On non-monotonic strategic reasoning*

Emiliano Catonini†

September 2019

Strong-Δ-Rationalizability (Battigalli 2003, Battigalli and Siniscalchi 2003) introduces first-order belief restrictions in forward induction reasoning. Without actual 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 Strong Rationalizability. I show that Strong-Δ-Rationalizability refines Strong Rationalizability in terms of outcomes when the restrictions correspond to belief in an outcome (distribution). Moreover, under such restrictions, the *epistemic priority* between rationality and restrictions is irrelevant for the predicted outcomes.

Keywords: Strong Rationalizability, Strong-Δ-Rationalizability, Path Restrictions, Epistemic Priority, Order Independence, Backward Induction.

*I want to thank Pierpaolo Battigalli, Carlo Cusumano, Andres Perea, two anonymous referees, 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 of Strong-Δ-Rationalizability are reassuringly compatible with *common strong belief in rationality*, as captured by Strong Rationalizability.

It turns out that, in all games with observable actions,[1] the set of outcomes predicted by Strong-Δ-Rationalizability is included in the set of strongly rationalizable outcomes when the restrictions corresponds to initial belief in an outcome (distribution), or in a set of outcomes that all receive positive probability.[2] 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. [25, 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 paths is interpreted as optimistic beliefs about the opponents' reaction, rather than disbelief that the opponents will stay on path themselves. 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

[1] i.e., games where, allowing for simultaneous moves, every player knows the current history of the game.

[2] 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.

refinements related to strategic stability have indeed been motivated informally with the idea that a deviator is expected to aim for a higher payoff than in equilibrium. Path restrictions have been used to make this idea precise: see Battigalli and Siniscalchi [8] for the Iterated Intuitive Criterion (Cho and Kreps [12]), Catonini [10] for "equilibrium paths that can be upset by a convincing deviation" (Osborne [15]), and (with different but related techniques) Sobel et al. [25] for Divine Equilibrium (Banks and Sobel [1]).

Why do path restrictions refine the set of strongly rationalizable outcomes? Note that, also under path restrictions, Strong-Δ-Rationalizability forces 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. A more general result highlights that the outcome inclusion between Strong-Δ-Rationalizability and Strong Rationalizability is preserved as long as the restrictions "never bite off-path". With this, I refer to restrictions that exclude belief systems only according to the probabilities they assign to the opponents' behavior along the paths that survive all steps of Strong-Δ-Rationalizability. Roughly speaking, under such restrictions, a strongly rationalizable path could be abandoned only if, at a step of reasoning, a deviation outside of the strongly rationalizable paths was profitable for all the beliefs about the continuation play; but in this case, some of these beliefs would be compatible with common strong belief in rationality, contradicting that the deviation is not strongly rationalizable. On the contrary, in presence of off-path restrictions, the deviation may occur only because the beliefs are constrained on a particular continuation play, which may not be compatible with some order of 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, 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, proven also by Perea [21]: 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,[3] which they prove being order independent in terms of outcomes in all games with perfect recall. Here, like in the recent work of Perea ([22], [21]), I do not use dominance characterizations.

Finally, I use the order independence result to show that Strong Rationalizability refines (in terms of outcomes) a generalization of backward induction to games without perfect information, Backwards Extensive-Form Rationalizability of Penta ([18], [19]). Perea [21] provides the same result with the Backwards Dominance Procedure (Perea [20]). Perea [22] had already shown that Strong Rationalizability refines backward induction in games with perfect information, to shed new light on their outcome *equivalence* in absence of relevant ties — a result originally proven by Battigalli [3], and then by Heifetz and Perea [13] in a more direct way. I derive 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. Section 5 presents the results on order independence and backward induction. Section 6 provides an example and the sketch of the proof of the workhorse lemma. In Section 7 I elaborate on similarities and differences between my approach and Perea's. The Appendix contains the proof of the workhorse lemma.

2 Preliminaries

Primitives of the game.[4] 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 $\overline{H} \subseteq \cup_{t=1,...,T}\overline{A}^t \cup \left\{h^0\right\}$ 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 $h' \prec h$.[5] Let $Z := \{z \in \overline{H} : \nexists h \in \overline{H}, z \prec h\}$ be the set of terminal histories (henceforth, *outcomes*

[3] 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 [24].

[4] The main notation is almost entirely borrowed from Osborne and Rubinstein [16].

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

or *paths*),[6] and $H := \overline{H} \backslash Z$ the set of non-terminal histories (henceforth, just *histories*). Second property: For every $h \in H$, there is a *Cartesian* set of actions profile $A(h)$ such that $(h, a) \in H$ such that $(h, a) \in \overline{H}$ if and only if $a \in A(h)$. For each $i \in I$, let $A_i(h)$ denote the set of actions available at h, so that $\times_{i \in I} A_i(h) = A(h)$.[7] 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

$$H(\overline{S}_J) := \left\{h \in H : \overline{S}_J(h) \neq \emptyset\right\}$$

denote the set of histories compatible with $\overline{S}_J$. For any $h = (h', a) \in \overline{H} \backslash \left\{h^0\right\}$, 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, [23]; henceforth, CPS).

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 $\overline{S}_{-i} \subseteq S_{-i}(h')$,*

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

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

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

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 say that a CPS strongly believes a sequence of events $(\overline{S}^q_{-i})^\infty_{q=0}$ when it strongly believes each event in the sequence. 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 $h \in H(s_i)$,*[8] *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 workhorse 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 $S^h_{i,0} = S^h_i$;

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 $s^h_i \in \rho_i(\mu^h_i) \subseteq S^h_{i,\infty}$.

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

Note three things. First, the fact that s_i^h is a sequential best reply to some μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^n$ does not imply that $s_i^h \in S_i^{n+1}$, and vice versa; this allows to encompass first-order belief restrictions and "slow" elimination orders. Second, EP2 allows $S_n^h = S_{n+1}^h$, and still $S_\infty^h \subsetneq S_{n+1}^h$; that is, the eliminations can stop for all players and then restart. Third, $S_{i,n}^h = \emptyset$ implies $S_{i,m}^h = \emptyset$ for all $m > n$, but does not imply $S_{j,\infty}^h = \emptyset$ 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" $((S_{i,q}^h(\widehat{h})|\widehat{h})_{i\in I})_{q=0}^\infty$ of an elimination procedure in a subgame $\Gamma(\widehat{h})$ $(\widehat{h} \succ h)$.

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

Proof: see the Appendix.

At each step of reasoning q and for each player i, the truncation $S_{i,q}^h(\widehat{h})|\widehat{h}$ is constructed by taking the strategies in $S_{i,q}^h$ that allow $\widehat{h}$ $(S_{i,q}^h(\widehat{h}))$, and restricting their domain to the histories that weakly follow $\widehat{h}$; that is, the histories of the subgame $\Gamma(\widehat{h})$. It will be important to keep in mind that a strategy $s_i^{\widehat{h}}$ can be eliminated from $S_{i,n}^h(\widehat{h})|\widehat{h}$ "exogenously": it may be a sequential best reply to some $\mu_i^{\widehat{h}}$ that strongly believes $(S_{-i,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$, and yet not belong to $S_{i,n+1}^h(\widehat{h})|\widehat{h}$, because no μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^n$ and induces $\mu_i^{\widehat{h}}$ in $\Gamma(\widehat{h})$ has sequential best replies that allow $\widehat{h}$.

The workhorse lemma of the paper claims the outcome inclusion between two elimination procedures, $((\overline{S}_{i,q}^h)_{i\in I})_{q=0}^\infty$ and $((S_{i,q}^h)_{i\in I})_{q=0}^\infty$, with the following feature. For each player i and each strategy $\overline{s}_i^h \in \overline{S}_{i,\infty}^h$, fix a CPS $\overline{\mu}_i^h(\overline{s}_i^h)$ that satisfies EP3; i.e., it strongly believes $(\overline{S}_{-i,q}^h)_{q=0}^\infty$, and $\overline{s}_i^h \in \rho_i(\overline{\mu}_i^h(\overline{s}_i^h)) \subseteq \overline{S}_{i,\infty}^h$. Say that a CPS μ_i^h "mimics" $\overline{\mu}_i^h(\overline{s}_i^h)$ along the paths predicted by $\overline{S}_\infty^h$ when, at every history along these paths, μ_i^h and $\overline{\mu}_i^h(\overline{s}_i^h)$ assign the same probabilities to (the opponents playing compatibly with) each of these paths: $\mu_i^h(S_{-i}(z)|\widetilde{h}) = \overline{\mu}_i^h(s_i^h)(S_{-i}(z)|\widetilde{h})$ for all $\widetilde{h} \in H(\overline{S}_\infty^h)$ and all $z \in \zeta(\overline{S}_\infty^h)$. Suppose that, for every step $m \in \mathbb{N}$, the following is true: if a CPS μ_i^h mimics some $\overline{\mu}_i^h(\overline{s}_i^h)$ along $\zeta(\overline{S}_\infty^h)$ *and* strongly believes $(\overline{S}_{-i,q}^h)_{q=0}^{m-1}$, then its sequential best replies survive the step of elimination of the first procedure: $\rho_i(\mu_i^h) \subseteq \overline{S}_{i,m}^h$; if a CPS μ_i^h mimics some $\overline{\mu}_i^h(\overline{s}_i^h)$ along $\zeta(\overline{S}_\infty^h)$ *and* strongly believes $(S_{-i,q}^h)_{q=0}^{m-1}$, then its sequential best replies survive the step of elimination of the second procedure: $\rho_i(\mu_i^h) \subseteq S_{i,m}^h$. Under these two conditions, the lemma claims that the second procedure predicts a superset $\zeta(S_\infty^h) \supseteq \zeta(\overline{S}_\infty^h)$ of the outcomes predicted by the first.

Lemma 1 *Fix $h \in H$ and two elimination procedures $((\overline{S}_{i,q}^h)_{i\in I})_{q=0}^\infty$, $((S_{i,q}^h)_{i\in I})_{q=0}^\infty$.*

For every $i \in I$, fix a map $\overline{\mu}_i^h : \overline{S}_{i,\infty}^h \to \Delta_i^{H^h}(S_{-i}^h)$ such that, for each $\overline{s}_i^h \in \overline{S}_{i,\infty}^h$, $\overline{\mu}_i^h(\overline{s}_i^h)$ strongly believes $(\overline{S}_{-i,q}^h)_{q=0}^{\infty}$, and $\overline{s}_i^h \in \rho_i(\overline{\mu}_i^h(\overline{s}_i^h)) \subseteq \overline{S}_{i,\infty}^h$.
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 the Appendix. In Section 6, I provide an example of outcome inclusion between two relevant elimination procedures, Strong-Δ-Rationalizability and Strong Rationalizability, and I illustrate the main intuition for it. Then, I sketch the proof of Lemma 1 in its generality.

4 Belief-restrictions and monotonicity

In this section, I am going to focus on 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 $((R_{i,q})_{i\in I})_{q=0}^{\infty}$ such that for every $n \in \mathbb{N}$, $i \in I$, and $s_i \in R_{i,n}$, there is μ_i that strongly believes $(R_{-i,q})_{q=0}^{n-1}$ with $s_i \in \rho_i(\mu_i)$.*[9]

Definition 6 *For each $i \in I$, fix $\Delta_i \subseteq \Delta^H(S_{-i})$. Strong-$\Delta$-Rationalizability is the elimination procedure $((R_{i,q}^{\Delta})_{i\in I})_{q=0}^{\infty}$ such that, for every $n \in \mathbb{N}$, $i \in I$, and $s_i \in S_i$, $s_i \in R_{i,n}^{\Delta}$ if and only if $s_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta_i$ that strongly believes $(R_{-i,q}^{\Delta})_{q=0}^{n-1}$.*[10]

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

1. *$(R_q^S)_{q=0}^M = (R_q)_{q=0}^M$, where M is the smallest $n \geq 0$ such that $R_{n+1} = R_n$;*

[9] The present definition of Strong Rationalizability is the one of Battigalli [4].
[10] The present definition of Strong-Δ-Rationalizability is the one of Battigalli and Prestipino [6].

2. *for every* $n > M$, $i \in I$, *and* $s_i \in S_i$, $s_i \in R^S_{i,n}$ *if and only if* $s_i \in \rho_i(\mu_i)$ *for some* $\mu_i \in \Delta_i$ *that strongly believes* $(R^S_{-i,q})^{n-1}_{q=0}$.[11]

Consider first-order belief restrictions $(\Delta_i)_{i\in I}$ with the following feature: 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 $h \in H(R^\Delta_\infty)$ to the (opponents playing compatibly with the) strongly-Δ-rationalizable paths $z \in \zeta(R^\Delta_\infty)$: $\mu_i(S_{-i}(z)|h)$. Then, Strong-Δ-Rationalizability satisfies the hypotheses of Lemma 1 as first elimination procedure. Strong Rationalizability, being an unconstrained procedure, saves at each step all the sequential best replies to every CPS that strongly believes in the previous steps, regardless of whether the CPS satisfies (1) or not. Therefore, Strong Rationalizability trivially satisfies the hypotheses of Lemma 1, and it can be taken as second elimination procedure. The desired outcome inclusion with respect to belief restrictions that "do not end up off-path" obtains.

Lemma 2 *For each* $i \in I$, *fix* $\Delta_i \subseteq \Delta^H(S_{-i})$. *Suppose that for each* $\overline{\mu}_i \in \Delta_i$ *and* $\mu_i \in \Delta^H(S_{-i})$,

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

Then, $\zeta(R^\Delta_\infty) \subseteq \zeta(R_\infty)$.

Proof. Fix $i \in I$. For each $s_i \in R^\Delta_{i,\infty}$, by Definition 6 there is $\overline{\mu}_i \in \Delta_i$ that strongly believes $(R^\Delta_{-i,q})^\infty_{q=0}$ such that $s_i \in \rho_i(\overline{\mu}_i) \subseteq R^\Delta_{i,\infty}$. For each μ_i that satisfies (1) with $\overline{\mu}_i$, by (2) we have $\mu_i \in \Delta_i$. So, for each $n \in \mathbb{N}$, if μ_i strongly believes $(R^\Delta_{-i,q})^n_{q=0}$, then $\rho_i(\mu_i) \subseteq R^\Delta_{i,n+1}$. Thus, A0 is satisfied for Strong-Δ-Rationalizability, while, as observed, it is always satisfied for Strong Rationalizability. Hence, by Lemma 1, $\zeta(R^\Delta_\infty) \subseteq \zeta(R_\infty)$. ■

Lemma 2 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. Beside the theoretical insight, though, Condition 2 is of little practical use to establish which restrictions guarantee the outcome inclusion: Whether the restrictions end up off-path or not has to be assessed with respect to the final output of Strong-Δ-Rationalizability itself.

[11] Selective Rationalizability in [9] is initialized with R_∞ and strong belief in $(R_{-i,q})^\infty_{q=0}$. It is easy to see that Definition 7 is equivalent in finite games.

Moreover, in [9] Selective Rationalizability is defined under the hypothesis of *independent rationalization*. That is, a valid μ_i is required to strongly believe $(R^{R\Delta}_{j,q})^{n-1}_{q=0}$ for all $j \neq i$, in place of just $(R^{R\Delta}_{-i,q})^{n-1}_{q=0}$. However, this hypothesis is immaterial for the result on Selective Rationalizability of this paper (Theorem 2) — a proof is available upon request.

Consider then first-order belief restrictions that correspond to the belief in a set of outcome distributions with the same support. To start, fix a set $\mathcal{Z} \subset \Delta(Z)$ of probability measures with the same support $\overline{Z}$ that can be induced by a *product* measure $\times_{i\in I}\nu_i \in \Delta(S)$ ($\nu_i \in \Delta(S_i)$) over strategy profiles — the focus on product measures is of course restrictive (also for the possible supports $\overline{Z}$), but it simplifies the exposition. Every player i initially believes that the opponents will play compatibly with an outcome distribution in $\mathcal{Z}$. With this, I define the notion of "path restrictions".

Definition 8 *Fix a set $\mathcal{Z} \subset \Delta(Z)$ of outcome distributions with support $\overline{Z} \subset Z$ that can be induced by a product measure over strategy profiles. The corresponding path restrictions of player i are given by the set of CPS's*

$$\Delta_i := \left\{\mu_i \in \Delta^H(S_{-i}) : \exists \eta \in \mathcal{Z}, \exists \nu_i \in \Delta(S_i), \forall z \in \overline{Z}, \nu_i(S_i(z))\mu_i(S_{-i}(z)|h^0) = \eta(z)\right\}.$$

If $\mathcal{Z}$ is a singleton, the corresponding path restrictions represent the belief in an outcome distribution; if $\overline{Z}$ is a singleton, they represent the belief in a specific path z: $\mu_i(S_{-i}(z)|h^0) = 1$.[12] When $\overline{Z}$ is not a singleton and $\mathcal{Z}$ is the set of all probability measures with support $\overline{Z}$, the corresponding path restrictions constitute a "restricted full support" condition with respect to a subset of outcomes.

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 player i eliminates all the strategies that are compatible with some path $z \in \overline{Z}$ (that is, $R^\Delta_{i,n} \cap S_i(z) = \emptyset$), it is easy to see that $R^\Delta_{j,n+1} = \emptyset$ for each $j \neq i$.

Remark 2 *Under path restrictions, if $R^\Delta_\infty \neq \emptyset$, then $\overline{Z} \subseteq \zeta(R^\Delta_\infty)$.*

Then, the restrictions never "end up off-path", and Lemma 2 can be applied.

Theorem 1 *Fix path restrictions $(\Delta_i)_{i\in I}$. We have $\zeta(R^\Delta_\infty) \subseteq \zeta(R_\infty)$.*

Proof. If $R^\Delta_\infty = \emptyset$, $\zeta(R^\Delta_\infty) \subseteq \zeta(R_\infty)$ is trivially true, so suppose $R^\Delta_\infty \neq \emptyset$. Fix $i \in I$, $\overline{\mu}_i \in \Delta_i$, and $\mu_i \in \Delta^H(S_{-i})$. By Definition 8, there are $\eta \in \mathcal{Z}$ and $\nu_i \in \Delta(S_i)$ such that $\nu_i(S_i(z))\overline{\mu}_i(S_{-i}(z)|h^0) = \eta(z)$ for all $z \in \overline{Z}$. By Remark 2, $\overline{Z} \subseteq \zeta(R^\Delta_\infty)$. Hence, we have

$$\begin{aligned}
\left(\forall \widetilde{h} \in H(R^\Delta_\infty), \forall z \in \zeta(R^\Delta_\infty), \mu_i(S_{-i}(z)|\widetilde{h}) = \overline{\mu}_i(S_{-i}(z)|\widetilde{h})\right) &\Rightarrow \\
\left(\forall z \in \overline{Z}, \mu_i(S_{-i}(z)|h^0) = \overline{\mu}_i(S_{-i}(z)|h^0)\right) &\Rightarrow \\
\left(\forall z \in \overline{Z}, \nu_i(S_i(z))\mu_i(S_{-i}(z)|h^0)) = \eta(z)\right) &\Rightarrow \quad (\mu_i \in \Delta_i).
\end{aligned}$$

[12] 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. A formal proof is available upon request.

Thus, (2) holds, and Lemma 2 yields $\zeta(R^{\Delta}_{\infty}) \subseteq \zeta(R_{\infty})$. ■

Corollary 1 *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, $\zeta(R^{\Delta}_{\infty}) \subseteq \zeta(R_{\infty})$.*

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

Theorem 2 *Fix path restrictions $(\Delta_i)_{i\in I}$. Then $\zeta(R^{\Delta}_{\infty}) = \zeta(R^{S}_{\infty})$.*

Proof. I show that $\zeta(R^{\Delta}_{\infty}) \subseteq \zeta(R^{S}_{\infty})$ — the proof of the opposite inclusion is identical. If $R^{\Delta}_{\infty} = \emptyset$, the inclusion holds trivially, so suppose $R^{\Delta}_{\infty} \neq \emptyset$. Fix $i \in I$. For each $s_i \in R^{\Delta}_{i,\infty}$, by Definition 6 there is $\overline{\mu}_i \in \Delta_i$ that strongly believes $(R^{\Delta}_{-i,q})^{\infty}_{q=0}$ such that $s_i \in \rho_i(\overline{\mu}_i) \subseteq R^{\Delta}_{i,\infty}$. For each μ_i that satisfies (1) with $\overline{\mu}_i$, proceeding like in the proof of Theorem 1, we find that $\mu_i \in \Delta_i$. So, for each $n \in \mathbb{N}$, if μ_i strongly believes $(R^{\Delta}_{-i,q})^{n}_{q=0}$, then $\rho_i(\mu_i) \subseteq R^{\Delta}_{i,n+1}$, and if μ_i strongly believes $(R^{S}_{-i,q})^{n}_{q=0}$, then $\rho_i(\mu_i) \subseteq R^{S}_{i,n+1}$. Thus, A0 is satisfied for both Strong-Δ-Rationalizability and Selective Rationalizability. Hence, Lemma 1 yields $\zeta(R^{\Delta}_{\infty}) \subseteq \zeta(R^{S}_{\infty})$. ■

Corollary 2 *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 $\zeta(R^{\Delta}_{\infty}) = \zeta(R^{S}_{\infty})$.*

An example of outcome inclusion between Strong-Δ-Rationalizability with path restrictions and Strong Rationalizability is provided in Section 6.

5 Order independence and backward induction

For any unconstrained elimination procedure, exactly as for Strong Rationalizability, A0 holds. An unconstrained elimination procedure is what I referred to in the Introduction as an order of 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 the iterated elimination of never sequential best replies in terms of outcomes obtains.

Theorem 3 *For every unconstrained elimination procedure $((S_{i,q})_{i\in I})^{\infty}_{q=0}$, $\zeta(S_{\infty}) = \zeta(R_{\infty})$.*

Proof. Any two unconstrained elimination procedures, taken in both orders, 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 by Penta ([18], [19]) and Perea ([20]) in the following way. Starting from the bottom of game, an action of a player at a history shall be eliminated when it is not "folding-back optimal" under any conjecture over the surviving actions of the opponents at the same history and at the future histories. Here I adopt Backwards Extensive-Form Rationalizability of Penta [18], because it is formulated in the language of extensive-form rationalizability, i.e., as a procedure of elimination of strategies that are not sequentially optimal for any viable conditional probability system. The following is a simplification of Penta's definition for games with complete information.

Definition 9 *Backwards Extensive-Form Rationalizability is a sequence* $((R^B_{i,q})_{i\in I})^{\infty}_{q=0}$ *where, for every* $i \in I$*,*

BR1 $R^B_{i,0} = S_i$*;*

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

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

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

Condition BR2.(ii) does not make a distinction between strategies in $R^B_{-i,n-1}$ that allow h or not; for this reason, it does not capture forward induction reasoning. Condition BR2.(i) requires s_i to be optimal given $\mu_i(\cdot|h)$ from any history h onwards, and not only when $h \in H(s_i)$ (cf. Definition 2 of sequential best reply). This entails that players' strategies are further refined also at histories that are not allowed anymore by some player. However, being such histories off-path, each step of Backwards Extensive-Form Rationalizability is outcome-equivalent to a step of an unconstrained elimination procedure. Moreover, Backwards Extensive-Form Rationalizability stops when the unconstrained procedure may not yet be allowed to, because EP3 is not satisfied.

Lemma 3 *Let* N *be the smallest* n *such that* $R^B_n = R^B_{n+1}$*. There exists an unconstrained elimination procedure* $((S_{i,q})_{i\in I})^{\infty}_{q=0}$ *such that for each* $n = 1, ..., N$*,*

$$S_{i,n} = \left\{ s_i \in S_{i,n-1} : \exists s'_i \in R^B_{i,n}, \forall h \in H(R^B_n) \cap H(s_i), s_i(h) = s'_i(h) \right\}.$$

Proof: see the Appendix.

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

Theorem 4 *Every strongly rationalizable outcome is a backwards extensive-form rationalizable outcome:* $\zeta(R_\infty) \subseteq \zeta(R_\infty^B)$.

Proof. Immediate from Lemma 3 and Theorem 3. ■

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

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

6 Example and sketch of the proof of the workhorse lemma

In this section, I provide an example of outcome inclusion between Strong-Δ-Rationalizability with path restrictions and Strong Rationalizability, and I illustrate the rough intuition behind it. Then, I generalize these ideas to sketch the proof of the workhorse lemma.

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	$0,0$	$0,0$	$0,3$

For brevity, I am going to use plans of actions (also known as reduced strategies) in place of strategies: given Definition 2, if a strategy is a sequential best reply to a CPS, all other strategies that correspond to the same plan of actions are as well. The plans of actions of Ann will be written as S, $N.U$, $N.M$, $N.D$; likewise for Bob.

Strong Rationalizability goes as follows. At the first step, Ann eliminates $N.D$, which is dominated by $1/2(N.U) + 1/2(N.M)$ in the subgame. At the second step, Bob eliminates $E.R$, now dominated by $1/2(E.L) + 1/2(E.C)$ in the subgame. At the third step, Ann eliminates $N.M$, dominated by $N.U$ in the subgame. At the fourth step, Bob eliminates $E.C$, dominated by $E.L$ in the subgame. The final output is $R_\infty = \{S, N.U\} \times \{W, E.L\}$.

Now, fix path $z := (N, W)$; thus, $S_A(z) = \{N.U, N.M, N.D\}$ and $S_B(z) = \{W\}$. The corresponding path restrictions are defined as

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

Strong-Δ-Rationalizability goes as follows. At the first step, Ann eliminates S, which is dominated by N under belief in W, and $N.D$, which is dominated by $1/2(N.U)+1/2(N.M)$ in the subgame; Bob eliminates $E.L$ and $E.C$, because they are both dominated by W under belief in N. At the second step, Ann eliminates $N.U$, now dominated by $N.M$ in the subgame, and Bob eliminates $E.R$, now dominated by W. The final output is: $R_\infty^\Delta = \{(N.M, W)\}$.

Note that $\zeta(R_\infty^\Delta) = \{z\} \subseteq \zeta(R_\infty)$, although $R_\infty^\Delta \cap R_\infty = \emptyset$. Why does the outcome inclusion hold, despite of the disjoint strategy sets? 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 a deviation to E. This is far from guaranteed, precisely because Strong Rationalizability and Strong-Δ-Rationalizability depart off the strongly-Δ-rationalizable path: the former ends up predicting (U, L), the latter M for Ann and R as the best rationalizable action 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 with the belief in N *any* belief after E that is compatible with the step of reasoning. Thus, if he abandons W, he can combine E with any action that can be justified after E as well. Therefore, if Bob abandons W at step n, the set of action profiles $R^{n-1}((N, E))|(N, E)$ 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 workhorse lemma generalizes these intuitions to all elimination procedures, predicting a set of paths, with many possible deviations, followed by a non-static subgame. That deviations can be followed by a dynamic game is the real challenge that the workhorse lemma has to face; generalizing in all other dimensions complicates the exposition but

does not present interesting challenges. Therefore, I sketch here the proof of the workhorse lemma in its generality, but with particular focus on how the issue of dynamic subgames is tackled.

Essentially, for a subgame that follows a hypothetical deviation along the second procedure (Procedure 2) from the paths predicted by the first procedure (Procedure 1), the task is to generate an *extensive form* best response set (Battigalli and Friedenberg [5]) of substrategies that make the deviation profitable, and prove that its sub-paths should have survived the truncation of the Procedure 1 in the subgame, a contradiction to the fact that the subgame follows a deviation from the paths predicted by Procedure 1 (henceforth, just "the paths"). As long as the paths survive also Procedure 2, players can form beliefs that, along the paths, mimic 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 the paths can be abandoned is that, at some step $n+1$, for all these beliefs, a player finds a particular deviation *outside* of the 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 they followed the 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 in 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. I want to show that they survive the truncation of Procedure 1 in the subgame, the desired contradiction. Note first that believing in these subpaths incentivizes our player to deviate from the paths also along Procedure 1. Given this, if the subgame is a static game, it is easy to observe that they do survive — see the example above. If the subgame has length higher than 1, then suppose by induction that the lemma is true in games of smaller length than the one under analysis — 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 induction hypothesis, the truncation of Procedure 1 induces a superset of those subpaths.

7 Discussion - Comparison with Perea ([21], [22])

To facilitate the comparison between my methodology and Perea's, I refer to the order independence problem tackled by Perea in [21].[13] 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 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 $\overline{S}_i$,[14] which excludes the presence of such deviation moves. Then,

[13] 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.

[14] 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

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 coincides with the maximal elimination order under the ρ^z operator;[15]

- show that Selective Rationalizability coincides with an elimination order under the ρ^z operator that coincides with Strong Rationalizability until convergence, and then proceeds at "full speed";[16]

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 $\widehat{S}_{-i}$.

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

[16] 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

- show that ρ^z is *monotone on reachable histories* (Perea [22]);

- invoke Theorem 3.2 in Perea [22] 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 [22]). 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

8.1 Proofs omitted from Sections 4 and 5

Proof of Remark 1. That a truncation satisfies EP1 and EP2 is obvious.

To show EP3, fix $i \in I$ and $s_i^{\widehat{h}} \in S_{i,\infty}^h(\widehat{h})|\widehat{h}$. I want to find $\mu_i^{\widehat{h}}$ that strongly believes $(S_{-i,q}^h(\widehat{h})|\widehat{h}))_{q=0}^{\infty}$ such that $s_i^{\widehat{h}} \in \rho_i(\mu_i^{\widehat{h}}) \subseteq S_{i,\infty}^h(\widehat{h})|\widehat{h}$.

Fix $s_i^h \in S_{i,\infty}^h(\widehat{h})$ such that $s_i^h|\widehat{h} = s_i^{\widehat{h}}$. Since $((S_{i,q}^h)_{i\in I})_{q=0}^{\infty}$ satisfies EP3, there is μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^{\infty}$ such that $s_i^h \in \rho_i(\mu_i^h) \subseteq S_{i,\infty}^h$. Construct $\mu_i^{\widehat{h}}$ as follows: for each $\widetilde{h} \succeq \widehat{h}$, define $\mu_i^{\widehat{h}}(\cdot|\widetilde{h})$ as the pushforward of $\mu_i^h(\cdot|\widetilde{h})$ through the map that associates each $s_{-i}^h \in S_{-i}^h(\widehat{h})$ with $s_{-i}^h|\widehat{h}$. It is easy to see that $\mu_i^{\widehat{h}}$ strongly believes $(S_{-i,q}^h(\widehat{h})|\widehat{h}))_{q=0}^{\infty}$. For each $\widetilde{h} \succeq \widehat{h}$, $\mu_i^{\widehat{h}}(\cdot|\widetilde{h})$ and $\mu_i^h(\cdot|\widetilde{h})$ induce the same distribution over terminal histories when coupled with any $\widetilde{s}_i^{\widehat{h}}$ and $\widetilde{s}_i^h$ such that $\widetilde{s}_i^h|\widehat{h} = \widetilde{s}_i^{\widehat{h}}$. Therefore, that s_i^h is a continuation best reply to $\mu_i^{\widehat{h}}(\cdot|\widetilde{h})$ implies that $s_i^{\widehat{h}}$ is a continuation best reply to $\mu_i^{\widehat{h}}$. Thus, $s_i^{\widehat{h}} \in \rho_i(\mu_i^{\widehat{h}})$.

Finally, fix a sequential best reply $\widetilde{s}_i^{\widehat{h}}$ to $\mu_i^{\widehat{h}}$. It is easy to see that the strategy $\widetilde{s}_i^h$ with $\widetilde{s}_i^h|\widehat{h} = \widetilde{s}_i^{\widehat{h}}$ and $\widetilde{s}_i^h(\widetilde{h}) = s_i^h(\widetilde{h})$ for each $\widetilde{h} \in H(s_i^h)\backslash H^{\widehat{h}}$ is a sequential best reply to μ_i^h. Thus, by $\rho_i(\mu_i^h) \subseteq S_{i,\infty}^h$, $\widetilde{s}_i^h \in S_{i,\infty}^h(\widehat{h})$, and then $\widetilde{s}_i^{\widehat{h}} \in S_{i,\infty}^h(\widehat{h})|\widehat{h}$. ■

Proof of Lemma 3. 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. Now I show it is unconstrained. Fix $n = 1, ..., N$ and, if $n > 1$, suppose by way of induction that for each $m = 1, ..., n-1$, $i \in I$, and μ_i that strongly believes $(S_{-i,q})_{q=0}^{m-1}$, we have $\rho_i(\mu_i) \subseteq S_{i,m}$. Fix μ_i that strongly believes $(S_{-i,q})_{q=0}^{n-1}$. I show that $\rho_i(\mu_i) \subseteq S_{i,n}$.

procedures, not even under path restrictions. However, for path restrictions, outcome equivalences hold. A proof is available upon request.

By definition of $S_{-i,n-1}$, I can construct μ_i' that satisfies BR2.(ii) such that

$$\mu_i'(S_{-i}(z)|h) = \mu_i(S_{-i}(z)|h), \quad \forall h \in H\left(S_{n-1}\right), \forall z \in \zeta\left(S_{n-1}\right). \tag{3}$$

For each $s_i' \in \rho_i(\mu_i')$, there is a realization-equivalent s_i'' that satisfies BR2.(i), so that $s_i'' \in R_{i,n}^B \subseteq R_{i,n-1}^B$. Then, $\zeta\left(\rho_i(\mu_i') \times R_{-i,n-1}^B\right) \subseteq \zeta\left(R_{n-1}^B\right)$. By definition of S_{n-1}, $\zeta\left(S_{n-1}\right) = \zeta\left(R_{n-1}^B\right)$. Thus, (i) $\zeta\left(\rho_i(\mu_i') \times R_{-i,n-1}^B\right) \subseteq \zeta\left(S_{n-1}\right)$. For each $s_i \in \rho_i(\mu_i)$, by the induction hypothesis we have $s_i \in S_{i,n-1}$. Thus, (ii) $\zeta\left(\rho_i(\mu_i) \times S_{-i,n-1}\right) \subseteq \zeta\left(S_{n-1}\right)$. Together with (3), (i) and (ii) imply that 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 $h \in H\left(S_{n-1}\right)$. As observed, there is $s_i'' \in R_{i,n}^B$ realization-equivalent to s_i', thus $s_i''(h) = s_i'(h) = s_i(h)$ for all $h \in H\left(S_{n-1}\right) \cap H(s_i)$. Hence, by definition of $S_{i,n}$, we have $s_i \in S_{i,n}$, as desired. ■

8.2 Proof of the workhorse lemma

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

For any $\widehat{h} \succeq h$, $(s_j^h)_{j\in I} \in S^h$, $(s_j^{\widehat{h}})_{j\in I} \in S^{\widehat{h}}$, $\mu_i^h \in \Delta^{H^h}(S_{-i}^h)$, $\mu_i^{\widehat{h}} \in \Delta^{H^{\widehat{h}}}(S_{-i}^{\widehat{h}})$, $\widehat{Z} \subseteq Z^{\widehat{h}}$, and $J \subseteq I$, let:

- $s_J^h =^{\widehat{Z}} s_J^{\widehat{h}}$ if for each $z \in \widehat{Z}$ and $\widetilde{h}$ with $\widehat{h} \preceq \widetilde{h} \prec z$, $s_J^h(\widetilde{h}) = s_J^{\widehat{h}}(\widetilde{h})$;
- $\mu_i^h =^{\widehat{Z}} \mu_i^{\widehat{h}}$ if for each $z \in \widehat{Z}$ and $\widetilde{h}$ with $\widehat{h} \preceq \widetilde{h} \prec z$, $\mu_i^h(S_{-i}^h(z)|\widetilde{h}) = \mu_i^{\widehat{h}}(S_{-i}^{\widehat{h}}(z)|\widetilde{h})$;
- $s_J^h =^{\widehat{h}} s_J^{\widehat{h}}$ and $\mu_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$ if, respectively, $s_J^h =^{Z^{\widehat{h}}} s_J^{\widehat{h}}$ and $\mu_i^h =^{Z^{\widehat{h}}} \mu_i^{\widehat{h}}$.

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

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

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

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

As anticipated, it is always possible to construct a map $\widehat{\mu}_i^h$ that satisfies IH1 at step $n+1$ by doing, for each $\overline{s}_i^h \in \overline{S}_{i,\infty}^h$ and at each $h \in H(\overline{S}_\infty^h)$, the pushforward of $\overline{\mu}_i^h(\overline{s}_i^h)(\cdot|h)$ through the map $\times_{j\neq i}\widehat{s}_j^h$ constructed at step n. The problem could be that, for some $l \in I$ and some $\overline{s}_l^h \in \overline{S}_{l,\infty}^h$, every μ_l^h that satisfies IH1 at step $n+1$ does not justify a strategy that imitates $\overline{s}_l^h$ 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 $s_l^h =^{\overline{Z}} \overline{s}_l^h$.[17]

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 mimics $\overline{\mu}_l^h(\overline{s}_l^h)$ 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

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

be the set of 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 $\mu_l^{\widehat{h}}(\cdot|\widehat{h})$).[18]

[17] 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.

[18] 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 $(\overline{S}_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^m$.

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 $\rho_l(\mu_l^h) \cap S_l^h(\widehat{h}) \neq \emptyset$;

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 $\rho_l(\mu_l^h) \cap S_l^h(\widehat{h}) \neq \emptyset$.[19]

C1 with $m = n$ is the result described in words after NIH. 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 end of this 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 such belief after $\widehat{h}$, (ii) strongly believes $(S_{-l,q}^h)_{q=0}^n$, (iii) mimics $\overline{\mu}_l^h(\overline{s}_l^h)$ along $\overline{Z}$, and (iv) incentivizes player l to deviate towards $\widehat{h}$.[20] But all such beliefs (one for each $\widehat{h} \in D_l$) *can* be induced by the same CPS μ_l^h that strongly believes $(S_{-l,q}^h)_{q=0}^n$ and mimics $\overline{\mu}_l^h(\overline{s}_l^h)$ along $\overline{Z}$, thus satisfying (i)-(ii)-(iii). So, μ_l^h violates (iv) for every $\widehat{h} \in D_l$. But then, under μ_l^h player l has no incentive to do any deviation from $\overline{Z}$, and so there is a sequential best reply to μ_l^h that imitates $\overline{s}_l^h$ along $\overline{Z}$, contradicting NIH. The proof that such μ_l^h can be constructed is based on the following idea. By the induction hypothesis, for any player $i \neq l$, the strategies in $S_{i,n}^h$ that imitate those in $\overline{S}_{i,\infty}^h$ 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 $S_{-l,n}^h(\widehat{h})|\widehat{h}$ that player l has to believe in. Thus, there are strategies in $S_{-l,n}^h$ that imitate those in $\overline{S}_{-l,\infty}^h$ 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 $S_{l,n+1}^h(\widehat{h})|\widehat{h}$ contains all the sequential best replies to all the beliefs in $M_n^{\widehat{h}}$, i.e., to all $\mu_l^{\widehat{h}}$ that strongly believe $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$. The same holds for $S_{i,n}^h(\widehat{h})|\widehat{h}$ with $i \neq l$, because by the induction

[19] The statement must hold vacously for some $p \in \mathbb{N}$ (i.e., $\overline{M}_p^{\widehat{h}} = \emptyset$): since $\widehat{h} \notin H(\overline{S}_{l,\infty}^h)$, there cannot be μ_l^h that strongly believes $(\overline{S}_{-l,q}^h)_{q=0}^\infty$ such that $\rho_l(\mu_l^h) \cap S_l^h(\widehat{h}) \neq \emptyset$, because $\rho_l(\mu_l^h) \subset \overline{S}_{l,\infty}^h$ by A0.

[20] 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.

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 1 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 length. 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 $\Gamma(h)$ is a static game, $D_l(\overline{S}^h_\infty) = \emptyset$, 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 length 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 $m \leq n$, $\overline{S}^{\widehat{h}}_{i,m} = S^h_{i,m}(\widehat{h})|\widehat{h}$; 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 $s^{\widehat{h}}_i \in \rho_i(\mu^{\widehat{h}}_i)$. I want to show that $((\overline{S}^{\widehat{h}}_{i,q})_{i\in I})_{q\geq 0}$ is an elimination procedure with non-empty output, and that it satisfies A0.

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 $\widetilde{h} \not\geq \widehat{h}$.[21] Thus, $\rho_i(\mu^h_i) \cap S^h_i(\widehat{h}) \neq \emptyset$, and by $\mu^h_i =^{\overline{Z}} \overline{\mu}^h_i(\overline{s}^h_i)$ and A0 (referred to $((S^h_{j,q})_{j\in I})_{q\geq 0}$), $\rho_i(\mu^h_i) \subseteq S^h_{i,m}$. So, $\rho_i(\mu^{\widehat{h}}_i) \subseteq \overline{S}^{\widehat{h}}_{i,m}$.

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 $\rho_l(\mu^h_l) \cap S^h_l(\widehat{h}) \neq \emptyset$. By A0, $\rho_l(\mu^h_l) \subseteq S^h_{l,n}$. So, $\rho_l(\mu^{\widehat{h}}_l) \subseteq \overline{S}^{\widehat{h}}_{l,n+1} \subseteq \overline{S}^{\widehat{h}}_{l,n}$.

Then, for every $i \in I$ and $\mu^{\widehat{h}}_i$ that strongly believes $(\overline{S}^{\widehat{h}}_{-i,q})^n_{q=0}$,[22] $\rho_i(\mu^{\widehat{h}}_i) \subseteq \overline{S}^{\widehat{h}}_{i,n} \neq \emptyset$. So, $\overline{S}^{\widehat{h}}_{i,n} \supseteq \overline{S}^{\widehat{h}}_{i,n+1} \neq \emptyset$. Therefore, $((\overline{S}^{\widehat{h}}_{i,q})_{i\in I})_{q\geq 0}$ is an elimination procedure with $\overline{S}^{\widehat{h}}_\infty \neq \emptyset$.

[21] The construction is shown explicitly by Lemma 4 in the Appendix.

[22] 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.

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

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 $(\overline{S}_{-i,q}^{\widehat{h}})_{q=0}^{m-1}$, $\rho_i(\mu_i^{\widehat{h}}) \subseteq \overline{S}_{i,m}^{\widehat{h}}$. Thus, $((\overline{S}_{i,q}^{\widehat{h}})_{i\in I})_{q\geq 0}$ satisfies A0.

Define now $((S_{i,q}^{\widehat{h}})_{i\in I})_{q\geq 0}$ as $((\overline{S}_{i,q}^h(\widehat{h})|\widehat{h})_{i\in I})_{q\geq 0}$. By Remark 1 it is an elimination procedure. I want to show that it satisfies A0.

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

For every $m \in \mathbb{N}$, $\overline{\mu}_l^{\widehat{h}}$ that strongly believes $(\overline{S}_{-l,q}^{\widehat{h}})_{q=0}^{\infty}$, and $\mu_l^{\widehat{h}} =^{\zeta(\overline{S}_{\infty}^{\widehat{h}})} \overline{\mu}_l^{\widehat{h}}$ that strongly believes $(S_{-l,q}^{\widehat{h}})_{q=0}^{m-1}$, by the same argument as above, $\mu_l^{\widehat{h}} \in \overline{M}_m^{\widehat{h}}$. Thus, by C2 there exists $\mu_l^h =^{\overline{Z}} \overline{\mu}_l^h(\overline{s}_l^h)$ that strongly believes $(\overline{S}_{-l,q}^h)_{q=0}^{m-1}$ such that $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$ and $\rho_l(\mu_l^h) \cap S_l^h(\widehat{h}) \neq \emptyset$. By A0, $\rho_l(\mu_l^h) \subseteq \overline{S}_{l,m}^h$. So, $\rho_l(\mu_l^{\widehat{h}}) \subseteq S_{l,m}^{\widehat{h}}$.

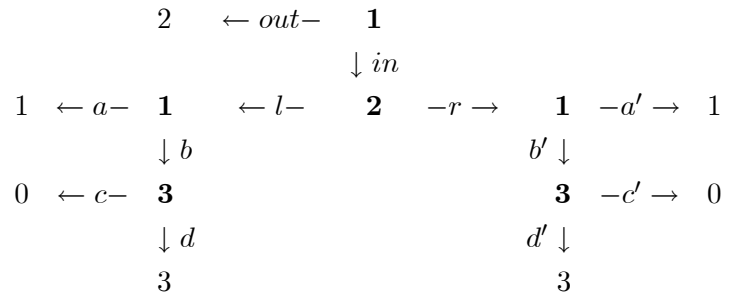
Then, for every $m \in \mathbb{N}$, $i \in I$, $\overline{\mu}_i^{\widehat{h}}$ that strongly believes $(\overline{S}_{-i,q}^{\widehat{h}})_{q=0}^{\infty}$ and $\mu_i^{\widehat{h}} =^{\zeta(\overline{S}_{\infty}^{\widehat{h}})} \overline{\mu}_i^{\widehat{h}}$ that strongly believes $(S_{-i,q}^{\widehat{h}})_{q=0}^{m-1}$,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

As assumed by induction on the length of the game, 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 $\widehat{h} \in D_l(\overline{S}_{\infty}^h)$. ■

8.2.1 Proof of Claim 1.

The main challenges for the proof of Claim 1 derive from the following issue. Fix an elimination procedure $((S_{i,q}^h)_{i\in I})_{q\geq 0}$ 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 $S_{-i,n}^h, ..., S_{-i,0}^h$. Fix two unordered histories $\widehat{h}, \overline{h} \in H(\widehat{s}_i^h) \cap H(\overline{s}_i^h)$ (that is, $\overline{h} \not\geq \widehat{h}$ and $\widehat{h} \not\geq \overline{h}$). Is there always a CPS that strongly believes $S_{-i,n}^h, ..., S_{-i,0}^h$ 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 unforeseen 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 contingencies, 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 4 *Fix an elimination procedure $((S_{i,q}^h)_{i\in I})_{q\geq 0}$, $n\in\mathbb{N}$, $i\in I$, $\widehat{h}\in H^h$, and μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^{n-1}$ such that $\mu_i^h(S_{-i}^h(\widehat{h})|p(\widehat{h}))=0$. Fix $s_i^h\in\rho_i(\mu_i^h)\cap S_i^h(\widehat{h})$, $\mu_i^{\widehat{h}}$ that strongly believes $(S_{-i,q}^h(\widehat{h})|\widehat{h})_{q=0}^{n-1}$, and $s_i^{\widehat{h}}\in\rho_i(\mu_i^{\widehat{h}})$.*

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

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

Proof.

Fix a map $\varsigma : S_{-i}^{\widehat{h}}\to S_{-i}^h$ such that for each $s_{-i}^{\widehat{h}}\in S_{-i}^{\widehat{h}}$, $\varsigma(s_{-i}^{\widehat{h}}) =^{\widehat{h}} s_{-i}^{\widehat{h}}$ and $\varsigma(s_{-i}^{\widehat{h}})\in S_{-i,m}^h(\widehat{h})$ for all $m\geq 0$ with $s_{-i}^{\widehat{h}}\in S_{-i,m}^h(\widehat{h})|\widehat{h}$. Since ς is injective, we can construct an array of probability measures $\widetilde{\mu}_i^h=(\widetilde{\mu}_i^h(\cdot|\widetilde{h}))_{\widetilde{h}\in H^h}$ on S_{-i}^h as $\widetilde{\mu}_i^h(\cdot|\widetilde{h})=\mu_i^h(\cdot|\widetilde{h})$ for all $\widetilde{h}\notin H^{\widehat{h}}$

and $\widetilde{\mu}_i^h(\varsigma(s_{-i}^{\widehat{h}})|\widetilde{h}) = \mu_i^{\widehat{h}}(s_{-i}^{\widehat{h}}|\widetilde{h})$ for all $\widetilde{h} \in H^{\widehat{h}}$ and $s_{-i}^{\widehat{h}} \in S_{-i}^{\widehat{h}}$. From the definition of ς, it immediately follows that $\widetilde{\mu}_i^h(S_{-i}(\widetilde{h})|\widetilde{h}) = 1$ for all $\widetilde{h} \in H^h$, that $\widetilde{\mu}_i^h$ strongly believes $(S_{-i,q}^h)_{q=0}^{n-1}$, and that $\widetilde{\mu}_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$. Finally, since $\widetilde{\mu}_i^h(S_{-i}(\widehat{h})|p(\widehat{h})) = 0$, $\widetilde{\mu}_i^h$ satisfies the chain rule.

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

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

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

Proof.

I am going to show that for each $i \neq l$ and $\overline{s}_i^h \in \overline{S}_{i,\infty}^h$, and for each 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., $\widetilde{s}_i^h =^{Z\backslash\cup_{\widehat{h}\in\widehat{D}} Z^{\widehat{h}}} \overline{s}_i^h$) and $\widetilde{s}_i^h =^{\widehat{h}} \varsigma(\widehat{h})$ for all $\widehat{h} \in \widehat{D}$.[23] Redistributing the probability from each $\overline{s}_i^h$ to the corresponding $\widetilde{s}_i^h$'s in the supports of $\overline{\mu}_l^h(\overline{s}_l^h)$, one obtains 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 $(\widetilde{S}_{-i,q}^h(\widehat{h})|\widehat{h})_{q=0}^{m-1}$ ([24]) such that $\varsigma(\widehat{h}) \in \rho_i(\mu_i^{\widehat{h}})$. Thus, by Lemma 4, there exist (i) $\widetilde{\mu}_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$ 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 (ii) $\widetilde{s}_i^h \in \rho_i(\widetilde{\mu}_i^h)$ 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 $\widetilde{h} \notin H^{\widehat{h}}$. Iterating for each $\widehat{h} \in \widehat{D}$, we obtain (i) $\widetilde{\mu}_i^h =^{Z\backslash\cup_{\widehat{h}\in\widehat{D}} Z^{\widehat{h}}} \mu_i^h$ that strongly believes $(\widetilde{S}_{-i,q}^h)_{q=0}^{m-1}$ such that

[23] 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 $\widehat{h} \in D_l$.

[24] 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, because $\overline{S}_{l,\infty}^h(\widehat{h}) = \emptyset$).

$\widetilde{\mu}_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$ for all $\widehat{h} \in \widehat{D}$, and (ii) $\widetilde{s}_i^h \in \rho_i(\widetilde{\mu}_i^h)$ such that $\widetilde{s}_i^h =^{Z \setminus \cup_{\widehat{h} \in \widehat{D}} Z^{\widehat{h}}} s_i^h$ (thus $\widetilde{s}_i^h =^{\overline{Z}} s_i^h$) 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 $\widehat{h} \in D$, 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 5, 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 $\overline{\mu}_l^{\widehat{h}}$, $\rho_l(\mu_l^h) \cap S_l^h(\widehat{h}) = \emptyset$ (which holds by the contrapositive hypothesis) implies $\rho_l(\widetilde{\mu}_l^h) \cap S_l^h(\widehat{h}) = \emptyset$. 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 5, 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, $\rho_l(\mu_{l,j}^h) \cap S_l^h(h^j) = \emptyset$ as well. For all $\widetilde{h} \notin H^{h^j}$ and $z \notin Z^{h^j}$, $\mu_{l,j}^h(S_{-l}(z)|\widetilde{h}) = \mu_{l,j-1}^h(S_{-l}(z)|\widetilde{h})$. Then, $\rho_l(\mu_{l,j}^h) = \rho_l(\mu_{l,j-1}^h)$. Inductively, $\rho_l(\mu_{l,k}^h) = \rho_l(\overline{\mu}_l^h)$.

So, we have:

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

ii) $\overline{s}_l^h \in \rho_l(\mu_{l,k}^h)$;

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

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

v) $\widetilde{\mu}_l^h =^{\overline{Z}} \overline{\mu}_l^h =^{\overline{Z}} \mu_{l,k}^h$, and $\overline{\mu}_l^h$ strongly believes $\overline{S}_{-l,\infty}^h$.

By (v), we obtain the following two facts: (a) if player l deviates out of $\overline{Z}$, under $\widetilde{\mu}_l^h$ and $\mu_{l,k}^h$ she expects the same distribution over histories in D_l and terminal histories that

immediately follow the deviation; (b) if she does not deviate out of $\overline{Z}$, under $\widetilde{\mu}_l^h$ and $\mu_{l,k}^h$ she expects the same outcome distribution. By (i), we have that (c) player l has no incentive to deviate out of $\overline{Z}$ under $\mu_{l,k}^h$. By (iv) and (a), deviations that can only lead to histories in $\overline{D}$ (or terminal histories) bring a lower expected payoff under $\widetilde{\mu}_l^h$ than under $\mu_{l,k}^h$. Hence, by (b) and (c), such deviations are suboptimal under $\widetilde{\mu}_l^h$ as well. By (iii), under $\widetilde{\mu}_l^h$, deviations that may lead to a history in D are suboptimal too. Hence, (d) player l has no incentive to deviate out of $\overline{Z}$ under $\widetilde{\mu}_l^h$ as well. By (b), (c), and (d), we have that for each $\widetilde{h} \in H(\overline{S}_\infty^h)$, $\widehat{r}_l(\widetilde{\mu}_l^h, \widetilde{h}) = \widehat{r}_l(\mu_{l,k}^h, \widetilde{h})$. Then, by (ii), there exists $s_l^h \in \rho_l(\widetilde{\mu}_l^h)$ such that $s_l^h(\widetilde{h}) = \overline{s}_l^h(\widetilde{h})$ for all $\widetilde{h} \in H(\overline{S}_\infty^h)$. ■

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*, **114**, 2019, 101-117.

[10] Catonini, E., "Self-Enforcing Agreements and Forward Induction Reasoning", working paper, 2019.

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

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

[21] Perea, A., "Order Independence in Dynamic Games", 2018, working paper.

[22] 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.

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

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

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