INNOCENZO GASPARINI INSTITUTE FOR ECONOMIC RESEARCH

Institutional Members: CEPR, NBER and Università Bocconi

WORKING PAPER SERIES

Monotonicity and Robust Implementation Under Forward-Induction Reasoning

Pierpaolo Battigalli, Emiliano Catonini

Working Paper n. 711

This Version: November, 2024

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

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

November 2024

Abstract

In sequential games, the set of paths consistent with rationality and forward-induction reasoning may change non-monotonically when adding transparent restrictions on players' beliefs (Battigalli & Friedenberg, *Theor. Econ.* 2012). Yet, we prove that—in an incomplete-information environment—predictions become sharper when the restrictions only concern initial beliefs about types. This implies that strong rationalizability for games with payoff uncertainty characterizes the path-predictions of forward-induction reasoning across all possible restrictions on players' initial hierarchies of exogenous beliefs. The result allows us to solve an open problem in implementation theory: the implementation of social choice functions through sequential mechanisms under forward-induction reasoning—which considerably expands the realm of implementable functions compared with simultaneous mechanisms (Mueller, *J. Econ. Theory* 2016)—is indeed robust in the sense of Bergemann and Morris (*Theor. Econ.* 2009).

Keywords: Incomplete information; Forward induction; Strong rationalizability; Path-monotonicity; Robust implementation; Sequential Mechanisms.

*We thank Carlo Andreatta, Alessandro Cherubin, Nicodemo De Vito, Drew Fudenberg, Shuige Liu, Viola Sigismondi, and Nicolas Sourisseau for useful comments. Pierpaolo Battigalli gratefully acknowledges financial support from ERC grant 101142844 TRAITS-GAMES.

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

Forward-induction reasoning disciplines players' belief revision in sequential games when they observe unexpected moves by the co-players. The basic ingredient is *strong belief in rationality*: when an unexpected move is consistent with the assumption that the co-player is rational (i.e., a subjective expected payoff maximizer) such assumption is maintained and shapes beliefs about the co-player's private information and/or future moves. *Common strong belief in rationality* adds on this by further assuming that, when an unexpected move is consistent with the co-player being rational and strongly believing in rationality, then also this higher-level assumption is maintained and shapes beliefs accordingly; yet higher-level assumptions are then considered (Battigalli & Siniscalchi 2002). It is well known that the strategy profiles and paths consistent with rationality and common strong belief in rationality may change non-monotonically with respect to transparent restrictions on players' beliefs (e.g., Battigalli & Friedenberg 2012). This is due to the so-called non-monotonicity of strong belief, which we will explain in detail below. Yet, we prove that, *in an incomplete-information environment, predictions become sharper when the restrictions only concern* ***exogenous*** *beliefs, i.e., initial beliefs about types.* Specifically, our main theorem states that the path-predictions of **strong directed rationalizability**—the solution concept characterizing the behavioral implications of rationality and common strong belief in rationality in sequential games with payoff uncertainty—are monotone with respect to restrictions on exogenous beliefs, i.e., more restrictive assumptions about such beliefs yield (weakly) more restrictive paths-of-play implications.[2] This monotonicity theorem implies that **strong rationalizability**, the "belief-free" version of directed strong rationalizability,[3] characterizes

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

[2] **"Directed rationalizability"** is the map from belief restrictions to the resulting rationalizable strategies (Battigalli & Friedenberg 2012; Battigalli, Catonini & De Vito 2024, Ch.s 8, 15). **"Strong"** refers to the reliance on the strong belief concept.

[3] In complete-information environments, this solution concept used to be called "extensive-form

the path-predictions of forward-induction reasoning across all possible restrictions on exogenous beliefs, a result that we apply to the theory of robust full implementation.

The power of sequential mechanisms to implement social choice functions (scf's) was first explored for *complete-information* environments, that is, under the assumption that agents' preferences, or payoff-types—even if they are unknown to the planner—are common knowledge among the agents.[4] In particular, drawing on work by Abreu & Matsushima (1992), Glazer & Perry (1996) proved that a very large class of social choice functions can be virtually implemented[5] by means of perfect-information sequential mechanisms, if players reason by backward induction and play the unique subgame perfect equilibrium. If the domain of preference profiles is finite, this is equivalent to strongly rationalizable virtual implementation, because backward and forward-induction reasoning yield the same path of play in generic finite games with complete and perfect information (Battigalli 1997, Battigalli & Siniscalchi 2002).

In environments with *incomplete information*, agents' behavior in a game form depends on their beliefs about each other's types. Just like in complete-information environments the planner is not assumed to know agents' commonly known payoff-types, in environments with incomplete information the planner may not know what hierarchies of exogenous beliefs the agents can hold and conceive. Formally, the planner may be uncertain about the relevant exogenous type structure, e.g., whether agents' beliefs are derived from a common prior on the domain of preference profiles and what it is (see Harsanyi 1967-68 and Mertens & Zamir 1985). In compliance with the Wilson's doctrine,[6] Bergemann & Morris (2009) analyze **robust implementation**, that is, the possibility to implement an scf independently of the exogenous type structure. They show that robust (virtual) full implementation of scf's by means of *static* (i.e., simultaneous-moves) 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 *sequen-*

rationalizability" (Pearce 1984, Battigalli, 1997). We avoid this terminology because there are different 'legitimate' formalizations of the rationalizability idea in the extensive-form analysis of sequential games. See, e.g., Battigalli & De Vito (2021) and the relevant references therein.

[4] See Moore & Repullo (1988), Ch. 10 of Osborne & Rubinstein (1994) and the references therein.

[5] Virtual implementation of an scf means that, for each type profile, the outcome predicted by the solution concept can be made arbitrarily close to the outcome prescribed by the scf. See the cited references.

[6] 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."

tial mechanisms and assuming that agents reason by *forward induction*—as captured by strong rationalizability—yields a very significant expansion of the implementable scf's. Yet, due to the aforementioned non-monotonicity of strong belief, it was not known whether implementation in strongly rationalizable strategies is robust to considering contextual restrictions on agents' exogenous interactive beliefs about each other's types.

Our game-theoretic result allows to solve this open problem. Although strongly rationalizable strategies may change non-monotonically when adding restrictions on exogenous beliefs, only the induced paths of play matter for the implementation of scf's. Since we prove that the set of possible paths under stronger restrictions on exogenous beliefs is weakly included in the one obtained with weaker or no restrictions, it follows that *strongly rationalizable implementation is robust* in the aforementioned sense.

The rest of paper is organized as follows. Section 2 provides a heuristic analysis and additional background. Section 3 contains the game-theoretic preliminaries. Section 4 states and explains the main theorem. Section 5 applies this result to the analysis of Bayesian games. Section 6 applies our game-theoretic results to the robust implementation problem. Section 7 discusses extensions. The Appendix collects proofs of key claims and lemmas that are omitted from the main body of the paper.

2 Heuristic analysis and detailed background

In this section, we first illustrate intuitions and difficulties behind our main monotonicity result by means of a heuristic analysis of an example (2.1). Next, we set the stage for the robust-implementation implications of our result, explaining its connection to rationalizability in static mechanisms (2.2) and sequential mechanisms (2.3).

2.1 Strong rationalizability, heuristic analysis of an example

Strong rationalizability is the iterated elimination, for each payoff-type of each player, of the strategies that are not sequential best replies to belief systems which assign probability 1, as long as possible, to the co-players' strategies that survive the previous elimination steps.[7] Strong directed rationalizability works in the same way, except

[7]There exist several versions of strong rationalizability, which differ by the adopted notion of sequential optimality (e.g., weak vs. strong sequential optimality), by the kind of belief systems (e.g.,

that the set of possible belief systems for each type is restricted exogenously and not just through iterated reasoning. The kind of belief restrictions we analyze in this paper only pertain to the initial beliefs about the payoff-types of the co-players. The elimination procedure is parameterized by the profile $\Delta = (\Delta_i)$ of restricted sets of beliefs. For any *fixed* Δ, we obtain a solution called "**strong Δ-rationalizability**." In the following example, we illustrate the two elimination procedures, the belief restrictions we use, and the main hurdle towards proving our general monotonicity result. Since we have not yet introduced all the required formal concepts, the analysis is necessarily heuristic and based on intuition.

Example 1 Consider a signaling game between players *1* (sender) and 2 (receiver) where the set of possible payoff-types θ_1 of the sender is $\Theta_1 = \{x, y, z\}$, the set of messages/signals is $M = S_1 = \{\ell, r\}$, and the sets of feasible reactions of the receiver are $\mathcal{A}_2(\ell) = \{a, b\}$ after message ℓ and $\mathcal{A}_2(r) = \{c, d, e\}$ after message r. Thus, the receiver's strategies are $S_2 = \mathcal{A}_2(\ell) \times \mathcal{A}_2(r)$, whereas the sender's strategies and signals coincide.[8] The payoffs are as follows:

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* (i.e., no restrictions on exogenous beliefs).
1. The first step of elimination follows from mere rationality. We can only eliminate message r for type x, as it is dominated by message ℓ. Thus, we write

$$S_1^1(x) = \{\ell\}$$

for the set of messages/signals consistent with rationality for type x. [Since no strategy of player 2 is eliminated in the first step, it follows that in even (odd) steps only

regular vs. complete conditional probability systems), and by how beliefs/behaviors are restricted at information sets that are inconsistent with the surviving strategies of some co-player. *None of these differences matters for our analysis.* The same applies to strong directed rationalizability, given the kind of belief restrictions we consider. See Battigalli, Catonini & Manili (2023) and references therein.

[8] We take the interim perspective: players are "born with their types" and do not make type-contingent plans.

eliminations for player 2 (player *1*) are possible.]

2. (*Player* 2) The optimal behavior of the receiver depends on his belief system $\mu_2 = (\mu_2(\cdot|\varnothing), \mu_2(\cdot|\ell), \mu_2(\cdot|r))$, where $\mu_2(\cdot|\varnothing)$ is the initial belief about the sender's type-message pair, and for each $m = \ell, r$, if $\mu_2(\Theta \times \{m\}\,|\varnothing) > 0$, then the belief $\mu_2(\cdot|m)$ after observing message m is derived from the initial belief by conditioning (in other words, μ_2 satisfies the chain rule).[9] At the second step of the elimination procedure, the initial belief is assumed to assign probability 1 to the type-message pairs that survived the first step:

$$\mu_2\left(\cup_{\theta_1 \in \Theta_1} \{\theta_1\} \times S_1^1(\theta_1)\,|\varnothing\right) = 1.$$

The same applies to the each belief $\mu_2(\cdot|m)$ provided that $m \in S_1^1(\theta_1)$ for some $\theta_1 \in \Theta_1$. Thus, here we have $\mu_2((x,r)\,|r) = 0$. In words, by strong belief in the sender's rationality, after observing message r the receiver concludes that the sender is not of type x—this is an instance of forward-induction reasoning. Given this, action e is never a best reply. Hence,

$$S_2^2 = \{a.c, b.c, a.d, b.d\},$$

where, for example, $a.c$ denotes the strategy choosing a after ℓ and c after r.

3. (*Player 1*) For type y, action r is not a best reply to any belief over S_2^2. Thus, $S_1^3(y) = \{\ell\}$.

4. (*Player* 2) Every belief system of the receiver must now assign probability 1 to type z after message r. Thus, $S_2^4 = \{a.c, b.c\}$.

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

No remaining strategy of the receiver can be eliminated. So, the result after arbitrarily many steps is:

$$\begin{aligned} \forall \theta_1 \;&\in\; \Theta_1,\, S_1^\infty(\theta_1) = \{\ell\}, \\ S_2^\infty \;&=\; \{a.c, b.c\}. \end{aligned}$$

It follows that the *strongly rationalizable paths* are (ℓ, a) and (ℓ, b) for every state (sender's type).

[9] We let $\varnothing$ denote the empty sequence of actions, i.e., the root of the leader-follower game tree.

Now consider the following restrictions on the exogenous beliefs of the receiver:[10] Let Δ_2 collect the belief systems 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, message r is eliminated for type x, so we write

$$S_1^{\Delta,1}(x) = S_1^1(x) = \{\ell\}.$$

But now, some strategies of the receiver are also eliminated. By the chain rule, every belief system $\mu_2 \in \Delta_2$ 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, the receiver best replies with a after ℓ and/or with c after r: $S_2^{\Delta,1} = \{a.c, b.c, a.d, a.e\}$.
Δ,***2.*** As in strong rationalizability, action e is never a best reply given r; hence, strategy $a.e$ of the receiver is eliminated:

$$S_2^{\Delta,2} = \{a.c, b.c, a.d\}.$$

Moreover, for type z, r is dominated by ℓ w.r.t. strategies in $S_2^{\Delta,1}$; so, $S_1^{\Delta,2}(z) = \{\ell\}$.
Δ,***3.*** For type y, r is dominated by ℓ over $S_2^{\Delta,2}$; thus,

$$S_1^{\Delta,3}(y) = \{\ell\}.$$

Moreover, every belief system of the receiver must now assign probability 1 to type y given message r. So,

$$S_2^{\Delta,3} = \{a.d\}.$$

We pinned down one strategy for each type of each player:

$$\forall \theta_1 \in \Theta_1,\ S_1^{\Delta,\infty}(\theta_1) = \{\ell\},$$
$$S_2^{\Delta,\infty} = \{a.d\}.$$

[10] Since there is only one type of receiver, we do not consider restrictions on the sender's exogenous beliefs. Formally, $\Theta_2 = \{\bar{\theta}_2\}$, a singleton, and $\Delta_1 = \{\mu^1 : \mu^1(\{\bar{\theta}_2\} \times S_2) = 1\}$. With this, the pair of restricted sets of beliefs $\Delta = (\Delta_1, \Delta_2)$ is determined by Δ_2.

The *strongly* Δ*-rationalizable path* is (ℓ, a) for every state. Consistently with our main result (Theorem 1), this is one of the two strongly rationalizable paths. Note, however, that the strongly Δ-rationalizable reaction of the receiver to r is d, whereas the strongly rationalizable one was c. ▲

In the example, as we will prove in full generality, strong directed rationalizability with restrictions on the initial beliefs about types refines strong rationalizability in terms of paths, for each possible state (profile of types). However, the strongly Δ-rationalizable *strategy* of the receiver is not strongly rationalizable. In particular, the implications about off-path behavior change non-monotonically after introducing the belief restrictions. Thus, strong directed rationalizability does not refine strong rationalizability in terms of strategies. For this reason, our path-monotonicity result cannot be proven with a straightforward induction argument.

2.2 Robust implementation, static mechanisms

To set the stage, we first explain the conceptual connection between robust implementation and rationalizability, 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, $\mathcal{T}$ 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 subclass 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 is 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).[11] 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.[12] However, the planner often ignores the contextual features represented 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 the same mechanism $\mathcal{M}$ fully implements 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). Since this paper is only concerned with different forms of full implementation, from now on 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 in Rationality* (RCBR).[13] 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.[14] In particular,

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

[12] 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 $\mathbf{Z}^{\Gamma^b}(\cdot)$.

[13] 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.

[14] See Battigalli (2003), 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.

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 by an induction argument that common belief in rationality and in contextual restrictions on exogenous interactive beliefs (which yields rationalizability in Bayesian games) 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 implementation. But Bergemann & Morris (2009) show that—within the domain of static mechanisms—even this form of implementation 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.

2.3 Robust implementation, 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), because there are different versions of rationalizability for sequential games characterizing the behavioral implications of different specifications of "common belief in rationality."[15] 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 equilibrium) 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.[16]

As mentioned in the Introduction and intuitively explained in 2.1, a stronger and more interesting notion 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 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" (Battigalli & Siniscalchi 2002). 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 significantly 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 $\gamma \geq 0$.

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

[16] See Mueller (2016) and (2020).

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 inductive argument based on the monotonicity of probability-1 (initial) belief, because *strong belief is not monotone:* As illustrated by Example 1, 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. When contextual considerations (e.g., social norms) also shape endogenous beliefs about behavior, unlike Example 1, it is easy to show that the set of induced *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, i.e., endogenous beliefs. Our main theorem shows that this is not by chance, or lack of trying to find counterexamples: *the set of* state-dependent *strongly rationalizable paths of play is* (always nonempty and) *monotone w.r.t. restrictions on exogenous beliefs.* 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. Now suppose that the relevant hierarchies of initial beliefs on the payoff-relevant uncertainty are adequately represented by a particular type structure $\mathcal{T}$ *à la* Harsanyi. It is without loss of generality to write the Harsanyi types as $t_i = (\theta_i, e_i)$ where coordinate e_i affects hierarchical exogenous beliefs, but does not affect payoffs. Appending $\mathcal{T}$ to $(\mathcal{M}, \mathcal{E})$

gives a sequential Bayesian game $\Gamma^{\mathrm{b}} = (\mathcal{M}, \mathcal{E}, \mathcal{T})$. In our analysis, we make the transition from game with payoff uncertainty Γ to Bayesian game Γ^{b} in two steps. First, we "duplicate" types by replacing each set Θ_i of payoff-types with a set $T_i = \Theta_i \times E_i$, and we note that the solution concept is invariant to such duplications: a pair (θ_i, s_i) is strongly rationalizable if and only if $((\theta_i, e_i), s_i)$ is strongly rationalizable in the "belief-free" game with duplicated types for each $e_i \in E_i$. Next we obtain a type structure $\mathcal{T}$ by adding belief maps $(\beta_i : T_i \to \Delta(T_{-i}))_{i \in I}$ to such game with duplicated types, which corresponds to specific restrictions on exogenous beliefs in this game: for each type $t_i = (\theta_i, e_i)$, the set of possible exogenous beliefs is the singleton $\{\beta_i(t_i)\}$. With this, our theorem implies that, for all Bayesian games Γ^{b} obtained by appending a Harsanyi type structure to Γ, $\emptyset \neq \mathbf{Z}^{\Gamma^{\mathrm{b}}}(\theta, e) \subseteq \mathbf{Z}^{\Gamma}(\theta)$, where $\mathbf{Z}^{\Gamma^{\mathrm{b}}}(\theta, e)$ is the set of strongly rationalizable paths at state (θ, e) in Γ^{b}. Therefore, $g\left(\mathbf{Z}^{\Gamma^{\mathrm{b}}}(\theta, e)\right) \approx \{f(\theta)\}$ for all such games Γ^{b} and states (θ, e) to an arbitrary degree of precision.

3 Preliminaries[17]

In this section we formally introduce the basic incomplete-information framework (3.1), systems of beliefs and sequential best replies (3.2), and the adopted solution concept, strong directed rationalizability (3.3).

3.1 Multistage games with payoff uncertainty

We consider the following *finite* multistage game with *observed actions* and *payoff uncertainty*.[18] 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) $\mathcal{A}(h) = \times_{i \in I} \mathcal{A}_i(h)$

[17] The formalism is based on the (still incomplete) draft of textbook *Game Theory: Analysis of Strategic Thinking* by Battigalli, Catonini, & De Vito. Chapter 15 of the book analyzes solution concepts for multistage games with incomplete information.

[18] See the discussion in Section 7 for extensions to imperfectly observed actions and infinite games.

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.*[19]

Each player i *knows* 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, or states of nature. 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 state θ.[20] Thus, a multistage game with payoff uncertainty and observed 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. If $|\mathcal{A}_i(h)| > 1$, then player i is active at nonterminal history h. If $|\mathcal{A}_i(h)| = 1$, player i is inactive and the unique element of $\mathcal{A}_i(h)$ can be thought of as a waiting action. If there is only one active player for each $h \in H$, then Γ features perfect (albeit incomplete) information, i.e., there are no simultaneous moves and (by the observed actions assumption) past moves are perfectly observed. In the analysis of examples, we omit to mention the waiting actions of inactive players.

We interpret each function u_i as the composition of a parameterized **utility function** $v_i : \Theta \times Y \to \mathbb{R}$ (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 $z = \zeta(s)$.[21] To ease notation, it is convenient to extend the path function to domain

[19] See the discussion of this assumption in Section 7.

[20] Since the profile of payoff-types θ determines each payoff function $u_{i,\theta} : Z \to \mathbb{R}$, we are implicitly assuming that there is distributed knowledge of the payoff-relevant state of nature. This assumption is made only to simplify the notation.

[21] Define recursively whether a history h is induced by a given strategy profile s: the empty

$\Theta \times S$ and codomain $\Theta \times Z$ 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.[22]

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
$\bar{\zeta} : \Theta \times S \to \Theta \times Z$	extended path function
$U_i = u_i \circ \bar{\zeta} : \Theta \times S \to \mathbb{R}$	(param.) strategic-form payoff function of i

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.

[22] Note that $S(h) = \times_{j \in I} S_j(h)$ for every history $h \in \bar{H}$.

3.2 Beliefs and best replies

We model the beliefs of each player i as the play unfolds by means of **conditional probability systems** (CPSs, Renyi, 1955): observed history h reveals that the set of possible type-strategy profiles of the co-players is $\Theta_{-i} \times S_{-i}(h)$; thus, we consider arrays of conditional beliefs $\mu_i = (\mu_i(\cdot|\Theta_{-i} \times S_{-i}(h)))_{h \in H}$ over such profiles, abbreviated in $\mu_i = (\mu_i(\cdot|h))_{h \in H}$. The set of CPSs of player i, denoted $\Delta^H(\Theta_{-i} \times S_{-i})$, is the subset of arrays of beliefs $\mu_i \in (\Delta(\Theta_{-i} \times S_{-i}))^H$ such that, for every $h \in H$, $\mu_i(\Theta_{-i} \times S_{-i}(h)|h) = 1$ and the *chain rule* holds, that is, for all $h, h' \in H$ and $E \subseteq \Theta_{-i} \times S_{-i}(h')$,

$$S_{-i}(h') \subseteq S_{-i}(h) \Longrightarrow \mu_i(E|h) = \mu_i(E|h')\,\mu_i(\Theta_{-i} \times S_{-i}(h')|h).$$

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.

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 $\theta_i \in \Theta_i$,

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

With this, we introduce profiles $\Delta = (\Delta_{i,\theta_i})_{i \in I, \theta_i \in \Theta_i}$ of type-dependent subsets of CPSs such that, for all $i \in I$ and $\theta_i \in \Theta_i$,[23]

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

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 follows. Let $\mathcal{H}_i(s_i) = \{h \in H : s_i \in S_i(h)\}$ denote the set of non-terminal histories that can occur if s_i is played. With this,

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

By standard dynamic programming arguments, $r_{i,\theta_i}(\mu_i) \neq \emptyset$ for all payoff-types θ_i

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

and CPSs μ_i.[24]

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 $h \in \mathcal{H}_i(\bar{s}_i)$, 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 $s_i \in S_i(h)$,

$$\mathbb{E}_{\mu_i(\cdot|h)}(U_i(\theta_i, \bar{s}_i, \cdot)) \geq \mathbb{E}_{\mu_i(\cdot|h)}(U_i(\theta_i, s_i, \cdot)).$$

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 $h \in \mathcal{H}_i(\bar{s}_i)$.

3.3 Strong (directed) rationalizability

As informally explained in the Introduction, our forward-induction analysis hinges on the notion of "strong belief." For each event $E_{-i} \subseteq \Theta_{-i} \times S_{-i}$ (e.g., that co-players' behavior is consistent with rationality), we say that a CPS 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:

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

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.*

[24] See Battigalli, Catonini & Manili (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.

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).[25]

Fix a profile $\Delta = (\Delta_{i,\theta_i})_{i\in I,\theta_i\in\Theta_i}$ of subsets of CPSs (see 3.2). Also, for each player $i \in I$ and event $E_{-i} \subseteq \Theta_{-i} \times S_{-i}$, let $\Delta^H_{\text{sb}}(E_{-i})$ denote the **set of CPSs μ_i that strongly believe** E_{-i}, and let $\Sigma_i^{\Delta,0} = \Theta_i \times S_i$. Then, for each $n > 0$, define the set of **strongly Δ-n-rationalizable** type-strategy pairs of i as

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

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

$$S_i^{\Delta,n}(\theta_i) = \left(\Sigma_i^{\Delta,n}\right)_{\theta_i} = \left\{s_i : (\theta_i, s_i) \in \Sigma_i^{\Delta,n}\right\},$$

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

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

Finally, let

$$\begin{aligned} \Sigma_i^{\Delta,\infty} &= \cap_{n>0} \Sigma_i^{\Delta,n}, \\ \Sigma^{\Delta,\infty} &= \times_{i\in I} \Sigma_i^{\Delta,\infty} \end{aligned}$$

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

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

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 by strategic reasoning, one can

[25] These articles use the term "(strong) Δ-rationalizability." Recall that 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.

prove by induction the following result:

Lemma 1 *(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 Δ-Rationalizability boils down to **Strong Rationalizability** (Pearce 1984, Battigalli 1997), which characterizes the behavioral implications of *Rationality and Common Strong Belief in Rationality* (Battigalli & Siniscalchi, 2002). We omit the superscript Δ to denote Strong Rationalizability: Σ_i^∞ (Σ_i^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 (n-)rationalizable strategies of θ_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 $\mathcal{Z}\left(\Sigma^{\Delta,\infty}\right) = \mathrm{proj}_Z \bar{\zeta}\left(\Sigma^{\Delta,\infty}\right)$ and the set of strongly Δ-rationalizable paths **at state of nature** θ is the section $\bar{\zeta}\left(\Sigma^{\Delta,\infty}\right)_\theta = \zeta\left(S^{\Delta,\infty}(\theta)\right)$.

The signaling game informally analyzed in Section 2.1 illustrates the formalism and concepts introduced in this section.

Example 2 Consider again the signaling game of Example 1. Game Γ is a two-stage game with perfect information, with

$$\begin{aligned}
\Theta_1 &= \{x, y, z\},\ \Theta_2 = \left\{\bar{\theta}_2\right\},\ \textit{(a singleton)},\\
H &= \{\varnothing, (\ell), (r)\},\ Z = \{(\ell, a), (\ell, b), (r, c), (r, d), (r, e)\},\ \bar{H} = H \cup Z,\\
\mathcal{A}_1(\varnothing) &= \{\ell, r\},\ \mathcal{A}_2(\ell) = \{a, b\},\ \mathcal{A}_2(r) = \{c, d, e\},
\end{aligned}$$

and the type-dependent payoff functions $u_i : \Theta \times Z \to \mathbb{R}$ are described by the following tables:

$u_1(\cdot, \ell, \cdot), u_2(\cdot, \ell, \cdot)$	a	b
$\theta_1 = x$	*3* 1	*1* 0
$\theta_1 = y$	*1* 0	*1* 1
$\theta_1 = z$	*3* 1	*1* 0

$u_1(\cdot, r, \cdot), u_2(\cdot, r, \cdot)$	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

Since player 2 is uninformed and inactive in the first stage, Σ_2 is isomorphic to S_2 and $\Delta^H(\Theta_2 \times S_2)$ is isomorphic to $\Delta(S_2)$ (by the chain rule and $S_2(\ell) = S_2(r) = S_2$). We intuitively explained in Example 1 how strong directed rationalizability works in this game. Thus, we only list below the formal result for each step using the notation introduced above. Without belief restrictions, we have:

$$\begin{aligned}
\Sigma_1^1 &= \{(x,\ell),(y,\ell),(y,r),(z,\ell),(z,r)\} \text{ (thus, } S_1^1(x) = \{\ell\}\text{), } \Sigma_2^1 = S_2; \\
\Sigma_1^2 &= \Sigma_1^1,\ \Sigma_2^2 = \{a,b\} \times \{c,d\}; \\
\Sigma_1^3 &= \{(x,\ell),(y,\ell),(z,\ell),(z,r)\}, \text{ (thus, } S_1^3(y) = \{\ell\}\text{), } \Sigma_2^3 = \Sigma_2^2; \\
\Sigma_1^4 &= \Sigma_1^3,\ \Sigma_2^4 = \{a.c, b.c\}; \\
\Sigma_1^5 &= \Theta_1 \times \{\ell\} \text{ (thus, } S_1^5(\theta_1) = \{\ell\} \text{ for all } \theta_1\text{), } \Sigma_2^5 = \Sigma_2^4; \\
\Sigma_1^\infty &= \Theta_1 \times \{\ell\},\ \Sigma_2^\infty = \{a.c, b.c\}; \\
\bar{\zeta}(\Sigma^\infty) &= \Theta \times (\{\ell\} \times \{a,b\}), \text{ thus } \zeta\left(S^{\Delta,\infty}(\theta)\right) = \{\ell\} \times \{a,b\} \text{ for all } \theta \in \Theta.
\end{aligned}$$

Since Θ_2 is a singleton, we can only have restrictions on the exogenous beliefs of player 2. Formalizing Example 1, let

$$\Delta_2 = \left\{\mu_2 \in \Delta^H(\Theta_1 \times S_1) : \left(marg_{\Theta_1}\mu_2(\cdot|\varnothing)\right)(z) = 1\right\}$$

denote the set of CPSs that intially assign probability 1 to type $\theta_1 = z$. With this, strong Δ-rationalizability yields:

$$\begin{aligned}
\Sigma_1^{\Delta,1} &= \{(x,\ell),(y,\ell),(y,r),(z,\ell),(z,r)\},\ \Sigma_2^{\Delta,1} = \{a.c, b.c, a.d, a.e\}; \\
\Sigma_1^{\Delta,2} &= \{(x,\ell),(y,\ell),(y,r),(z,\ell)\},\ \Sigma_2^{\Delta,2} = \{a.c, b.c, a.d\}; \\
\Sigma_1^{\Delta,3} &= \Theta_1 \times \{\ell\},\ \Sigma_2^{\Delta,3} = \{a.d\}; \\
\Sigma_1^{\Delta,\infty} &= \Theta_1 \times \{\ell\},\ \Sigma_2^{\Delta,\infty} = \{a.d\}; \\
\bar{\zeta}\left(\Sigma^{\Delta,\infty}\right) &= \Theta \times \{(\ell,a)\}, \text{ thus } \zeta\left(S^{\Delta,\infty}(\theta)\right) = \{(\ell,a)\} \text{ for all } \theta \in \Theta.
\end{aligned}$$

Thus $\mathcal{Z}\left(\Sigma^{\Delta,\infty}\right) \subset \mathcal{Z}(\Sigma^\infty)$; but $\Sigma_2^{\Delta,\infty} \not\subseteq \Sigma_2^\infty$, actually $\Sigma_2^{\Delta,\infty} \cap \Sigma_2^\infty = \emptyset$. ▲

4 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, 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* $(\theta, s) \in \Sigma^{\Delta,\infty} \neq \emptyset$, *there exists* $s' \in S$ *such that* $(\theta, s') \in \Sigma^{\infty}$ *and* $\zeta(s) = \zeta(s')$.

The assumption that the belief restrictions 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 co-player'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* conditional beliefs about the co-player's type.

Given that, as we saw in Examples 1 and 2, the two elimination procedures may induce completely disjoint off-path behaviors, proving path-monotonicity is hard. Our proof is based on a kind of double-induction argument.[26]

4.1 Proof of theorem 1

Non-emptiness follows from Lemma 1. Here we only focus on path-inclusion. Since comparing directly strong rationalizability and strong Δ-rationalizability is difficult,

[26] The techniques we use have common elements with the techniques used by Perea (2018) and Catonini (2020) in complete-information games to prove, respectively, an *order-independence* result for strong rationalizability and an *outcome-monotoncity* result for directed rationalizability with respect to initial belief restrictions about the path of play. In particular, like Perea (2018), we decompose the problem of comparing two very different elimination procedures into a chain of pairwise comparisons between more similar procedures, and the proof of a key claim (Claim 4 in the proof of Theorem 1) draws on Catonini (2020).

we construct a sequence of elimination procedures that gradually transform strong Δ-rationalizability into strong rationalizability, 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: $\Sigma^{K-1} \subset \Sigma^K = \Sigma^\infty$ ($\subset$ denotes *strict* inclusion). Note that K is well defined because the game is finite. 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 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; thus, suppose that strong rationalizability deletes some pair (θ_i, s_i) for at least one player i, so that $K > 0$.

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

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

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 $n \in \{1, ..., k\}$ 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 Procedure k coincide with strong rationalizability: $\mathrm{X}_k^n = \Sigma^n$ 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, Procedure k 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 $(\Sigma^n)_{n=0}^\infty$ 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 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 consecutive 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 "$\cdot|_{\widehat{H}}$" 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

$$\mathcal{H}(\mathrm{X}) = \{h \in H : \exists\, (\theta, s) \in \mathrm{X}, h \prec \zeta(s)\}$$

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

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

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 $n = k$. From now on, let $n = k+1$; this will bring us to formulate the induction hypothesis of the formal proof with the appropriate indexes.

At step $n - 1 = k$, P adopts the belief restrictions and Q does not, so:

$$\mathrm{P}^{n-1} \subseteq \mathrm{Q}^{n-1}. \tag{3}$$

At step $n = k+1$ 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:

$$\mathrm{P}^n|_{\mathcal{H}(\mathrm{P}^{n-1})} \subseteq \mathrm{Q}^n|_{\mathcal{H}(\mathrm{P}^{n-1})}. \tag{4}$$

At step $n+1 = k+2$, the comparison becomes more complex.

First: Is this step of procedure P still more restrictive than Q regarding beliefs at histories in $\mathcal{H}(\mathrm{P}^{n-1})$ about the co-players' types and moves at those histories, as expression (4) seems to suggest?

The answer is yes, but only thanks to the assumption that restrictions only concern exogenous beliefs. Restrictions on the beliefs about the endogenous/strategic uncertainty could allow player i to believe in some $(\theta_{-i}, s_{-i}) \in \mathrm{P}^n_{-i}$, but not in any counterpart $\left(\theta_{-i}, s'_{-i}\right) \in \mathrm{Q}^n_{-i}$ with $s_{-i}|_{\mathcal{H}(\mathrm{P}^{n-1})} = s'_{-i}|_{\mathcal{H}(\mathrm{P}^{n-1})}$. The role of restricting only the *initial* beliefs is more subtle. Strong belief in P^n_{-i} and in Q^n_{-i} may induce, by forward-induction reasoning, different beliefs about θ_{-i} at some history $h' \in \left(\mathcal{H}(\mathrm{P}^n_{-i}) \cap \mathcal{H}(\mathrm{Q}^n_{-i})\right) \setminus \mathcal{H}(\mathrm{P}^{n-1})$. If there were restrictions on such beliefs at h', it could well be that some of the beliefs derived from Q^n_{-i} would be incompatible with the restrictions. Via the chain rule, this could also rule out some beliefs at some $h \in \mathcal{H}(\mathrm{P}^{n-1})$ such that $h \prec h'$.

But this is not the end of the story. Strong belief in Q^n_{-i} may be more restrictive, or "differently restrictive," compared with strong belief in P^n_{-i} regarding behavior outside of $\mathcal{H}(\mathrm{P}^{n-1})$. This is because the inclusion of equation (4) is restricted to $\mathcal{H}(\mathrm{P}^{n-1})$. Thus, strong belief in Q^n_{-i} may rule out some belief about the reactions of the co-players to a deviation of i from $\mathcal{H}(\mathrm{P}^{n-1})$ which is instead allowed by strong belief in P^n_{-i}. Example 2 may help to understand this point.[27] Belief in $\Sigma_2^4 = \{a.c, b.c\}$ imposes belief in reaction c after a deviation from the unique on-path signal induced by $\Sigma_1^{\Delta,3} = \Theta_1 \times \{\ell\}$. By contrast, belief in $\Sigma_2^{\Delta,4} = \{a.d\}$ imposes belief in reaction d. With this, it is conceivable that there might be a deviation from one of the paths consistent with P^{n+1} that player i expects to lead out of $\mathcal{H}(\mathrm{P}^{n-1})$ and always be strictly profitable under strong belief in Q^n_{-i}. This is what makes it hard to prove that

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

[27] The example compares directly strong Δ-rationalizability and strong rationalizability, which cannot formally take the role of procedures P and Q, but it still displays the possible relationship between P and Q that we are illustrating.

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

Here is where the similarity between the two procedures comes into play: $\mathcal{H}(\mathrm{P}^{n-1}) \supseteq \mathcal{H}(\mathrm{Q}^n)$ 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 $\mathrm{Q}^{n-1} \subseteq \mathrm{Q}^{n-2}$ and Q and P coincide up to step $n-2$,

$$\mathrm{Q}^{n-1} \subseteq \mathrm{P}^{n-2}.$$

Next, we consider step n of Q and step $n-1$ 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 (4)) to obtain

$$\mathrm{Q}^n|_{\mathcal{H}(\mathrm{Q}^{n-1})} \subseteq \mathrm{P}^{n-1}|_{\mathcal{H}(\mathrm{Q}^{n-1})}. \tag{6}$$

Thus, since $\mathcal{H}(\mathrm{Q}^{n-1}) \supseteq \mathcal{H}(\mathrm{Q}^n)$, we have $\mathcal{H}(\mathrm{P}^{n-1}) \supseteq \mathcal{H}(\mathrm{Q}^n)$, as we wanted to show.

Proving (6) was easy because we could rely on the inclusion $\mathrm{Q}^{n-1} \subseteq \mathrm{P}^{n-2}$, which is stated for complete strategies. But to continue and prove

$$\mathrm{P}^{n+2}|_{\mathcal{H}(\mathrm{P}^{n+1})} \subseteq \mathrm{Q}^{n+2}|_{\mathcal{H}(\mathrm{P}^{n+1})},$$

we need $\mathcal{H}(\mathrm{P}^n) \supseteq \mathcal{H}(\mathrm{Q}^{n+1})$, that is, we need

$$\mathrm{Q}^{n+1}|_{\mathcal{H}(\mathrm{Q}^n)} \subseteq \mathrm{P}^n|_{\mathcal{H}(\mathrm{Q}^n)}, \tag{7}$$

and to prove this we run into the same complications we had for (5). However, recall that we were able to prove (5) after showing that $\mathcal{H}(\mathrm{P}^{n-1}) \supseteq \mathcal{H}(\mathrm{Q}^n)$; we can prove (7) in the same way, with the roles of the two procedures flipped, because $\mathcal{H}(\mathrm{Q}^{n-1}) \supseteq \mathcal{H}(\mathrm{P}^{n-1})$ by (3).

At this point, considering any $n \geq k$, it should be clear that if take induction

hypotheses of the kind

$$\mathrm{P}^n|_{\mathcal{H}(\mathrm{P}^{n-1})} \subseteq \mathrm{Q}^n|_{\mathcal{H}(\mathrm{P}^{n-1})}, \tag{8}$$

$$\mathrm{Q}^n|_{\mathcal{H}(\mathrm{Q}^{n-1})} \subseteq \mathrm{P}^{n-1}|_{\mathcal{H}(\mathrm{Q}^{n-1})}, \tag{9}$$

then we can use them to prove the next iteration of (9), namely

$$\mathrm{Q}^{n+1}|_{\mathcal{H}(\mathrm{Q}^n)} \subseteq \mathrm{P}^n|_{\mathcal{H}(\mathrm{Q}^n)}, \tag{10}$$

and we can use (8) and (10) to prove the next iteration of (8).

Now we formulate this two-fold induction hypothesis for the formal proof. For every $n \geq k$,

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 $h \in \mathcal{H}(\mathrm{X}^{n-1}_{k-1})$ (thus, step n of Procedure $k-1$ path-refines step n of Procedure k);

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 $h \in \mathcal{H}(\mathrm{X}^{n-1}_k)$ (thus, step n of Procedure k path-refines step $n-1$ of Procedure $k-1$);

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 $(\theta, s) \in \mathrm{X}^K_0 \supseteq \Sigma^{\Delta,\infty}$, there exists $s' \in S$ such that $(\theta, s') \in \mathrm{X}^K_K = \Sigma^\infty$ 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 by way of induction, following the strategy we outlined above. The formal proofs of Claims 1-4 stated below are deferred to the Appendix.

Basis steps

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

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

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

Inductive steps

The proofs of the two inductive steps, IH1(n)-IH2(n)⇒IH1($n+1$) and IH1(n)-IH2(n)⇒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 of IH1(n)-IH2(n)⇒IH2($n+1$). We relegate the proof of IH1(n)-IH2(n)⇒IH1($n+1$), which uses the previously obtained IH2($n+1$), to the supplemental appendix.

Inductive step, part 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 $\widetilde{\mu}_i^{(\theta_i,s_i)} \in \cap_{m=0}^{n-1}\Delta_{\mathrm{sb}}^{H}(\mathrm{X}_{k-1,-i}^{m}) \cap \Delta_{i,\theta_i}$ and of a strategy $\widetilde{s}_i^{(\theta_i,s_i)} \in r_{i,\theta_i}(\widetilde{\mu}_i^{(\theta_i,s_i)}) \subseteq \mathrm{X}_{k-1,i}^{n}$ such that $\widetilde{s}_i^{(\theta_i,s_i)}(h) = s_i(h)$ for all $h \in \mathcal{H}(\mathrm{X}_{k}^{n})$. To ease notation, in what follows we do not make the dependence of both the CPS and the strategy on the fixed pair (θ_i, s_i), so we write $\widetilde{s}_i = \widetilde{s}_i^{(\theta_i,s_i)}$ and $\widetilde{\mu}_i = \widetilde{\mu}_i^{(\theta_i,s_i)}$. 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($n+1$).

The construction of $\widetilde{\mu}_i$ and $\widetilde{s}_i$ will be based on four claims for which we provide formal proofs in the Appendix; here, before each claim, we only provide the main ingredients of its proof.

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 that are consistent with step n of Procedure k. Consistently with notation used for sets of nonterminal histories and in Example 2, for any $\mathrm{X} \subseteq \Theta \times S$, we let

$$\mathcal{Z}(\mathrm{X}) = \mathrm{proj}_{Z}\bar{\zeta}(\mathrm{X}) = \{z \in Z : \exists (\theta, s) \in \mathrm{X}, \zeta(s) = z\}$$

denote the set of possible paths given X.

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* $h \in \mathcal{H}(\mathrm{X}_{k}^{n}) \cap \mathcal{H}_i(s_i)$,

$$\forall (\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}_{k}^{n}), \quad \widetilde{\mu}_i(\{\theta_{-i}\} \times S_{-i}(z)|h) = \mu_i(\{\theta_{-i}\} \times S_{-i}(z)|h). \tag{11}$$

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

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

In what follows, we will also use the following implication of standard dynamic programming arguments.[28]

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 of procedure k and mimics $\widetilde{s}_i$ at each $h \in \mathcal{H}\left(\mathrm{X}_k^n\right) \cap \mathcal{H}_i(\widetilde{s}_i') = \mathcal{H}\left(\mathrm{X}_k^n\right) \cap \mathcal{H}_i(\widetilde{s}_i)$. Now note that, by equation (11) and the fact that μ_i strongly believes $\mathrm{X}_{k,-i}^n$, every strategy s_i' that does not leave the paths induced by profiles in X_k^n yields the same expected payoff under $\widetilde{\mu}_i(\cdot|h)$ and $\mu_i(\cdot|h)$ for every $h \in \mathcal{H}\left(\mathrm{X}_k^n\right) \cap \mathcal{H}_i(s_i')$. Obviously, $s_i, \widetilde{s}_i' \in \mathrm{proj}_{S_i}\mathrm{X}_{k,i}^n$ do not leave those paths, and since $\widetilde{s}_i$ mimics $\widetilde{s}_i'$ as described above, $\widetilde{s}_i$ does not leave those paths either. But then, for each $h \in \mathcal{H}\left(\mathrm{X}_k^n\right) \cap \mathcal{H}_i(s_i) \cap \mathcal{H}_i(\widetilde{s}_i)$, the fact that s_i and $\widetilde{s}_i$ are continuation best replies to (respectively) μ_i and $\widetilde{\mu}_i$ at h implies that they are also continuation best replies (respectively) to $\widetilde{\mu}_i$ and μ_i at h. To extend this claim to every $h \in \mathcal{H}\left(\mathrm{X}_k^n\right) \cap \mathcal{H}_i(s_i)$, we need to make sure that h is also reached by some sequential best reply $\widetilde{s}_i$ to $\widetilde{\mu}_i$; for this, we just need an inductive application of Claim 3, from the initial history and moving downwards.

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

By Claim 3 with $\widetilde{H} = \mathcal{H}(\mathrm{X}_k^n) \cap \mathcal{H}_i(s_i)$ and Claim 4, there exists $\widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}_i)$ such that $\widetilde{s}_i(h) = s_i(h)$ for all $h \in \mathcal{H}(\mathrm{X}_k^n)$. (For each $h \in \mathcal{H}(\mathrm{X}_k^n) \backslash \mathcal{H}_i(s_i)$, since $h \notin \mathcal{H}_i(\widetilde{s}_i)$, we can always set $\widetilde{s}_i(h) = s_i(h)$ because we use the weak notion of sequential best reply which only refers to histories consistent with the candidate strategy.) From equation (2) it follows that $\{\theta_i\} \times r_{i,\theta_i}(\widetilde{\mu}_i) \subseteq \mathrm{X}_{k-1,i}^n$. Thus, $\widetilde{s}_i \in \mathrm{X}_{k-1,i}^n$.

[28] We provide such arguments in the Appendix: see Lemma 3.

5 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**[29] 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

$$u_i^{\mathrm{e}}\left((\theta_j, e_j)_{j \in I}, z\right) = u_i\left((\theta_j)_{j \in I}, z\right)$$

for all $(\theta_j, e_j)_{j \in I} \in \times_{j \in I} T_j$ 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 co-players' 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

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

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

[29] The term "elaboration" was introduced by Fudenberg *et al.* (1988) with a related, but different meaning: They added payoff types to define incomplete-information perturbations, whereas we add a payoff-irrelevant component to existing payoff types.

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.

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

$$\Sigma_i^{\mathrm{e},n} = \left\{(t_i, s_i) : \exists \mu_i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\Sigma_{-i}^{\mathrm{e},m}), s_i \in r_{i,t_i}^{\mathrm{e}}(\mu_i)\right\},$$

where,

$$r_{i,t_i}^{\mathrm{e}}(\mu_i) = \left\{\bar{s}_i : \forall h \in \mathcal{H}_i(\bar{s}_i), \bar{s}_i \in \arg\max_{s_i \in S_i(h)} \mathbb{E}_{\mu_i(\cdot|h)}\left(u_i^{\mathrm{e}}(t_i, \cdot, \zeta(s_i, \cdot))\right)\right\}$$

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:

$$S_i^{\mathrm{e},n}(t_i) = \left(\Sigma_i^{\mathrm{e},n}\right)_{t_i} = \left\{s_i : (t_i, s_i) \in \Sigma_i^{\mathrm{e},n}\right\}.$$

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

Lemma 2 *Fix any elaboration* Γ^{e} *of* Γ. *For all* $i \in I$, $n \in \mathbb{N}_0$, *and* $(\theta_i, e_i) \in T_i$, $S_i^{\mathrm{e},n}(\theta_i, e_i) = S_i^n(\theta_i)$.

Now require that the belief system (CPS) μ_i that justifies a pair (t_i, s_i) 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 $i \in I$, $\Sigma_i^{\mathrm{b},0} = T_i \times S_i$, and for each $n \in \mathbb{N}$

$$\Sigma_i^{\mathrm{b},n} = \left\{(t_i, s_i) : \exists \mu_i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\Sigma_{-i}^{\mathrm{b},m}), \mathrm{marg}_{T_{-i}} \mu_i(\cdot|\varnothing) = \beta_i(t_i), s_i \in r_{i,t_i}^{\mathrm{b}}(\mu_i)\right\},$$

where $r_{i,t_i}^{\mathrm{b}}(\mu_i) = r_{i,t_i}^{\mathrm{e}}(\mu_i)$ (defined above) for each $\mu_i \in \Delta^H(T_{-i} \times S_{-i})$, because $u_i^{\mathrm{b}} =$

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

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

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 for all $i \in I$. With this, 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 2 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 $(\theta, e, s) \in \Sigma^{\mathrm{b},\infty} \neq \emptyset$, there exists $s' \in S$ such that $(\theta, s') \in \Sigma^\infty$ and $\zeta(s) = \zeta(s')$.*

6 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 observed 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

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

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 $s \in S^{\infty}(\theta) \neq \emptyset$, $\|g(\zeta(s)) - f(\theta)\| < \varepsilon$.*[30]

Bergemann & Morris (2009) introduce the notion of *robust implementation*, which requires the mechanism to implement the social choice function for any exogenous restrictions on players' collectively coherent hierarchies of beliefs about types, such

[30] 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 1 that $S^{\infty}(\theta) \neq \emptyset$ for all $\theta \in \Theta$.

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.[31] 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 $s \in S^{\mathrm{b},\infty}(t) \neq \emptyset$, $\|g(\zeta(s)) - f(\theta)\| < \varepsilon$. ■

7 Discussion and extensions

In this section we consider some limitations of our analysis and we discuss possible extensions and related conceptual issues.

[31] Interim correlated rationalizbaility is the appropriate notion of rationalizability for Bayesian games. See Bergemann & Morris (2012) and the relevant references therein.

7.1 Imperfectly observed actions

Our results extend to finite sequential games with imperfectly observed actions, as long as perfect recall holds. Our arguments go through by replacing nonterminal histories $h \in H$ with information sets $h_i \in H_i$ for each player i. In particular, perfect recall allows to preserve the dynamic consistency of subjective expected utility maximization and the factorization of the sets of strategy profiles consistent with any given information set h_i as $S(h_i) = S_i(h_i) \times S_{-i}(h_i)$, which are key elements of our analysis. However, from the perspective of mechanism design, perfect recall as defined in traditional game theory[32] is a *hybrid* property of information partitions that should be "unpacked," separating the information reaching players as per the rules specified by the mechanism from the mnemonic abilities of the agents playing the game, which are personal traits just like their preferences. As shown in Battigalli & Generoso (2024), such separation is both possible and conceptually useful: information partitions can be *derived* from primitive elements describing the rules of the game on the one hand, and mnemonic abilities on the other hand. Perfect recall of information partitions (with the implicit assumption that it is commonly known that it holds) obtains if either (1) (it is commonly known that) the relevant agents have perfect memory, a personal feature, or (2) the commonly known game rules are such that moving players are always reminded of the pieces of information that previously reached them and the actions they took. In both cases, information sets h_i correspond to *personal histories* of signals/messages received and actions taken by i. Even assuming that other forms of cognitive rationality hold (e.g., logical and introspective abilities), it is conceivable that agents may have memory lapses that interfere with implementation goals. In this case, the designer can "enforce" perfect recall by (2).

7.2 Infinite type sets and infinite horizon

We consider finite multistage games with payoff uncertainty and—in our analysis of implementation—finite Bayesian elaborations of such games. In particular, we assume that the horizon and the sets of payoff types Θ_i and Harsanyi types T_i are finite. Note that the analysis of strong rationalizability in Battigalli (2003) and its epistemic foundation in Battigalli & Tebaldi (2019) allow for a continuum of types and an infinite horizon, provided that some regularity assumptions hold, e.g., that type

[32] As well as the observed-actions assumption of this paper.

sets are compact metric spaces, feasible actions sets $\mathcal{A}_i(h)$ are finite for all $h \in H$ and $i \in I$, and payoff functions are continuous in the obvious product topology. However, the proof of Theorem 1 relies on the fact that, in a finite game with payoff uncertainty, the procedure of elimination of type-strategy pairs (θ_i, s_i) ends after finitely many steps. Even assuming a finite game with payoff uncertainty, its Bayesian elaborations could be infinite, allowing for an infinite set of possible (hierarchies of) exogenous beliefs. In this case, our proof of Theorem 2 (which yields the robust implementation Corollary 1) does not apply, because it adapts the proof of Theorem 1. Nonetheless, we conjecture that our results can be extended to games with infinite type sets and infinite horizon according to the following sketch of proof.

Introduce a sequence of elimination procedures of type-strategy pairs (θ_i, s_i) where each procedure k introduces the belief restrictions at step $k+1$. Thus, as in the proof of Theorem 1, procedure 0 is strong Δ-rationalizability, but the sequence of procedures is now *infinite* and no procedure coincides with strong rationalizability. Despite this, we conjecture that the inductive hypotheses can be formulated and proven as in the proof of Theorem 1. If so, for each strong Δ-rationalizable pair (θ_i, s_i), "diagonalizing over procedures," IH1 guarantees the existence of an "equivalent" pair $\left(\theta_i, s_i^k\right)$ at each step k of procedure k, which coincides with step k of strong rationalizability. Under the stated assumptions, the strategy set $S_i = \times_{h \in H} \mathcal{A}_i(h)$ is compact. Thus, the sequence $\left(\theta_i, s_i^k\right)_{k>0}$ admits a converging subsequence; call its limit (θ_i, s_i^*). For each k, $S_i^k(\theta_i)$ is compact and contains $S_i^{k+1}(\theta_i), S_i^{k+2}(\theta_i), ...$ So, it contains $s_i^{k+1}, s_i^{k+2}, ...$ and s_i^*. This means that (θ_i, s_i^*) survives all steps of strong rationalizability.

8 Appendix

This section contains ancillary results and 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).

8.1 Dynamic programming and forward consistency

We use the following dynamic programming result. First recall from Section 3.2 that a strategy $\bar{s}_i$ is a continuation best reply (from h) to conditional belief $\mu_i(\cdot|h) \in$

$\Delta\left(\Theta_{-i} \times S_{-i}\left(h\right)\right)$ for type θ_i if, for every $s_i \in S_i(h)$,

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

Lemma 3 *Fix a CPS μ_i, a type θ_i, and a strategy s_i. If, for every $h \in \mathcal{H}_i(s_i)$, 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, $s_i \in r_{i,\theta_i}\left(\mu_i\right)$.*

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

$$\mathbb{E}_{\mu_i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, \bar{s}_i, \cdot)\right) > \mathbb{E}_{\mu_i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i, \cdot)\right). \tag{12}$$

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 $\mu_i\left(\Theta_{-i} \times S_{-i}\left(z\right)|\bar{h}\right) > 0$ and all $\left(\theta_{-i}, s_{-i}\right) \in \Theta_{-i} \times S_{-i}\left(z\right)$, write

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

so that

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

With this, letting $\bar{A}_{-i} = \left\{a_{-i} : \mu_i\left(\Theta_{-i} \times S_{-i}\left(\bar{h}, a_{-i}\right)|\bar{h}\right) > 0\right\}$, we can make the following decomposition:

$$\mathbb{E}_{\mu_i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i', \cdot)\right) = \sum_{a_{-i} \in \bar{A}_{-i}} \mu_i\left(\Theta_{-i} \times S_{-i}\left(\bar{h}, a_{-i}\right)|\bar{h}\right) \mathbb{E}_{\mu_i\left(\cdot|\left(\bar{h}, \left(s_i\left(\bar{h}\right), a_{-i}\right)\right)\right)}\left(U_i(\theta_i, s_i', \cdot)\right).$$

By choice of $\bar{h}$, s_i is a continuation best reply to each $\mu_i(\cdot|h)$ with $h = \left(\bar{h}, \left(s_i\left(\bar{h}\right), a_{-i}\right)\right) \in H$, and by $s_i'\left(\bar{h}\right) = s_i\left(\bar{h}\right)$, $s_i' \in S_i(h)$ as well. 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 $a_{-i} \in \bar{A}_{-i}$ (the other action profiles in $\mathcal{A}_{-i}\left(\bar{h}\right)$ do not affect expected payoff calculations). It follows that

$$\mathbb{E}_{\mu_i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i, \cdot)\right) \geq \mathbb{E}_{\mu_i\left(\cdot|\bar{h}\right)}\left(U_i(\theta_i, s_i', \cdot)\right). \tag{13}$$

Equations (12) and (13) 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' is not a continuation best reply to $\mu_i\left(\cdot|\bar{h}\right)$. ■

The omitted parts of the proof of Theorem 1 require to construct CPSs that strongly believe some key events. It turns out that it is simpler to construct a "forward-consistent belief system" (Battigalli, Catonini & Manili 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 $\hat{\mu}_i = (\hat{\mu}_i(\cdot|h))_{h\in H} \in \left(\Delta\left(\Theta_{-i} \times S_{-i}\right)\right)^H$ 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 $E_{-i} \subseteq \Theta_{-i} \times S_{-i}(h')$,

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

The forward chain rule is weaker than the chain rule, because, as noted in Section 3.2, $S_{-i}(h') \subseteq S_{-i}(h)$ does not imply $h \preceq h'$. The definition of "strong belief" for a forward-consistent belief system is the same as for a CPS: belief system $\hat{\mu}_i$ **strongly believes** E_{-i} if

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

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

Lemma 4 (Battigalli, Catonini & 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 $h \in \mathcal{H}_i(s_i)$.*

8.2 Omitted parts of the proof of Theorem 1

8.2.1 Proof of Claim 1

We construct an array of beliefs $\hat{\mu}_i = (\hat{\mu}_i(\cdot|h))_{h \in H}$ such that, for each $h \in H$:

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

F1. for all h' such that $h \prec h'$,

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

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

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

F4. if $h \in \mathcal{H}(\mathrm{X}^n_k) \cap \mathcal{H}_i(s_i)$,

$$\forall(\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}^n_k), \;\; \hat{\mu}_i(\{\theta_{-i}\} \times S_{-i}(z)|h) = \mu_i(\{\theta_{-i}\} \times S_{-i}(z)|h). \tag{15}$$

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 4, 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 $h \in \mathcal{H}_i(s_i)$. By $\widetilde{\mu}_i(\cdot|\varnothing) = \hat{\mu}_i(\cdot|\varnothing)$, F3, and $\mu_i \in \Delta_{i,\theta_i}$, we get $\widetilde{\mu}_i \in \Delta_{i,\theta_i}$. Finally, for every $h \in \mathcal{H}(\mathrm{X}^n_k) \cap \mathcal{H}_i(s_i)$, $\widetilde{\mu}_i(\cdot|h) = \hat{\mu}_i(\cdot|h)$ and F4 yield (11).

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 all $h \in \mathcal{H}(\mathrm{X}^{n-1}_k)$. With this, define a map $\widetilde{\eta} : \Theta_{-i} \times S_{-i} \to \Theta_{-i} \times S_{-i}$ as follows:

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

For each $h \in \mathcal{H}\left(\mathrm{X}_k^n\right)$, define $\hat{\mu}_i\left(\cdot|h\right)$ as the $\widetilde{\eta}$-pushforward (image measure) of $\mu_i\left(\cdot|h\right)$. For future reference, observe that

$$\hat{\mu}_i\left(\mathrm{X}_{k-1,-i}^{n-1}|h\right) = \mu_i\left(\widetilde{\eta}^{-1}(\mathrm{X}_{k-1,-i}^{n-1})|h\right) \geq \mu_i\left(\mathrm{X}_{k,-i}^{n}|h\right) = 1, \tag{16}$$

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

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

For each $h \in \widetilde{H}$, let $p^*(h)$ denote the longest $\bar{h} \prec h$ with $\bar{h} \in \mathcal{H}\left(\mathrm{X}_k^n\right)$ 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 $\bar{\mu}_i \in \cap_{m=0}^{n-1}\Delta_{\mathrm{sb}}^H(\mathrm{X}_{k-1,-i}^m)$, and for each $h \in H \backslash \left(\mathcal{H}\left(\mathrm{X}_k^n\right) \cup \widetilde{H}\right) =: \hat{H}$, 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 $h \in \mathcal{H}\left(\mathrm{X}_k^n\right)$, equation (16) yields $\hat{\mu}_i\left(\mathrm{X}_{k-1,-i}^{n-1}|h\right) = 1$. For each $h \in \widetilde{H}$, equation (16) yields $\hat{\mu}_i\left(\mathrm{X}_{k-1,-i}^{n-1}|p^*(h)\right) = 1$, from which $\hat{\mu}_i\left(\mathrm{X}_{k-1,-i}^{n-1}|h\right) = 1$ follows by construction. For each $h \in \hat{H}$ and $m = 0, ..., n-1$, if $h \in \mathcal{H}(\mathrm{X}_{k-1,-i}^m)$, $\hat{\mu}_i\left(\mathrm{X}_{k-1,-i}^{m}|h\right) = 1$ follows from $\hat{\mu}_i\left(\cdot|h\right) = \bar{\mu}_i\left(\cdot|h\right)$ and $\bar{\mu}_i \in \Delta_{\mathrm{sb}}^H(\mathrm{X}_{k-1,-i}^m)$.

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

$$\hat{\mu}_i(\{\theta_{-i}\} \times S_{-i}(h')|h) = \mu_i(\{\theta_{-i}\} \times S_{-i}(h')|h), \tag{17}$$

which yields: condition (15) when $h' \in \mathcal{Z}(\mathrm{X}_k^n)$, thus F4; F3 when h and h' coincide with the initial history; and, for future reference,

$$\hat{\mu}_i(\Theta_{-i} \times S_{-i}(h')|h) = \mu_i(\Theta_{-i} \times S_{-i}(h')|h). \tag{18}$$

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

$$\widetilde{\eta}^{-1}\left(\{\theta_{-i}\} \times S_{-i}(h')\right) = \{\theta_{-i}\} \times S_{-i}(h'). \tag{19}$$

Fix first $s_{-i} \in S_{-i}$ such that $(\theta_{-i}, s_{-i}) \in \widetilde{\eta}^{-1}(\{\theta_{-i}\} \times S_{-i}(h'))$. 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 $\widetilde{h} \in \mathcal{H}(\mathrm{X}_k^{n-1})$, so in particular for each $\widetilde{h} \prec h'$, given that $h' \in \mathcal{H}(\mathrm{X}_k^n) \cup \mathcal{Z}(\mathrm{X}_k^n)$. Hence, $s'_{-i} \in S_{-i}(h')$ 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 $\widetilde{h} \in \mathcal{H}(\mathrm{X}_k^{n-1})$, so in particular for each $\widetilde{h} \prec h'$, given that $h' \in \mathcal{H}(\mathrm{X}_k^n) \cup \mathcal{Z}(\mathrm{X}_k^n)$. Hence, $s_{-i} \in S_{-i}(h')$ implies $s'_{-i} \in S_{-i}(h')$, which means $(\theta_{-i}, s_{-i}) \in \widetilde{\eta}^{-1}(\{\theta_{-i}\} \times S_{-i}(h'))$.

Finally, we show that $\hat{\mu}_i$ satisfies F0 and F1. For each $h \in \mathcal{H}(\mathrm{X}_k^n)$, since $\mu_i(\Theta_{-i} \times S_{-i}(h)|h) = 1$, equation (18) with $h' = h$ yields F0. For each $h \in \widetilde{H}$, F0 follows by conditioning. For each $h \in \hat{H}$, F0 holds by $\hat{\mu}_i(\cdot|h) = \bar{\mu}_i(\cdot|h)$.

For F1, equation (14) holds 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: $h \in \mathcal{H}(\mathrm{X}_k^n)$. If $h' \in \mathcal{H}(\mathrm{X}_k^n)$, 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, $\bar{h} \in \mathcal{H}(\mathrm{X}_k^n)$. For each $E \subseteq \Theta_{-i} \times S_{-i}(\bar{h})$, by construction of $\hat{\mu}_i$ and equation (18), 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 (19) 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,

$$\hat{\mu}_i(E|\bar{h})\hat{\mu}_i(\Theta_{-i} \times S_{-i}(\bar{h})|h) = \hat{\mu}_i(E|h). \tag{20}$$

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 holds by definition of $\hat{\mu}_i(E'|h')$ and the second and third equalities follow from equation (20) with $\bar{h} = p^*(h')$ and $E = \Theta_{-i} \times S_{-i}(h')$ for the second equality, $E = E'$ for the third. □

8.2.2 Proof of Claim 2

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

8.2.3 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 3 that $\widetilde{s}_i \in r_{i,\theta_i}(\widetilde{\mu}_i)$. □

8.2.4 Proof of Claim 4

First note that $\mathcal{H}(\text{X}_k^n) \cap \mathcal{H}_i(s_i)$ is closed with respect to prefixes (predecessors): for each $h \in \mathcal{H}(\text{X}_k^n) \cap \mathcal{H}_i(s_i)$ each prefix $h' \prec h$ belongs to $\mathcal{H}(\text{X}_k^n) \cap \mathcal{H}_i(s_i)$. So, suppose by way of induction that Claim 4 holds for every $h' \prec h$, which 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 $\zeta(\widetilde{s}_i, \widetilde{s}_{-i}) \in \mathcal{Z}(\text{X}_k^n)$ for every $(\theta_{-i}, \widetilde{s}_{-i}) \in \text{supp}\widetilde{\mu}_i(\cdot|h)$. So, fix $(\theta_{-i}, \widetilde{s}_{-i}) \in \text{supp}\widetilde{\mu}_i(\cdot|h)$. Note that $\{\theta_i\} \times r_{i,\theta_i}(\widetilde{\mu}_i) \subseteq \text{X}_{k-1,i}^n$, and hence $\widetilde{s}_i \in \text{proj}_{S_i}\text{X}_{k-1,i}^n$. So, by IH1($n$) there exists $\widetilde{s}_i' \in \text{proj}_{S_i}\text{X}_{k,i}^n$ such that $\widetilde{s}_i'(h) = \widetilde{s}_i(h)$ for every $h \in \mathcal{H}(\text{X}_{k-1}^{n-1})$.[33] Fix $\left(\theta_{-i}, \widetilde{s}_{-i}'\right) \in \widetilde{\eta}^{-1}((\theta_{-i}, \widetilde{s}_{-i})) \subseteq \text{X}_{k,-i}^n$—it exists because $\widetilde{\mu}_i(\cdot|h) = \hat{\mu}_i(\cdot|h)$ and $\hat{\mu}_i(\cdot|h)$ is the $\widetilde{\eta}$-pushforward of $\mu_i(\cdot|h)$ (see the proof of Claim 1). By definition of $\widetilde{\eta}$, $\zeta(\widetilde{s}_i', \widetilde{s}_{-i}') \in \mathcal{Z}(\text{X}_k^n)$. For every $\widetilde{h} \prec \zeta(\widetilde{s}_i', \widetilde{s}_{-i}')$, we have $\widetilde{h} \in \mathcal{H}(\text{X}_k^n) \subseteq$

[33] 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 $H(\text{X}_{k-1}^{n-1})$ and not just those in $H(\text{X}_{k-1}^n)$.

$\mathcal{H}(\mathrm{X}_k^{n-1})$, hence $\widetilde{s}_{-i}(\widetilde{h}) = \widetilde{s}'_{-i}(\widetilde{h})$ by definition of $\widetilde{\eta}$. Claim 2 gives $\mathcal{H}(\mathrm{X}_k^n) \subseteq \mathcal{H}(\mathrm{X}_{k-1}^{n-1})$, therefore $\widetilde{s}_i(\widetilde{h}) = \widetilde{s}'_i(\widetilde{h})$ as well. It follows that $\zeta(\widetilde{s}_i, \widetilde{s}_{-i}) = \zeta(\widetilde{s}'_i, \widetilde{s}'_{-i}) \in \mathcal{Z}(\mathrm{X}_k^n)$.

For each $(\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}_k^n)$, 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 (11), $\widetilde{s}_i$ induces the same probability over each $(\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}_k^n)$ 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 $\Theta_{-i} \times (Z \backslash \mathcal{Z}(\mathrm{X}_k^n))$ is zero: as we have previously shown, for each $(\theta_{-i}, \widetilde{s}_{-i}) \in \mathrm{supp}\widetilde{\mu}_i(\cdot|h)$, $\zeta(\widetilde{s}_i, \widetilde{s}_{-i}) \in \mathcal{Z}(\mathrm{X}_k^n)$. 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 $\zeta(s_i, s_{-i}) \in \mathcal{Z}(\mathrm{X}_k^n)$. 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 $\widetilde{\mu}_i(\cdot|h)$. $\square$

8.3 Proof of Lemma 2

The statement is trivially true for $n = 0$. Suppose by way of induction that it is true for each $m < n$; 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 $\mu_i \in \cap_{m=0}^{n-1} \Delta_{\mathrm{sb}}^H(\Sigma_{-i}^m)$ such that $\bar{s}_i \in r_{i,\theta_i}(\mu_i)$. Define $\mu_i^{\mathrm{e}} \in (\Delta(T_{-i} \times S_{-i}))^H$ as follows: for all $h \in H$, $s_{-i} \in S_{-i}(h)$, $(\theta_{-i}, e_{-i}) \in T_{-i}$,

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

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

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

[4] Battigalli, P., and N. De Vito (2021): "Beliefs, Plans and Perceived Intentions in Dynamic Games," *Journal of Economic Theory*, 195, 105283.

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

[11] Battigalli, P., Catonini, E., and De Vito, N. (2024): *Game Theory: Analysis of Strategic Thinking*, typescript.

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

[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] Fudenberg, D., D. Kreps, and D.K. Levine (1988): "On the Robustness of Equilibrium Refinements," *Journal of Economic Theory,* 44, 354-380.

[20] Glazer, J., and M. Perry (1996): "Virtual Implementation in Backwards Induction," *Games and Economic Behavior*, 15, 27-32.

[21] 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.

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

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

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

[25] Osborne, M., and A. Rubinstein (1994): *A Course in Game Theory.* Cambridge MA: MIT Press.

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

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

[28] 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.

Supplemental Appendix of "Monotonicity and Robust Implementation under Forward Induction Reasoning."

November 20, 2024

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

0.1 Proof of part IH1 of the inductive step 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 Procedure $k-1$ path-refines step $n+1$ of Procedure k. 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 $h \in \mathcal{H}(\mathrm{X}_{k-1}^{n})$.

By definition of $\mathrm{X}_{k-1,i}^{n+1}$ (cf. eq. (2) in the main text), 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 $h \in \mathcal{H}\left(\mathrm{X}_{k-1}^{n}\right) \cap \mathcal{H}_i(s_i)$,

$$\forall (\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}_{k-1}^{n}), \quad \hat{\mu}_i(\{\theta_{-i}\} \times S_{-i}(z)|h) = \mu_i(\{\theta_{-i}\} \times S_{-i}(z)|h). \tag{S.A}$$

Proof. We construct an array of beliefs $\widetilde{\mu}_i = (\widetilde{\mu}_i(\cdot|h))_{h\in H}$ as follows. By IH1(n), for every $(\theta_{-i}, s_{-i}) \in \mathrm{X}^n_{k-1,-i}$, there exists a profile $(\widetilde{s}_j^{(\theta_j,s_j)})_{j\neq i} \in S_{-i}$ such that $(\theta_j, \widetilde{s}_j^{(\theta_j,s_j)})_{j\neq i} \in \mathrm{X}^n_{k,-i}$ and, for each $j \neq i$, $\widetilde{s}_j^{(\theta_j,s_j)}(h) = s_j(h)$ for all $h \in \mathcal{H}(\mathrm{X}^{n-1}_{k-1})$. With this, define a map $\hat{\eta} : \Theta_{-i} \times S_{-i} \to \Theta_{-i} \times S_{-i}$ as follows:

$$\forall\,(\theta_{-i}, s_{-i}) \in (\Theta_{-i} \times S_{-i}), \quad \hat{\eta}\,(\theta_{-i}, s_{-i}) = \begin{cases} (\theta_j, \widetilde{s}_j^{(\theta_j,s_j)})_{j\neq i} & \text{if } (\theta_{-i}, s_{-i}) \in \mathrm{X}^n_{k-1,-i} \\ (\theta_{-i}, s_{-i}) & \text{otherwise} \end{cases}.$$

For each $h \in \mathcal{H}\left(\mathrm{X}^n_{k-1}\right)$, define $\widetilde{\mu}_i(\cdot|h)$ as the $\hat{\eta}$-pushforward (image measure) of $\mu_i(\cdot|h)$. Now define

$$\widetilde{H} = \left\{h \in H\backslash\mathcal{H}\left(\mathrm{X}^n_{k-1}\right) : \exists \bar{h} \in \mathcal{H}\left(\mathrm{X}^n_{k-1}\right), \bar{h} \prec h, \widetilde{\mu}_i\left(\Theta_{-i} \times S_{-i}(h)|\bar{h}\right) > 0\right\}.$$

For each $h \in \widetilde{H}$, let $p^*(h)$ denote the longest $\bar{h} \prec h$ with $\bar{h} \in \mathcal{H}\left(\mathrm{X}^n_{k-1}\right)$ such that $\widetilde{\mu}_i\left(\Theta_{-i} \times S_{-i}(h)|\bar{h}\right) > 0$, and derive $\widetilde{\mu}_i(\cdot|h)$ by conditioning $\widetilde{\mu}_i(\cdot|p^*(h))$. To conclude the construction, fix $\bar{\mu}_i \in \cap_{m=0}^n \Delta^H_{\mathrm{sb}}(\mathrm{X}^m_{k,-i})$, and for each $h \in H\backslash\left(\mathcal{H}\left(\mathrm{X}^n_{k-1}\right) \cup \widetilde{H}\right) =: \hat{H}$, let $\widetilde{\mu}_i(\cdot|h) = \bar{\mu}_i(\cdot|h)$. The proof that $\widetilde{\mu}_i$ is a forward-consistent belief system with the desired properties, and that it can be transformed into the desired CPS $\hat{\mu}_i$ satisfying

$$\forall h \in \mathcal{H}_i(s_i), \quad \hat{\mu}_i(\cdot|h) = \widetilde{\mu}_i(\cdot|h), \tag{1}$$

is the same as in the proof of Claim 1 in part IH2 of the inductive step, so we omit it. □

Claim 2-bis: $\mathcal{H}(\mathrm{X}^n_{k-1}) \subseteq \mathcal{H}(\mathrm{X}^n_k)$.

Proof. Fix $\hat{s} \in \mathrm{proj}_S \mathrm{X}^n_{k-1}$. By IH1$(n)$, there exists $\hat{s}' \in \mathrm{proj}_S \mathrm{X}^n_k$ such that $\hat{s}'(\hat{h}) = \hat{s}(\hat{h})$ for every $\hat{h} \in \mathcal{H}(\mathrm{X}^{n-1}_{k-1}) \supseteq \mathcal{H}(\mathrm{X}^n_{k-1})$. Thus, $\zeta(\hat{s}) = \zeta(\hat{s}') \in \mathcal{Z}(\mathrm{X}^n_k)$. □

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 3 that $\hat{s}_i \in r_{i,\theta_i}(\hat{\mu}_i)$. □

Now fix $\hat{\mu}_i$ as per Claim 1-bis. From the definition of $\mathrm{X}_{k,i}^{n+1}$ (cf. eq. (2) in the main text), 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 $h \in \mathcal{H}(\mathrm{X}_{k-1}^n)$. By Claim 3-bis with $\hat{H} = \mathcal{H}(\mathrm{X}_{k-1}^n) \cap \mathcal{H}_i(s_i)$, this is a consequence of the following result. (For each $h \in \mathcal{H}(\mathrm{X}_{k-1}^n) \backslash \mathcal{H}_i(s_i)$, since $h \notin \mathcal{H}_i(\hat{s}_i)$, we can always set $\hat{s}_i(h) = s_i(h)$ because we use a notion of sequential best reply which only refers to the histories that are consistent with the candidate strategy.)

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

Proof. First note that $\mathcal{H}\left(\mathrm{X}_{k-1}^n\right) \cap \mathcal{H}_i(s_i)$ is closed with respect to prefixes (predecessors): for each $h \in \mathcal{H}\left(\mathrm{X}_{k-1}^n\right) \cap \mathcal{H}_i(s_i)$ each prefix $h' \prec h$ belongs to $\mathcal{H}\left(\mathrm{X}_{k-1}^n\right) \cap \mathcal{H}_i(s_i)$. 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 $\zeta(\hat{s}_i, \hat{s}_{-i}) \in \mathcal{Z}(\mathrm{X}_{k-1}^n)$ for every $(\theta_{-i}, \hat{s}_{-i}) \in \mathrm{supp}\hat{\mu}_i(\cdot|h)$. So, fix $(\theta_{-i}, \hat{s}_{-i}) \in \mathrm{supp}\hat{\mu}_i(\cdot|h)$. Note that $\{\theta_i\} \times r_{i,\theta_i}(\hat{\mu}_i) \subseteq \mathrm{X}_{k,i}^{n+1}$, and hence $\hat{s}_i \in \mathrm{proj}_{S_i}\mathrm{X}_{k,i}^{n+1}$. So, 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 $h \in \mathcal{H}(\mathrm{X}_k^n)$. Fix $\left(\theta_{-i}, \hat{s}_{-i}'\right) \in \hat{\eta}^{-1}((\theta_{-i}, \hat{s}_{-i})) \subseteq \mathrm{X}_{k-1,-i}^n$—it exists by equation (1) and construction of $\widetilde{\mu}_i(\cdot|h)$. Obviously, $\zeta(\hat{s}_i', \hat{s}_{-i}') \in \mathcal{Z}(\mathrm{X}_{k-1}^n)$. For every $\hat{h} \prec \zeta(\hat{s}_i', \hat{s}_{-i}')$, we have $\hat{h} \in \mathcal{H}(\mathrm{X}_{k-1}^n) \subseteq \mathcal{H}(\mathrm{X}_{k-1}^{n-1})$, hence $\hat{s}_{-i}(\hat{h}) = \hat{s}_{-i}'(\hat{h})$ by construction of $\hat{\eta}$. Claim 2-bis gives $\mathcal{H}(\mathrm{X}_{k-1}^n) \subseteq \mathcal{H}(\mathrm{X}_k^n)$, therefore $\hat{s}_i(\hat{h}) = \hat{s}_i'(\hat{h})$ as well. It follows that $\zeta(\hat{s}_i, \hat{s}_{-i}) = \zeta(\hat{s}_i', \hat{s}_{-i}') \in \mathcal{Z}(\mathrm{X}_{k-1}^n)$.

For each $(\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}_{k-1}^n)$, 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 (S.A), $\hat{s}_i$ induces the same probability over each $(\theta_{-i}, z) \in \Theta_{-i} \times \mathcal{Z}(\mathrm{X}_{k-1}^n)$ 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 $\Theta_{-i} \times \left(Z \backslash \mathcal{Z}(\mathrm{X}_{k-1}^n)\right)$ is zero: as we have previously shown, for each $(\theta_{-i}, \hat{s}_{-i}) \in \mathrm{supp}\hat{\mu}_i(\cdot|h)$, $\zeta(\hat{s}_i, \hat{s}_{-i}) \in \mathcal{Z}(\mathrm{X}_{k-1}^n)$. 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 $\zeta(s_i, s_{-i}) \in \mathcal{Z}(\mathrm{X}_{k-1}^n)$. 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 $\hat{\mu}_i(\cdot|h)$. □

0.2 No path-monotonicity under restrictions on 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: $\Sigma^\infty = \{(0, In), (1, Out)\} \times \{\ell\}$. 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,

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

(a restriction on the *endogenous* beliefs of player 2).

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: $\Sigma^{\Delta,\infty} = \{(0, Out), (1, Out)\} \times \{c\}$. It follows that the only strongly Δ-rationalizable path is (Out).