reply for θ_i to $\hat{\mu}^i(\cdot|h)$. □

7.2 No path-monotonicity under restrictions to endogenous beliefs: an example

Consider the signalling game with $\Theta_1 = \{0, 1\}$, $A_1 = \{In, Out\}$, $A_2 = \{\ell, c, r\}$ and payoffs specified by the following table:

Payoffs of *1* and 2:

after In	ℓ	c	r
$\theta_1 = 0$	*1* 1	*-1* 0	*0* -1
$\theta_1 = 1$	*0* 0	*-1* 1	*1* -1

after Out	end
$\theta_1 = 0$	*0.5* *
$\theta_1 = 1$	*0.5* *

We first analyze the game with strong rationalizability (that is, without belief restrictions), which can be computed by iterated conditional dominance. Note that in this game there is a one-one correspondence between actions and strategies. For each step, only one action/strategy for (only one type of) only one player is eliminated:

1. r is the only conditionally dominated action and it is eliminated.
2. Given this, type $\theta_1 = 1$ expects to get at most 0 from In, which is eliminated for this type.
3. Player 2 rationalizes In assuming that it was chosen by type $\theta_1 = 0$ (forward induction), therefore c is eliminated.
4. Finally, type $\theta_1 = 0$ expects In to yield payoff 1; thus, Out is eliminated for this type.

To conclude, Out is the only strongly rationalizable action/strategy for type $\theta_1 = 1$, In is the only strongly rationalizable action/strategy for type $\theta_1 = 0$, and ℓ is the only strongly rationalizable action/strategy for player 2: $C_{\text{sb}}^{\infty} = \{(0, In), (1, Out)\} \times \{\ell\}$.

Thus, the type-dependent strongly rationalizable paths are

if $\theta_1 = 0$	$z = (In, \ell)$,
if $\theta_1 = 1$	$z = (Out)$.

Next we consider directed rationalizability assuming that (only) the following is transparent: player 2 becomes certain of type $\theta_1 = 1$ upon observing In, that is,

$$\Delta_2 = \left\{ \bar{\mu}^2 \in \Delta^H(\Theta_1 \times S_1) : \mu_2((1, In) \,|(In)) = 1 \right\}.$$

1. Δ. Both ℓ and r are eliminated in Step 1 of directed rationalizability because of the assumed belief-restriction.

2. Δ. Given this, In is eliminated for *both* types of player 1. This *makes it impossible to rationalize* In.

Hence, the only strongly Δ-rationalizable action/strategy of both types of player 1 is Out, and the only strongly Δ-rationalizable action/strategy of player 2 is c: $C_{\text{sb}}^{\Delta,\infty} = \{(0, Out), (1, Out)\} \times \{c\}$. It follows that the only strongly Δ-rationalizable path is (Out).