On non-monotonic strategic reasoning*

Emiliano Catonini†

January 2019

Strong-Δ-Rationalizability (Battigalli 2003, Battigalli and Siniscalchi 2003) is a prominent and widely applied solution concept that introduces first-order belief restrictions in forward induction reasoning. In absence of restrictions, it coincides with Strong Rationalizability (Battigalli and Siniscalchi 2002). These solution concepts are based on the notion of *strong belief* (Battigalli and Siniscalchi 2002). The non-monotonicity of strong belief implies that the predictions of Strong-Δ-Rationalizability can be incompatible with the predictions of Strong Rationalizability. Here I show that Strong-Δ-Rationalizability refines Strong Rationalizability in terms of outcomes as long as the restrictions have no bite off-path. So, Strong-Δ-Rationalizability yields a subset of strongly rationalizable outcomes when the restrictions correspond to the belief in an outcome, in an outcome distribution, or in a set of outcomes that all receive positive probability. Moreover, under such restrictions, the epistemic priority between beliefs in rationality and beliefs in the restrictions is irrelevant: the outcomes predicted by Strong-Δ-Rationalizability and Selective Rationalizability (Catonini 2019) coincide. The workhorse lemma behind these results allows to show also the order independence of the "iterated elimination of never sequential best replies" (of which Strong Rationalizability is the maximal elimination order), and that Strong Rationalizability refines Backward Induction. The outcome equivalence of Strong Rationalizability and Backward Induction in perfect information games with no relevant ties (Battigalli 1997) follows.

Keywords: Strong-Δ-Rationalizability, Strong Rationalizability, First-order Belief Restrictions, Epistemic Priority, Order Independence, Backward Induction.

*I thank Pierpaolo Battigalli and Andres Perea for useful conversations, three anonymous referees of LOFT 2018, and all the attendants of my presentation at the conference. The study has been funded within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) and by the Russian Academic Excellence Project '5-100.

†Higher School of Economics, ICEF, emiliano.catonini@gmail.com

1 Introduction

Strong Rationalizability (Battigalli and Siniscalchi [7]) is a form of extensive-form rationalizability (Pearce [17]) based on the notion of *strong belief*. Concretely, it is the iterated elimination of "never sequential best replies" to belief systems that assign probability 1, as long as possible, to opponents' strategies that survive the previous step of the procedure. Strong-Δ-Rationalizability (Battigalli [4], Battigalli and Siniscalchi [8]) introduces first-order belief restrictions in the same reasoning scheme: only belief systems in an exogenously given set are allowed at all steps.

It is well-known that the introduction of belief restrictions can let the elimination procedure depart completely from Strong Rationalizability. This is due to the non-monotonicity of strong belief: strong belief in an event does not imply strong belief in a larger event. Even in a perfect information game without relevant ties, the introduction of first-order belief restrictions can induce completely different outcomes with respect to the unique strongly rationalizable/backward induction one (see, e.g., the introductory example in Catonini [9]). Are there interesting restrictions under which Strong-Δ-Rationalizability refines the set of strongly rationalizable outcomes? Under such restrictions, the predictions are reassuringly robust to "restricted" and "unrestricted" forward induction reasoning, as captured, respectively, by Strong-Δ-Rationalizability and Strong Rationalizability.

It turns out that in all games with observable actions (i.e. games where, allowing for simultaneous moves, every player knows the current history of the game) the set of outcomes predicted by Strong-Δ-Rationalizability is included in the set of strongly rationalizable outcomes as long as the restrictions "never bite off-path". With this, I refer to restrictions that exclude belief systems only based on the probabilities they assign to opponents' behavior along the paths that survive all steps of Strong-Δ-Rationalizability. So, off-path restrictions are responsible for the general non-monotonicity of Strong-Δ-Rationalizability. The reason is the following. Suppose that at some step of reasoning, a history where the move of a player is object of the opponents' belief restrictions ends up off-path (think for instance of a threat by our player that the opponents believe in). Then, the opponents will not verify whether their belief restrictions regarding the move of our player are compatible with higher orders of belief rationality, because the off-path behavior is not refined any longer by the elimination procedure (so, the threat keeps being believed and may sustain a path that is not compatible with common strong belief in rationality).

Beside the theoretical insight, though, this broad condition for outcome monotonicity is of little practical use: one cannot check it without actually performing Strong-Δ-Rationalizability. Yet, an important class of restrictions always satisfies this condition: initial common belief in an outcome (distribution), or in a set of outcomes that all receive

positive probability.[1] I will refer to both as "path restrictions". Path restrictions are important both for theory and applications. Agreements among real players often specify only one or more outcomes to achieve and fall through if a player deviates — i.e., they do not specify off-path behavior; see Catonini [10]. Or, if the source of the belief restrictions is learning, players are likely to have a rich record of observations of the path of play, but limited or no experience of what would happen off-path — a motivation used by Sobel et. al. [23, page 310]. Theoretically, path restrictions can be used to test the compatibility of an outcome distribution with a kind of forward induction reasoning, whereby a deviation from the path(s) is interpreted as an attempt of the deviator to improve her payoff with respect to the expected payoff under the outcome distribution. In [9] I elaborate on this use of path restrictions and on the connection between this kind of strategic reasoning and strategic stability (Kohlberg and Mertens [14]). Equilibrium refinements related to strategic stability are often motivated informally with the idea that deviations signal optimistic beliefs about the reactions, rather than disbelief in the equilibrium path(s). Path restrictions have been used to make this idea precise. Battigalli and Siniscalchi justify the Iterated Intuitive Criterion (Cho and Kreps [12]) precisely with non-emptiness of Strong-Δ-Rationalizability under belief in the equilibrium outcome distribution. In [10] I provide analogous motivation to rule out "equilibrium paths that cannot be upset by a convincing deviation" (Osborne [15]). Sobel et al. [23] justify the Iterated Intuitive Criterion and an amended version of Divine Equilibrium (Banks and Sobel [1]) with an iterated elimination procedure that, in signaling games, coincides with Strong-Δ-Rationalizability under belief in the equilibrium outcome distribution. However, by the non-monotonicity of strong belief, it is not clear whether an outcome distribution that is compatible with these refinements or with Strong-Δ-Rationalizability is also compatible with the best rationalization principle (Battigalli [2]) — i.e., with the idea that players always ascribe to the opponents the highest order of belief in rationality that is compatible with their observed behavior. Note that Strong-Δ-Rationalizability, also under path restrictions, allows players to give up orders of belief in rationality that are per se compatible with the observed behavior, in order to keep orders of belief in the path. The monotonicity result of this paper (which can be easily extended to signaling games) shows that, nonetheless, the outcomes compatible with this form of forward induction reasoning are also compatible with common strong belief in rationality.

In [9] I define an elimination procedure with first-order belief restrictions, Selective Rationalizability, that refines Strong Rationalizability, thus guarantees common strong belief in rationality. Selective Rationalizability is based on the idea that all orders of belief in rationality receive epistemic priority over all orders of belief in the restrictions. For instance,

[1] These are two extremes of the spectrum of sets of outcome distributions with the same support. The whole spectrum will be considered by the formal analysis.

when a player displays behavior that cannot be rational under her restrictions, Selective Rationalizability requires to keep the belief that the player is rational (if per se compatible with the observed behavior) and drop the belief that her restrictions hold, while Strong-Δ-Rationalizability captures the opposite epistemic priority choice. Selective Rationalizability and Strong-Δ-Rationalizability can even predict non-empty disjoint outcome sets for the same restrictions (see [9]). But in light of the main monotonicity result, one expects the two solution concepts to give the same predictions under path restrictions. I prove that this is the case. Hence, path restrictions have the further advantage of giving predictions that are robust to this epistemic priority choice.

The workhorse lemma of the paper also yields the following result, also proven by Perea [19]: in games with observable actions, the iterated deletion of never sequential best replies under the strong belief operator (of which Strong Rationalizability is the maximal elimination order) is order independent in terms of predicted outcomes. Chen and Micali [11] characterize Strong Rationalizability with the maximal iterated elimination of *distinguishably dominated* strategies,[2] which they prove being order independent in terms of outcomes in all games with perfect recall. Here, like in the recents work of Perea ([20], [19]), I do not use dominance characterizations.

Finally, I use the order independence result to show that Strong Rationalizability refines a generalization of backward induction to games without perfect information, Backwards Extensive Form Rationalizability of Penta [18]. Perea [19] provides an analogous result with the "backwards dominance procedure". Perea [20] had shown this already for games with perfect information and no relevant ties, where the backward induction outcome is unique, to shed new light on the outcome equivalence between backward induction and forward induction, a result originally proven by Battigalli [3] and then by Heifetz and Perea [13] in a more direct way. I provide this classical result as a corollary as well.

Section 2 introduces the formal framework for the analysis. Section 3 defines elimination procedures and introduces the workhorse lemma. Section 4 presents the results on outcome monotonicity with respect to first-order belief restrictions and outcome equivalence with respect to the epistemic priority choice, along with an example. Section 5 presents the results on order independence and backward induction. Section 6 provides the proof of the workhorse lemma. In Section 7 I elaborate on similarities and differences between my approach and Perea's. The Appendix collects the proofs omitted from the main text.

[2] They do so by proving the equivalence between distinguishable and conditional dominance, whereby the maximal iterated elimination of conditionally eliminated strategies was proven to be equivalent to extensive-form rationalizability by Shimoji and Watson [22].

2 Preliminaries

Primitives of the game.[3] Let I be the finite set of *players*. For any profile of sets $(X_i)_{i\in I}$ and any subset of players $\emptyset \neq J \subseteq I$, I write $X_J := \times_{j\in J} X_j$, $X := X_I$, $X_{-i} := X_{I\setminus\{i\}}$. Let $(\overline{A}_i)_{i\in I}$ be the finite sets of *actions* potentially available to each player. Let $\overline{H} \subseteq \cup_{t=1,...,T} \overline{A}^t \cup \{h^0\}$ be the set of histories, where $h^0 \in \overline{H}$ is the empty, initial history and T is the finite horizon. The set $\overline{H}$ must have the following properties. First property: For any $h = (a^1, ..., a^t) \in \overline{H}$ and $l < t$, it holds $h' = (a^1, ..., a^l) \in \overline{H}$, and I write $h' \prec h$.[4] Let $Z := \{z \in \overline{H} : \nexists h \in \overline{H}, z \prec h\}$ be the set of terminal histories (henceforth, *outcomes* or *paths*)[5], and $H := \overline{H}\setminus Z$ the set of non-terminal histories (henceforth, just *histories*). Second property: For every $h \in H$, there exists a non-empty set $A_i(h) \subseteq \overline{A}_i$ for each $i \in I$,[6] such that $(h, a) \in \overline{H}$ if and only if $a \in A(h)$. For each $i \in I$, let $u_i : Z \to \mathbb{R}$ be the *payoff function*. The list $\Gamma = \langle I, \overline{H}, (u_i)_{i\in I}\rangle$ is a *finite game with complete information and observable actions*.

Derived objects. A *strategy* of player i is an element of $\times_{h\in H} A_i(h)$. Let S_i denote the set of all strategies of i. A strategy *profile* $s \in S$ naturally induces a unique outcome $z \in Z$. Let $\zeta : S \to Z$ be the function that associates each strategy profile with the induced outcome. For any $h \in \overline{H}$, the set of strategies of i compatible with h is:

$$S_i(h) := \{s_i \in S_i : \exists z \succeq h, \exists s_{-i} \in S_{-i}, \zeta(s_i, s_{-i}) = z\}.$$

For any subset of players $J \subseteq I$ and any $\overline{S}_J \subset S_J$, let $\overline{S}_J(h) := S_J(h) \cap \overline{S}_J$. Let $H(\overline{S}_J) := \{h \in H : \overline{S}_J(h) \neq \emptyset\}$ denote the set of histories compatible with $\overline{S}_J$. For any $h = (h', a) \in \overline{H}\setminus\{h^0\}$, let $p(h)$ denote the immediate predecessor h' of h.

Since the game has observable actions, each history $h \in H$ is the root of a subgame $\Gamma(h)$. If $h \neq h^0$, all the objects in $\Gamma(h)$ will be denoted with h as superscript, except for histories and outcomes, which will be identified with histories and outcomes of the whole game, and not redefined as shorter sequences of action profiles. For any $h \in H$, $s_i^h \in S_i^h = \times_{h'\succeq h} A_i(h')$, and $\widehat{h} \in H^h = \{h' \in H : h \preceq h'\}$, $s_i^h|\widehat{h}$ will denote the strategy $s_i^{\widehat{h}} \in S_i^{\widehat{h}}$ such that $s_i^{\widehat{h}}(\widetilde{h}) = s_i^h(\widetilde{h})$ for all $\widetilde{h} \succeq \widehat{h}$. For any $\overline{S}_i^h \subseteq S_i^h$, $\overline{S}_i^h|\widehat{h}$ will denote the set of all strategies $s_i^{\widehat{h}} \in S_i^{\widehat{h}}$ such that $s_i^{\widehat{h}} = s_i^h|\widehat{h}$ for some $s_i^h \in \overline{S}_i^h$.

Beliefs. In this dynamic framework, beliefs are modeled as Conditional Probability Systems (Renyi, [21]; henceforth, CPS).

[3] The main notation is almost entirely taken from Osborne and Rubinstein [16].

[4] Then, $\overline{H}$ endowed with the precedence relation $\prec$ is a tree with root h^0.

[5] "Path" will be used with emphasis on the moves, and "outcome" with emphasis on the end-point of the game.

[6] When player i is not truly active at history h, $A_i(h)$ consists of just one "wait" action.

Definition 1 *Fix $i \in I$. An array of probability measures $(\mu_i(\cdot|h))_{h\in H}$ over co-players' strategies S_{-i} is a Conditional Probability System if for all $h \in H$, $\mu_i(S_{-i}(h)|h) = 1$, and for all $h' \succ h$ and $\overline{S}_{-i} \subseteq S_{-i}(h')$,*

$$\mu_i(\overline{S}_{-i}|h) = \mu_i(S_{-i}(h')|h) \cdot \mu_i(\overline{S}_{-i}|h').$$

The set of all CPS's on S_{-i} is denoted by $\Delta^H(S_{-i})$.

For any subset of opponents' strategies $\overline{S}_{-i} \subseteq S_{-i}$, I say that a CPS $\mu_i \in \Delta^H(S_{-i})$ *strongly believes* $\overline{S}_{-i}$ if, for all $h \in H(\overline{S}_{-i})$, $\mu_i(\overline{S}_{-i}|h) = 1$. I fix the following convention: $H(\emptyset) = \emptyset$. With this, the empty set is always strongly believed, because the condition is vacuously satisfied.

Rationality. I consider players who reply rationally to their conjectures. Rationality here means that players, at every history, choose an action that maximizes expected payoff given the belief about how the opponents will play and the expectation to play rationally again in the continuation of the game. By standard arguments, this is equivalent to playing a *sequential best reply* to the CPS.

Definition 2 *Fix $\mu_i \in \Delta^H(S_{-i})$. A strategy $s_i \in S_i$ is a sequential best reply to μ_i if for every $h \in H(s_i)$,*[7] *s_i is a continuation best reply to $\mu_i(\cdot|h)$, i.e., for every $\widetilde{s}_i \in S_i(h)$,*

$$\sum_{s_{-i}\in S_{-i}(h)} u_i(\zeta(s_i, s_{-i}))\mu_i(s_{-i}|h) \geq \sum_{s_{-i}\in S_{-i}(h)} u_i(\zeta(\widetilde{s}_i, s_{-i}))\mu_i(s_{-i}|h).$$

The set of sequential best replies to μ_i is denoted by $\rho_i(\mu_i)$. For each $h \in H$, the set of continuation best replies to $\mu_i(\cdot|h)$ is denoted by $\widehat{r}_i(\mu_i, h)$.

3 Elimination procedures and the main lemma

I provide a very general notion of elimination procedure for a subgame $\Gamma(h)$, which encompasses all the procedures I am ultimately interested in, or that will be used in the proofs.

Definition 3 *Fix $h \in H$. An elimination procedure in $\Gamma(h)$ is a sequence $((S^h_{i,q})_{i\in I})^\infty_{q=0}$ where, for every $i \in I$,*

EP1 $S^h_{i,0} = S^h_i$;

[7] It would be immaterial for the analysis to require s_i to be optimal also at the histories precluded by itself.

EP2 $S^h_{i,n-1} \supseteq S^h_{i,n}$ *for all* $n \in \mathbb{N}$*;*

EP3 for every $s^h_i \in S^h_{i,\infty} = \cap_{n\in\mathbb{N}} S^h_{i,n}$*, there exists* μ^h_i *that strongly believes* $(S^h_{-i,q})^\infty_{q=0}$ *such that* $s^h_i \in \rho_i(\mu^h_i) \subseteq S^h_{i,\infty}$*.*

Note three things. First, that $s^h_i \in \rho_i(\mu^h_i)$ for some μ^h_i that strongly believes $(S^h_{-i,q})^n_{q=0}$ does not imply that $s^h_i \in S^{n+1}_i$, and vice versa; this allows to encompass first-order belief restrictions and "slow" elimination orders. Second, EP2 allows $S^h_n = S^h_{n+1}$, and still $S^h_\infty \subsetneq S^h_{n+1}$; that is, the eliminations can stop for all players and then restart. Third, $S^h_{i,n} = \emptyset$ implies $S^h_{i,m} = \emptyset$ for all $m > n$, but does not imply $S^h_{j,\infty} = \emptyset$ for $j \neq i$: as already established, the empty set is always strongly believed, hence EP3 can be satisfied for j. These three facts allow Definition 3 to encompass the "truncation" $((S^h_{i,q}(\widehat{h})|\widehat{h})_{i\in I})^\infty_{q=0}$ of an elimination procedure in a subgame $\Gamma(\widehat{h})$ ($\widehat{h} \succ h$).

Remark 1 *For every elimination procedure* $((S^h_{i,q})_{i\in I})^\infty_{q=0}$ *and every* $\widehat{h} \succ h$*,* $((S^h_{i,q}(\widehat{h})|\widehat{h})_{i\in I})^\infty_{q=0}$ *is an elimination procedure.*

It will be important to keep in mind that a strategy $s^{\widehat{h}}_i$ can be eliminated from $S^h_{i,n}(\widehat{h})|\widehat{h}$ "exogenously": it may be a sequential best reply to some $\mu^{\widehat{h}}_i$ that strongly believes $(S^h_{-i,q}(\widehat{h})|\widehat{h})^n_{q=0}$, but no μ^h_i that strongly believes $(S^h_{-i,q})^n_{q=0}$ and corresponds to $\mu^{\widehat{h}}_i$ in $\Gamma(\widehat{h})$ induces player i to allow $\widehat{h}$.

The workhorse lemma of the paper claims the outcome inclusion between two elimination procedures, $((\overline{S}^h_{i,q})_{i\in I})^\infty_{q=0}$ and $((S^h_{i,q})_{i\in I})^\infty_{q=0}$, with the following feature. For each player i and each strategy $\overline{s}^h_i \in \overline{S}^h_{i,\infty}$, fix a CPS $\overline{\mu}^h_i(\overline{s}^h_i)$ that satisfies EP3; i.e., it strongly believes $(\overline{S}^h_{-i,q})^\infty_{q=0}$ and justifies $\overline{s}^h_i$: $\overline{s}^h_i \in \rho_i(\overline{\mu}^h_i(\overline{s}^h_i)) \subseteq \overline{S}^h_{i,\infty}$. Consider a CPS μ^h_i that "mimicks" $\overline{\mu}^h_i(\overline{s}^h_i)$ along the paths predicted by the first procedure; that is, at every history $\widetilde{h} \in H(\overline{S}^h_\infty)$, μ^h_i and $\overline{\mu}^h_i(\overline{s}^h_i)$ assign the same probabilities to (the opponents playing compatibly with) each $z \in \zeta(\overline{S}^h_\infty)$: $\mu^h_i(S_{-i}(z)|\widetilde{h}) = \overline{\mu}^h_i(s^h_i)(S_{-i}(z)|\widetilde{h})$. Suppose that, for each $m \in \mathbb{N}$, if every such μ^h_i strongly believes $(\overline{S}^h_{-i,q})^{m-1}_{q=0}$, then $\rho_i(\mu^h_i) \subseteq \overline{S}^h_{i,m}$, and if μ^h_i strongly believes $(S^h_{-i,q})^{m-1}_{q=0}$, then $\rho_i(\mu^h_i) \subseteq S^h_{i,m}$. If this is true, the lemma claims that the second procedure predicts a superset $\zeta(S^h_\infty) \supseteq \zeta(\overline{S}^h_\infty)$ of the outcomes predicted by the first.

Lemma 1 *Fix* $h \in H$ *and two elimination procedures* $((\overline{S}^h_{i,q})_{i\in I})^\infty_{q=0}$*,* $((S^h_{i,q})_{i\in I})^\infty_{q=0}$*.*
For every $i \in I$*, fix a map* $\overline{\mu}^h_i : \overline{S}^h_{i,\infty} \to \Delta^{H^h}_i(S^h_{-i})$ *such that, for each* $\overline{s}^h_i \in \overline{S}^h_{i,\infty}$*,* $\overline{\mu}^h_i(\overline{s}^h_i)$ *strongly believes* $(\overline{S}^h_{-i,q})^\infty_{q=0}$*, and* $\overline{s}^h_i \in \rho_i(\overline{\mu}^h_i(\overline{s}^h_i)) \subseteq \overline{S}^h_{i,\infty}$*.*
Suppose that the two procedures satisfy the following property:

A0 *for every $i \in I$, $\overline{s}_i^h \in \overline{S}_{i,\infty}^h$, $m \in \mathbb{N}$, and for every μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^{m-1}$ (resp., $(\overline{S}_{-i,q}^h)_{q=0}^{m-1}$) and satisfies*

$$\mu_i^h(S_{-i}(z)|\widetilde{h}) = \overline{\mu}_i^h(\overline{s}_i^h)(S_{-i}(z)|\widetilde{h}) \quad \forall \widetilde{h} \in H(\overline{S}_\infty^h), \forall z \in \zeta(\overline{S}_\infty^h), \tag{1}$$

we have $\rho_i(\mu_i^h) \subseteq S_{i,m}^h$ (resp., $\rho_i(\mu_i^h) \subseteq \overline{S}_{i,m}^h$).

Then, $\zeta(\overline{S}_\infty^h) \subseteq \zeta(S_\infty^h)$.

The proof of the lemma is in Section 6. I provide here an overview, whose main ideas are also illustrated in the next section with an example. Take the paths induced by the first procedure (Procedure 1). As long as they survive also the second procedure (Procedure 2), players can form beliefs that, along the paths, mimick the beliefs that justify the output of Procedure 1. The sequential best replies to these beliefs survive Procedure 2 by A0. Then, the only way one of these paths can disappear is that at some step $n+1$, for all the beliefs over the previous steps that mimick a specific one, a player finds a deviation *outside* of these paths more profitable, no matter what she believes the reactions of the opponents to the deviation will be. The opponents may be surprised by the deviation, hence they may react with any continuation plan that survives until step n. This is because until step n all players followed all paths while believing that the others would follow them as well. Suppose for simplicity that the deviator has a deterministic belief as to which subgame the deviation will lead to. So, the truncation of Procedure 2 in the subgame at step n features, both for the deviator and the opponents, all the sequential best replies to all the beliefs over step n itself. Then, the output of the following, auxiliary elimination procedure for the subgame is non-empty: it coincides with the truncation of Procedure 2 until step n, and iteratively eliminates the substrategies that are never sequential best replies afterwards. Take the subpaths induced by this auxiliary procedure. We want to show that they should have survived also the truncation of Procedure 1 in the subgame, which implies that the subgame is reached at the end of Procedure 1, contradicting that it follows a deviation from the paths predicted by Procedure 1. Note first that these subpaths, as long as they survive, incentivize the deviation also along the first procedure. Now, if the subgame is a static game, i.e., it has depth 1, it is easy to see that they survive — see the example of the next section. If the subgame has depth higher than 1, then suppose by induction that the lemma is true in games of smaller depth than $\Gamma(h)$ — as a basis step, it is easy to see that the lemma holds in static games. By the absence of off-path restrictions (i.e., by A0 for Procedures 1 and 2) and by the deviation incentives, the truncation of Procedure 1 in the subgame and the auxiliary procedure both satisfy A0 with respect to the subpaths induced by the auxiliary procedure, thus with inverted roles with respect to Procedures 1

and 2 that generated them. Then, by the lemma, the truncation of the Procedure 1 induces a non-empty superset of those subpaths, which leads to the desired contradiction.

4 Belief-restrictions and monotonicity

In this section, I am going to consider the following elimination procedures (for the whole game).

Definition 4 *An elimination procedure* $((S_{i,q})_{i\in I})_{q=0}^{\infty}$ *is "unconstrained" when for every* $n \in \mathbb{N}$, $i \in I$, *and* μ_i *that strongly believes* $(S_{-i,q})_{q=0}^{n-1}$, $\rho_i(\mu_i) \subseteq S_{i,n}$.

Definition 5 *Strong Rationalizability is the unconstrained elimination procedure* $((S_i^q)_{i\in I})_{q=0}^{\infty}$ *such that for every* $n \in \mathbb{N}$, $i \in I$, *and* $s_i \in S_{i,n}$, *there is* μ_i *that strongly believes* $(S_{-i,q})_{q=0}^{n-1}$ *with* $s_i \in \rho_i(\mu_i)$.[8]

Definition 6 *For each* $i \in I$, *fix* $\Delta_i \subseteq \Delta^H(S_{-i}^h)$. *Strong-*$\Delta$*-Rationalizability is the elimination procedure* $((S_{i,\Delta}^q)_{i\in I})_{q=0}^{\infty}$ *such that, for every* $n \in \mathbb{N}$, $i \in I$, *and* $s_i \in S_i$, $s_i \in S_{i,\Delta}^n$ *if and only if* $s_i \in \rho_i(\mu_i)$ *for some* $\mu_i \in \Delta_i$ *that strongly believes* $(S_{-i,\Delta}^q)_{q=0}^{n-1}$.[9]

Definition 7 *For each* $i \in I$, *fix* $\Delta_i \subseteq \Delta^H(S_{-i}^h)$. *Selective Rationalizability is the elimination procedure* $((S_{i,R\Delta}^q)_{i\in I})_{q=0}^{\infty}$ *such that:*

1. $(S_{R\Delta}^q)_{q=0}^{M} = (S^q)_{q=0}^{M}$, *where* M *is the smallest* $n \geq 0$ *such that* $S^{n+1} = S^n$;
2. *for every* $n > M$, $i \in I$, *and* $s_i \in S_i$, $s_i \in S_{i,R\Delta}^n$ *if and only if* $s_i \in \rho_i(\mu_i)$ *for some* $\mu_i \in \Delta_i$ *that strongly believes* $(S_{-i,R\Delta}^q)_{q=0}^{n-1}$.[10]

Consider first-order belief restrictions $(\Delta_i)_{i\in I}$ with the following characteristic: for each player i and CPS μ_i, all that matters to determine whether μ_i belongs to Δ_i are the probabilities assigned at the strongly-Δ-rationalizable histories $h \in H(S_{\Delta}^{\infty})$ to the (opponents playing compatibly with the) strongly-Δ-rationalizable paths $z \in \zeta(S_{\Delta}^{\infty})$: $\mu_i(S_{-i}(z)|h)$. Then, Strong-Δ-Rationalizability satisfies the hypotheses of Lemma 1 as first elimination procedure, whereas Strong Rationalizability, being an unconstrained procedure, obviously satisfies the hypotheses of Lemma 1 as second elimination procedure. The desired outcome inclusion with respect to belief restrictions that "do not end up off-path" obtains.

[8] The present definition of Strong Rationalizability is the one of Battigalli [4].

[9] The present definition of Strong-Δ-Rationalizability is the one of Battigalli and Prestipino [6].

[10] Selective Rationalizability in [9] is initialized with $S_{R\Delta}^0 = S^{\infty}$ and strong belief in $(S_{-i}^q)_{q=0}^{\infty}$. The present definition complies with EP1 and is equivalent in finite games. Moreover, in [9] Selective Rationaliability is defined under the assumption of *independent rationalization*. That is, a valid μ_i is required to strongly believe $(S_{j,R\Delta}^q)_{q=0}^{n-1}$ for all $j \neq i$, in place of just $(S_{-i,R\Delta}^q)_{q=0}^{n-1}$. However, this assumption is immaterial for the result on Selective Rationalizability of this paper (Theorem 4).

Theorem 1 *For each $i \in I$, fix $\Delta_i \subseteq \Delta^H(S_{-i})$ and suppose that for each $\mu_i \in \Delta_i$ and $\mu_i' \in \Delta^H(S_{-i})$,*

$$\left(\forall \widetilde{h} \in H(S_\Delta^\infty), \forall z \in \zeta(S_\Delta^\infty), \mu_i'(S_{-i}(z)|\widetilde{h}) = \mu_i(S_{-i}(z)|\widetilde{h})\right) \Rightarrow \left(\mu_i' \in \Delta_i\right). \tag{2}$$

Then, $\zeta(S_\Delta^\infty) \subseteq \zeta(S^\infty)$.

Proof. For each $i \in I$ and $s_i \in S_{i,\Delta}^\infty$, fix $\overline{\mu}_i(s_i) \in \Delta_i$ that strongly believes $(S_{-i,\Delta}^q)_{q=0}^\infty$ such that $s_i \in \rho_i(\overline{\mu}_i(s_i))$. For each μ_i' that satisfies (1) with $\overline{\mu}_i(s_i)$, by (2) we have $\mu_i' \in \Delta_i$. So, if μ_i' strongly believes $(S_{-i,\Delta}^q)_{q=0}^n$, $\rho_i(\mu_i') \subseteq S_{i,\Delta}^{n+1}$. Thus, A0 of Lemma 1 is satisfied for Strong-Δ-Rationalizability, while it is trivially satisfied for Strong Rationalizability. Hence, by Lemma 1, $\zeta(S_\Delta^\infty) \subseteq \zeta(S^\infty)$. ■

As discussed in the Introduction, Theorem 1 provides insight on what can determine the non-monotonicity of the predicted outcome set with respect to the belief restrictions: the presence of off-path restrictions. Yet, it is of little help in determining ex-ante which belief restrictions preserve the predictions of common strong belief in rationality and which do not. This is because whether the restrictions are off-path or not has to be assessed with respect to the final output of Strong-Δ-Rationalizability itself.

Consider now first-order belief restrictions that correspond to the belief in a set of outcome distributions with the same support. To start, consider the set of all product measures $\nu = \times_{i \in I} \nu_i \in \Delta(S)$ $(\nu_i \in \Delta(S_i))$ over strategy profiles that induce a distribution over outcomes with a given support $\overline{Z} \subset Z$ — the focus on product measures is of course restrictive (also for the possible supports $\overline{Z}$), but it simplifies the exposition. Formally,

$$\Delta_{\overline{Z}} = \left\{\nu = \times_{i \in I} \nu_i \in \Delta(S) | \nu(S(z)) > 0 \Leftrightarrow z \in \overline{Z}\right\}.$$

Fix a non-empty subset $\overline{\Delta}_{\overline{Z}}$ of $\Delta_{\overline{Z}}$. Every player i initially believes that the opponents will play compatibly with an outcome distribution induced by some $\nu \in \overline{\Delta}_{\overline{Z}}$. So, the belief restrictions of player i are

$$\Delta_i := \left\{\mu_i \in \Delta^H(S_{-i}) | \exists \nu \in \overline{\Delta}_{\overline{Z}}, \forall z \in \overline{Z}, \mu_i(S_{-i}(z)|h^0) = \left(\times_{j \neq i} \nu_j\right)(S_{-i}(z))\right\}.$$

If $\overline{\Delta}_{\overline{Z}}$ is a singleton, these restrictions correspond to the belief in an outcome distribution; if $\overline{Z}$ is a singleton, they correspond to the belief in a specific path z: $\mu_i(S_{-i}(z)|h^0) = 1$.[11] For this reason, I will generically refer to restrictions constructed as above as "path restrictions",

[11] It can be easily proven that, for Strong-Δ-Rationalizability and Selective Rationalizability, initial belief in $S_{-i}(z)$ is equivalent to strong belief in $S_j(z)$ for all $j \neq i$ (the *belief in the (path) agreement* as modeled in [10]). The reason is that after a deviation from the path by a player different than j, believing that j would have kept complying with the path is not restrictive for the expected behavior of j after the deviation.

and to $\overline{Z}$ as the "support" of the restrictions. When $\overline{Z}$ is not a singleton and $\overline{\Delta}_{\overline{Z}} = \Delta_{\overline{Z}}$, these restrictions correspond to a "restricted full support" condition with respect to a subset of paths.

Under path restrictions, when Strong-Δ-Rationalizability yields a non-empty set, all the paths in $\overline{Z}$ must be strongly-Δ-rationalizable: if at some step n a player i eliminates all the strategies compatible with a path $z \in \overline{Z}$, that is, $S^n_{i,\Delta} \cap S_i(z) = \emptyset$, it is easy to see that $S^{n+1}_{j,\Delta} = \emptyset$ for each $j \neq i$. Then, (2) holds and Theorem 1 can be applied.

Theorem 2 *Fix path restrictions $(\Delta_i)_{i\in I}$ with support $\overline{Z}$. Then, $\zeta(S^\infty_\Delta) \subseteq \zeta(S^\infty)$.*

Proof. If $S^\infty_\Delta = \emptyset$, $\zeta(S^\infty_\Delta) \subseteq \zeta(S^\infty)$ is trivially true, so suppose $S^\infty_\Delta \neq \emptyset$. This implies that for each $z \in \overline{Z}$, $z \in \zeta(S^\infty_\Delta)$. So, for each $i \in I$, $\mu_i \in \Delta_i$, and $\mu'_i \in \Delta^H(S_{-i})$, we have

$$\begin{aligned}\left(\forall \widetilde{h} \in H(S^\infty_\Delta), \forall z \in \zeta(S^\infty_\Delta), \mu'_i(S_{-i}(z)|\widetilde{h}) = \mu_i(S_{-i}(z)|\widetilde{h})\right) \quad &\Rightarrow \\ \left(\forall z \in \overline{Z}, \mu'_i(S_{-i}(z)|h^0) = \mu_i(S_{-i}(z)|h^0)\right) \quad &\Rightarrow \quad \left(\mu'_i \in \Delta_i\right).\end{aligned}$$

Thus, (2) holds, and Theorem 1 yields $\zeta(S^\infty_\Delta) \subseteq \zeta(S^\infty)$. ■

Corollary 3 *Fix $z \in Z$. Let Δ_i be the set of all $\mu_i \in \Delta^H(S_{-i})$ such that $\mu_i(S_{-i}(z)|h^0) = 1$. Then, $\zeta(S^\infty_\Delta) \subseteq \zeta(S^\infty)$.*

Also Selective Rationalizability eventually saves only strategies that are sequential best replies under beliefs in the restricted set. Therefore, for path restrictions, A0 in Lemma 1 holds for Selective Rationalizability and Strong-Δ-Rationalizability regardless of the roles assigned to the two procedures. Then, the outcome equivalence of the two procedures under path restrictions obtains.

Theorem 4 *Fix path restrictions $(\Delta_i)_{i\in I}$ with support $\overline{Z}$. Then $\zeta(S^\infty_\Delta) = \zeta(S^\infty_{R\Delta})$.*

Proof. I show that $\zeta(S^\infty_\Delta) \subseteq \zeta(S^\infty_{R\Delta})$ — the proof of the opposite inclusion is identical. If $S^\infty_\Delta = \emptyset$, it is trivially true, so suppose $S^\infty_\Delta \neq \emptyset$. This implies that for each $z \in \overline{Z}$, $z \in \zeta(S^\infty_\Delta)$. For each $i \in I$, and $s_i \in S^\infty_{i,\Delta}$, $s_i \in \rho_i(\overline{\mu}_i)$ for some $\overline{\mu}_i \in \Delta_i$. For each μ_i with $\mu_i(S_{-i}(z')|h^0) = \overline{\mu}_i(S_{-i}(z')|h^0)$ for all $z' \in \zeta(S^\infty_\Delta)$, we have $\mu_i(S_{-i}(z)|h^0) = \overline{\mu}_i(S_{-i}(z)|h^0)$ for each $z \in \overline{Z}$, because $z \in \zeta(S^\infty_\Delta)$. Then, $\mu_i \in \Delta_i$. So, if μ_i strongly believes $(S^q_{-i,\Delta})^n_{q=0}$, $\rho_i(\mu_i) \subseteq S^{n+1}_{i,\Delta}$, and if μ_i strongly believes $(S^q_{-i,R\Delta})^n_{q=0}$, $\rho_i(\mu_i) \subseteq S^{n+1}_{i,\Delta}$. Thus, A0 of Lemma 1 is satisfied for both Strong-Δ-Rationalizability and Selective Rationalizability. Hence, $\zeta(S^\infty_\Delta) \subseteq \zeta(S^\infty_{R\Delta})$. ■

Corollary 5 *Fix $z \in Z$. Let Δ_i be the set of all $\mu_i \in \Delta^H(S_{-i})$ such that $\mu_i(S_{-i}(z)|h^0) = 1$. Then $\zeta(S^\infty_\Delta) = \zeta(S^\infty_{R\Delta})$.*

Finally, I provide an example of Strong Rationalizability and Strong-Δ-Rationalizability, and I illustrate the rough intuition for their outcome inclusion under path restrictions. Consider the following game.

$A\backslash B$	W	E
N	$2,2$	$\cdot-$
S	$0,0$	$2,2$

$\rightarrow$

$A\backslash B$	L	C	R
U	$1,1$	$1,0$	$0,0$
M	$0,0$	$0,1$	$1,0$
D	$0,0$	$0,0$	$0,3$

Strong Rationalizability goes as follows. At the first step, Ann eliminates $N.D$. At the second step, Bob eliminates $E.R$. At the third step, Ann eliminates $N.M$. At the fourth step, Bob eliminates $E.C$. The final output is $S^{\infty} = (S, N.U) \times (W, E.L)$.

Now, let Δ_i be the set of CPS's that initially believe in opponents' strategies that follow the path $z := (N, W)$:

$$\Delta_i := \left\{ \mu_i \in \Delta_i^H(S_{-i}) : \mu_i(S_{-i}(z)|h^0) = 1 \right\}, \ i = A, B.$$

Strong-Δ-Rationalizability goes as follows. At the first step, Ann eliminates S and $N.D$, and Bob eliminates $E.L$ and $E.C$. At the second step, Ann eliminates $N.U$ and Bob eliminates $E.R$. The final output is: $S_{\Delta}^{\infty} = \{(N.M, W)\}$.

Note that $\zeta(S_{\Delta}^{\infty}) = \{z\} \subseteq \zeta(S^{\infty})$, although $S_{\Delta}^{\infty} \cap S^{\infty} = \emptyset$. Why is it the case? At every step of Strong Rationalizability, Ann can play N as long as Bob may play W. For Bob, it is a bit more complicated. To play W, he needs at the same time to believe in N and have a belief for the subgame that deters the deviation. This is far from guaranteed, precisely because Strong Rationalizability and Strong-Δ-Rationalizability depart off the strongly-Δ-rationalizable path(s): the former ends up predicting (U, L), the latter M for Ann and R as the last action to be eliminated for Bob. But note two things. First: Ann can play N while believing in W and thus being surprised by E. So, after E, she must come up with a new belief, and then, at each step of reasoning, she can combine N with any action that can be justified after E. Second: for Bob to abandon W at a step of reasoning, he must be willing to deviate to E for every belief after E he can associate with the belief in N. By the argument about Ann, he can associate *any* belief after E compatible with step n with the belief in N. So, if W is abandoned at step $n+1$, the set of action profiles $S^n((N, E))|(N, E)$ must feature all the best replies of both players to beliefs over the set. This allows to refine the set by iteratively eliminating dominated actions and find a non-empty best response set. Now, by the absence of off-path restrictions and by the same combination arguments, this best response set would survive and keep supporting E throughout Strong-Δ-Rationalizability. But E does *not* survive Strong-Δ-Rationalizability. This tells us that Bob cannot abandon

W in Strong Rationalizability.

The main challenge for the proof of the workhorse lemma is that the subgame that follows the supposed deviation may not be a static game. But then, we can generate an *extensive form* best response set (Battigalli and Friedenberg [5]) in the subgame that follows the deviation, and prove that its paths should survive the truncation of the other elimination procedure in the subgame. The proof uses the workhorse lemma itself for smaller games as sketched at the end of the previous section. Other issues, such as the multiplicity of the possible deviations (and of the subgames that may follow) and of the paths induced by the first procedure, complicate the exposition but do not present interesting challenges.

5 Order independence and backward induction

In absence of belief restrictions, that is, for unconstrained elimination procedures, A0 trivially holds. An unconstrained elimination procedure is what I referred to in the Introduction as an order of iterated elimination of never sequential best replies. Thus, using Lemma 1 in both directions with Strong Rationalizability and any other unconstrained elimination procedure, the order independence of iterated elimination of never sequential best replies in terms of predicted outcomes obtains.

Theorem 6 *For any unconstrained elimination procedure* $((S_{i,q})_{i\in I})_{q=0}^{\infty}$, $\zeta(S_\infty) = \zeta(S^\infty)$.

Proof. Any two unconstrained elimination procedures, taken in both orders, obviously satisfy A0. The results follows then from Lemma 1. ■

In games with observable actions, the well-known backward induction procedure for games with perfect information has been generalized as follows. Starting from the bottom of game, an action of a player at a history is eliminated when it is not "folding-back optimal" against any conjecture over the surviving actions of the opponents at the same history and at the future histories. Penta [18] has translated backward induction for games with observable actions in the language of extensive-form rationalizability, i.e., as a procedure of elimination of strategies that are not sequentially optimal for any appropriate conditional probability system. Penta's Backwards Extensive-Form Rationalizability is simplified here for games with complete information.

Definition 8 *Backwards Extensive-Form Rationalizability is a sequence* $((S^q_{i,BR})_{i\in I})_{q=0}^{\infty}$ *where, for every* $i \in I$,

BR1 $S^0_{i,BR} = S_i$;

BR2 for each $n \in \mathbb{N}$ *and* $s_i \in S_i$, $s_i \in S_{i,BR}^n$ *if and only if there exists* $\mu_i \in \Delta^H(S_{-i})$ *such that, for each* $h \in H$,

(i) *there is* $\widetilde{s}_i \in \widehat{r}_i(\mu_i, \widetilde{h})$ *such that* $\widetilde{s}_i|h = s_i|h$;

(ii) *for each* $\widetilde{s}_{-i} \in S_{-i}(h)$ *with* $\mu_i(\widetilde{s}_{-i}|h) > 0$, *there is* $s_{-i} \in S_{-i,BR}^{n-1}$ *such that* $\widetilde{s}_{-i}|h = s_{-i}|h$.

Condition BR2.(ii) does not make a distinction between strategies in $S_{-i,BR}^{n-1}$ that allow h or not; for this reason, forward induction is not captured. Condition BR2.(i) requires s_i to be a continuation best reply not only at each $h \in H(s_i)$, as for sequential best replies of Definition 2, but also at each $h \notin H(s_i)$. This also induces to refine players' strategies further at histories that are no more allowed by some player. Nonetheless, Backwards Extensive-Form Rationalizability is outcome-equivalent to an *unfinished*, unconstrained elimination procedure.

Lemma 2 *Let* N *be the smallest* n *such that* $S_{BR}^n = S_{BR}^{n+1}$. *There exists an unconstrained elimination procedure* $((S_{i,q})_{i \in I})_{q=0}^{\infty}$ *such that for each* $n \leq N$,

$$S_n = \left\{ s \in S : \exists s' \in S_{BR}^n, \forall h \in H(S_{BR}^n), s(h) = s'(h) \right\}.$$

Hence, Backwards Extensive Rationalizability predicts a superset of the outcomes predicted by Strong Rationalizability.

Theorem 7 *Every strongly rationalizable outcome is a backwards extensive-form rationalizable outcome:* $\zeta(S^\infty) \subseteq \zeta(S_{BR}^\infty)$.

Proof. Immediate from Lemma 2 and Theorem 6. ■

In perfect information games without relevant ties, the backward induction outcome is unique. Thus, the following obtains.

Corollary 8 (Battigalli, [3]) *In every perfect information game without relevant ties, Strong Rationalizability and backward induction yield the same unique outcome.*

6 Proof of the workhorse lemma

The proof of Lemma 1 is by induction. Let $\overline{Z} := \zeta(\overline{S}^h_\infty)$. The induction hypothesis claims that each $S^h_{i,n}$ contains strategies that imitate those in $\overline{S}^h_{i,\infty}$ along the $\overline{Z}$ paths. This implies the desired outcome inclusion, and it also allows to construct beliefs for step $n+1$ that satisfy (1): for each $\overline{s}^h_i \in \overline{S}^h_{i,\infty}$, for each $\widetilde{h} \in H(\overline{S}^h_\infty)$, one can substitute in the supports of $\overline{\mu}^h_i(\overline{s}^h_i)(\cdot|\widetilde{h})$ the strategies in $\overline{S}^h_{-i,\infty}$ with their imitations in $S^h_{-i,n}$, and complete the new μ^h_i as to strongly believe $(S^h_{-i,q})^n_{q=0}$. By A0, $\rho_i(\mu^h_i) \subseteq S^h_{i,n+1}$. If player i has no incentive to move out of $\overline{Z}$ under μ^h_i, there is a sequential best reply to it that imitates $\overline{s}^h_i$ along $\overline{Z}$: at any $\widetilde{h} \in H(\overline{S}^h_\infty)$, with any $s_i \in S_i(\widetilde{h})$ such that $\zeta(\{s_i\} \times \overline{S}^h_{-i,\infty}(\widetilde{h})) \subseteq \overline{Z}$, $\mu^h_i(\cdot|\widetilde{h})$ and $\overline{\mu}^h_i(\overline{s}^h_i)(\cdot|\widetilde{h})$ induce the same outcome distribution, hence justify the same moves if other strategies are suboptimal. But there is no guarantee of this. Before tackling this fundamental issue, let us formalize the induction hypothesis. It will come in handy to formalize it in terms also of the beliefs that mimick the $\overline{\mu}^h_i(\overline{s}^h_i)$'s along $\overline{Z}$, beside the strategies that imitate the $\overline{s}^h_i$'s. To shorten the formulation of these concepts, I introduce some additional notation.

For any $\widehat{h} \succeq h$, $(s^h_j)_{j\in I} \in S^h$, $(s^{\widehat{h}}_j)_{j\in I} \in S^{\widehat{h}}$, $\mu^h_i \in \Delta^{H^h}(S^h_{-i})$, $\mu^{\widehat{h}}_i \in \Delta^{H^{\widehat{h}}}(S^{\widehat{h}}_{-i})$, $\widehat{Z} \subseteq Z^{\widehat{h}}$, and $J \subseteq I$, let:

- $s^h_J =^{\widehat{Z}} s^{\widehat{h}}_J$ if for each $z \in \widehat{Z}$ and $\widetilde{h}$ with $\widehat{h} \preceq \widetilde{h} \prec z$, $s^h_J(\widetilde{h}) = s^{\widehat{h}}_J(\widetilde{h})$;
- $\mu^h_i =^{\widehat{Z}} \mu^{\widehat{h}}_i$ if for each $z \in \widehat{Z}$ and $\widetilde{h}$ with $\widehat{h} \preceq \widetilde{h} \prec z$, $\mu^h_i(S^h_{-i}(z)|\widetilde{h}) = \mu^{\widehat{h}}_i(S^{\widehat{h}}_{-i}(z)|\widetilde{h})$;
- $s^h_J =^{\widehat{h}} s^{\widehat{h}}_J$ and $\mu^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$ if, respectively, $s^h_J =^{Z^{\widehat{h}}} s^{\widehat{h}}_J$ and $\mu^h_i =^{Z^{\widehat{h}}} \mu^{\widehat{h}}_i$.

Induction Hypothesis (n): for each $i \in I$, there exist maps $\widehat{\mu}^h_i : \overline{S}^h_{i,\infty} \to \Delta^{H^h}(S^h_{-i})$ and $\widehat{s}^h_i : \overline{S}^h_{i,\infty} \to S^h_i$ such that for each $\overline{s}^h_i \in \overline{S}^h_{i,\infty}$:

IH1. $\widehat{\mu}^h_i(\overline{s}^h_i)$ strongly believes $(S^h_{-i,q})^{n-1}_{q=0}$, and $\widehat{\mu}^h_i(\overline{s}^h_i) =^{\overline{Z}} \overline{\mu}^h_i(\overline{s}^h_i)$ (i.e., $\widehat{\mu}^h_i(\overline{s}^h_i)$ satisfies (1));

IH2. $\widehat{s}^h_i(\overline{s}^h_i) =^{\overline{Z}} \overline{s}^h_i$ and $\widehat{s}^h_i(\overline{s}^h_i) \in \rho_i(\widehat{\mu}^h_i(\overline{s}^h_i))$ (so, by A0, $\widehat{s}^h_i(\overline{s}^h_i) \in S^h_{i,n}$)).

Basis step (1): for all $i \in I$, the Induction Hypothesis holds with $\widehat{\mu}^h_i(\cdot) = \overline{\mu}^h_i(\cdot)$ and $\widehat{s}^h_i(\cdot)$ the identity map.

As anticipated, it is always possible to construct a map $\widehat{\mu}^h_i$ that satisfies IH1 at step $n+1$ by doing, for each $\overline{s}^h_i \in \overline{S}^h_{i,\infty}$ and at each $h \in H(\overline{S}^h_\infty)$, the pushforward of $\overline{\mu}^h_i(\overline{s}^h_i)(\cdot|h)$ through the map $\times_{j\neq i}\widehat{s}^h_j$ constructed at step n. The problem could be that, for some $l \in I$ and some $\overline{s}^h_l \in \overline{S}^h_{l,\infty}$, every μ^h_l that satisfies IH1 at step $n+1$ does not justify a strategy that imitates $\overline{s}^h_l$ along $\overline{Z}$.

Negation of the induction hypothesis at step n+1:

NIH. there exist $l \in I$ and $\overline{s}_l^h \in \overline{S}_{l,\infty}^h$ such that for every $\mu_l^h =^{\overline{Z}} \overline{\mu}_l^h(\overline{s}_l^h)$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$, there is no $s_l^h \in \rho_l(\mu_l^h)$ with $s_l^h =^{\overline{Z}} \overline{s}_l^h$.[12]

I am going to claim that, if this was the case, there would be a unilateral deviation by player l out of $\overline{Z}$ with the following property: every belief over the reactions of the opponents compatible with step n is also induced by a CPS that mimicks $\overline{\mu}_l^h(\overline{s}_l^h)$ along $\overline{Z}$ and incentivizes player l to do that particular deviation. From here, I will eventually arrive to the conclusion that some reactions that justify the deviation should have survived throughout $((\overline{S}_{i,q}^h)_{i\in I})_{q=0}^\infty$ as well, a contradiction.

Additional notation is needed. Let

$$D_l := \{\widetilde{h} \in H(\overline{S}_{-i}^h) \backslash H(\overline{S}_\infty^h) : p(\widetilde{h}) \in H(\overline{S}^h)\}$$

be the histories that immediately follow a unilateral deviation by player l from the paths. For every $\widehat{h} \in D_l$ and $m \in \mathbb{N}$, call $M_m^{\widehat{h}}$ (resp., $\overline{M}_m^{\widehat{h}}$) the set of all $\mu_l^{\widehat{h}}$ that strongly believe $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^m$ (resp., $(\overline{S}_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^m$) with the following property: there exists $\overline{\mu}_l^{\widehat{h}}$ that strongly believes $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$ such that the initial expected payoff of player l under $\overline{\mu}_l^{\widehat{h}}$ (i.e., under $\overline{\mu}_l^{\widehat{h}}(\cdot|\widehat{h})$) is not higher than under $\mu_l^{\widehat{h}}$ (i.e., under $\mu_l^{\widehat{h}}(\cdot|\widehat{h})$).[13]

Claim 1 *There exists $\widehat{h} \in D_l$ such that:*

C1. for every $m \leq n$ and $\mu_l^{\widehat{h}} \in M_m^{\widehat{h}}$, there exists $\mu_l^h =^{\overline{Z}} \overline{\mu}_l^h(\overline{s}_l^h)$ that strongly believes $(S_{-l,q}^h)_{q=0}^m$ such that $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$ and $\rho_l(\mu_l^h)(\widehat{h}) \neq \emptyset$;

C2. for every $p \in \mathbb{N}$ and $\mu_l^{\widehat{h}} \in \overline{M}_p^{\widehat{h}}$, there exists $\mu_l^h =^{\overline{Z}} \overline{\mu}_l^h(\overline{s}_l^h)$ that strongly believes $(\overline{S}_{-l,q}^h)_{q=0}^p$ such that $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$ and $\rho_l(\mu_l^h)(\widehat{h}) \neq \emptyset$.[14]

C1 with $m = n$ is the result described in words above. C1 also extends the claim to the previous steps of reasoning, focusing on the beliefs about the reactions that are at least as optimistic as those of step $n+1$. C2 extends the claim to the other procedure and all steps of reasoning. It will become clear later why these additional claims are needed.

The proof of Claim 1, deferred to the appendix, is by contraposition. I illustrate it for C1 with $m = n$; it can be easily extended to all other cases. Suppose that for every $\widehat{h} \in D_l$, there is a belief over $S_{-l,n}^h(\widehat{h})|\widehat{h}$ for which there is no CPS in $\Gamma(h)$ that (i) induces it after $\widehat{h}$,

[12] Note that, to be rigorous, no $\mu_l^h =^{\overline{Z}} \overline{\mu}_l^h(\overline{s}_l^h)$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$ is assumed to exist yet.

[13] Note: $\overline{\mu}_l^{\widehat{h}}$ strongly believes $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$ also when $\mu_l^{\widehat{h}}$ strongly believes $(\overline{S}_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^m$.

[14] Since $\widehat{h} \notin H(\overline{S}_\infty^h)$, the statement must hold vacously for some $p \in \mathbb{N}$ (i.e. $\overline{M}_p^{\widehat{h}} = \emptyset$).

(ii) strongly believes $(S^h_{-l,q})^n_{q=0}$, (iii) mimicks $\overline{\mu}^h_l(\overline{s}^h_l)$ along $\overline{Z}$, and (iv) incentivizes player l to deviate towards $\widehat{h}$.[15] But all such beliefs *can* be induced by the same CPS μ^h_l that strongly believes $(S^h_{-l,q})^n_{q=0}$ and mimicks $\overline{\mu}^h_l(\overline{s}^h_l)$ along $\overline{Z}$, thus satisfying (i)-(ii)-(iii). So, μ^h_l violates (iv) for every $\widehat{h} \in D_l$. But then, under μ^h_l player l has no incentive to do any deviation from $\overline{Z}$, and so there is a sequential best reply to μ^h_l that imitates $\overline{s}^h_l$ along $\overline{Z}$, contradicting NIH. The proof that such μ^h_l can be constructed is based on the following idea. By the induction hypothesis, for any player $i \neq l$, the strategies in $S^h_{i,n}$ that imitate those in $\overline{S}^h_{i,\infty}$ along $\overline{Z}$ are sequential best replies to CPS's that assign probability 0 to deviations by the opponents from $\overline{Z}$ until they happen. Being surprised by each deviation, player i must come up with a new belief afterwards. For each $\widehat{h} \in D_l$, these new beliefs can justify the strategies in $S^h_{-l,n}(\widehat{h})|\widehat{h}$ that player l has to believe in. Thus, there are strategies in $S^h_{-l,n}$ that imitate those in $\overline{S}^h_{-l,\infty}$ along $\overline{Z}$ and react to player l's deviation in any way that is compatible with step n. These strategies support the required combination of beliefs.

Now it is time to use the negation of the induction hypothesis and Claim 1 to arrive to a contradiction and thus prove the lemma. C1 with $m = n$ and A0 imply that $S^h_{l,n+1}(\widehat{h})|\widehat{h}$ contains all the sequential best replies to all the beliefs in $M^{\widehat{h}}_n$, i.e., to all $\mu^{\widehat{h}}_l$ that strongly believe $(S^h_{-l,q}(\widehat{h})|\widehat{h})^n_{q=0}$. The same holds for $S^h_{i,n}(\widehat{h})|\widehat{h}$ with $i \neq l$, because by the induction hypothesis player i can allow $\widehat{h}$ while assigning probability 0 to reaching it until it is actually reached, and then she can come up with any new belief and best reply to it. So, for each player i, $S^h_{i,n}(\widehat{h})|\widehat{h}$ contains all the sequential best replies to CPS's that strongly believe $(S^h_{-i,q}(\widehat{h})|\widehat{h})^n_{q=0}$. Then, we can refine $S^h_n(\widehat{h})|\widehat{h}$ with an iterated deletion of never sequential best replies (from n onwards) and obtain an elimination procedure with non-empty output. Consider now the truncation of $((\overline{S}^h_{i,q})_{i\in I})^\infty_{q=0}$ after $\widehat{h}$. By Remark 3 it is an elimination procedure as well. If we conclude that it yields a non-empty output, we contradict that $\widehat{h}$ is a deviation from $\overline{Z}$. How to do that? By showing two things. First, that the lemma holds in games of smaller depth. This can be assumed by induction, because the lemma obviously holds in static games. (The formal argument will be that $\widehat{h}$ cannot exist in static games.) Second, the two elimination procedures in $\Gamma(\widehat{h})$, *with inverted roles* with respect to those in $\Gamma(h)$ that originated them, satisfy A0. Here is where the rest of Claim 1 kicks in. The whole argument is now formalized.

Proof that the negation of the induction hypothesis leads to contradiction.

If $\Gamma(h)$ has depth 1, $D_l(\overline{S}^h_\infty) = \emptyset$, so the existence of $\widehat{h}$ by Claim 1 is already a contradiction. This allows to assume by way of induction that Lemma 1 holds in games of smaller depth than $\Gamma(h)$, thus in $\Gamma(\widehat{h})$. Define $((\overline{S}^{\widehat{h}}_{i,q})_{i\in I})_{q\geq 0}$ as follows: for every $i \in I$ and $m \leq n$,

[15] In presence of probabilistic beliefs along the paths, player l can be unsure as to which $h \in D_l$ will realize after the deviation, but this is immaterial for the argument.

$\overline{S}^{\widehat{h}}_{i,m} = S^h_{i,m}(\widehat{h})|\widehat{h}$; for every $m > n$, $s^{\widehat{h}}_i \in \overline{S}^{\widehat{h}}_{i,m}$ if and only if there exists $\mu^{\widehat{h}}_i$ that strongly believes $(\overline{S}^{\widehat{h}}_{-i,q})^{m-1}_{q=0}$ such that $s^{\widehat{h}}_i \in \rho_i(\mu^{\widehat{h}}_i)$.

For every $i \neq l$, since $\widehat{h} \in D_l(\overline{S}^h_\infty)$, $\overline{S}^h_{i,\infty}(\widehat{h}) \neq \emptyset$. So, fix $\overline{s}^h_i \in \overline{S}^h_{i,\infty}(\widehat{h})$. For every $m \leq n$, the Induction Hypothesis provides $\widehat{s}^h_i(\overline{s}^h_i) \in S^h_{i,m}(\widehat{h})$ and $\widehat{\mu}^h_i(\overline{s}^h_i)$ that strongly believes $(S^h_{-i,q})^{m-1}_{q=0}$. Note that $\widehat{\mu}^h_i(\overline{s}^h_i) =^{\overline{Z}} \overline{\mu}^h_i(\overline{s}^h_i)$ implies $\widehat{\mu}^h_i(\overline{s}^h_i)(S^h_{-i}(\widehat{h})|p(\widehat{h})) = 0$. Hence, for every $\mu^{\widehat{h}}_i$ that strongly believes $(\overline{S}^{\widehat{h}}_{-i,q})^{m-1}_{q=0}$, I can construct $\mu^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$ that strongly believes $(S^h_{-i,q})^{m-1}_{q=0}$ such that $\mu^h_i(\cdot|\widetilde{h}) = \widehat{\mu}^h_i(\overline{s}^h_i)(\cdot|\widetilde{h})$ for all $\widetilde{h} \not\geq \widehat{h}$.[16] Thus, $\rho_i(\mu^h_i)(\widehat{h}) \neq \emptyset$, and by $\mu^h_i =^{\overline{Z}} \overline{\mu}^h_i(\overline{s}^h_i)$ and A0, $\rho_i(\mu^h_i) \subseteq S^h_{i,m}$. So, $\rho_i(\mu^{\widehat{h}}_i) \subseteq \overline{S}^{\widehat{h}}_{i,m} \neq \emptyset$.

Fix $\mu^{\widehat{h}}_l$ that strongly believes $(\overline{S}^{\widehat{h}}_{-l,q})^n_{q=0}$; trivially, $\mu^{\widehat{h}}_l \in M^{\widehat{h}}_n$. Hence, by C1, there exists $\mu^h_l =^{\overline{Z}} \overline{\mu}^h_l(\overline{s}^h_l)$ that strongly believes $(S^h_{-l,n})^n_{q=0}$ such that $\mu^h_l =^{\widehat{h}} \mu^{\widehat{h}}_l$ and $\rho_l(\mu^h_l)(\widehat{h}) \neq \emptyset$. By A0, $\rho_l(\mu^h_l) \subseteq S^h_{l,n}$. So, $\rho_l(\mu^{\widehat{h}}_l) \subseteq \overline{S}^{\widehat{h}}_{l,n+1} \subseteq \overline{S}^{\widehat{h}}_{l,n} \neq \emptyset$.

Then, for every $i \in I$ and $\mu^{\widehat{h}}_i$ that strongly believes $(\overline{S}^{\widehat{h}}_{-i,q})^n_{q=0}$,[17] $\rho_i(\mu^{\widehat{h}}_i) \subseteq \overline{S}^{\widehat{h}}_{i,n} \neq \emptyset$. So, $\overline{S}^{\widehat{h}}_{i,n} \supseteq \overline{S}^{\widehat{h}}_{i,n+1}$, and $((\overline{S}^{\widehat{h}}_{i,q})_{i\in I})_{q\geq 0}$ is an elimination procedure with $\overline{S}^{\widehat{h}}_\infty \neq \emptyset$.

Fix $m \leq n$, $\overline{\mu}^{\widehat{h}}_l$ that strongly believes $(\overline{S}^{\widehat{h}}_{-l,q})^\infty_{q=0}$, and $\mu^{\widehat{h}}_l =^{\zeta(\overline{S}^{\widehat{h}}_\infty)} \overline{\mu}^{\widehat{h}}_l$ that strongly believes $(\overline{S}^{\widehat{h}}_{-l,q})^{m-1}_{q=0}$. Since (i) $\rho_l(\overline{\mu}^{\widehat{h}}_l) \times \overline{S}^{\widehat{h}}_{-l,\infty} \subseteq \overline{S}^{\widehat{h}}_\infty$, (ii) $\overline{\mu}^{\widehat{h}}_l$ strongly believes $\overline{S}^{\widehat{h}}_{-l,\infty}$, and (iii) $\mu^{\widehat{h}}_l =^{\zeta(\overline{S}^{\widehat{h}}_\infty)} \overline{\mu}^{\widehat{h}}_l$, player l initially expects a non lower payoff under $\mu^{\widehat{h}}_l$ than under $\overline{\mu}^{\widehat{h}}_l$. So, since $\overline{\mu}^{\widehat{h}}_l$ strongly believes $(\overline{S}^{\widehat{h}}_{-l,q})^n_{q=0} = (S^h_{-l,q}(\widehat{h})|\widehat{h})^n_{q=0}$, $\mu^{\widehat{h}}_l \in M^{\widehat{h}}_m$. Thus, by C1 there exists $\mu^h_l =^{\overline{Z}} \overline{\mu}^h_l(\overline{s}^h_l)$ that strongly believes $(S^h_{-l,q})^{m-1}_{q=0}$ such that $\mu^h_l =^{\widehat{h}} \mu^{\widehat{h}}_l$ and $\rho_l(\mu^h_l)(\widehat{h}) \neq \emptyset$. By A0, $\rho_l(\mu^h_l) \subseteq S^h_{l,m}$. So $\rho_l(\mu^{\widehat{h}}_l) \subseteq \overline{S}^{\widehat{h}}_{l,m}$.

Then, for every $m \in \mathbb{N}$, $i \in I$, $\overline{\mu}^{\widehat{h}}_i$ that strongly believes $(\overline{S}^{\widehat{h}}_{-i,q})^\infty_{q=0}$ and $\mu^{\widehat{h}}_i =^{\zeta(\overline{S}^{\widehat{h}}_\infty)} \overline{\mu}^{\widehat{h}}_i$ that strongly believes $(\overline{S}^{\widehat{h}}_{-i,q})^{m-1}_{q=0}$, $\rho_i(\mu^{\widehat{h}}_i) \subseteq \overline{S}^{\widehat{h}}_{i,m}$. Thus, $((\overline{S}^{\widehat{h}}_{i,q})_{i\in I})_{q\geq 0}$ satisfies A0.

Define now $((S^{\widehat{h}}_{i,q})_{i\in I})_{q\geq 0}$ as $((\overline{S}^h_{i,q}(\widehat{h})|\widehat{h})_{i\in I})_{q\geq 0}$. By Remark 1 it is an elimination procedure.

Fix $i \neq l$ and $m \in \mathbb{N}$. Since $\overline{\mu}^h_i(\overline{s}^h_i)$ strongly believes $\overline{S}^h_{-i,\infty}$, we have $\overline{\mu}^h_i(\overline{s}^h_i)(S^h_{-i}(\widehat{h})|p(\widehat{h})) = 0$. Hence, for every $\mu^{\widehat{h}}_i$ that strongly believes $(S^{\widehat{h}}_{-i,q})^{m-1}_{q=0}$, I can construct $\mu^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$ that strongly believes $(\overline{S}^h_{-i,q})^{m-1}_{q=0}$ such that $\mu^h_i(\cdot|\widetilde{h}) = \overline{\mu}^h_i(\overline{s}^h_i)(\cdot|\widetilde{h})$ for all $\widetilde{h} \not\geq \widehat{h}$. Thus, $\rho_i(\mu^h_i)(\widehat{h}) \neq \emptyset$, and by $\mu^h_i =^{\overline{Z}} \overline{\mu}^h_i(\overline{s}^h_i)$ and A0, $\rho_i(\mu^h_i) \subseteq \overline{S}^h_{i,m}$. So, $\rho_i(\mu^{\widehat{h}}_i) \subseteq S^{\widehat{h}}_{i,m} \neq \emptyset$.

For every $m \in \mathbb{N}$, $\overline{\mu}^{\widehat{h}}_l$ that strongly believes $(\overline{S}^{\widehat{h}}_{-l,q})^\infty_{q=0}$, and $\mu^{\widehat{h}}_l =^{\zeta(\overline{S}^{\widehat{h}}_\infty)} \overline{\mu}^{\widehat{h}}_l$ that strongly believes $(S^{\widehat{h}}_{-l,q})^{m-1}_{q=0}$, by the same argument as above, $\mu^{\widehat{h}}_l \in \overline{M}^{\widehat{h}}_m$. Thus, by C2 there exists $\mu^h_l =^{\overline{Z}} \overline{\mu}^h_l(\overline{s}^h_l)$ that strongly believes $(\overline{S}^h_{-l,q})^{m-1}_{q=0}$ such that $\mu^h_l =^{\widehat{h}} \mu^{\widehat{h}}_l$ and $\rho_l(\mu^h_l)(\widehat{h}) \neq \emptyset$. By

[16] The construction is shown explicitly by Lemma 3 in the Appendix.

[17] For $i \neq l$, observe that strong belief in $(\overline{S}^{\widehat{h}}_{-i,q})^n_{q=0}$ trivially implies strong belief in $(\overline{S}^{\widehat{h}}_{-i,q})^{n-1}_{q=0}$, the condition used above.

A0, $\rho_l(\mu_l^h) \subseteq \overline{S}_{l,m}^h$. So, $\rho_l(\mu_l^{\widehat{h}}) \subseteq S_{l,m}^{\widehat{h}} \neq \emptyset$.

Then, for every $m \in \mathbb{N}$, $i \in I$, $\overline{\mu}_i^{\widehat{h}}$ that strongly believes $(\overline{S}_{-i,q}^{\widehat{h}})_{q=0}^{\infty}$ and $\mu_i^{\widehat{h}} =^{\zeta(\overline{S}_{\infty}^{\widehat{h}})} \overline{\mu}_i^{\widehat{h}}$ that strongly believes $(S_{-i,q}^{\widehat{h}})_{q=0}^{m-1}$, $\rho_i(\mu_i^{\widehat{h}}) \subseteq S_{i,m}^{\widehat{h}}$. Thus, $((S_{i,q}^{\widehat{h}})_{i\in I})_{q\geq 0}$ satisfies A0. As argued, Lemma 1 holds in $\Gamma(\widehat{h})$. Hence, $\zeta(S_{\infty}^{\widehat{h}}) \supseteq \zeta(\overline{S}_{\infty}^{\widehat{h}}) \neq \emptyset$. But this contradicts $\widehat{h} \in D_l(\overline{S}_{\infty}^h)$. ■

7 Discussion - Comparison with Perea ([19], [20])

To facilitate the comparison between my methodology and Perea's, I refer to the order independence problem tackled by Perea in [19].[18] The fundamental problem is the need to overcome the non-monotonicity of the strong belief, which implies that delaying the elimination of some strategy can provoke the elimination of another strategy that would have not been eliminated otherwise.

Consider two nested, Cartesian sets of strategy profiles, $\overline{S} = \times_{i\in I}\overline{S}_i \subset \widehat{S} = \times_{i\in I}\widehat{S}_i$. Fix a player $i \in I$ and consider the sets $\widehat{S}_i^*, \overline{S}_i^*$ of sequential best replies to CPS's that strongly believe, respectively, $\widehat{S}_{-i}$ and $\overline{S}_{-i}$. By non-monotonicity of strong belief, it needs not be the case that $\overline{S}_i^* \subset \widehat{S}_i^*$. However, for every CPS that strongly believes $\overline{S}_{-i}$, there is one that strongly believes $\widehat{S}_{-i}$ and is identical to the first at all histories compatible with $\overline{S}$: at such histories ($H(\overline{S})$), the first CPS has to give probability 1 to $\overline{S}_{-i}$, and we have $\overline{S}_{-i} \subset \widehat{S}_{-i}$. Then, for every $\overline{s}_i \in \overline{S}_i^*$, there is $s_i \in \widehat{S}_i^*$ that is identical to $\overline{s}_i$ at each $h \in H(\overline{S})$. So, if we focus on the paths induced by $\overline{S}$, $\widehat{S}^*$ must induce a (weakly) larger subset of them with respect to $\overline{S}^*$. If $\overline{S}$ has been obtained by iterated elimination of never sequential best replies, we have $\zeta(\overline{S}^*) \subseteq \zeta(\overline{S})$, and then $\zeta(\overline{S}^*) \subseteq \zeta(\widehat{S}^*)$ as well.

However, if one wants to iterate further and find the sequential best replies to CPS's that strongly believe $\widehat{S}_{-i}^*$ and $\overline{S}_{-i}^*$, there is the problem that these two sets are no more nested. So, let us restart with two sets $\overline{S}, \widehat{S}$ where the projection of $\overline{S}$ on $H(\overline{S})$ is smaller than that of $\widehat{S}$ (on $H(\overline{S})$ as well). Now, $\widehat{S}_{-i}$ may feature fewer reactions than $\overline{S}_{-i}$ to a deviation by player i from the paths induced by $\overline{S}$, and this may induce i to leave one of

[18] Perea formulates the problem as order independence of the iterated elimination under the strong belief *reduction* operator, i.e., the elimination at each step of some strategies that are not sequential best replies to CPS's that strongly believe in the opponents' strategies that survived the last step, without memory of the previous steps. This is probably the most interesting formulation of the problem, because slow reduction orders can be seen as heuristic procedures, while keeping memory of all steps makes more sense for the maximal elimination order, because it reflects common strong belief in rationality. Since the main object of this paper is monotonicity with respect to belief restrictions and not order independence, the requirement that the final strategies be sequential best replies to CPS's that strongly believe in all the previous steps has been kept for convenience. This makes an iterated deletion of never sequential best replies as defined in this paper not necessarily an order of elimination under the strong belief reduction operator, and vice versa, although both maximal elimination orders coincide with Strong Rationalizability. However, this subtle difference is immaterial for this discussion.

these paths under strong belief in $\widehat{S}_{-i}$, but not under strong belief in $\overline{S}_{-i}$. The challenge is proving that this possibility does not arise when $\overline{S}, \widehat{S}$ have been derived from the iterated elimination of never sequential best replies. Perea does it by restricting the definition of "*monotonicity on reachable histories*" to sets $\overline{S}, \widehat{S}$ where the projection of $\widehat{S}_i^*$ on $H(\overline{S})$ is smaller than that of $\overline{S}_i$,[19] which excludes the presence of such deviation moves. Then, he divides the order independence problem into a chain of pairwise comparisons between iterated eliminations $(\overline{S}_n)_{n\geq 0}, (\widehat{S}_n)_{n\geq 0}$ that are identical up to the step m, after which the first becomes maximal, and so does the second one step later. In this way, for each $n > m$, $\overline{S}_n$ and $\widehat{S}_n$, and with inverted roles $\overline{S}_n$ and $\widehat{S}_{n+1}$ are shown to satisfy the requirements. In this paper, I observe that if the deviation by player i depicted above was to arise, there would be a extensive-form best response set of the subgame that follows the deviation which justifies it; but then, the paths induced by this set and the deviation should have survived also the procedure that generated $\overline{S}$, a contradiction.

In my view, the approach of Perea is better suited than the approach of this paper for *order independence* problems, because it teaches why delaying some eliminations does not matter. The approach of this paper is inspired by, and designed for, *outcome monotonicity* problems, and it aims to shed light on their roots by comparing directly maximal elimination procedures under belief restrictions, showing when and why deviations from the paths induced by the more restrictive procedure cannot occur in the less restrictive one. The identification of sets of outcome distributions with the same support as class of restrictions that preserve outcome monotonicity was indeed triggered by this view and by the workhorse lemma that captures the main conceptual insight. However, the outcome equivalence of Strong-Δ-Rationalizability with Selective Rationalizability under such restrictions, which implies the outcome inclusion with Strong Rationalizability, can also be seen as an order independence problem (while the workhorse lemma in its generality *cannot*). Indeed, in finite games, Selective Rationalizability can be seen as a slow elimination order of strategies that are not sequential best replies under the belief restrictions, where the first steps coincide with Strong Rationalizability. So, focusing for simplicity on a single path z, one could:

- define a reduction operator ρ^z that takes a Cartesian set of strategy profiles $\overline{S} = \times_{i\in I}\overline{S}_i$ and, for each player i, returns the strategies in $\overline{S}_i$ that are sequential best replies to a CPS that strongly believes $\overline{S}_{-i}$ and initially believes in $S_{-i}(z)$;

- invoke Proposition 1 in Battigalli and Prestipino [6] to claim that Strong-Δ-Rationalizability

[19] Perea works with a *reduction* operator, thus not really with $\widehat{S}_i^*$ but with $\widehat{S}_i^* \cap \widehat{S}_i$, which in principle might be poorer than needed (or even empty). This is why Perea imposes in the definition of monotonocity under reachable histories that $\widehat{S}$ has been derived from an iterated application of the operator. For the strong belief operator, via combination arguments similar to those of this paper, this ensures the every projection of a strategy in $\widehat{S}_i$ on the paths induced by $\overline{S}$ can be associated in $\widehat{S}_i$ with off-path behavior that best replies to beliefs over $\widehat{S}_{-i}$.

coincides with the maximal elimination order under the ρ^z operator;[20]

- show that Selective Rationalizability coincides with an elimination order under the ρ^z operator that coincides with Strong Rationalizability until convergence, and then procedes at "full speed";[21]

- show that ρ^z is *monotone on reachable histories* (Perea [20]);

- invoke Theorem 3.2 in Perea [20] to claim the outcome equivalence of Selective Rationalizability and Strong-Δ-Rationalizability.

That ρ^z is monotone on reachable histories is so far only a conjecture, but I expect the proof to follow the same lines of the proof of Perea that the strong belief operator is monotone on reachable histories (Theorem 3.1 in [20]). Exploiting the existing results and proofs, the roadmap above would probably result into a shorter proof of the outcome equivalence and outcome monotonicity results under path restrictions.

8 Appendix

Proof of Remark 1. EP1 and EP2 are obvious. To prove EP3, note the following. For every $i \in I$ and $s_i^{\widehat{h}} \in S_{i,\infty}^h(\widehat{h})|\widehat{h}$, there exists $s_i^h \in S_{i,\infty}^h$ such that $s_i^h|\widehat{h} = s_i^{\widehat{h}}$. By EP3 for $((S_{i,q}^h)_{i\in I})_{q=0}^{\infty}$, there exists μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^{\infty}$ such that $s_i^h \in \rho_i(\mu_i^h) \subseteq S_{i,\infty}^h$. Thus, the pushforward $\mu_i^{\widehat{h}}$ of $(\mu_i^h(\cdot|\widetilde{h}))_{\widetilde{h}\in H^{\widehat{h}}}$ through the map $s_{-i}^h \mapsto s_{-i}^h|\widehat{h}$ strongly believes $(S_{-i,q}^h(\widehat{h})|\widehat{h})_{q=0}^{\infty}$. Clearly $s_i^{\widehat{h}} \in \rho_i(\mu_i^{\widehat{h}})$. Finally, fix $\overline{s}_i^{\widehat{h}} \in \rho_i(\mu_i^{\widehat{h}})$. Define $\overline{s}_i^h$ as $\overline{s}_i^h(\widetilde{h}) = s_i^h(\widetilde{h})$ for all $\widetilde{h} \not\succeq \widehat{h}$ and $\overline{s}_i^h|\widehat{h} = \overline{s}_i^{\widehat{h}}$ for all $\widetilde{h} \succeq \widehat{h}$. Clearly $\overline{s}_i^h \in \rho_i(\mu_i^h)$. Thus, $\overline{s}_i^{\widehat{h}} \in S_{i,\infty}^h(\widehat{h})|\widehat{h}$. ■

Proof of Lemma 2. Define $\left((S_{i,n})_{i\in I}\right)_{n=0}^{N}$ as in the statement of the lemma, and for each $n > N$ and $i \in I$, let $s_i \in S_{i,n}$ if and only if there exists μ_i that strongly believes $(S_{-i,q})_{q=0}^{n-1}$ such that $s_i \in \rho_i(\mu_i)$. It is immediate to see that $((S_{i,q})_{i\in I})_{q=0}^{\infty}$ is an elimination procedure. To show that it is unconstrained, fix $n \leq N$ and suppose by way of induction that for each $m < n$, $i \in I$, and μ_i that strongly believes $(S_{-i,q})_{q=0}^{m-1}$, we have $\rho_i(\mu_i) \subseteq S_{i,m}$ (it is vacously true for $m = 0$). Fix μ_i that strongly believes $(S_{-i,q})_{q=0}^{n-1}$. I show that $\rho_i(\mu_i) \subseteq S_{i,n}$. By definition of $S_{-i,n-1}$, I can construct μ_i' that satisfies BR2.(ii) such that for all $h \in H(S_{n-1})$ and $z \in \zeta(S_{n-1})$, $\mu_i(S_{-i}(z)|h) = \mu_i'(S_{-i}(z)|h)$. For each $s_i' \in \rho_i(\mu_i')$, there is a realization equivalent s_i'' that satisfies BR2.(i), so that $s_i'' \in S_{i,BR}^n \subseteq$

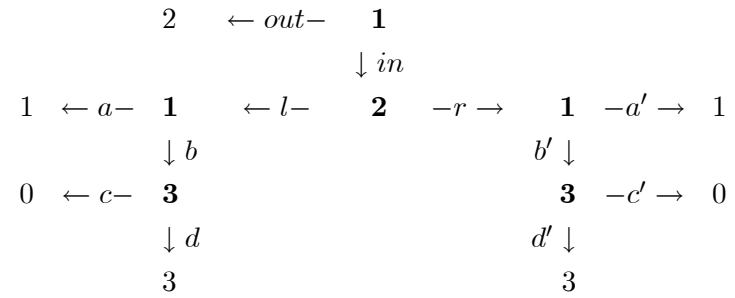
[20] It is easy to see that path restrictions are *closed under composition* (Battigalli and Prestipino [6]).

[21] Under the hypotheses of *strategic independence* (Battigalli [2]), or just *independent rationalization* (Catonini [9]), Strong-Δ-Rationalizability and Selective Rationalizability cannot be written as reduction procedures, not even under path restrictions. However, for path restrictions, outcome equivalences hold. A proof is available upon request.

$S^{n-1}_{i,BR}$. For each $s_i \in \rho_i(\mu_i)$, by the induction hypothesis we have $s_i \in S_{i,n-1}$. Then, we have $\zeta\left(\rho_i(\mu_i') \times S^{n-1}_{-i,BR}\right) \subseteq \zeta\left(S^{n-1}_{BR}\right)$, $\zeta\left(\rho_i(\mu_i) \times S_{-i,n-1}\right) \subseteq \zeta\left(S_{n-1}\right)$, and by construction $\zeta\left(S_{n-1}\right) = \zeta\left(S^{n-1}_{BR}\right)$. Thus, for each $s_i \in \rho_i(\mu_i)$, there is $s_i' \in \rho_i(\mu_i')$ such that $s_i(h) = s_i'(h)$ for all $h \in H\left(S_{n-1}\right)$. Since there is $s_i'' \in S^n_{i,BR}$ realization equivalent to s_i', so that $s_i''(h) = s_i'(h) = s_i(h)$ for all $h \in H\left(S_{n-1}\right) \cap H(s_i)$, by definition of $S_{i,n}$ we have $s_i \in S_{i,n}$. Thus, $\rho_i(\mu_i) \subseteq S_{i,n}$. ■

Proof of Claim 1.

The main challenges for the proof of Claim 1 derive from the following issue. Fix an elimination procedure $((S^h_{i,q})_{i\in I})_{q\geq 0}$ and a player $i \in I$. Consider two sequential best replies $\widehat{s}_i^h, \overline{s}_i^h$ to two different CPS's that strongly believe $S^h_{-i,n}, ..., S^h_{-i,0}$. Fix two unordered histories $\widehat{h}, \overline{h} \in H(\widehat{s}_i^h) \cap H(\overline{s}_i^h)$ (that is, $\overline{h} \not\succeq \widehat{h}$ and $\widehat{h} \not\succeq \overline{h}$). Is there always a CPS that strongly believes $S^h_{-i,n}, ..., S^h_{-i,0}$ and a sequential best reply to it $s_i^h \in S_i(\widehat{h}) \cap S_i(\overline{h})$ such that $s_i^h|\widehat{h} = \overline{s}_i^h|\widehat{h}$ and $s_i^h|\overline{h} = \widehat{s}_i^h|\overline{h}$? No. The reason is the following: Player i may allow $\widehat{h}$ and $\overline{h}$ either because she is confident that $\widehat{h}$ will be reached and she has optimistic beliefs after $\widehat{h}$, or because she is confident that $\overline{h}$ will be reached and she has optimistic beliefs after $\overline{h}$. If $\widehat{s}_i^h$ is optimal under the first conjecture and $\overline{s}_i^h$ is optimal under the second conjecture, $\widehat{s}_i^h|\overline{h}$ and $\overline{s}_i^h|\widehat{h}$ may be "emergency plans" for unpredicted contingencies, where the beliefs do not justify the choice to allow $\overline{h}$ and $\widehat{h}$ in the first place. This can be seen already from the set of justifiable strategies of a player. The following is a simplified version of the game in Figure 4 in Battigalli [3], provided by Gul and Reny. The payoffs are of player 1.



Player 1 will rationally plan $in.a.b'$ if she first expects r and d', and then c once she gets surprised by l. Similarly, player 1 can rationally plan $in.b.a'$. However, player 1 cannot rationally plan $in.a.a'$: the best payoff she can get is lower than the outside option. Actions a and a' are emergency plans for unforeseen contigencies, and best respond to beliefs that do not justify playing in in the first place.

But in the proof of Claim 1, I will combine a player's behavior along a set of expected paths with her reactions to opponents' deviations from those paths. Differently from the

example above, all these unforeseen contingencies are allowed by our player under the same rational plan of following those paths. Then, the combination is always possible. To begin, the first lemma formalizes the basic intuition that, after being surprised, a player can come up with any new belief, and thus can combine any possible reaction to the surprise with any plan she had if the surprise had not taken place.

Lemma 3 *Fix an elimination procedure* $((S^h_{i,q})_{i\in I})_{q\geq 0}$, $n\in\mathbb{N}$, $i\in I$, $\widehat{h}\in H^h$, *and* μ^h_i *that strongly believes* $(S^h_{-i,q})_{q=0}^{n-1}$ *such that* $\mu^h_i(S^h_{-i}(\widehat{h})|p(\widehat{h}))=0$. *Fix* $s^h_i\in\rho_i(\mu^h_i)\cap S^h_i(\widehat{h})$, $\mu^{\widehat{h}}_i$ *that strongly believes* $(S^h_{-i,q}(\widehat{h})|\widehat{h})_{q=0}^{n-1}$, *and* $s^{\widehat{h}}_i\in\rho_i(\mu^{\widehat{h}}_i)$.

Consider the unique $\widetilde{s}^h_i =^{\widehat{h}} s^{\widehat{h}}_i$ *such that for every* $\widetilde{h}\notin H^{\widehat{h}}$, $\widetilde{s}^h_i(\widetilde{h})=s^h_i(\widetilde{h})$.

There exists $\widetilde{\mu}^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$ *that strongly believes* $(S^h_{-i,q})_{q=0}^{n-1}$ *such that* $\widetilde{\mu}^h_i(\cdot|\widetilde{h})=\mu^h_i(\cdot|\widetilde{h})$ *for all* $\widetilde{h}\notin H^{\widehat{h}}$, *and* $\widetilde{s}^h_i\in\rho_i(\widetilde{\mu}^h_i)$ *(so,* $\rho_i(\mu^h_i)(\widehat{h})\neq\emptyset$ *implies* $\rho_i(\widetilde{\mu}^h_i)(\widehat{h})\neq\emptyset$*).*

Proof.

Fix a map $\varsigma: S^{\widehat{h}}_{-i}\to S^h_{-i}$ such that for each $s^{\widehat{h}}_{-i}\in S^{\widehat{h}}_{-i}$, $\varsigma(s^{\widehat{h}}_{-i}) =^{\widehat{h}} s^{\widehat{h}}_{-i}$ and $\varsigma(s^{\widehat{h}}_{-i})\in S^h_{-i,m}(\widehat{h})$ for all $m\geq 0$ with $s^{\widehat{h}}_{-i}\in S^h_{-i,m}(\widehat{h})|\widehat{h}$. Since ς is injective, we can construct an array of probability measures $\widetilde{\mu}^h_i=(\widetilde{\mu}^h_i(\cdot|\widetilde{h}))_{\widetilde{h}\in H^h}$ on S^h_{-i} as $\widetilde{\mu}^h_i(\cdot|\widetilde{h})=\mu^h_i(\cdot|\widetilde{h})$ for all $\widetilde{h}\notin H^{\widehat{h}}$ and $\widetilde{\mu}^h_i(\varsigma(s^{\widehat{h}}_{-i})|\widetilde{h})=\mu^{\widehat{h}}_i(s^{\widehat{h}}_{-i}|\widetilde{h})$ for all $\widetilde{h}\in H^{\widehat{h}}$ and $s^{\widehat{h}}_{-i}\in S^{\widehat{h}}_{-i}$. From the definition of ς, it immediately follows that $\widetilde{\mu}^h_i(S_{-i}(\widetilde{h})|\widetilde{h})=1$ for all $\widetilde{h}\in H^h$, that $\widetilde{\mu}^h_i$ strongly believes $(S^h_{-i,q})_{q=0}^{n-1}$, and that $\widetilde{\mu}^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$. Finally, since $\widetilde{\mu}^h_i(S_{-i}(\widehat{h})|p(\widehat{h}))=0$, $\widetilde{\mu}^h_i$ satisfies the chain rule.

Fix $\widetilde{h}\in H(\widetilde{s}^h_i)\backslash H^{\widehat{h}}=H(s^h_i)\backslash H^{\widehat{h}}$. If $\widetilde{h}\prec\widehat{h}$, by $\mu^h_i(S^h_{-i}(\widehat{h})|p(\widehat{h}))=0$ we have $\mu^h_i(S^h_{-i}(\widehat{h})|\widetilde{h})=0$, and for every $s^h_{-i}\notin S^h_{-i}(\widehat{h})$, $\zeta(s^h_i,s^h_{-i})=\zeta(\widetilde{s}^h_i,s^h_{-i})$. If $\widetilde{h}\not\prec\widehat{h}$, for every $s^h_{-i}\in S^h_{-i}(\widetilde{h})$, $\widehat{h}\notin H(s^h_i,s^h_{-i})$, so $\zeta(s^h_i,s^h_{-i})=\zeta(\widetilde{s}^h_i,s^h_{-i})$. Hence $s^h_i\in\widehat{r}_i(\mu^h_i,\widetilde{h})$ implies $\widetilde{s}^h_i\in\widehat{r}_i(\mu^h_i,\widetilde{h})=\widehat{r}_i(\widetilde{\mu}^h_i,\widetilde{h})$. Fix $\widetilde{h}\in H(\widetilde{s}^h_i)\cap H^{\widehat{h}}=H(s^{\widehat{h}}_i)$. For every $s^{\widehat{h}}_{-i}\in S^{\widehat{h}}_{-i}$, $\widetilde{\mu}^h_i(\varsigma(s^{\widehat{h}}_{-i})|\widetilde{h})=\mu^{\widehat{h}}_i(s^{\widehat{h}}_{-i}|\widetilde{h})$. For every $\widehat{s}^h_i\in S^h_i(\widehat{h})$, $\zeta(\widehat{s}^h_i|\widehat{h},s^{\widehat{h}}_{-i})=\zeta(\widehat{s}^h_i,\varsigma(s^{\widehat{h}}_{-i}))$. So, $\widetilde{s}^h_i|\widehat{h}=s^{\widehat{h}}_i\in\widehat{r}_i(\mu^{\widehat{h}}_i,\widetilde{h})$ implies $\widetilde{s}^h_i\in\widehat{r}_i(\widetilde{\mu}^h_i,\widetilde{h})$. ■

The same combination argument is now formalized from the point of view of the deviator and her beliefs, for different deviations from the same set of paths. I will refer directly to the context of Claim 1.

Lemma 4 *Let* $((\widetilde{S}^h_{i,q})_{i\in I})_{q\geq 0}$ *denote* $((S^h_{i,q})_{i\in I})_{q\geq 0}$ *(resp.,* $((\overline{S}^h_{i,q})_{i\in I})_{q\geq 0}$*). Fix* $m\leq n$ *(resp.,* $m\in\mathbb{N}$*) and* $\widehat{D}\subseteq D_l$. *For every* $\widehat{h}\in\widehat{D}$, *fix* $\widetilde{\mu}^{\widehat{h}}_l$ *that strongly believes* $(\widetilde{S}^h_{-l,q}(\widehat{h})|\widehat{h})_{q=0}^m$.

There exists $\widetilde{\mu}^h_l =^{\overline{Z}} \overline{\mu}^h_l(\overline{s}^h_l)$ *(resp.,* $\widetilde{\mu}^h_l =^{Z\backslash\cup_{\widehat{h}\in\widehat{D}}Z^{\widehat{h}}} \overline{\mu}^h_l(\overline{s}^h_l)$*) that strongly believes* $(\widetilde{S}^h_{-l,q})_{q=0}^m$ *such that* $\widetilde{\mu}^h_l =^{\widehat{h}} \widetilde{\mu}^{\widehat{h}}_l$ *for all* $\widehat{h}\in\widehat{D}$.

Proof.

I am going to show that for every $i \neq l$ and $\overline{s}_i^h \in \overline{S}_{i,\infty}^h$, and for every map $\varsigma : \widehat{h} \in \widehat{D} \mapsto s_i^{\widehat{h}} \in \widetilde{S}_{i,m}^h(\widehat{h})|\widehat{h}$, there exists $\widetilde{s}_i^h \in \widetilde{S}_{i,m}^h$ such that $\widetilde{s}_i^h =^{\overline{Z}} \overline{s}_i^h$ (resp., $\widetilde{s}_i^h =^{Z\setminus\cup_{\widehat{h}\in\widehat{D}}Z^{\widehat{h}}} \overline{s}_i^h$) and $\widetilde{s}_i^h =^{\widehat{h}} \varsigma(\widehat{h})$ for all $\widehat{h} \in \widehat{D}$.[22] Using all such $\widetilde{s}_i^h$'s, it is easy to construct the desired $\widetilde{\mu}_l^h$.

Drawing from the induction hypothesis of the proof of Lemma 1 (resp., from EP3), let μ_i^h and s_i^h denote $\widehat{\mu}_i^h(\overline{s}_i^h)$ and $\widehat{s}_i^h(\overline{s}_i^h)$ (resp., $\overline{\mu}_i^h(\overline{s}_i^h)$ and $\overline{s}_i^h$). Thus, μ_i^h strongly believes $(\widetilde{S}_{-i,q}^h)_{q=0}^{m-1}$ and $s_i^h \in \rho_i(\mu_i^h)$. Fix $\widehat{h} \in \widehat{D} \cap H(\overline{s}_i^h)$. Since $\mu_i^h =^{\overline{Z}} \overline{\mu}_i^h(\overline{s}_i^h)$ and $\overline{\mu}_i^h(\overline{s}_i^h)$ strongly believes $\overline{S}_{-i,\infty}^h$, $\mu_i^h(S_{-i}^h(\widehat{h})|p(\widehat{h})) = 0$. Since $\varsigma(\widehat{h}) \in \widetilde{S}_{i,m}^h(\widehat{h})|\widehat{h}$, there exists $\mu_i^{\widehat{h}}$ that strongly believes $(\widetilde{S}_{-i,q}^h(\widehat{h})|\widehat{h})_{q=0}^{m-1}$ ([23]) such that $\varsigma(\widehat{h}) \in \rho_i(\mu_i^{\widehat{h}})$. Thus, by Lemma 3, there exist $\widetilde{\mu}_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$ that strongly believes $(\widetilde{S}_{-i,q}^h)_{q=0}^{m-1}$ such that $\widetilde{\mu}_i^h(\cdot|\widetilde{h}) = \mu_i^h(\cdot|\widetilde{h})$ for all $\widetilde{h} \notin H^{\widehat{h}}$, and $\widetilde{s}_i^h \in \rho_i(\widetilde{\mu}_i^h)$ such that $\widetilde{s}_i^h =^{\widehat{h}} \varsigma(\widehat{h})$ and $\widetilde{s}_i^h(\widetilde{h}) = s_i^h(\widetilde{h})$ for all $\widetilde{h} \notin H^{\widehat{h}}$ (hence, $\widetilde{s}_i^h =^{\overline{Z}} \overline{s}_i^h$). Iterating for each $\widehat{h} \in \widehat{D}$, we obtain $\widetilde{\mu}_i^h =^{Z\setminus\cup_{\widehat{h}\in\widehat{D}}Z^{\widehat{h}}} \mu_i^h$ that strongly believes $(\widetilde{S}_{-i,q}^h)_{q=0}^{m-1}$ such that $\widetilde{\mu}_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$ for all $\widehat{h} \in \widehat{D}$, and $\widetilde{s}_i^h \in \rho_i(\widetilde{\mu}_i^h)$ such that $\widetilde{s}_i^h =^{Z\setminus\cup_{\widehat{h}\in\widehat{D}}Z^{\widehat{h}}} s_i^h$ and $\widetilde{s}_i^h =^{\widehat{h}} \varsigma(\widehat{h})$ for all $\widehat{h} \in \widehat{D}$. By A0, $\widetilde{s}_i^h \in \widetilde{S}_{i,m}^h$. ■

Now the proof of Claim 1 can be formalized.

Proof of Claim 1

Suppose by contraposition that there is a partition $(D, \overline{D})$ of D_l such that for every $\widehat{h} \in D_l$, there exist $m(\widehat{h}) \leq n$ and $\mu_l^{\widehat{h}} \in M_{m(\widehat{h})}^{\widehat{h}}$ that violate C1, and for every $\widehat{h} \in \overline{D}$ there exist $m(\widehat{h}) \in \mathbb{N}$ and $\mu_l^{\widehat{h}} \in \overline{M}_{m(\widehat{h})}^{\widehat{h}}$ that violate C2. (Note: D or $\overline{D}$ may be empty.) For each $\widehat{h} \in D_l$, fix $\overline{\mu}_l^{\widehat{h}}$ that strongly believes $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$ under which player l expects a non higher payoff than under $\mu_l^{\widehat{h}}$. Let $\overline{\mu}_l^h := \overline{\mu}_l^h(\overline{s}_l^h)$. By Lemma 4, there exists $\widetilde{\mu}_l^h =^{\overline{Z}} \overline{\mu}_l^h$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$ such that for every $\widehat{h} \in D_l$, $\widetilde{\mu}_l^h =^{\widehat{h}} \overline{\mu}_l^{\widehat{h}}$. I want to show that there exists $s_l^h \in \rho_l(\widetilde{\mu}_l^h)$ such that $s_l^h =^{\overline{Z}} \overline{s}_l^h$, contradicting NIH.

Fix $\widehat{h} \in D$. Substitute $\overline{\mu}_l^{\widehat{h}}$ with $\mu_l^{\widehat{h}}$ in the construction of $\widetilde{\mu}_l^h$ and obtain a new $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$ that strongly believes $(S_{-l,q}^h)_{q=0}^{m(\widehat{h})}$ with $\mu_l^h(S_{-l}(z)|\widetilde{h}) = \widetilde{\mu}_l^h(S_{-l}(z)|\widetilde{h})$ for all $\widetilde{h} \notin H^{\widehat{h}}$ and $z \notin Z^{\widehat{h}}$. Since player l expects a non lower payoff against $\mu_l^{\widehat{h}}$ than against $\overline{\mu}_l^{\widehat{h}}$, $\rho_l(\mu_l^h)(\widehat{h}) = \emptyset$ (which holds by the contrapositive hypothesis) implies $\rho_l(\widetilde{\mu}_l^h)(\widehat{h}) = \emptyset$. So, $H(\rho_l(\widetilde{\mu}_l^h)) \cap D = \emptyset$.

Write $\overline{D} = \{h^1, ..., h^k\}$ where $m(h^1) \geq ... \geq m(h^k)$. By Lemma 4, for each $j = 1, ..., k$, there exists $\mu_{l,j}^h =^{Z^h\setminus\cup_{t=1}^j Z^{h^t}} \overline{\mu}_l^h$ that strongly believes $(\overline{S}_{-l,q}^h)_{q=0}^{m(h^j)}$ such that $\mu_{l,j}^h =^{h^t} \mu_l^{h^t}$ for all $1 \leq t \leq j$. Let $\mu_{l,0}^h := \overline{\mu}_l^h$. Fix $j = 1, ..., k$ and suppose to have shown that $\rho_l(\mu_{l,j-1}^h) = \rho_l(\overline{\mu}_l^h)$. Then, $\rho_l(\mu_{l,j-1}^h) \cap S_l^h(h^j) = \emptyset$. By the contrapositive hypothesis,

[22] For $((S_{i,q}^h)_{i\in I})_{q\geq 0}$, the map ς is well defined because by the induction hypothesis of the proof of Lemma 1, $S_{i,m}^h(\widehat{h}) \neq \emptyset$ for all $\widehat{h} \in D_l$.

[23] Here is where the convention that every CPS strongly believes the empty set comes in handy: $\widetilde{S}_{-i,q}^h(\widehat{h})$ can be empty ($\overline{S}_{-i,q}^h(\widehat{h})$ certainly is for sufficiently high q).

$\rho_l(\mu^h_{l,j}) \cap S^h_l(h^j) = \emptyset$ as well. For all $\widetilde{h} \notin H^{h^j}$ and $z \notin Z^{h^j}$, $\mu^h_{l,j}(S_{-l}(z)|\widetilde{h}) = \mu^h_{l,j-1}(S_{-l}(z)|\widetilde{h})$. Then, $\rho_l(\mu^h_{l,j}) = \rho_l(\mu^h_{l,j-1})$. Inductively, $\rho_l(\mu^h_{l,k}) = \rho_l(\overline{\mu}^h_l)$.

So, we have:

i) $\rho_l(\mu^h_{l,k}) \cap D_l = \emptyset$;

ii) $\overline{s}^h_l \in \rho_l(\mu^h_{l,k})$;

iii) $H(\rho_l(\widetilde{\mu}^h_l)) \cap D = \emptyset$;

iv) for each $\widehat{h} \in \overline{D}$, player l initially expects a non lower payoff under $\mu^{\widehat{h}}_l$ than under $\overline{\mu}^{\widehat{h}}_l$, and recall that $\widetilde{\mu}^h_l =^{\widehat{h}} \overline{\mu}^{\widehat{h}}_l$ and $\mu^h_{l,k} =^{\widehat{h}} \mu^{\widehat{h}}_l$;

v) $\widetilde{\mu}^h_l =^{\overline{Z}} \mu^h_{l,k} =^{\overline{Z}} \overline{\mu}^h_l$ and $\overline{\mu}^h_l$ strongly believes $\overline{S}^h_{-l,\infty}$.

By (v), when player l deviates out of $\overline{Z}$, she expects the same distribution over (only) histories in D_l (and terminal histories) under $\widetilde{\mu}^h_l$ and $\mu^h_{l,k}$, and when she does not deviate out of $\overline{Z}$, she expects the same outcome distribution. By (i), player l has no incentive to deviate out of $\overline{Z}$ under $\mu^h_{l,k}$. By (iii), under $\widetilde{\mu}^h_l$, deviations that may lead to a history in D are suboptimal for player l, and by (iv) deviations that can only lead to histories in $\overline{D}$ bring a lower expected payoff than under $\mu^h_{l,k}$. So, under $\widetilde{\mu}^h_l$, player l will not want to deviate towards $\overline{D}$ either ($H(\rho_l(\widetilde{\mu}^h_l)) \cap \overline{D} = \emptyset$), and for each $\widetilde{h} \in H(\overline{S}^h_\infty)$ we have $\widehat{r}_l(\widetilde{\mu}^h_l, \widetilde{h}) = \widehat{r}_l(\mu^h_{l,k}, \widetilde{h})$. Then, by (ii), there exists $s^h_l \in \rho_l(\widetilde{\mu}^h_l)$ such that $s^h_l(\widetilde{h}) = \overline{s}^h_l(\widetilde{h})$ for all $\widetilde{h} \in H(\overline{S}^h_\infty)$. ■

References

[1] Banks, J. S. and J. Sobel, "Equilibrium Selection in Signaling Games," *Econometrica,* 55(3) (1987), 647-661.

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

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

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

[5] Battigalli, P. and A. Friedenberg, "Forward induction reasoning revisited", *Theoretical Economics*, **7**, 2012, 57-98.

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

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

[8] Battigalli P. and M. Siniscalchi, "Rationalization and Incomplete Information," *The B.E. Journal of Theoretical Economics*, **3(1)**, 2003, 1-46.

[9] Catonini, E., "Rationalizability and Epistemic Priority Orderings", *Games and Economic Behavior*, 2019, forthcoming.

[10] Catonini, E., "Self-Enforcing Agreements and Forward Induction Reasoning", working paper, 2017.

[11] Chen, J., and S. Micali, "The order independence of iterated dominance in extensive games", *Theoretical Economics*, **8**, 2013, 125-163.

[12] Cho I.K. and D. Kreps, "Signaling Games and Stable Equilibria", *Quarterly Journal of Economics*, **102**, 1987, 179-222.

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

[14] Kohlberg, E. and J.F. Mertens, "On the Strategic Stability of Equilibria", *Econometrica*, **54**, 1986, 1003-1038.

[15] Osborne, M., "Signaling, Forward Induction, and Stability in Finitely Repeated Games", *Journal of Economic Theory*, **50**, 1990, 22-36.

[16] Osborne, M. J. and A. Rubinstein, "A Course in Game Theory", 1994, Cambridge, Mass.: MIT Press.

[17] Pearce, D., "Rational Strategic Behavior and the Problem of Perfection", *Econometrica*, **52**, 1984, 1029-1050.

[18] Penta, A., "Backward Induction Reasoning in Games with Incomplete Information", 2011, working paper.

[19] Perea, A., "Order Independence in Dynamic Games", working paper, 2017.

[20] Perea, A., "Why Forward Induction leads to the Backward Induction outcome: a new proof for Battigalli's theorem", *Games and Economic Behavior*, **110**, 2018, 120–138.

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

[22] Shimoji, M. and J. Watson, "Conditional dominance, rationalizability, and game forms", *Journal of Economic Theory*, **83(2)**, 1998, 161-195.

[23] Sobel, J., L. Stole, I. Zapater, "Fixed-Equilibrium Rationalizability in Signaling Games," *Journal of Economic Theory*, **52**, 1990, 304-331.