Non-binding agreements and forward induction reasoning*

Emiliano Catonini†

November 2015

In dynamic games, players may observe a deviation from a pre-play, possibly incomplete, non-binding agreement before the game is over. The attempt to rationalize the deviation may lead players to revise their beliefs about co-players' behavior in the continuation of the game. This instance of forward induction reasoning is based on interactive beliefs about not just rationality but also the compliance with the agreement itself. I study the effects of such rationalization on the self-enforceability of the agreement. Accordingly, outcomes of the game are deemed to be implementable by some agreement or not. Conclusions depart substantially from what the equilibrium refinement tradition suggests. A non subgame perfect equilibrium may represent a self-enforcing agreement, while a subgame perfect equilibrium may not. The incompleteness of the agreement can be crucial to implement an equilibrium outcome. However, every game possesses an outcome which is compatible with both forward induction and subgame perfection.

Keywords: Agreements, Self-Enforceability, Forward Induction, Strong-Δ-Rationalizability, Selective Rationalizability, Stability.

J.E.L. Classification: C72, C73, D86.

*I am grateful to Pierpaolo Battigalli for introducing me to the mysteries of epistemic game theory. Thank you also to Adam Brandenburger, Alfredo Di Tillio, Amanda Friedenberg, Mattia Landoni, Elena Manzoni, Burkhard Schipper, Dimitrios Tsomocos, Yi-Chun Chen, Xiao Luo, Satoru Takahashi, Amanda Jakobsson, Madhav Aney, Shurojit Chatterji and Atsushi Kajii for precious suggestions.

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

1 Introduction

When the players of a dynamic game can communicate before the game starts, they are likely to exploit this opportunity to reach a possibly incomplete agreement[1] about how to play. In most cases, the context allows them to reach only a non-binding agreement, which cannot be enforced by a court of law. The only way a non-binding agreement can affect the behavior of players is through the beliefs it is able to induce in their minds. This paper sheds light on which agreements players can believe in and, among them, which agreements players will comply with. Moreover, in an implementation perspective, the paper investigates which outcomes of the game can be ensured by *some* agreement. The paper will not deal with the pre-play bargaining phase. Yet, the evaluation of their credibility has a clear and strong feedback on which agreements are likely to be reached.

In this paper I take the view that players will believe in the agreement only if this is compatible with reasonable assumptions about rationality,[2] beliefs in rationality and their interaction with the beliefs in the agreement of all orders. Ann will believe in the agreement only if Bob may comply with it in case he is rational, he believes in the agreement, he believes that Ann is rational and believes in the agreement (which may add non-agreed upon restrictions on what Bob expects Ann to do), and so on. Moreover, I take the view that deviations, or more generally past actions, are not interpreted as mistakes but as intentional choices. Suppose that for Bob, in case he is rational and believes in the agreement, some move makes sense only if he plans to play a certain action thereafter. Ann, upon observing such move, will believe that Bob will play that action (and Bob may use the move to signal this). This instance of forward induction reasoning is based not just on the belief in Bob's rationality but also on its interaction with the belief that Bob believes in the agreement. Example 2 in Section 2 provides a case in point. Consider now a move that Bob, if he is rational and believes in the agreement, cannot find profitable whatever he plays thereafter. Example 1 in Section 2 provides a situation of this kind. Then Ann cannot mantain both beliefs that Bob is rational and that he believes in the agreement. Which of the

[1] The mathematical representation of agreements in this paper can be given also different interpretations. For instance, the agreement can represent a set of public announcements (from a subset of players)

[2] The notion of rationality employed in this paper imposes expected utility maximization but it does not impose by itself any restriction on beliefs. See Section 3 for details.

two she will (try to)[3] mantain determines the choice of the appropriate rationalizability concept to capture these lines of strategic reasoning. *Strong-Δ-rationalizability* (Battigalli, [5]; Battigalli and Siniscalchi, [11]) captures the hypothesis that the beliefs in the agreement are given higher *epistemic priority* than the beliefs in rationality; that is, in a contingency which some player would not have allowed under the beliefs in the agreement and rationality of order, respectively, n and $n-1$,[4] the co-players abandon the belief in rationality of order n.[5] When instead the co-players want to retain all orders of belief in rationality that are per se compatible with the observed behavior and rather drop the orders of belief in the agreement that are at odds with them, *selective rationalizability* is the appropriate tool (see Catonini [13]). Since the agreement originates from just cheap talk among players, I suggest that selective rationalizability is the appropriate solution concept in most situations. However, for robustness purposes and theoretical insight, it will be useful to evaluate agreements with both rationalizability concepts. Strong-Δ-rationalizability and selective rationalizability deliver either an empty set (if believing in the agreement is not compatible with the strategic reasoning hypotheses) or the possible behavioral implications of the agreement (otherwise). Some behavioral implications of a credible agreement may be inconsistent with the agreement itself: a player may or may not comply with the agreement depending on her conjectures where the incomplete agreement and strategic reasoning do not uniquely pin down the moves of the co-players.[6]

For notational simplicity and for the proofs of some results, the focus is restricted to the class of finite games with complete information and observable actions.[7] However, the methodology can be applied to all dynamic games with a countable set of non-terminal histories[8] and perfect recall, hence possibly infinite horizon.[9] Which

[3] The observed behavior may contradict also some order of belief in rationality per se.

[4] For $n = 1$ I mean rationality itself. The first-order-belief in the agreement has bite only when associated with rationality. Likewise, the belief in the agreement of order n has bite only when associated with the belief in rationality of order $n-1$. See Section 3 for details.

[5] All the beliefs in the agreement of higher orders can be mantained, since they have no behavioral consequence without the belief in rationality of the lower order, so they cannot be falsified by observation. This remark is due to Battigalli and Prestipino [8].

[6] Or when indifferences apply.

[7] Games where every player always knows the current history of the game, i.e. - allowing for truly simultaneous moves - information sets are singletons. For instance, all repeated games with perfect monitoring are games with observable actions.

[8] The limitation to a countable set of non-terminal histories allows to use Conditional Probability Systems (see Section 3), which require a countable set of conditioning events.

[9] Battigalli and Prestipino [8] have the most general analysis of strong-Δ-rationalizability in finite

agreements are *credible* and will be complied with? Which outcomes of the game can be achieved through some agreement? To answer these questions, the concepts of *self-enforceability* (of agreements) and *implementability* (of outcomes) are defined.

First, a very natural class of agreements is analyzed. A *path agreement* does not attempt to prescribe behavior at contingencies that follow a violation of the agreement itself. Thus, it just corresponds to agreeing on an outcome to be achieved. Example 2 in Section 2 analyzes a path agreement. Independently of the epistemic priority hypothesis, only for a strongly rationalizable,[10] subgame perfect equilibrium (henceforth, SPE) outcome, the corresponding path agreement can be self-enforcing, but (unless it is the only one) not even its credibility is guaranteed.

What about non-subgame perfect equilibrium outcomes? The corresponding path agreement may be credible too. Moreover, a non subgame perfect equilibrium outcome may be implementable with appropriate off-the-path threats (whereas a SPE one may not). Example 1 in Section 2 is a case in point. For two-players game, I provide a full characterization of implementable outcomes: an outcome is implementable under epistemic priority to rationality (resp., to the agreement) if and only if it is induced by a "strict"[11] equilibrium in strongly rationalizable (resp., sequentially rational) strategies. The incompleteness of the agreement can be crucial to implement an equilibrium outcome, also when players are more than two. The last examples of Section 4.1 and Section 4.2 illustrate how.

At this point, one may wonder whether there always exists (the support of) a SPE outcome (distribution) that can be implemented by some agreement. The answer is negative under the assumption that players do not agree on mixed actions. However, a preliminary investigation in this direction leads to a result of broader interest: in every game with observable actions, there are always a SPE and an equilibrium in strongly rationalizable strategies which induce the same distribution over outcomes.[12] That is, backward induction (plus equilibrium reasoning)[13] and forward induction (as

games: incomplete information, imperfectly observable actions, chance moves. Selective rationalizability can be extended in the same direction too.

[10] Strong rationalizability is a modification of Extensive Form Rationalizability (Pearce, [27]), which Battigalli and Siniscalchi [10] use to characterize the behavioral implications of Rationality and Common Strong Belief in Rationality.

[11] i.e.,without best replies to the equilibrium conjecture which would induce a different outcome.

[12] In games with perfect information and no relevant ties, strong rationalizability yields as unique outcome the SPE one, as proved by Battigalli and Siniscalchi [9] and Heifetz and Perea [23].

[13] Chen and Micali [14] already proved that backward induction without equilibrium reasoning, i.e. purely as an elimination procedure, has an overlap with forward induction in terms of predicted

captured by strong rationalizability)[14] never give disjoint predictions.

While the rationalizability literature provides concepts and tools for the analysis, the equilibrium refinement literature was the first to introduce forward induction considerations into equilibrium reasoning. Stable equilibria by Kohlberg and Mertens [24] (henceforth K&M) is a set-valued solution concept that captures instances of forward induction triggered here by a path agreement. Osborne [26] shows that *equilibrium paths that can be upset by a convincing deviation*[15] are not stable. Govindan and Wilson [20] refine sequential equilibria with a weaker notion of forward induction.[16] However, these works have two common shortcomings. First, they never question subgame perfection as a must-have for a "strategically stable" solution.[17] Second, the strategic reasoning that leads to play (or reject) such equilibria is unclear (or limited). The rationalizability approach adopted in this paper, which is backed by epistemic foundations, allows to eliminate both shortcomings. First, there is no constraint about how precisely and on which kind of equilibrium or disequilibrium behavior players agree. Second, there is transparency about which particular agreements, beliefs and epistemic assumptions trigger different lines of reasoning.

In this sense, this work can also be interpreted as the axiomatic realization of a program akin to K&M (see [24], p. 1020).[18] Epistemic priority to rationality implies,

outcomes. In an earlier paper, Battigalli [4] proves the same result in a smaller class of games.

[14]Strong-rationalizability captures "unrestricted" forward induction, based on the beliefs in rationality only. The existence of such Nash allows to easily claim the existence, at least in 2-players games, of an implementable SPE outcome distribution with agreements on mixed actions. In this sense, also "restricted" forward induction based on the agreement and subgame perfection do not give disjoint predictions. In 2-players games, Govindan and Wilson [20] prove the existence of sequential equilibria which capture a weak notion of forward induction based on the beliefs in the equilibrium path, see footnote 16.

[15]Particular sequences of equilibria of the stage game of a finitely repeated game. I characterize them as non credible agreements in the Online Appendix.

[16]In the refined sequential equilibria, players believe in strategies of the opponent which are best replies to a weakly sequential equilibrium with the same outcome distribution, but without restrictions on beliefs. Thus, beliefs in rationality and in the path above the second order are not captured. (I conjecture that second order beliefs are, whereas the authors observe explicitely strong belief in rationality only.)

[17]In some cases, the actual realization of K&M's ideal program departs from subgame perfection. Yet, in the attempt to capture it, valuable equilibria are disregarded: all the non-SPE outcomes implemented in the examples of this paper are unstable, although they also pass the credibility test, so they are compatible with forward induction a la K&M (see this paragraph). K&M regard the inability to imply subgame perfection as a weakness of stability, and "hope that in the future some appropriately modified definition of stability will, in addition, imply connectedness and backwards induction." This paper suggests the opposite direction.

[18]K&M write: "We agree that an ideal way to discuss which equilibria are stable, and to delineate

in generic games, iterated weak dominance.[19] Implementable outcomes are proved (and not assumed) to be admissible and Nash. Full-fledged forward induction reasoning is captured and clarified.[20] Agreements on *pure* actions provide clear motivation and intuitive implementation, whereas stability requires hard-to-guess mixed strategies even for the most intuitive outcomes. Finally, I do not disregard non subgame perfect equilibrium outcomes. Players are not necessarily required to coordinate on expected behavior after a deviation. A credible threat must be optimal against a plan of the deviator which is compatible with what the deviation signals (after all possible steps of reasoning) about her future intentions. In the equilibrium literature, forward induction is based on the following interpretation of a deviation: the deviator believed in the path (i.e. that the co-players would have followed the path), but does not believe in the threat. But then, that the threat is a best reply to a plan of the deviator which is a best reply to the threat itself is of no additional value. Moreover, threats compatible with both forward induction and subgame perfection typically do not exist (see the first example of Section 4.1): if a threat compatible with forward induction is made explicit, then that *another* threat is subgame perfect is of no additional value either. If the threat is not explicit, subgame perfection will be obtained as a result and not as an assumption (see Theorem 2), but very few SPE paths are self-enforcing. Selective rationalizability is based on an agnostic interpretation of deviations: when a rational player displays disbelief in the agreement, the co-players are free to assume that the agreement has been believed to any partial extent compatible with the observed behavior.[21] In addition, a "theoretical" rather than "literal" use of path agreements, in the fashion of the Iterated Intuitive Criterion by Cho and Kreps [15], allows to evaluate the robustness of threats against the interpretation above (see Example 2 in Section 2 for clarifications).[22] The implementation of all the outcomes

this common feeling, would be to proceed axiomatically. However, we do not yet feel ready for such an approach; we think the discussion in this section will abundantly illustrate the difficulties involved." Thirty years later, the achievements of epistemic game theory allow to overcome many of these difficulties.

[19] In generic games, iterated weak dominance is equivalent to strong rationalizability (see Battigalli and Siniscalchi [10]), of which selective rationalizability is a refinement.

[20] In K&M, forward induction is defined on the normal form. This makes it hard to understand to what extent forward induction reasoning is captured, especially at information sets that do not immediately follow a unilateral deviation from the equilibrium path, and for many steps of reasoning.

[21] Strong-Δ-rationalizability assumes instead that the agreement is always fully believed.

[22] Selective rationalizability can be modified to capture a finer epistemic priority ordering, with the beliefs in the compliance with the path in between the beliefs in rationality and the beliefs in the threats. This would automatize such robustness control.

in the examples of this paper is robust in this sense.

To explain intuitively these ideas, Section 2 discusses two examples (formally solved in the Online Appendix). In the first, players can profitably agree on a non subgame perfect equilibrium of the game, also when they assign priority to rationality. In the second, players would like to achieve a desirable SPE outcome, but forward induction reasoning makes the corresponding agreement not credible. Section 3 introduces the theoretical framework and the analytic tools for the formal treatment of Section 4. Section 5 concludes with directions for future research. The proofs of the theorems where observability of actions and finiteness are actually used are postponed to the Appendix.

2 Main examples

Example 1 In a city, two parties can form a coalition for the election of the mayor if they choose the same positioning on a few issues. If they both choose a *R*adical positioning, their coalition will win and split equally a surplus of 10. If they both choose a *M*oderate positioning, their coalition will win and the surplus to split grows to 12. The problem is that Party 2 may be tempted to take a radical positioning even if Party 1 chooses a moderate one. In this case, P_2 would not win at the first round but would acceed to the second round (a ballot) against a third candidate. In the last public speeches, P_1 can ask its voters to *S*upport P_2 or the *A*lternative candidate and P_2 can make a political *O*ffer to P_1's voters, which costs 2, or *N*ot. P_2 wins for sure if P_1 supports and with probability 1/2 if not but the offer is made. P_1 earns 2 if the supported candidate wins, unless P_2 wins without making the offer.

$P_1 \backslash P_2$	M	R
M	$(6,6)$	$\cdot-$
R	$(0,0)$	$(5,5)$

$- - - \quad - \rightarrow$

$P_1 \backslash P_2$	N	O
A	$(2,0)$	$(1,4)$
S	$(0,10)$	$(2,8)$

The subgame has only one equilibrium, where A and N are played with probability 1/3. Since the payoff of P_2 in this equilibrium is higher than 6, (M,M) is not a SPE outcome. Yet, it is the outcome of a Nash equilibrium where P_1 plays M and A and P_2 plays M and N.

The two parties meet before declaring their positioning to the public and try to reach an agreement. Can they credibly agree on the Nash equilibrium? The answer is yes and it is robust to the two different epistemic priority assumptions. Suppose that P_2 deviates to R. Is it credible that P_1 replies with A? A rational P_2 believing in the agreement would not have played R. Hence P_1 must either believe that P_2 does not believe in the agreement or that it is irrational. If P_1 drops the belief that P_2 is rational (epistemic priority to the agreement), it can expect any move and A is a best reply to N. If P_2 believes that P_1 reasons in this way, it can believe that P_1 would reply to the deviation with A. The agreement is credible and, once believed, players will play (M, M), consistently with the agreement. If P_1 drops the belief that P_2 believes in the agreement (epistemic priority to rationality), it can expect any rational move: $(R.N)$ is a best reply to all conjectures that put probability 1 on S and A is a best reply to N in the subgame.[23] Again, if P_2 believes that P_1 reasons in this way, it can believe that P_1 would reply to the deviation with A; so the agreement is credible and players will play (M, M). That is, the complete agreement on the Nash profile is self-enforcing. To keep the games simple, situations where the incompleteness of the agreement is instead necessary to implement a desirable Nash outcome are provided in the last examples of Section 4.1 and 4.2.

Note that the threat is credible also if P_2, after the deviation, is assumed to be rational and have believed in the path (i.e. that P_1 would have played M) although not in the threat. Rationalizing the deviation in this way, P_1 can expect P_2 to play N and hence best reply with A. Expecting this rationalization, P_2 can believe in A and hence not deviate.

The agreement on the SPE path (R, R) is self-enforcing. In other cases, forward induction reasoning based on the beliefs in the path may rule out the off-the-path beliefs that are necessary to prevent a deviation.[24] This is the case in Example 2.

Example 2 The duopolists of the cola market, A and B, have to decide their marketing strategy before two big sport events, for which the population will gather

[23] Also $(R.O)$ is rational. But S is best reply to O and N is best reply to S, so we can run in circle and the incentive to play A remains compatible with the beliefs in rationality of all orders.

[24] Forward induction arguments can also go in favour of the agreement. Suppose that after a unilateral deviation from a path agreement, the deviator can get a higher payoff than under the path only if the co-player plays a certain action. Suppose that this action is best reply only to an action that prevents the deviator to get a higher payoff than under the path. Then this action will not be played and the possibility of deviation is ruled out via forward induction.

in front of the television. There are 10 million buyers: 2 of them are somewhat loyal to brand A, 2 of them are somewhat loyal to brand B, the others just follow the advertisments before the event (if both or none brands do it, they split equally). At the current price, each million of buyers brings a profit of 1. Advertising costs 2. There is also another marketing strategy, which consists of a discount in the supermarkets before the event. Every buyer switches to the brand making discounts, but the profit drops to 0.2 per million buyers.

The game is a twice repeated prisoner dilemma with a punishment action, because advertisement (D) is a best reply to both no advertisment (C) and advertisment, while the aggressive discount campaign (P) is best reply only to the other firm doing the same, and the profits of both firms fall anyway.

$A\backslash B$	C	D	P
C	$5,5$	$2,6$	$0,2$
D	$6,2$	$3,3$	$0,2$
P	$2,0$	$2,0$	$1,1$

There is a SPE where the two firms collude in the first stage. Suppose that the two marketing directors, Ann and Bob, agree not to advertise their products before the first event and to do it before the second event. It is understood that the agreement falls through if it was violated for the first event. All this sums up to agreeing on the SPE path.

The agreement does not rule out punishment P after a deviation; therefore, it seems possible (although not guaranteed) that players fear it and comply with the path. Instead, by forward induction reasoning, punishment is actually ruled out and thus the path agreement is not credible. Bob, if he is rational and believes that Ann will comply with the agreement, will play D in the first stage only if he does not expect P in the second. Ann, if she believes this and observes D, will then (by forward induction) expect Bob to play D again and play D herself. Bob, if he believes that Ann would interpret the deviation in this way, will actually play D in the first stage. Ann, if she believes that Bob expects such interpretation, will predict the deviation and not believe in the agreement.

Two objections may be raised at this point.

First, Ann could interpret the deviation as follows: "Bob believed that I would

have not complied with the agreement, and best replied by not complying himself." But then, if the beliefs of Ann are dynamically consistent (as I will require), she must believe that Bob does not trust her from the start: the deviation of Bob, assuming he is rational, is not at odds with the belief that Ann complies with the agreement.

Second, anticipating that the path agreement would not work, players could explicitely threaten the punishments beforehand. In the next section I propose a few reasons why this may not happen. However, even if players reach the complete agreement, the non credibility of the path agreement could make the belief in the complete agreement fall for the following reason. Suppose that after a deviation, the co-players pose themselves the following dilemma: the deviator did not believe in the agreement on path (i.e. she expected someone else to deviate) or does not believe in our post-deviation threat? If the co-players rationalize the deviation in the second way, the instances of forward induction reasoning triggered by a path agreement take place.

Battigalli and Siniscalchi ([11] and [12]) show that the Iterated Intuitive Criterion test is equivalent to non-emptiness of strong-Δ-rationalizability under the restrictions that correspond to the equilibrium outcome distribution. Theorem 1 in Section 4.1 will prove that non-emptiness of selective rationalizability and of strong-Δ-rationalizability are equivalent for path restrictions. Therefore, the "credibility of the path agreement" test can be seen as a generalization[25] of the Iterated Intuitive Criterion, and provides for it the motivation above. Moreover, only strongly rationalizable outcomes can pass the test (see Proposition 1). The test is applied to "equilibrium paths that can be upset by a convincing deviation" in the Online Appendix, and it gives negative response. When the test gives positive response, it also provides conjectures under which players comply with the path and which are compatible with the rationalization of deviations discussed above. From such conjectures, it *may* be possible to derive an agreement that implements the outcome. This possibility is discussed in Section 4.2.

[25] Once extended to outcome distributions and games with incomplete information.

3 Agreements, beliefs and strategic reasoning

3.1 Preliminaries

Let I be the finite set of *players*. For any collection $(X_i)_{i\in I}$ and $J \subseteq I$, I write $X_J := \times_{j\in J} X_j$, $X := X_I$ and $X_{-i} := X_{I\setminus\{i\}}$; for any profile $x \in X$, I write $x_J := \text{Proj}_{X_J} x$.

Let $(\overline{A}_i)_{i\in I}$ be the finite sets of *actions* potentially available to each player and $\overline{H} \subseteq \cup_{t=1,...,T} \overline{A}^t \cup \{\emptyset\}$, where T is the finite horizon of the game, the set of *histories*, with the following properties:

- $h^0 := \{\emptyset\} \in \overline{H}$ (root of the game);
- for any $h = (a^1, ..., a^t) \in \overline{H}$ and $l < t$, $h' = (a^1, ..., a^l) \in \overline{H}$ and I write $h' \prec h$;[26]
- calling $Z := \{z \in \overline{H} : \forall h \in \overline{H}, z \not\prec h\}$ the set of *terminal histories* and $H := \overline{H}\setminus Z$ the set of *non-terminal histories*, for every $i \in I$ there exists a non-empty-valued correspondence $A_i : H \rightrightarrows \overline{A}_i$ such that for every $h \in H$, $(h, a) \in \overline{H}$ if and only if $a \in A(h)$.

Finally, for every $i \in I$, let $u_i : Z \to \mathbb{R}$ be the payoff function; $\Gamma = \left\langle I, \overline{H}, (u_i)_{i\in I} \right\rangle$ is a *finite game with complete information and observable actions*.

Now I can derive the following objects. A *strategy* is a function $s_i : H \to \overline{A}_i$ such that for every $h \in H$, $s_i(h) \in A_i(h)$. The set of all strategies is denoted by S_i. The set of strategies that are compatible with a history $h = (a^1, ..., a^t) \in \overline{H}$ is defined as:

$$S_i(h) := \left\{ s_i \in S_i : s_i(h^0) = a_i^1 \wedge \forall l < t, s_i((a^1, ..., a^l)) = a_i^{l+1} \right\}.$$

For any $\overline{S}_i \subset S_i$, $\overline{S}_i(h) := S_i(h) \cap \overline{S}_i$ and $\overline{S}_i[h] := \{a_i \in \overline{A}_i : \exists s_i \in \overline{S}_i(h), s_i(h) = a_i\}$.

On the other hand, the set of histories that are compatible with a set of strategy (sub-)profiles $\overline{S}_J \subseteq S_J$ is defined as $\overline{H}(\overline{S}_J) := \left\{h \in \overline{H} : \overline{S}_J(h) \neq \emptyset\right\}$. Then, the sets of terminal histories and non-terminal histories that are compatible with $\overline{S}_J$ are respectively $\zeta(\overline{S}_J) := \overline{H}(\overline{S}_J) \cap Z$ and $H(\overline{S}_J) := \overline{H}(\overline{S}_J) \cap H$.

Given a strategy $s_i \in S_i$, the set of realization equivalent strategies is $[s_i] := \{s_i' \in S_i : \forall h \in H(s_i), s_i'(h) = s_i(h)\}$. The immediate predecessor of a history $h \in \overline{H}\setminus\{h^0\}$ is denoted by $p(h)$.[27]

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

[27] The history h' such that $h = (h', a)$ for some $a \in \overline{A}$.

In games with observable actions, every player always knows the current history of the game. That is, *information sets* are singletons and coincide with histories. I will use the term information set instead of history when it is the appropriate one in games where the two do not coincide. Note moreover that to simplify notation every player is required to play an action at every history: when a player is not truly active at a history, her set of feasible actions consists of just one "wait" action.

3.2 Agreements, complete agreements and path agreements

I consider pre-play, non-binding *agreements* among players of the following form.

Definition 1 (Agreement) *An agreement is a profile of correspondences* $e = (e_i : H \rightrightarrows \overline{A}_i)_{i \in I}$, *such that for every* $i \in I$ *and* $h \in H$, $\emptyset \neq e_i(h) \subseteq A_i(h)$.

An agreement specifies at every information set the *pure* actions among which players are expected to choose. Note that the agreement can restrict the behavior of players also at information sets that are supposedly precluded by the agreement itself. When this is not the case and the agreement allows to reach only one outcome of the game, I call it a *path agreement*.

Definition 2 (Path Agreement) *Fix* $z = (a^1, ..., a^t) \in Z$. *A path agreement on* z *is an agreement* $e = (e_i)_{i \in I}$ *such that for every* $i \in I$, $e_i(h^0) = a_i^1$; *for every* $l < t$, $e_i((a^1, ..., a^l)) = a_i^{l+1}$; *for every* $h \not< z$, $e_i(h) = A_i(h)$.

Path agreements are particularly interesting for many reasons. First, they correspond to a very natural kind of agreement: choosing an outcome as a goal. Second, punishments are not explicitely threatened in many real-life situations: discussing what to do in case the partner defects is an inconvenient way to start a relationship. Third, players may anticipate that punishments would not be believed: the belief in an agreement that has already been violated is very likely to fall. Or, punishments may be believed only if compatible with the rationalization of deviations from the path discussed in the previous section. Both for its self-enforceability as real agreement and for its credibility as test, a path agreement has an additional desirable property: conclusions are robust to the epistemic priority choice (see Theorem 1).

A *complete agreement* assigns one action to each player at each information set. Thus, it can correspond to an entire (subgame perfect) equilibrium of the game.

The wide range of agreements between path and complete will turn out to be useful too. The next subsection provides the tools for the evaluation of any agreement.

3.3 Beliefs, rationality and rationalizability

The elementary belief that an agreement may be able to induce is that the co-players will comply with it. This belief can be represented as a restricted set of conjectures about which strategies the co-players are going to play. In this dynamic framework, beliefs are modeled as Conditional Probability Systems (Renyi, [29]; henceforth, CPS). Here I define CPS's directly for the problem at hand.

Definition 3 *A Conditional Probability System on* $(S_{-i}, (S_{-i}(h))_{h\in H})$ *is a mapping* $\mu(\cdot|\cdot) : 2^{S_{-i}} \times \{S_{-i}(h)\}_{h\in H} \to [0,1]$ *satisfying the following axioms:*

CPS-1 for every $C \in (S_{-i}(h))_{h\in H}$, $\mu(\cdot|C)$ *is a probability measure on* S_{-i}*;*

CPS-2 for every $C \in (S_{-i}(h))_{h\in H}$, $\mu(C|C) = 1$*;*

CPS-3 for every $E \in 2^{S_{-i}}$ *and* $C, D \in (S_{-i}(h))_{h\in H}$, *if* $E \subseteq D \subseteq C$, *then* $\mu(E|C)$=$\mu(E|D)\mu(D|C)$.

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

For brevity, the conditioning events will be indicated with just the information set, which represents all the information acquired by players through observation.

I say that players believe in the agreement if at every information set they believe in what the agreement prescribes at the subsequent ones. In the Online Appendix, an example shows the important reason why players are not required to believe in strategies that comply with the agreement at information sets which have not been reached and cannot be reached anymore.

Definition 4 (Belief in the agreement) *Fix an agreement* $e = (e_i)_{i\in I}$. *For every* $i \in I$, *let* Δ_i^e *be the set of all* $\mu_i = (\mu_i(\cdot|h))_{h\in H} \in \Delta^H(S_{-i})$ *such that for every* $h \in H$*:*

$$\text{Supp}\mu_i(\cdot|h) \subseteq \left\{(s_j)_{j\neq i} \in S_{-i}(h) : \forall j \neq i, \forall \widetilde{h} \succeq h, s_j(\widetilde{h}) \in e_j(\widetilde{h})\right\}.$$

Note that any two players have the same restrictions about any third player. For a path agreement on $z \in Z$, Δ_i^e coincides with the set of $\mu_i \in \Delta^H(S_{-i})$ such that $\mu_i(S_{-i}(z)|h^0) = 1$: path agreements will be treated keeping this specification in mind.

For any agreement $e = (e_i)_{i \in I}$ and each player $i \in I$, Δ_i^e is compact:[28] this allows to claim that selective rationalizability and strong-Δ-rationalizability with $(\Delta_i^e)_{i \in I}$ have the epistemic characterizations of [13] and, for instance, [8].

I consider players who reply rationally to their conjectures. By rationality I mean that players, at every information set, choose an action that maximizes expected utility given the conjecture about how co-players will play and the expectation to reply rationally again in the continuation of the game. This is equivalent (see Battigalli [3]) to playing a *sequential best reply* to the CPS.

Definition 5 *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 $h \in H(s_i)$, s_i is a continuation best reply to $\mu_i(\cdot|h)$, i.e. for every $\widetilde{s}_i \in S_i(h)$,*

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

I say that a strategy s_i is *rational* if it is a sequential best reply to some $\mu_i \in \Delta^H(S_{-i})$. The set of sequential best replies to μ_i is denoted by $\rho(\mu_i)$. The set of best replies to a conjecture $\nu_i \in \Delta(S_{-i})$ in the normal form of the game is denoted by $r(\nu_i)$. Note that if $s_i \in \rho(\mu_i)$, $[s_i] \subseteq \rho(\mu_i)$, and if $s_i \in r(\nu_i)$, $[s_i] \subseteq r(\nu_i)$. Moreover, the following relationship between continuation and sequential best replies holds: if s_i is a continuation best reply to $\mu_i(\cdot|h)$ and $h \in H(\rho(\mu_i))$, there exists $\widehat{s}_i \in \rho(\mu_i)(h)$ such that for every $\widetilde{h} \succeq h$, if $\widetilde{h} \in H(\text{Supp}\mu_i(\cdot|h))$, then $\widehat{s}_i(\widetilde{h}) = s_i(\widetilde{h})$. The reason is that a continuation best reply at h has to be a continuation best reply at every history h' that can be reached with positive probability from h, because the conjecture at h' is the conjecture at h conditional on h'.

Here I take the view that players refine further their conjectures through strategic reasoning based on beliefs in rationality and beliefs in the belief in the agreement. If players assign *epistemic priority to (the beliefs in) the agreement* (see the Introduction and [13]), the appropriate solution concept is *strong-Δ-rationalizability*. I translate its ultimate definition by Battigalli and Prestipino [8] in the framework of this paper.

Definition 6 (strong-Δ-rationalizability) *Fix an agreement $e = (e_i)_{i \in I}$.*

[28] For each $h \in H$, the set of probability measures that assign probability 1 to a particular subset of $S_{-i}(h)$ is compact in $\Delta(S_{-i})$, endowed with the topology of weak convergence. Thus, the set of arrays of probability measures that believe in the remainder of the agreement at every information set is compact in $(\Delta(S_{-i}))^H$, endowed with the product topology. Since Δ_i^e is the intersection between such set and the compact set $\Delta^H(S_{-i})$, it is compact itself.

(Step 0) *For every* $i \in I$, *let* $S^0_{i,\Delta^e} = S_i$.

(Step $n > 0$) *For every* $i \in I$ *and* $s_i \in S_i$, *let* $s_i \in S^n_{i,\Delta^e}$ *if and only if there exists* $\mu_i \in \Delta^e_i$ *such that:*

D1 $s_i \in \rho(\mu_i)$

D2 $\forall q = 0, ..., n-1$, $\forall h \in H$, $S^q_{-i,\Delta^e}(h) \neq \emptyset \Rightarrow \mu_i(S^q_{-i,\Delta^e}|h) = 1$ *(i.e.* μ_i strongly believes S^q_{-i,Δ^e}*);*

Finally let $S^\infty_{i,\Delta^e} = \cap_{n \geq 0} S^n_{i,\Delta^e}$. *The profiles in* $S^\infty_{\Delta^e}$ *are called strongly-*Δ*-rationalizable.*

Strong rationalizability (Battigalli and Siniscalchi, [10]) can be seen as a special case of strong-Δ-rationalizability where no restriction applies ($\Delta^e_i = \Delta^H(S_{-i})$) and it will be denoted by dropping the subscript Δ^e.

The n-th step of strong-Δ-rationalizability captures the first n assumptions below:

1. players are rational and believe that co-players will comply with the agreement (and believe that everyone else will comply with the agreement, and so on);
2. players believe that 1 holds as long as not contradicted by observation (i.e. at every information set that can be reached if 1 holds);
3. players believe that 1 and 2 hold as long as not contradicted by observation;
4. ...

The sentence in brackets in assumption 1 means that players commonly believe in the agreement at every information set, regardless of the compatibility with beliefs in rationality of any order. This is not a necessary assumption to characterize strong-Δ-rationalizability: the belief in the agreement of order n has no behavioral implication once the belief in rationality of order $n-1$ is dropped. Both epistemic characterizations can be found in [8].

Strong-Δ-rationalizability does two things.

First, it constitutes a compatibility test for the belief in the agreement with the strategic reasoning hypotheses. If strong-Δ-rationalizability delivers an empty set, the agreement does not pass the test. This happens when a player at some step allows an information set only with strategies that do not comply with the remainder

of the agreement, hence they are not compatible with the first-order-belief restrictions of the co-players.

Second, if the agreement passes the test, the strongly-Δ-rationalizable strategy profiles coincide with the behavioral implications of rationality, (common) belief in the agreement, and common strong belief in, jointly, rationality and (common) belief in the agreement.

If players assign *epistemic priority to rationality* (see the Introduction and [13]), the appropriate solution concept is *selective rationalizability* [13].

Definition 7 (selective rationalizability) *Fix an agreement* $e = (e_i)_{i \in I}$. *Let* $((S_j^m)_{j \in I})_{m \geq 0}$ *be the strong rationalizability procedure.*

(Step 0) *For every* $i \in I$, *let* $S^0_{i,R\Delta^e} = S_i$.

(Step $n > 0$) *For every* $i \in I$ *and* $s_i \in S_i$, *let* $s_i \in S^n_{i,R\Delta^e}$ *if and only if there exists* $\mu_i \in \Delta_i^e$ *such that:*

S1 $s_i \in \rho(\mu_i)$;

S2 $\forall q = 0, ..., n-1$, $\forall h \in H$, $S^q_{-i,R\Delta^e}(h) \neq \emptyset \Longrightarrow \mu_i(S^q_{-i,R\Delta^e}|h) = 1$;

S3 $\forall q \geq 0$, $\forall h \in H$, $S^q_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(S^q_{-i}|h) = 1$.

Finally let $S^\infty_{i,R\Delta^e} = \cap_{n \geq 0} S^n_{i,R\Delta^e}$. *The profiles in* $S^\infty_{R\Delta^e}$ *are called selectively-rationalizable.*

The n-th step of selective rationalizability captures the first n assumptions below:

1. players believe in the agreement, are rational, believe that co-players are rational (as long as not contradicted by observation), and so on;
2. players believe that 1 holds as long as not contradicted by observation;
3. players believe that 1 and 2 hold as long as not contradicted by observation;
4. ...

The first step already captures common strong belief in rationality [10]. This is necessary to guarantee that at every information set players give epistemic priority to the beliefs in rationality of all order which are compatible with the information

set (see the Discussion section of [13] for details). Hence, selective rationalizability refines strong rationalizability by selecting the strategies that are compatible with the beliefs in the agreement and consequent forward induction reasoning.

Selective rationalizability accomplishes the same two tasks of strong-Δ-rationalizability. A non-empty set of selectively-rationalizable strategy profiles coincides with the behavioral implications of (i) rationality and common strong belief in rationality, (ii) belief in the agreement, and common strong belief in, jointly, (i) and (ii).

As already argued, it seems more plausible that players will assign epistemic priority to rationality in presence of a non-binding agreement. However, for robustness purposes and theoretical insight, the next section evaluates agreements with both rationalizability tools.

4 Self-enforceability and implementability

4.1 Credibility and self-enforceability

In order to evaluate a given agreement, two features have to be investigated. First, whether the agreement is credible or not. Second, if the agreement is credible, whether players will comply with it or not.

An agreement is *credible* if it passes the appropriate rationalizability test. For brevity, the generic rationalizability procedure will be indicated with the subscript e; it has to be replaced with Δ^e or $R\Delta^e$ to obtain the correct expressions under the chosen epistemic priority assumption.

Definition 8 (Credibility) *An agreement $e = (e_i)_{i \in I}$ is credible if $S_e^\infty \neq \emptyset$.*

Credibility does not imply that players will comply with the agreement, but only that they may do so. If players, once they refine their conjectures according to the agreement, always have the strict incentive to comply with it, the agreement is *self-enforcing.*

Definition 9 (Self-enforceability) *An agreement $e = (e_i)_{i \in I}$ is self-enforcing if it is credible and for every $i \in I$, $s_i \in S_{i,e}^\infty$ and $h \in H(S_e^\infty) \cap H(s_i)$, $s_i(h) \in e_i(h)$.*

The definition requires explicitely that *every* rationalizable strategy complies with the agreement at the information sets which can be reached under (itself and) the

rationalizable strategy profiles, so that violation of the agreement will actually occur. Credibility implies that players also believe in the compliance with the agreement at the other information sets, so that the desired behavioral consequences apply.

One might think that a self-enforcing agreement under priority to rationality will be, a fortiori, self-enforcing under priority to the agreement. An example in the appendix of [13] shows that this is not the case. Instead, for a path agreement credibility and self-enforceability are robust to the epistemic priority assumption.

Theorem 1 *A path agreement is credible/self-enforcing under priority to rationality if and only if it is credible/self-enforcing under priority to the agreement.*

Hence, all the conditions for credibility and self-enforceability of path agreements can be stated regardless of the epistemic priority assumption, and the proofs can rely on either selective rationalizability or strong-Δ-rationalizability.

The first condition claims that a path can be credible only if it is induced by some strongly rationalizable strategy profile.[29]

Proposition 1 *Fix $z \in Z$. If $z \notin \zeta(S^\infty)$, the corresponding path agr. is not credible.*

Proof. Selective rationalizability is a refinement of strong rationalizability. ■

Thus, only a strongly rationalizable outcome can pass the credibility test.[30]

This necessary condition for credibility becomes sufficient for self-enforceability when the path is the only strongly rationalizable one.

Proposition 2 *If $\zeta(S^\infty)$ is a singleton, the corresponding path agr. is self-enforcing.*

Proof. Let $\zeta(S^\infty) = \{z\}$. For every $i \in I$, $S_i^\infty \subseteq S_i(z)$. Thus, for every $j \in I$ and $\mu_j \in \Delta^H(S_{-j})$ that satisfies S3, $\mu_j(S_{-j}(z)|h^0) = 1$, so $\mu_j \in \Delta_j^e$. Hence, the restrictions are immaterial, so that selective rationalizability coincides with strong rationalizability after convergence.[31] ■

[29] Moreover, results of Battigalli and Siniscalchi [11] imply that only a self-confirming equilibrium outcome (Fudenberg and Levine [18]; called conjectural equilibrium in Battigalli [2]) can pass the credibility test. However, for implementability the outcome will need to be Nash.

[30] By virtue of Remark 7 in the Appendix, this property holds true if the test is extended to distributions over outcomes rather than just one outcome. The extension to incomplete information games would prove that the Iterated Intuitive Criterion selects strongly rationalizable outcomes.

[31] Selective rationalizability without actual restrictions could be empty when there exists $i \in I$ and $s_i \in S_i^\infty$ such that for every μ_i that strongly believes $(S_{-i}^n)_{n\geq 0}$, $s_i \notin \rho(\mu_i)$. This is never the case under finiteness or mild regularity conditions.

Analogously, while SPE paths can be self-enforcing or not, non SPE paths can be credible but not self-enforcing.

Theorem 2 *Fix $z \in Z$. If the corresponding path agreement is self-enforcing, then there exists a SPE*[32] $(\sigma_i)_{i \in I} \in \times_{i \in I} \Delta(S_i)$ *such that for every $i \in I$, $\sigma_i(S_i(z)) = 1$.*

Differently than for strongly rationalizable paths, when the game has only one SPE path, it needs not be self-enforcing.

$$
\begin{array}{ccccc}
 & A & - & - & \rightarrow \\
O & \downarrow & & I & \\
 & (4,\cdot) & & &
\end{array}
\qquad
\begin{array}{|c|c|c|c|}
\hline
A/B & L & C & R \\
\hline
U & (5,0) & (0,5) & (0,0) \\
\hline
M & (0,5) & (5,0) & (0,0) \\
\hline
D & (3,0) & (3,0) & (3,3) \\
\hline
\end{array}
$$

In the subgame, there is no equilibrium where both U and M are played: if Ann is indifferent between them, she prefers D. If U was played and M not, Bob would not play L, so U cannot be a best reply, and analogously for M. Thus, the unique equilibrium is (D, R) and it induces Ann to choose the outside option. Yet, Ann may be sufficiently confident of L or C to enter the subgame and play U or M.[33] This also implies that the subgame perfect threat R is not credible, being a best reply only to D. The path is however credible. It can be shown that a path is credible when it is the unique SPE path or, more generally, when it is robust to any choice of off-the-path equilibria:[34] this is the notion of credibility used by Gossner [19] in his work on incomplete codes.

When SPE paths are more than one, none of them may be even credible.[35] Consider the twofold repetition of the following game. The players must perform a task that yields a profit of 3 to each of them at the total effort cost of 2. If at least one

[32] SPE in mixed strategies are defined in the Appendix, using additional notation. For every SPE in mixed strategy there is a SPE in behavioral strategies that induces the same outcome distribution, thus the Theorem holds also with SPE in behavioral strategies.

[33] So that Bob in turn may reply with L or C and all 4 actions remain rationalizable.

[34] For a path not to be credible, some player at some step of the rationalizability procedure must have the incentive to abandon it for every reaction she may expect to her deviation. The inner recursive step in the proof of Lemma 10 in the Appendix shows that in such case a (probabilistic) SPE path of the subgame would have survived until that step. This shows that the non credible path cannot fall in the class I am considering.

[35] I conjecture that with agreements on mixed actions, a credible SPE outcome distribution exists. This would extend the main result of Govindan and Wilson [20] to full-fledged forward induction reasoning.

player works, the task is performed; if only one player works, she pays the total cost of effort; if they both work, they share the effort cost equally.

$A \backslash B$	$Work$	$FreeRide$
W	$2, 2$	$1, 3$
FR	$3, 1$	$0, 0$

No deterministic SPE path is even credible. If the path prescribes the same Nash in both stages, the unhappy player can signal with a deviation the intention to switch to the preferred equilibrium in the second stage. If the path prescribes to play one Nash in the first stage and the other Nash in the second stage, the player whose preferred equilibrium is played in the first stage can deviate from it to signal the intention to play it in the second stage. Similar paths have been already classified by Osborne [26] for 2-players, finitely-repeated games as *equilibrium paths that can be upset by a convincing deviation*; in the Online Appendix, the definition and the formal reasoning that makes them not credible are provided.

For complete agreements the equivalence between self-enforceability under priority to the agreement and under priority to rationality does not hold. When the epistemic priority falls on the agreement, a "strict" SPE is self-enforcing. I say that an equilibrium $s \in S$ is *strict* when for every $i \in I$, $r_i(s_{-i}) \subseteq S_i(\zeta(s))$. The proof of this result is in the Online Appendix.

When the epistemic priority falls on rationality, instead, not all SPE are credible agreements. A SPE outcome may not be strongly rationalizable, hence also not selectively rationalizable. Moreover, even if the outcome is strongly rationalizable, the SPE strategy profile may be not.[36] Yet, the outcome may be induced also by some Nash equilibrium in strongly rationalizable strategies. Or, like in the first example of Section 2, players may be interested in a non subgame perfect equilibrium. For Nash equilibria, also when the priority falls on the agreement, the complete agreement may not work, because an off-the-path threat may not be optimal against any clear-cut intention of the deviator or of a third player. Here I illustrate an example of the first

[36] Even in a perfect information game like the variation of the centipede in Reny [28], the agreement corresponding to the only SPE is not credible, because among the strongly rationalizable strategy profiles, despite inducing the SPE outcome, there is not the SPE one (forward induction and backward induction moves do not coincide)

kind; an example of the second kind is provided at the end of the next subsection.

$$\begin{array}{cccccc} & Ann & - & - & \longrightarrow & \\ O & \downarrow & & I & & \\ & (4,4) & & & & \end{array}$$

$A\backslash B$	L	C	R
U	$(9,0)$	$(8,3)$	$(1,2)$
D	$(8,3)$	$(9,0)$	$(0,2)$

In this game, O is not a SPE outcome because in any equilibrium of the subgame, R cannot be played with probability higher than $1/2$ (otherwise Ann would play U and Bob C). The Nash equilibria $(O.U, R)$ and $(O.D, R)$ are not self-enforcing agreements: whatever Ann promises to do in case she deviates to I, R is not a best reply. Yet, if Bob refuses to "listen" to Ann's post-deviation promises, the incomplete agreement on O and R is self-enforcing (regardless of the epistemic priority assumption).

Therefore, when players want to implement some equilibrium of the game and are prone to discuss and believe off-the-path threats, it is more appropriate to tackle the problem through the alternative perspective: can the outcome be implemented by *some* agreement? This issue is addressed in the next subsection.

4.2 Implementability

Taking the opposite, implementation perspective, which outcomes of the game can be achieved through *some* agreement? I say that a set of outcomes is *implementable* if for some agreement the appropriate rationalizability procedure delivers a non-empty subset of it.

Definition 10 (Implementability) *A set of outcomes $P \subset Z$ is implementable if there exists a credible agreement $e = (e_i)_{i \in I}$ such that $\zeta(S_e^\infty) \subseteq P$.*

In the Online Appendix I discuss the relationship between implementability and two much weaker but somehow related solution concepts: *Extensive Form Best Response Sets* (Battigalli and Friedenberg, [7]) and *Mutually acceptable courses of actions* (Greenberg et al., [21]).

Focusing on a single outcome, one should first check if the corresponding path agreement is self-enforcing.[37] If this is not the case, regardless of the epistemic priority assumption, implementability by loosening restrictions can be excluded.

[37] Given the results of Section 4.1, one can hope so only for a strongly rationalizable, SPE outcome.

Theorem 3 *Fix $z \in Z$ and the corresponding path agreement $e = (e_i)_{i \in I}$. If z is implemented by some agreement $e' = (e'_i)_{i \in I}$ such that for every $i \in I$ and $h \in H$, $e_i(h) \subseteq e'_i(h)$, then $e = (e_i)_{i \in I}$ is self-enforcing.*

Thus, if players want to reach a given outcome, without willing or trusting the possibility to coordinate in case of deviation, they cannot do any better than agreeing on the path. This is a sort of "revelation principle" for agreements design: leaving some mystery about on-the-path moves cannot be of any help for the goal.

Allowing for off-the-path restrictions, is it possible to implement a non-equilibrium outcome? The answer is negative under both epistemic priority assumptions. For the remainder of this section, the needed additional arguments for the priority to rationality/selective rationalizability case are provided in square brackets.

Proposition 3 *Fix $z \in Z$. If z is implemented by $e = (e_i)_{i \in I}$, then there exists a strict Nash equilibrium $s \in S_e^\infty$ such that $\zeta(s) = z$.*

Proof. Fix $i \in I$**.** By credibility there exists $\mu_i \in \Delta_i^e$ such that $\mu_i(S_{-i,e}^\infty | h^0) = 1$.[38] So, for every $j \neq i$, there exists $s_j \in S_{j,e}^\infty$ such that for every $h \in H$, $s_j(h) \in e_j(h)$. By implementability $s_j \in S_j(z)$.

Fix $\mu_i \in \Delta_i^e$ that strongly believes $(S_{-i,e}^n)_{n=0}^\infty$ [and $(S_{-i}^n)_{n=0}^\infty$] such that $\mu_i(s_{-i}|h^0) = 1$. Suppose by contradiction that $\emptyset \neq r(\mu_i(\cdot|h^0)) \backslash S_i(z) \ni \widetilde{s}_i$. Then there would exist $\widehat{s}_i \in \rho(\mu_i)$ such that for every $h \in H(\widetilde{s}_i) \cap H(s_{-i})$, $\widetilde{s}_i(h) = \widehat{s}_i(h)$,[39] so that $\widehat{s}_i \notin S_i(z)$ too, contradicting implementability. So $s_i \in r(s_{-i}) = r(\mu_i(\cdot|h^0)) = S_i(z)$. ■

The condition is not sufficient. In the last example, (R_1, L_2) is a strict Nash outcome that cannot be implemented by any agreement, because it is sustained by R_3 and R_4 but R_3 is not rational against R_4. If players are more than two, off-the-path threats of two different players, even if rational, may be incompatible with each other. With two players, still, a threat could be rational but not optimal against any clear-cut intention of the deviator. In the second case, there are clear conditions under which the problem can be solved by leaving the agreement incomplete (see the end of the previous subsection for an example).

[38] This may be false when there exist $i \in I$ and $s_i \in S_{i,e}^\infty$ such that for every μ_i that strongly believes $(S_{i,e}^n)_{n \geq 0}$, $s_i \notin \rho(\mu_i)$. This is never the case under finiteness or mild regularity conditions.

[39] See the relationship between continuation best replies and sequential best replies in Section 3.

Proposition 4 *Consider a 2-players game and a strict Nash equilibrium* (s_1,s_2)*. If* s_1 *and* s_2 *are rational [strongly rationalizable],* $z := \zeta(s)$ *is implementable.*

Proof. For every $i \in I$ and $h \in H(s_i)$, let $e_i(h) = s_i(h)$, otherwise, let $e_i(h) = A_i(h)$. Since s_i is rational $[s_i \in S_i^\infty]$, there exists $\mu_i \in \Delta^H(S_{-i})$ [that strongly believes $(S_{-i}^n)_{n=0}^\infty$][40] such that $s_i \in \rho(\mu_i)$. For every $\overline{\mu}_i \in \Delta_i^e$, $\overline{\mu}_i([s_{-i}]\,|h^0) = 1$ and for every $s'_{-i} \in [s_{-i}]$, since by strictness $r(s_{-i}) \subseteq S_i(z)$, $r(s'_{-i}) \subseteq S_i(z)$ too. Hence, $S_{i,e}^1 \subseteq S_i(z)$, yielding implementability in case of credibility. Then, for every $i \in I$ and $m \in \mathbb{N}$ such that $s \in S_e^{m-1}$, there exists $\widetilde{\mu}_i \in \Delta_i^e$ that strongly believes $(S_{-i,e}^n)_{n=0}^{m-1}$ [and $(S_{-i}^n)_{n=0}^\infty$] such that $\widetilde{\mu}_i(s_{-i}|h^0) = 1$ and for every $h \notin H(S_{-i}(z))$, $\widetilde{\mu}_i(\cdot|h) = \mu_i(\cdot|h)$. For every $h \prec z$, $\widetilde{\mu}_i(s_{-i}|h) = 1$ and $s_i \in r(s_{-i})$; for every $h \in H(s_i)\backslash H(S_{-i}(z))$, s_i is a continuation best reply to $\widetilde{\mu}_i(\cdot|h)$ by construction; else, $h \notin H(S_i(z))$, so $h \notin H(s_i)$. Thus, $s_i \in \rho(\widetilde{\mu}_i) \subseteq S_{i,e}^m$. Inductively, $s \in S_e^\infty \neq \emptyset$. ■

Hence, in 2-players games I can fully characterize the set of implementable outcomes with equilibria in *pure* strategies as follows.

Theorem 4 *Consider a 2-players game. An outcome* $z \in Z$ *is implementable under priority to the agreement [rationality] if and only if there exists a rational [strongly rationalizable] strict Nash equilibrium* $s \in S$ *such that* $\zeta(s) = z$*.*

Proof. The "if" direction coincides with Proposition 4. The "only if" direction coincides with Proposition 3 once observed that if $s \in S_{\Delta^e}^\infty$, s is rational and if $s \in S_{R\Delta^e}^\infty \subseteq S^\infty$, s is strongly rationalizable. ■

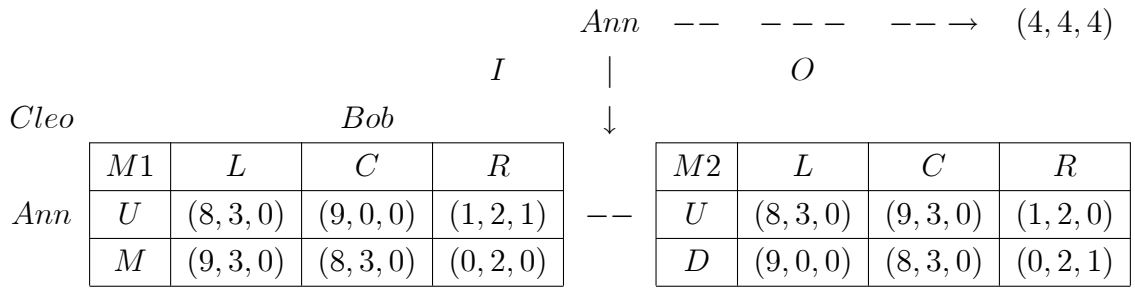
Then, since a rationalizable strategy is obviously rational, differently than for self-enforceability, implementability under priority to rationality implies implementability under priority to the agreement (at least in two-players games).

Corollary 5 *In a 2-players game, if an outcome is implementable under priority to rationality, then it is implementable under priority to the agreement.*

In games with more than two players, there is no equilibrium equivalent of implementable outcomes. Here I provide a 3-players game where a non subgame perfect

[40] Under piority to rationality, the existence is guaranteed by finiteness or mild regularity conditions. Otherwise, the Proposition can be restated requiring such CPS's to exist.

equilibrium outcome is implementable thanks to an incomplete agreement (and incompleteness is not only on the side of the deviator).



	$M1$	L	C	R
Ann	U	$(8,3,0)$	$(9,0,0)$	$(1,2,1)$
	M	$(9,3,0)$	$(8,3,0)$	$(0,2,0)$

	$M2$	L	C	R
U	$(8,3,0)$	$(9,3,0)$	$(1,2,0)$
D	$(9,0,0)$	$(8,3,0)$	$(0,2,1)$

Ann moves first and chooses between a "fair" outside option and a simultaneous moves game with Bob and Cleo, where Cleo chooses the matrix. In any equilibrium of the subgame, R cannot be played with probability higher than $1/2$, otherwise Ann would choose U and then Bob would switch to L. Thus, O is not a SPE outcome. Still, Bob can credibly threaten to play R. For this to hold true, it is crucial that Cleo does not reveal what she is going to play: for each matrix, Bob has a safe option that dominates R. If Ann plays I, either she is irrational or she has not believed in the threat. Under both interpretations, she may play U or M and Cleo may react with $M1$ or $M2$. Thus, it is credible that Bob will react with R.

The definition of implementability can be strenghtened with the path agreement credibility test. First, the outcome should pass the test. Second, the agreement that implements the outcome should be compatible with forward induction reasoning based on the beliefs in the path. A way to obtain this is to derive the agreement from the conjectures under which players remain on path at the end of the test. Yet, such conjectures may not be convertible into an agreement, because of: (correlated) randomizations; strategic incompatibility between the threats of different players; without observable actions, the disagreement of two deviators about the reaction to the deviation of a third player who cannot recognize who the deviator is. Strategic incompatibilities may be solved with an incomplete agreement. All the outcomes implemented in the examples are credible, and the threats are compatible with forward induction reasoning about the path.

If one insists on SPE outcomes, under priority to rationality an important preliminary question arises: is there always the support of a SPE outcome distribution among strongly rationalizable outcomes? The following theorem states that it is ac-

tually so[41] for the wide class of games with observable actions:[42] subgame perfection and strong rationalizability never give completely disjoint predictions.

Theorem 6 *Consider the set of strongly rationalizable strategy profiles S^∞. There exists a SPE $(\sigma_i)_{i\in I} \in \times_{i\in I}\Delta(S_i)$ and an equilibrium $(\widetilde{\sigma}_i)_{i\in I} \in \times_{i\in I}\Delta(S_i^\infty)$ such that for every $z \in Z$, $\prod_{i\in I}\sigma_i(S_i(z)) = \prod_{i\in I}\widetilde{\sigma}_i(S_i(z))$.*[43]

5 Directions for future research

The issue of compliance with agreements could be interestingly analyzed in psychological games (see for instance Battigalli and Dufwenberg [6]) or games with ambiguity averse players. Both ambiguity aversion and belief-dependent payoffs, like in the case of guilt-averse players, could sustain the self-enforceability a wider range of agreements, including path ones. Other psychological considerations could motivate formally the preference for path agreements, which do not involve the discussion of what to do in case someone defects.

The methodology developed here can be applied to a wide range of economic problems, from non-binding commitment of governments (see for instance Bassetto [1] on capital taxation) to dynamic collusion in oligopolies. On this topic, Harrington [22] documents instances of mutual partial understanding among firms whose implications can be understood with the methodology of this paper. Another application which is close in spirit to this work can be found in Gossner [19]. In his study of incomplete codes, he calls self-enforcing the equilibrium paths that can be sustained by any choice of equilibria in the off-the-path subgames. As already argued, such paths are at least credible in the sense of this paper (but, so far, using the finite horizon hypothesis).

Although the focus of the paper has been kept on complete information for interpretative and notational easiness, the methodology can be applied also to dynamic games with incomplete information. In an incomplete information environment, players reaching an agreement at the interim stage may reveal their types in the bargaining process or not. In the first case, first-order-beliefs restrictions could be extended to

[41] And these outcomes are all induced by a mixed rationalizable Nash, an important feature if one wants to go on proving implementability with probabilistic agreements.

[42] I conjecture that the result holds for all extensive form games with perfect recall.

[43] While $\times$ denotes a cartesian product, $\prod$ will always denote an algebraic product.

payoff-relevant types. In the second case, what the agreement suggests about co-players types would be embodied in the rationalizability procedure.

Finally, there are two complementary issues, already investigated in the literature, to which this analysis could be profitably connected. The first is the pre-play bargaining issue. This paper sheds light on which agreements are self-enforcing and which outcomes are implementable through agreements. How will players ultimately choose among them? Welfare considerations and a theory of bargaining could refine the answer. Dufwenberg, Servátka and Vadovic [16] propose an interesting approach to the pre-play bargaining issue. Miller and Watson [25], instead, analyze a problem of bargaining between players who can reach an agreement at every stage of a repeated game. When players can communicate during the game, the second issue is renegotiation proofness (see, for instance, [17]). Past moves and consequent forward induction considerations could influence the bargaining power of players in renegotiation.

6 Appendix

Additional notation (for a subgame $\Gamma(h)$ of $\Gamma = \langle I, \overline{H}, (u_i)_{i\in I}\rangle$).

Histories of $\Gamma(h)$ are identified for convenience by the histories of Γ following h, and not redefined as shorter lists of action profiles occuring *after* h.

- For any $h \in H$ and $i \in I$:
  - $H^h := \{\widetilde{h} \in H : \widetilde{h} \succeq h\}$, $Z^h := \{z \in Z : z \succ h\}$;
  - S_i^h is the set of all strategies $s_i^h : \widetilde{h} \in H^h \mapsto a_i \in A_i(\widetilde{h})$ of $\Gamma(h)$.
- for any $J \subseteq I$, $z \in H^h \cup Z^h$, $\widehat{h} \in H^h$ and subsets of strategies $(\overline{S}_j^h)_{j\in I}$:
  - $\overline{S}_J^h(z)$, $\overline{S}_J^h[\widehat{h}]$, $H(\overline{S}_J^h)$, $\Delta^{H^h}(S_{-i}^h)$, $\zeta(\cdot)$, $r(\cdot)$ and $\rho(\cdot)$ are defined like in Γ;
  - $D_i(\overline{S}^h) := \{\widetilde{h} \in H : \exists \overline{h} \in H(\overline{S}^h), \exists a_i \in A_i(\overline{h})\backslash \overline{S}_i^h[\overline{h}], \exists a_{-i} \in \overline{S}_{-i}^h[\overline{h}], \widetilde{h}{=}(\overline{h}, (a_i, a_{-i}))\}$;[44]
- for any $(s_j^h)_{j\in I} \in S^h$, $(s_j^{\widehat{h}})_{j\in I} \in S^{\widehat{h}}$, $\mu_i^h \in \Delta^{H^h}(S_{-i}^h)$, $\widetilde{\mu}_i^{\widehat{h}} \in \Delta^{H^{\widehat{h}}}(S_{-i}^{\widehat{h}})$ and $\widehat{H} \subseteq H^{\widehat{h}}$:
  - $s_i^h|\widehat{h}$ is the restriction of s_i^h to $H^{\widehat{h}}$, $s_J^h|\widehat{h} := (s_j^h|\widehat{h})_{j\in J}$, $\overline{S}_J^h|\widehat{h} := (\widetilde{s}_J^h|\widehat{h})_{\widetilde{s}_J^h \in \overline{S}_J^h}$;

[44] Set of histories that follow a unilateral deviation by player i from the histories induced by $\overline{S}^h$.

- $\widehat{r}(\mu_i^h, \widehat{h})$ is the set of continuation best replies to $\mu_i^h(\cdot|\widehat{h})$;
- $s_J^h =^{\widehat{H}} s_J^{\widehat{h}}$ if for every $\widetilde{h} \in \widehat{H}$, $s_J^h(\widetilde{h}) = s_J^{\widehat{h}}(\widetilde{h})$; $S_{J,\widehat{H}}^{h,s_J^{\widehat{h}}} := \{\widetilde{s}_J^h \in S_J^h : \widetilde{s}_J^h =^{\widehat{H}} s_J^{\widehat{h}}\}$;
- $\mu_i^h =^{\widehat{H}} \widetilde{\mu}_i^{\widehat{h}}$ if for every $\widetilde{h} \in \widehat{H}$ and $\widetilde{z} \succ p(\widetilde{z}) \in \widehat{H} \cap H^{\widetilde{h}}$, $\mu_i^h(S_{-i}^h(\widetilde{z})|\widetilde{h}) = \widetilde{\mu}_i^{\widehat{h}}(S_{-i}^{\widehat{h}}(\widetilde{z})|\widetilde{h})$;
- $s_J^h =^{\widehat{h}} s_J^{\widehat{h}}$ and $\mu_i^h =^{\widehat{h}} \widetilde{\mu}_i^{\widehat{h}}$ if, respectively, $s_J^h =^{H^{\widehat{h}}} s_J^{\widehat{h}}$ and $\mu_i^h =^{H^{\widehat{h}}} \widetilde{\mu}_i^{\widehat{h}}$.

I will often use the fact that $=^{\widehat{H}}$ is transitive and that if $\mu_i^h =^{\widehat{H}} \widetilde{\mu}_i^{\widehat{h}}$ and $\widetilde{H} \subseteq \widehat{H}$, then $\mu_i^h =^{\widetilde{H}} \widetilde{\mu}_i^{\widehat{h}}$.[45] Moreover, note that by CPS-3 (with $\widehat{h}$ as "intermediate" history):[46]

♠ if $\widetilde{H} \cap H^{\widehat{h}} = \emptyset$, $p(\widehat{h}) \in \widetilde{H}$, $\mu_i^h =^{\widetilde{H}} \widetilde{\mu}_i^h$, $\mu_i^h =^{\widehat{H}} \widetilde{\mu}_i^{\widehat{h}}$ and $\widetilde{\mu}_i^h =^{\widehat{H}} \widetilde{\mu}_i^{\widehat{h}}$, then $\mu_i^h =^{\widetilde{H} \cup \widehat{H}} \widetilde{\mu}_i^h$.

Rewrite selective rationalizability as $(S_i^0, ..., S_i^M, S_{i,R\Delta^e}^0, S_{i,R\Delta^e}^1, ...)_{i \in I}$, where $((S_i^q)_{i \in I})_{q \geq 0}$ is strong rationalizability and M is the smallest $q \in \mathbb{N}$ such that $S^q = S^{q+1} = S^\infty$ (it exists by finiteness of the game).

In a subgame, substrategies can be eliminated "exogenously" and not because they are not sequential best replies to any valid conjecture in the subgame. On the other hand, substrategies can survive even if the opponents do not reach the subgame anymore. Thus I define the following.

Definition 11 *Fix $h \in H$. A* reduction procedure *is a sequence $((S_{i,q}^h)_{i \in I})_{q \geq 0}$ where:*

- *for every $i \in I$, $S_{i,0}^h = S_i^h$;*
- *for every $n > 0$ and $s_i^h \in S_{i,n}^h$, $s_i^h \in S_{i,n-1}^h$ and there exists $\mu_i^h \in \Delta^{H^h}(S_{-i}^h)$ s. t.:*
  - *$s_i^h \in \rho(\mu_i^h)$;*
  - *for every $q < n$ and $\widetilde{h} \in H^h$, if $S_{-i,q}^h(\widetilde{h}) \neq \emptyset$, then $\mu_i^h(S_{-i,q}^h|\widetilde{h}) = 1$.*

Let M be the smallest $q \in \mathbb{N}$ such that $S_q^h = S_{q+1}^h$ and let $S_{i,\infty}^h := S_{i,M}^h$ (it exists by finiteness). Note that even if $S_{-i,n-1}^h$ is empty, $S_{i,n}^h$ needs not be empty. Still, for semplicity I will say that $s_i^h \in S_{i,n}^h$ is justified by a μ_i^h *that strongly believes* (henceforth, t.s.b.) $(S_{-i,q}^h)_{q=0}^{n-1}$. The definition of reduction procedure encompasses

[45] These properties clearly hold also when $=^{\widehat{H}}$ applies between strategies.

[46] By CPS-3, for every $\widetilde{h} \in \widetilde{H}$ and $\widetilde{z} \succ p(\widetilde{z}) \in \widehat{H} \cap H^{\widetilde{h}}$, $\mu_i^h(S_{-i}^h(\widetilde{z})|\widetilde{h}) = \mu_i^h(S_{-i}^h(\widehat{h})|\widetilde{h})\mu_i^h(S_{-i}^h(\widetilde{z})|\widehat{h}) = \widetilde{\mu}_i^h(S_{-i}^h(\widehat{h})|\widetilde{h})\widetilde{\mu}_i^{\widehat{h}}(S_{-i}^{\widehat{h}}(\widetilde{z})|\widehat{h}) = \widetilde{\mu}_i^h(S_{-i}^h(\widehat{h})|\widetilde{h})\widetilde{\mu}_i^h(S_{-i}^{\widehat{h}}(\widetilde{z})|\widehat{h}) = \widetilde{\mu}_i^h(S_{-i}^h(\widetilde{z})|\widetilde{h})$.

both strong-Δ-rationalizability and selective rationalizability (rewritten as before) and their implications in the subgames. Indeed, if $((S^h_{i,q})_{i\in I})_{q\geq 0}$ is a reduction procedure, $((S^h_{i,q}(\widehat{h})|\widehat{h})_{i\in I})_{q\geq 0}$ is a reduction procedure for any $\widehat{h}\in H^h$.

Some lemmata claim the survival of strategies, or conjectures over such strategies, which combine substrategies that have survived by assumption. The reason why such lemmata are needed is merely the following. Fix $\widehat{s}^h_i, \overline{s}^h_i \in S^h_{i,n}$ and $\widehat{h}, \overline{h} \in H(\widehat{s}^h_i)\cap H(\overline{s}^h_i)$ such that $\overline{h} \not\geq \widehat{h} \not\geq \overline{h}$: there needs not exist $s^h_i \in S^h_{i,n}(\widehat{h}) \cap S^h_{i,n}(\overline{h})$ such that $s^h_i|\widehat{h} = \overline{s}^h_i|\widehat{h}$ and $s^h_i|\overline{h} = \widehat{s}^h_i|\overline{h}$. The intuitive 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 certain expectations after $\widehat{h}$, or because she is confident that $\overline{h}$ will be reached and she has certain expectations after $\overline{h}$. If $\widehat{s}^h_i$ is best reply to the first conjecture and $\overline{s}^h_i$ is best reply to the second conjecture, $\widehat{s}^h_i|\overline{h}$ and $\overline{s}^h_i|\widehat{h}$ may be "emergency plans" for an unpredicted contingency, after which the expectations would not have justified the choice to allow $\overline{h}$ and $\widehat{h}$ in the first place.

I start with a lemma about the combination of substrategies. Consider a player who may be surprised by history $\widehat{h}$, in the sense that she may play a strategy that allows $\widehat{h}$ while believing that the co-players do not until $\widehat{h}$ is actually reached. This player can keep the same beliefs and the same strategy out of $\Gamma(\widehat{h})$, whatever she believes the co-players would do and hence however she may play after $\widehat{h}$. Being a rather intuitive result, the proof is relegated to the Online Appendix.

Lemma 1 *Fix a reduction procedure $((S^h_{i,q})_{i\in I})_{q\geq 0}$, $i\in I$, $n\in\mathbb{N}$, $\widehat{h}\in H^h$ and $\mu^h_i \in \Delta^{H^h}(S^h_{-i})$ t.s.b. $(S^h_{-i,q})^{n-1}_{q=0}$ such that $\mu^h_i(S^h_{-i}(\widehat{h})|p(\widehat{h})) = 0$. Fix $s^h_i \in \rho(\mu^h_i)$, $\mu^{\widehat{h}}_i$ t.s.b. $(S^h_{-i,q}(\widehat{h})|\widehat{h})^{n-1}_{q=0}$ and $s^{\widehat{h}}_i \in \rho(\mu^{\widehat{h}}_i)$.*

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

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

PROOF OF THEOREMS 1 AND 3

Fix a set of outcomes and call $\widehat{H}$ the set of histories of $\Gamma(h)$ that precede them. Suppose that $\mu^h_i =^{\widehat{H}} \overline{\mu}^h_i$ and that the co-players are not expected to deviate outside of $\widehat{H}$, i.e. they are expected to play compatibly with some of those outcomes at every history in $\widehat{H}$. If also the sequential best replies to μ^h_i and the sequential best reply to $\overline{\mu}^h_i$ do not deviate outside of $\widehat{H}$, then they mimic each other inside of $\widehat{H}$.

Lemma 2 *Fix $l \in I$, $h \in H$ and $\widehat{Z} \subseteq Z^h$. Let $\widehat{H} := \{\widetilde{h} \in H^h : \exists z \in \widehat{Z}, \widetilde{h} \prec z\}$. Fix $\mu_l^h =^{\widehat{H}} \overline{\mu}_l^h$ such that for every $\widetilde{\mu}_l^h \in \{\mu_l^h, \overline{\mu}_l^h\}$:*

$$\rho(\widetilde{\mu}_l^h) \subseteq \{\widetilde{s}_l^h \in S_l^h : \forall \widetilde{h} \in \widehat{H} \cap H(\widetilde{s}_l^h), \forall \widetilde{s}_{-l}^h \in \text{Supp}\widetilde{\mu}_l^h(\cdot|\widetilde{h}), H(\widetilde{s}_l^h, \widetilde{s}_{-l}^h) \subseteq \widehat{H}\}. \quad \text{(A0)}$$

Then for every $\overline{s}_l^h \in \rho(\overline{\mu}_l^h)$ there exists $s_l^h \in \rho(\mu_l^h)$ such that $s_l^h =^{\widehat{H}} \overline{s}_l^h$.

Proof. Fix $\overline{s}_l^h \in \rho(\overline{\mu}_l^h)$. For every $\widetilde{h} \notin \widehat{H}$ and $\overline{h} \succ \widetilde{h}$, $\overline{h} \notin \widehat{H}$. Hence there exists $s_l^h =^{\widehat{H}} \overline{s}_l^h$ such that for every $\widetilde{h} \notin \widehat{H}$, $s_l^h \in \widehat{r}(\mu_l^h, \widetilde{h})$.

Fix $\widetilde{h} \in H(s_l^h) \cap H(\rho(\mu_l^h)) \cap \widehat{H}$ and $\widetilde{s}_l^h \in \rho(\mu_l^h)(\widetilde{h})$. For every $h \preceq \overline{h} \prec \widetilde{h}$, $\overline{h} \in \widehat{H}$. Then, since $s_l^h =^{\widehat{H}} \overline{s}_l^h$, $\widetilde{h} \in H(\overline{s}_l^h)$. Hence $\overline{s}_l^h \in \widehat{r}(\overline{\mu}_l^h, \widetilde{h})$ and for every $\overline{s}_{-l}^h \in \text{Supp}\overline{\mu}_l^h(\cdot|\widetilde{h})$, by A0 $H(\overline{s}_l^h, \overline{s}_{-l}^h) \subseteq \widehat{H}$; so, since $s_l^h =^{\widehat{H}} \overline{s}_l^h$, $\zeta(s_l^h, \overline{s}_{-l}^h) = \zeta(\overline{s}_l^h, \overline{s}_{-l}^h)$. Thus $s_l^h \in \widehat{r}(\overline{\mu}_l^h, \widetilde{h})$ too and for every $z \in \zeta(\{s_l^h\} \times \text{Supp}\overline{\mu}_l^h(\cdot|\widetilde{h})) =: \overline{Z}$, $p(z) \in \widehat{H}$; so by $\mu_l^h =^{\widehat{H}} \overline{\mu}_l^h$, $\mu_l^h(S_{-l}^h(z)|\widetilde{h}) = \overline{\mu}_l^h(S_{-l}^h(z)|\widetilde{h}) =: \nu(z)$. Then s_l^h induces the same probability distribution $(\nu(z))_{z \in \overline{Z}}$[47] with $\overline{\mu}_l^h(\cdot|\widetilde{h})$ and $\mu_l^h(\cdot|\widetilde{h})$ over Z^h. For every $\widetilde{s}_{-l}^h \in \text{Supp}\mu_l^h(\cdot|\widetilde{h})$, by A0 $H(\widetilde{s}_l^h, \widetilde{s}_{-l}^h) \subseteq \widehat{H}$. Thus for every $z \in \zeta(\{\widetilde{s}_l^h\} \times \text{Supp}\mu_l^h(\cdot|\widetilde{h})) =: \widetilde{Z}$, $p(z) \in \widehat{H}$; so by $\mu_l^h =^{\widehat{H}} \overline{\mu}_l^h$, $\overline{\mu}_l^h(S_{-l}^h(z)|\widetilde{h}) = \mu_l^h(S_{-l}^h(z)|\widetilde{h}) =: \eta(z)$. Then $\widetilde{s}_l^h$ induces the same probability distribution $(\eta(z))_{z \in \widetilde{Z}}$ with $\mu_l^h(\cdot|\widetilde{h})$ and $\overline{\mu}_l^h(\cdot|\widetilde{h})$ over Z^h. Hence, since $s_l^h \in \widehat{r}(\overline{\mu}_l^h, \widetilde{h})$ and $\widetilde{s}_l^h \in \widehat{r}(\mu_l^h, \widetilde{h})$, $\widetilde{s}_l^h \in \widehat{r}(\overline{\mu}_l^h, \widetilde{h})$ and $s_l^h \in \widehat{r}(\mu_l^h, \widetilde{h})$ too.

So for every $\widetilde{h} \in H(s_l^h) \cap H(\rho(\mu_l^h))$, $s_l^h \in \widehat{r}(\mu_l^h, \widetilde{h})$. I show that $H(s_l^h) \backslash H(\rho(\mu_l^h)) = \emptyset$, so that $s_l^h \in \rho(\mu_l^h)$. Suppose not. Then there exists $\widehat{h} \in H(s_l^h) \backslash H(\rho(\mu_l^h))$ such that $p(\widehat{h}) \in H(s_l^h) \cap H(\rho(\mu_l^h))$. But since $s_l^h \in \widehat{r}(\mu_l^h, p(\widehat{h}))$, from the observation about the relationship between continuation and sequential best replies in Section 3, I can deduce that $\widehat{h} \in H(\rho(\mu_l^h))$ too. ■

From now on, whenever $\mu_l^h =^{\widehat{H}} \overline{\mu}_l^h$, there will exist a set of outcomes such that $\widehat{H}$ is the set of histories of $\Gamma(h)$ that precede them.

Lemma 3 deals with a similar situation as in Lemma 1, but from the perspective of the deviator. If the co-players may be surprised by different deviations from the same predicted behavior, the deviator can expect any combination of reactions. The Lemma is less general to target the particular setting in which it will be used. The Lemma is rather intuitive and the proof is mostly a tedious book-keeping exercise. Therefore, it is relegated to the Online Appendix.

[47] The probability distribution over Z induced by a strategy $s_i^h \in S_i^h(\widetilde{h})$ with $\mu_i^h(\cdot|\widetilde{h})$ is $(\mu_i^h(S_{-i}^h(z)|\widetilde{h}))_{z \in \zeta(\{s_i^h\} \times S_{-i}^h(\widetilde{h}))}$ (and probability 0 to every $z \notin \zeta(\{s_i^h\} \times S_{-i}^h(\widetilde{h}))$).

Lemma 3 *Fix a red. procedure* $((\widetilde{S}^h_{i,q})_{i\in I})_{q\geq 0}$*, subsets of strategies* $(\overline{S}^h_i)_{i\in I}$*,* $m \in \mathbb{N}$ *and* $l \in I$*. Let* $H^S := H(\overline{S}^h)$ *and* $D^S := D_l(\overline{S}^h)$*. For every* $i \neq l$*, suppose that there exists a map* $\overline{\overline{\mu}}^h_i : \overline{S}^h_i \to \Delta^{H^h}(S^h_{-i})$ *such that for every* $s^h_i \in \overline{S}^h_i$*,* $\overline{\overline{\mu}}^h_i(s^h_i)$ *s.bel.* $\overline{S}^h_{-i}$*, and:*

A1 there exist maps $\overline{\mu}^h_i : \overline{S}^h_i \to \Delta^{H^h}(S^h_{-i})$ *and* $\overline{s}^h_i : \overline{S}^h_i \to S^h_i$ *such that for every* $s^h_i \in \overline{S}^h_i$*,* $\overline{\mu}^h_i(s^h_i) =^{H^S} \overline{\overline{\mu}}^h_i(s^h_i)$ *strongly bel.* $(\widetilde{S}^h_{-i,q})^{m-1}_{q=0}$ *and* $\rho(\overline{\mu}^h_i(s^h_i)) \ni \overline{s}^h_i(s^h_i) =^{H^S} s^h_i$*;*

A2 for every $s^h_i \in \overline{S}^h_i$ *and* $\mu^h_i =^{H^S} \overline{\overline{\mu}}^h_i(s^h_i)$ *t.s.b.* $(\widetilde{S}^h_{-i,q})^{m-1}_{q=0}$*,* $\rho(\mu^h_i) \subseteq \widetilde{S}^h_{i,m}$*.*

Fix μ^h_l *t.s.b.* $(\widetilde{S}^h_{-l,q})^m_{q=0}$ *and* $\times_{i\neq l}(\overline{s}^h_i(s^h_i))_{s^h_i\in\overline{S}^h_i}$*. Fix* $\widetilde{D} \subseteq D^S$ *and for every* $\widehat{h} \in \widetilde{D}$*, fix* $\widetilde{\mu}^{\widehat{h}}_l$ *t.s.b.* $(\widetilde{S}^h_{-l,q}(\widehat{h})|\widehat{h})^m_{q=0}$*. Let* $H^* := H^h \setminus \cup_{\widehat{h}\in\widetilde{D}} H^{\widehat{h}}$*.*

There exists $\widetilde{\mu}^h_l =^{H^*} \mu^h_l$ *t.s.b.* $(\widetilde{S}^h_{-l,q})^m_{q=0}$ *such that for every* $\widehat{h} \in \widetilde{D}$*,* $\widetilde{\mu}^h_l =^{\widehat{h}} \widetilde{\mu}^{\widehat{h}}_l$*.*

For future reference, note that every $\overline{\overline{\mu}}^h_i$ t.s.b. $(\widetilde{S}^h_{-i,q})^\infty_{q=0}$ and $H(\widetilde{S}^h_\infty)$ satisfy A0.

Fix a set of strategy profiles $\overline{S}^h$ delivered by a reduction procedure. Suppose that until step n, each player i is willing to play strategies that mimic those in $\overline{S}^h_i$ along the paths induced by $\overline{S}^h$ while expecting the co-players to do the same. At step $n+1$, instead, some player l stops playing any strategy of hers that mimics a strategy $\widehat{s}^h_l$ in $\overline{S}^h_l$. Since at n the co-players may be surprised by any deviation, player l might expect them to play any combination of substrategies that survive n steps after the potential deviations. Hence, there must exist one particular deviation that player l prefers to mimicking $\widehat{s}^h_l$ whatever she may conjecture thereafter.

Lemma 4 *Fix reduction procedures* $((\overline{S}^h_{i,q})_{i\in I})_{q\geq 0}$ *and* $((S^h_{i,q})_{i\in I})_{q\geq 0}$*. For every* $i \in I$ *call* $\overline{S}^h_i := \overline{S}^h_{i,\infty}$ *and let* $\overline{\overline{\mu}}^h_i : \overline{S}^h_i \to \Delta^{H^h}(S^h_{-i})$ *be a map such that for every* $s^h_i \in \overline{S}^h_i$*,* $\overline{\overline{\mu}}^h_i(s^h_i)$ *strongly believes* $(\overline{S}^h_{-i,q})^\infty_{q=0}$ *and* $s^h_i \in \rho(\overline{\overline{\mu}}^h_i(s^h_i))$ *(so that* $(\overline{S}^h_n)_{n\geq 0}$ *satisfies A1 with* $\overline{\mu}^h_i(\cdot) = \overline{\overline{\mu}}^h_i(\cdot)$*). Let* $H^S := H(\overline{S}^h)$*. Fix* $n \in \mathbb{N}$*,* $l \in I$ *and* $\widehat{s}^h_l \in \overline{S}^h_l$ *such that:*[48]

A3 for every $i \in I$ *and* $m \leq n$*,* $(S^h_q)_{q\geq 0}$ *satisfies* A1*;*

A4 for every $i \in I$ *and* $m \in \mathbb{N}$*,* $(S^h_q)_{q\geq 0}$ *satisfies* A2*;*

A5 for every $i \in I$ *and* $m \in \mathbb{N}$*,* $(\overline{S}^h_q)_{q\geq 0}$ *satisfies* A2*;*

A6 for every $s^h_l =^{H^S} \widehat{s}^h_l$ *and* $\mu^h_l =^{H^S} \overline{\overline{\mu}}^h_l(\widehat{s}^h_l)$ *t.s.b.* $(S^h_{-l,q})^n_{q=0}$*,* $s^h_l \notin \rho(\mu^h_l)$*.*

[48] A3, A4 and A5 need not hold for $i = l$ to claim Lemma 3 and prove this Lemma. However, l has been included to reuse A3, A4 and A5 in the final proof of the theorems.

Let $D^S := D_l(\overline{S}^h)$. *For every* $\widehat{h} \in D^S$ *and* $m \in \mathbb{N}$ *call* $M^{\widehat{h}}_m$ *(resp.* $\overline{M}^{\widehat{h}}_m$*) the set of* $\mu^{\widehat{h}}_l$ *t.s.b.* $(S^h_{-l,q}(\widehat{h})|\widehat{h})^m_{q=0}$ *(resp.* $(\overline{S}^h_{-l,q}(\widehat{h})|\widehat{h})^m_{q=0}$*) such that there exist* $H_{\widehat{h}} \subseteq H^{\widehat{h}}$ *and* $\widehat{\mu}^{\widehat{h}}_l =^{H_{\widehat{h}}} \mu^{\widehat{h}}_l$ *t.s.b.* $(S^h_{-l,q}(\widehat{h})|\widehat{h})^n_{q=0}$ *which satisfy* A0.[49]

Then there exists $\widehat{h} \in D^S$ *such that for every* $m \leq n$, $p \in \mathbb{N}$ *and* $(\mu^{\widehat{h}}_l, \widetilde{\mu}^{\widehat{h}}_l) \in M^{\widehat{h}}_m \times \overline{M}^{\widehat{h}}_p$, *there exist (1)* $\mu^h_l =^{H^S} \overline{\overline{\mu}}^h_l(\widehat{s}^h_l)$ *t.s.b.* $(S^h_{-l,q})^m_{q=0}$ *and (2)* $\widetilde{\mu}^h_l =^{H^S} \overline{\overline{\mu}}^h_l(\widehat{s}^h_l)$ *t.s.b.* $(\overline{S}^h_{-l,q})^p_{q=0}$ *such that* $\mu^h_l =^{\widehat{h}} \mu^{\widehat{h}}_l$, $\widetilde{\mu}^h_l =^{\widehat{h}} \widetilde{\mu}^{\widehat{h}}_l$ *and* $\rho(\mu^h_l)(\widehat{h}) \neq \emptyset \neq \rho(\widetilde{\mu}^h_l)(\widehat{h})$.

Proof. Suppose by contraposition that there is a partition $(D, \overline{D})$ of D^S such that for every $\widehat{h} \in D$ there exist $m(\widehat{h}) \leq n$ and $\mu^{\widehat{h}}_l \in M^{\widehat{h}}_{m(\widehat{h})}$ that violate 1, and for every $\widehat{h} \in \overline{D}$ there exist $m(\widehat{h}) \in \mathbb{N}$ and $\mu^{\widehat{h}}_l \in \overline{M}^{\widehat{h}}_{m(\widehat{h})}$ that violate 2. For every $\widehat{h} \in D^S$ fix corresponding $\widehat{\mu}^{\widehat{h}}_l$ and $H_{\widehat{h}}$. Write $\overline{D} = \{h^1, ..., h^k\}$ where $m(h^1) \geq ... \geq m(h^k)$. Let $\overline{\overline{\mu}}^h_l := \overline{\overline{\mu}}^h_l(\widehat{s}^h_l)$.

Fix any $\mu^h_l =^{H^S} \overline{\overline{\mu}}^h_l$ t.s.b. $(S^h_{-l,q})^n_{q=0}$ and $\times_{i \neq l}(\overline{s}^h_i(s^h_i))_{s^h_i \in \overline{S}^h_i}$ (one clearly exists); let $H^* := H^h \backslash \cup_{\widehat{h} \in D^S} H^{\widehat{h}}$; by Lemma 3 there exist:

- $\widetilde{\mu}^h_l =^{H^*} \mu^h_l =^{H^S} \overline{\overline{\mu}}^h_l$ t.s.b. $(S^h_{-l,q})^n_{q=0}$ such that for every $\widehat{h} \in D^S$, $\widetilde{\mu}^h_l =^{\widehat{h}} \widehat{\mu}^{\widehat{h}}_l$;
- for every $\widehat{h} \in D$, $\widetilde{\mu}^h_{l,\widehat{h}} =^{H^*} \mu^h_l =^{H^S} \overline{\overline{\mu}}^h_l$ t.s.b. $(S^h_{-l,q})^{m(\widehat{h})}_{q=0}$ such that $\widetilde{\mu}^h_{l,\widehat{h}} =^{\widehat{h}} \mu^{\widehat{h}}_l =^{H_{\widehat{h}}} \widehat{\mu}^{\widehat{h}}_l$ and for every $\widetilde{h} \in D^S \backslash \{\widehat{h}\}$, $\widetilde{\mu}^h_{l,\widehat{h}} =^{\widetilde{h}} \widehat{\mu}^{\widetilde{h}}_l$, so that by ♠, $\widetilde{\mu}^h_{l,\widehat{h}} =^{H^h \backslash (H^{\widehat{h}} \backslash H_{\widehat{h}})} \widetilde{\mu}^h_l$;

Let $\widetilde{\mu}^h_{l,0} := \overline{\overline{\mu}}^h_l =: \mu^h_l$ and $\widetilde{H} := \cup_{\widehat{h} \in \overline{D}} H_{\widehat{h}} \cup H^S$; by Lemma 3 there exist:

- for every $1 \leq j \leq k$, $\widetilde{\mu}^h_{l,j} =^{H^h \backslash \cup^j_{t=1} H^{h^t}} \overline{\overline{\mu}}^h_l$ t.s.b. $(\overline{S}^h_{-l,q})^{m(h^j)}_{q=0}$ such that for every $1 \leq t \leq j$, $\widetilde{\mu}^h_{l,j} =^{h^t} \mu^{h^t}_l =^{H_{h^t}} \widehat{\mu}^{h^t}_l$; so by ♠, $\widetilde{\mu}^h_{l,j} =^{H^h \backslash H^{h^j}} \widetilde{\mu}^h_{l,j-1}$ and $\widetilde{\mu}^h_{l,k} =^{\widetilde{H}} \widetilde{\mu}^h_l$.

Fix $\widehat{h} \in D$ and let $\widehat{H} := H^h \backslash (H^{\widehat{h}} \backslash H_{\widehat{h}})$: $\widetilde{\mu}^h_l$ and $\widehat{H}$ satisfy A0 because $\widetilde{\mu}^h_l =^{\widehat{h}} \widehat{\mu}^{\widehat{h}}_l$;[50] $\widetilde{\mu}^h_{l,\widehat{h}}$ and $\widehat{H}$ satisfy A0 because by the contrapositive hypothesis $\rho(\widetilde{\mu}^h_{l,\widehat{h}})(\widehat{h}) = \emptyset$. Then by Lemma 2 $\rho(\widetilde{\mu}^h_l)(\widehat{h}) = \emptyset$ too.[51] So $H(\rho(\widetilde{\mu}^h_l)) \cap D = \emptyset$.

Now I show inductively that $H(\rho(\widetilde{\mu}^h_{l,k})) \cap D^S = \emptyset$. Fix $1 \leq j \leq k$ and suppose that $H(\rho(\widetilde{\mu}^h_{l,j-1})) \cap D^S = \emptyset$, which is true for $j = 1$ because $\rho(\overline{\overline{\mu}}^h_l) \subseteq \overline{S}^h_l$. Hence $\widetilde{\mu}^h_{l,j-1}$

[49] Note that: $\widehat{\mu}^{\widehat{h}}_l$ refers to the second procedure even when $\mu^{\widehat{h}}_l$ refers to the first; $\mu^{\widehat{h}}_l$ and $H_{\widehat{h}}$ need not satisfy A0.

[50] The sequential best replies to $\widetilde{\mu}^h_l$ that reach $\widehat{h}$, obviously mimic those to $\widehat{\mu}^{\widehat{h}}_l$ after $\widehat{h}$.

[51] Player l expects the same payoff against $\widetilde{\mu}^h_{l,\widehat{h}}$ and $\widetilde{\mu}^h_l$ without reaching $\widehat{h}$ and a non-higher payoff against $\widetilde{\mu}^h_l$ after reaching $\widehat{h}$. Hence, since she does not want to deviate under $\widetilde{\mu}^h_{l,\widehat{h}}$, she does not want to deviate under $\widetilde{\mu}^h_l$.

and $H^h \backslash H^{h^j}$ satisfy A0; $\widetilde{\mu}^h_{l,j}$ and $H^h \backslash H^{h^j}$ satisfy A0 because by the contrapositive hypothesis $\rho(\widetilde{\mu}^h_{l,j})(h^j) = \emptyset$; then by Lemma 2 $H(\rho(\widetilde{\mu}^h_{l,j})) \cap D^S = \emptyset$ too. Inductively, $H(\rho(\widetilde{\mu}^h_{l,k})) \cap D^S = \emptyset$.

Hence, $\widetilde{\mu}^h_{l,k}$ and $\widetilde{H}$ satisfy A0; $\widetilde{\mu}^h_l$ and $\widetilde{H}$ satisfy A0 because $H(\rho(\widetilde{\mu}^h_l)) \cap D = \emptyset$ and for every $\widehat{h} \in \overline{D}$, $\widetilde{\mu}^h_l =^{\widehat{h}} \widehat{\mu}^{\widehat{h}}_l$; then by Lemma 2 $H(\rho(\widetilde{\mu}^h_l)) \cap D^S = \emptyset$ too. Hence $\widetilde{\mu}^h_l$ and H^S satisfy A0; $\overline{\overline{\mu}}^h_l$ and H^S satisfy A0. So by Lemma 2 there exists $s^h_l \in \rho(\widetilde{\mu}^h_l)$ such that $s^h_l =^{H^S} \widehat{s}^h_l$, violating A6. ■

Proof of Theorem 1 and 3.

The idea is the following. Take the set delivered by strong-Δ-rationalizability or selective rationalizability for a given path agreement. Suppose that at some step n of the same procedure for a looser agreement or of the alternative procedure for the same agreement, a player l excludes to mimic a strategy of the set for some profitable deviation outside the histories induced by the set. More precisely, at step n of the second procedure A3 and A6 are satisfied. The absence of off-the-path restrictions allows to claim A4 and A5 and apply Lemma 4 (first statement), which together with Lemma 1 implies that both deviator and co-players will play at step n of the second procedure any sequential best reply to any CPS in the subgame. But then, with a reduction procedure, I can find an set of strategy profiles of the subgame which should have survived also in the first procedure, a contradiction. To prove this last statement, I can apply the same logics, using as second procedure the implications in the subgame of the first procedure and as first procedure the new reduction procedure. For the latter, A5 holds by construction, for the former A4 holds by the previous application of Lemma 4 (second statement). Thus, the problem is recursive and ends when the post-deviation subgame, by finiteness, is just a simultaneous moves game.

The same reasoning applies when the first procedure refers to a looser agreement and the second procedure to the path agreement in case the first procedure implements the path, so that the set it yields is compatible with the path restrictions.

Technically, the proof shows that A3, A4, A5 and A6 cannot hold together also for player l, in this particular structure, because otherwise the second statement of Lemma 4 would hold non vacously for every $p \in \mathbb{N}$, which contradicts $\widehat{h} \in D^S$.

So, let $(\overline{S}_q)^\infty_{q=0}$ be strong-Δ-rationalizability or selective rationalizability for a path agreement e on $z \in Z$. Let $(S_q)^\infty_{q=0}$ be any of the two for e or a looser agreement e'. I show that $\zeta(S_\infty) \supseteq \zeta(\overline{S}_\infty)$. Thus, $\zeta(S^\infty_{R\Delta^e}) \subseteq \zeta(S^\infty_{\Delta^e})$ and $\zeta(S^\infty_{R\Delta^e}) \supseteq \zeta(S^\infty_{\Delta^e})$ (Theorem 1) and $\zeta(S^\infty_{e'}) \supseteq \zeta(S^\infty_e)$. Vice versa, let $(\overline{S}_q)^\infty_{q=0}$ apply to e' and $(S_q)^\infty_{q=0}$ to e. I show

that if $\zeta(\overline{S}_\infty) = z$, still $\zeta(S_\infty) \supseteq \zeta(\overline{S}_\infty)$, so $\zeta(S^\infty_{e'}) \subseteq \zeta(S^\infty_e)$ and Theorem 3 holds.

The proof of $\zeta(S_\infty) \supseteq \zeta(\overline{S}_\infty)$ is common to all cases (see the next two footnotes). To better understand the proof, I suggest to read it first for $k = 0$, second for $k = 1$, and third for $k = 2$: for $k > 2$ the logics are the same as for $k = 2$.

Let $\overline{S}_\infty \neq \emptyset$, otherwise the inclusion is trivially verified. Set $k = 0$. For notational convenience, h^k will be substituted by just k in subscripts and superscripts.

RECURSIVE STEP k

If $k = 0$, h^0 is the root of the game. If $k > 0$, h^k is defined in Recursive Step $k-1$.

If $k = 0$, let $(\overline{S}^0_q)^\infty_{q=0} := (\overline{S}_q)^\infty_{q=0}$; else, $(\overline{S}^k_q)^\infty_{q=0}$ is defined in Recursive Step $k-1$.

If $k = 0$, let $(S^0_q)^\infty_{q=0} := (S_q)^\infty_{q=0}$; else, let $(S^k_q)^\infty_{q=0} := (\overline{S}^{k-1}_q(h^k)|h^k)^\infty_{q=0}$.

For every $i \in I$, let $\overline{\overline{\mu}}^k_i : \overline{S}^k_{i,\infty} \to \Delta^{H^k}(S^k_{-i})$ be a map such that for every $s^k_i \in \overline{S}^k_{i,\infty}$, $\overline{\overline{\mu}}^k_i(s^k_i)$ strongly believes $(\overline{S}^k_{-i,q})^\infty_{q=0}$, $s^k_i \in \rho(\overline{\overline{\mu}}^k_i(s^k_i))$, and, if $k = 0$, $\overline{\overline{\mu}}^0_i(s^0_i) \in \Delta^e_i$.[52]

I prove by induction that $\zeta(\overline{S}^k_\infty) \subseteq \zeta(S^k_\infty)$.

Premise (k=0): for every $n \in \mathbb{N}$, $i \in I$, $s_i \in \overline{S}_{i,\infty}$.and $\mu_i =^{H(\overline{S}_\infty)} \overline{\overline{\mu}}^0_i(s_i)$ t.s.b. $(S_{-i,q})^n_{q=0}$ or $(\overline{S}_{-i,q})^n_{q=0}$, by $\overline{\overline{\mu}}^0_i(s_i) \in \Delta^e_i$, $\overline{\overline{\mu}}^0_i(s_i)(S_{-i}(z)|h^0) = 1$, so $\mu_i(S_{-i}(z)|h^0) = 1$. Thus $\mu_i \in \Delta^e_i \subseteq \Delta^{e'}_i$. So A4 and A5 hold.

Premise (k>0): A4 holds by Fact $k-1$. A5 holds by Claim $k-1$. (To be read in Recursive step $k-1$ with k and $k-1$ in place of $k+1$ and k)

Inductive Hypothesis (n): $(S^k_q)^\infty_{q=0}$ satisfies A3 at n (so by A4 $\zeta(S^k_n) \supseteq \zeta(\overline{S}^k_\infty)$).

Basis step (1): for every $i \in I$, the I.H. holds with $\overline{\mu}^k_i(\cdot) = \overline{\overline{\mu}}^k_i(\cdot)$.

Inductive step (n+1).

Suppose by contradiction that the Inductive Hypothesis does not hold at $n+1$. Then A6 holds for some $l \in I$ and $\widehat{s}^k_l \in \overline{S}^k_{l,\infty}$. Lemma 4 yields $h^{k+1} \in D_l(\overline{S}^k_\infty)$. If $\Gamma(h^k)$ has depth[53] 1, h^{k+1} cannot exist and we have the desired contradiction (**Exit Rule**). Else, define $((\overline{S}^{k+1}_{i,q})_{i\in I})_{q\geq 0}$ as follows: for every $i \in I$ and $m \leq n$, $\overline{S}^{k+1}_{i,m} = S^k_{i,m}(h^{k+1})|h^{k+1}$; for every $m > n$, $s^{k+1}_i \in \overline{S}^{k+1}_{i,m}$ if and only if there exists μ^{k+1}_i t.s.b. $(\overline{S}^{k+1}_{-i,q})^{m-1}_{q=0}$ such that $s^{k+1}_i \in \rho(\mu^{k+1}_i)$.

For every $i \neq l$, since $h^{k+1} \in D_l(\overline{S}^k_\infty)$, $\emptyset \neq \overline{S}^k_{i,\infty}(h^{k+1}) \ni \widehat{s}^k_i$. For every $m \leq n$, the I. H. provides $\overline{s}^k_i(\widehat{s}^k_i) \in S^k_{i,m}(h^{k+1}) \neq \emptyset$ and $\overline{\mu}^k_i(\widehat{s}^k_i) =^{H(\overline{S}^k_\infty)} \overline{\overline{\mu}}^k_i(\widehat{s}^k_i)$ t.s.b. $(S^k_{-i,q})^{m-1}_{q=0}$ such that $\overline{\mu}^k_i(\widehat{s}^k_i)(S^k_{-i}(h^{k+1})|p(h^{k+1})) = 0$. Hence, by Lemma 1, for every μ^{k+1}_i t.s.b.

[52] In case $(\overline{S}_q)^\infty_{q=0} = (S^q_{e'})^\infty_{q=0}$, I assumed $\overline{S}_{-i,\infty} \subseteq S_{-i}(z)$; hence, for every μ_i t.s.b. $\overline{S}_{-i,\infty}$, $\mu_i \in \Delta^e_i$.

[53] The difference between the lenght T of the longest path $(a^1, ..., a^T) \succ h$ and the lenght of h.

$(\overline{S}^{k+1}_{-i,q})^{m-1}_{q=0}$, there exists $\mu^k_i =^{k+1} \mu^{k+1}_i$ t.s.b. $(S^k_{-i,q})^{m-1}_{q=0}$ such that $\mu^k_i =^{H(\overline{S}^k_\infty)} \overline{\overline{\mu}}^k_i(\widehat{s}^k_i)$ and $\rho(\mu^k_i)(h^{k+1}) \neq \emptyset$. By A4 $\rho(\mu^k_i) \subseteq S^k_{i,m}$. So $\rho(\mu^{k+1}_i) \subseteq \overline{S}^{k+1}_{i,m}$.

Fix μ^{k+1}_l t.s.b. $(\overline{S}^{k+1}_{-l,q})^n_{q=0}$: trivially $\mu^{k+1}_l \in M^{k+1}_n$. Hence by Lemma 4 (1) there exists $\widetilde{\mu}^k_l =^{H(\overline{S}^k_\infty)} \overline{\overline{\mu}}^k_l(\widehat{s}^k_l)$ t.s.b. $(S^k_{-l,n})^n_{q=0}$ such that $\widetilde{\mu}^k_l =^{k+1} \mu^{k+1}_l$ and $\rho(\widetilde{\mu}^k_l)(h^{k+1}) \neq \emptyset$. By A4, $\rho(\widetilde{\mu}^k_l) \subseteq S^k_{l,n}$. So $\rho(\mu^{k+1}_l) \subseteq \overline{S}^{k+1}_{l,n} \neq \emptyset$.

Hence for every $i \in I$ and μ^{k+1}_i t.s.b. $(\overline{S}^{k+1}_{-i,q})^n_{q=0}$, $\rho(\mu^{k+1}_i) \subseteq \overline{S}^{k+1}_{i,n} \neq \emptyset$, so that $\overline{S}^{k+1}_{i,n} \supseteq \overline{S}^{k+1}_{i,n+1}$ and $((\overline{S}^{k+1}_{i,q})_{i\in I})_{q\geq 0}$ is a reduction procedure with $\overline{S}^{k+1}_\infty \neq \emptyset$.

For every $m \leq n$ and $\widehat{\mu}^{k+1}_l$ t.s.b. $(\overline{S}^{k+1}_{-l,q})^\infty_{q=0}$, fix $\mu^{k+1}_l =^{H(\overline{S}^{k+1}_\infty)} \widehat{\mu}^{k+1}_l$ t.s.b. $(\overline{S}^{k+1}_{-l,q})^{m-1}_{q=0}$. Since $\widehat{\mu}^{k+1}_l$ and $H(\overline{S}^{k+1}_\infty)$ satisfy A0, by Lemma 4 (1) there exists $\widetilde{\mu}^k_l =^{H(\overline{S}^k_\infty)} \overline{\overline{\mu}}^k_l(\widehat{s}^k_l)$ t.s.b. $(S^k_{-l,q})^{m-1}_{q=0}$ such that $\widetilde{\mu}^k_l =^{k+1} \mu^{k+1}_l$ and $\rho(\widetilde{\mu}^k_l)(h^{k+1}) \neq \emptyset$. By A4, $\rho(\widetilde{\mu}^k_l) \subseteq S^k_{l,m}$. So $\rho(\mu^{k+1}_l) \subseteq \overline{S}^{k+1}_{l,m}$.

Then, for every $m \in \mathbb{N}$, $i \in I$, $\widehat{\mu}^{k+1}_i$ t.s.b. $(\overline{S}^{k+1}_{-i,q})^\infty_{q=0}$ and $\mu^{k+1}_i =^{H(\overline{S}^{k+1}_\infty)} \widehat{\mu}^{k+1}_i$ t.s.b. $(\overline{S}^{k+1}_{-i,q})^{m-1}_{q=0}$, $\rho(\mu^{k+1}_i) \subseteq \overline{S}^{k+1}_{i,m}$. (**Claim** k)

If $\zeta(\overline{S}^{k+1}_\infty) \subseteq \zeta(\overline{S}^k_\infty)$, then $h^{k+1} \in H(\overline{S}^k_\infty)$, the desired contradiction to claim that the I. H. holds for $n+1$. The goal of Recursive Step $k+1$ is precisely to prove $\zeta(\overline{S}^{k+1}_\infty) \subseteq \zeta(\overline{S}^k_\infty)$. Before, observe what follows.

For every $i \neq l$, $m \in \mathbb{N}$ and μ^{k+1}_i t.s.b. $(\overline{S}^k_{-i,q}(h^{k+1})|h^{k+1})^{m-1}_{q=0}$, by Lemma 1 there exists $\widetilde{\mu}^k_i =^{k+1} \mu^{k+1}_i$ t.s.b. $(\overline{S}^k_{-i,q})^{m-1}_{q=0}$ such that for every $\widetilde{h} \notin H^{k+1}$, $\widetilde{\mu}^k_i(\cdot|\widetilde{h}) = \overline{\overline{\mu}}^k_i(\widehat{s}^k_i)(\cdot|\widetilde{h})$ and $\rho(\widetilde{\mu}^k_i)(h^{k+1}) \neq \emptyset$.[54] By A5, $\rho(\widetilde{\mu}^k_i) \subseteq \overline{S}^k_{i,m}$.

For every $m \in \mathbb{N}$ and $\widehat{\mu}^{k+1}_l$ t.s.b. $(\overline{S}^{k+1}_{-l,q})^\infty_{q=0}$, fix $\mu^{k+1}_l =^{H(\overline{S}^{k+1}_\infty)} \widehat{\mu}^{k+1}_l$ t.s.b. $(\overline{S}^k_{-l,q}(h^{k+1})|h^{k+1})^{m-1}_{q=0}$. Since $\widehat{\mu}^{k+1}_l$ and $H(\overline{S}^{k+1}_\infty)$ satisfy A0, by Lemma 4 (2) there exists $\widetilde{\mu}^k_l =^{H(\overline{S}^k_\infty)} \overline{\overline{\mu}}^k_l(\widehat{s}^k_l)$ t.s.b. $(\overline{S}^k_{-l,q})^{m-1}_{q=0}$ such that $\widetilde{\mu}^k_l =^{k+1} \mu^{k+1}_l$ and $\rho(\widetilde{\mu}^k_l)(h^{k+1}) \neq \emptyset$. By A5 $\rho(\widetilde{\mu}^k_l) \subseteq \overline{S}^k_{l,m}$.

Then, for every $m \in \mathbb{N}$, $i \in I$, $\widehat{\mu}^{k+1}_i$ t.s.b. $(\overline{S}^{k+1}_{-i,q})^\infty_{q=0}$ and $\mu^{k+1}_i =^{H(\overline{S}^{k+1}_\infty)} \widehat{\mu}^{k+1}_l$ t.s.b. $(\overline{S}^k_{-l,q}(h^{k+1})|h^{k+1})^{m-1}_{q=0}$, $\rho(\mu^{k+1}_i) \subseteq \overline{S}^k_{i,m}(h^{k+1})|h^{k+1}$. (**Fact** k)

Note that the Exit Rule must be verified at some k, because the game has finite depth. Then $\zeta(\overline{S}^k_\infty) \subseteq \zeta(S^k_\infty)$, the desired contradiction to claim that $\zeta(\overline{S}^{k-1}_\infty) \subseteq \zeta(S^{k-1}_\infty)$. Proceding backwards, we obtain $\zeta(\overline{S}_\infty) \subseteq \zeta(S_\infty)$. ■

The proof can be employed also to show an interesting relationship between strong rationalizability, which can be seen as strong-Δ-rationalizability under the loosest possible agreement, and strong-Δ-rationalizability. The latter does not in general deliver a subset of the former in terms of strategy profiles, nor in terms of outcomes. Yet, it does deliver a subset of outcomes when at the end of the procedure, there is no

[54] Recall that: $\overline{\overline{\mu}}^k_i(s^k_i)$ s.b. $(\overline{S}^k_{-i,q})^\infty_{q=0}$; $\overline{\overline{\mu}}^k_i(s^k_i)(S^k_{-i}(h^{k+1})|p(h^{k+1})) = 0$; $s^k_i \in \rho(\overline{\overline{\mu}}^k_i(s^k_i))(h^{k+1}) \neq \emptyset$.

restriction left off the induced paths. For coherency with the framework, restrictions are expressed here in terms of an agreement, but the result holds through for more general restrictions (such as those deriving from an outcome distribution).

Remark 7 *Fix an agreement* $e = (e_i)_{i \in I}$. *If for every* $h \notin H(S^{\infty}_{\Delta^e})$ *and* $i \in I$, $e_i(h) = \emptyset$, *then* $\zeta(S^{\infty}_{\Delta^e}) \subseteq \zeta(S^{\infty})$.

PROOF OF THEOREMS 2 AND 6.

The proofs of theorems 2 and 6 are applications of Lemma 10, which is based on the following idea. For simplicity, think of SPE and equilibria in pure strategies. When I say that a SPE survives/is eliminated I mean that an equilibrium which induces the SPE path survives/is eliminated. Fix a reduction procedure and a collection of unordered subgames, each with an associated SPE, such that until step m: (A1) the associated SPE survive; (A2) the subgames follow unilateral deviations from equilibria which survive $m-1$ steps (used in next paragraph); (A3) players save the sequential best replies to CPS's against which they yield in expectation at least the payoff of a SPE in the class of SPE that induce in the subgames the associated SPE. I show that a SPE in the restricted class survives until step m. Suppose not. Take one of the SPE in the class that survive more steps, say $n-1$. There is a unilateral deviation from its path that the deviator takes at n whatever she conjectures thereafter. This is proved in Lemma 8, which resembles Lemma 4. The same holds for the co-players since they can be surprised by the deviation. But then (see next paragraph), in the new post-deviation subgame some SPE in the class of SPE that induce in the smaller old subgames the associated SPE survives until step n. Since the new subgame cannot coincide with or be contained in an old subgame (also shown in Lemma 8), substituting the smaller old subgames with the new subgame, the new class of SPE is smaller than the first. Then, iterating, m weakly decreases. So A1,2,3 still hold. Since the subgames keep enlarging, I obtain a contradiction.

I still need to prove that a SPE of the new subgame in the restricted class survives until step n. Suppose not. However the surviving substrategies feature an equilibrium of the subgame which induces the equilibria of the smaller old subgames (I prove this inside Lemma 10 using A1, A2 and the fact that players save all the sequential best replies to CPS's at step n). There is a unilateral deviation from the equilibrium path which the deviator takes if in the post-deviation subgame she expects at least the payoff of a SPE in the class of SPE that induce in the smaller old subgames the

associated SPE. This is proved in Lemma 9 which is the dual of Lemma 8. The same holds for the co-players since they can be surprised by the deviation. This yields A3 for the new post-deviation subgame: suppose that the survival of a SPE of the new subgame in the restricted class has been proved inductively on the depth of subgames. Since the new subgame cannot coincide with or be contained in an old subgame (also shown in Lemma 9), substituting the smaller old subgames with the new subgame, the new class of SPE is smaller than the first. Then, iterating, the contrapositive hypothesis still holds. Since the subgames keep enlarging, I obtain a contradiction.

In the proof of Lemma 10, the two iterative procedures described above constitute the outer and the inner recursive proofs for the inductive step. I will always refer to uncorrelated CPS's[55] and distributions.

Additional notation:
Fix $h \preceq \widehat{h}$, $i \in J \subseteq I$, $\mu_i^h \in \Delta^{H^h}(S_{-i}^h)$, $(\widetilde{\sigma}_j^h)_{j\in I} \in \times_{j\in I}\Delta(S_j^h)$, $(\sigma_j^{\widehat{h}})_{j\in I} \in \times_{j\in I}\Delta(S_j^{\widehat{h}})$:

- $\widetilde{\sigma}_J^h(\cdot) \in \Delta(S_J^h)$ is the product of the marginal distributions $(\widetilde{\sigma}_j^h)_{j\in J}$;[56]
- $H(\widetilde{\sigma}_J^h) := H(\text{Supp}\widetilde{\sigma}_J^h)$, $\widetilde{\sigma}_J^h[\widehat{h}] := (\text{Supp}\widetilde{\sigma}_J^h)[\widehat{h}]$, $D_i(\widetilde{\sigma}^h) := D_i(\text{Supp}\widetilde{\sigma}^h)$;
- $\widetilde{\sigma}_J^h|\widehat{h}$ is the product of $(\widetilde{\sigma}_j^h|\widehat{h})_{j\in J}$ and $\widetilde{\sigma}_i^h|\widehat{h} \in \Delta(S_i^{\widehat{h}})$ is def. for every $s_i^{\widehat{h}} \in S_i^{\widehat{h}}$ as
  - $(\widetilde{\sigma}_i^h|\widehat{h})(s_i^{\widehat{h}}) = \widetilde{\sigma}_i^h(\{s_i^h \in S_i^h(\widehat{h}) : s_i^h|\widehat{h} = s_i^{\widehat{h}}\})/\widetilde{\sigma}_i^h(S_i^h(\widehat{h}))$ if $\widehat{h} \in H(\widetilde{\sigma}_i^h)$,
  - $(\widetilde{\sigma}_i^h|\widehat{h})(s_i^{\widehat{h}}) = \widetilde{\sigma}_i^h(\{s_i^h \in S_i^h : s_i^h|\widehat{h} = s_i^{\widehat{h}}\})$ else;
- $\widetilde{\sigma}_J^h =^* \sigma_J^{\widehat{h}}$ if for every $z \succ \widehat{h}$ with $p(z) \in H(\sigma^{\widehat{h}})$ and $j \in J$, $(\widetilde{\sigma}_j^h|\widehat{h})(S_j^{\widehat{h}}(z))=\sigma_j^{\widehat{h}}(S_j^{\widehat{h}}(z))$;[57]
- $\widetilde{\sigma}_J^h =^{\widehat{h}} \sigma_J^{\widehat{h}}$ if for every $z \succ \widehat{h}$ and $j \in J$, $(\widetilde{\sigma}_j^h|\widehat{h})(S_j^{\widehat{h}}(z)) = \sigma_j^{\widehat{h}}(S_j^{\widehat{h}}(z))$;
- $\mu_i^h =^* \sigma_{-i}^{\widehat{h}}$ if $\mu_i^h(\cdot|\widehat{h}) =^* \sigma_{-i}^{\widehat{h}}$; $\mu_i^h =^{\widehat{h}} \sigma_{-i}^{\widehat{h}}$ if $\mu_i^h(\cdot|\widehat{h}) =^{\widehat{h}} \sigma_{-i}^{\widehat{h}}$;
- $\pi_i(\widetilde{\sigma}^h)$ is i's exp. payoff under $\widetilde{\sigma}^h$; $\pi(\widetilde{\sigma}_{-i}^h) := \max_{\widetilde{s}_i^h \in S_i^h} \sum_{s_{-i}^h \in \text{Supp}\widetilde{\sigma}_{-i}^h} u_i(\zeta(\widetilde{s}_i^h, s_{-i}^h))\widetilde{\sigma}_{-i}^h(s_{-i}^h)$;
- $\widetilde{\sigma}_{-i}^h|\widehat{h} \geq \sigma_{-i}^{\widehat{h}}$ if $\pi(\widetilde{\sigma}_{-i}^h|\widehat{h}) \geq \pi(\sigma_{-i}^{\widehat{h}})$; $\mu_i^h \geq \widetilde{\sigma}_{-i}^h$ if $\mu_i^h(\cdot|h) \geq \widetilde{\sigma}_{-i}^h$;

[55] A CPS μ_i^h is uncorrelated if for every $\widetilde{h} \in H^h$, $\mu_i^h(\cdot|\widetilde{h}) = \times_{j\neq i}\text{Marg}_{S_j^h}\mu_i^h(\cdot|\widetilde{h})$

[56] This is an exception to the rule of subscripts: $\widetilde{\sigma}_J^h$ is not a (sub-)profile of distributions but an uncorrelated joint distribution. Equilibria $(\widetilde{\sigma}_j^h)_{j\in I}$ will be represented as the joint uncorrelated distribution $\widetilde{\sigma}^h$ they induce, and then $\widetilde{\sigma}_J^h := \text{Marg}_{S_J^h}\widetilde{\sigma}^h$.

[57] Notice that z is not necessarily a terminal history. Notice also that only the histories whose predecessor is induced with positive probability by $\sigma^{\widehat{h}}$ and not all those compatible with $\sigma_J^{\widehat{h}}$ matter.

- $\widetilde{\sigma}^h$ is a SPE of $\Gamma(h)$ if for every $\widetilde{h} \in H^h$, $\widetilde{\sigma}^h|\widetilde{h}$ is an equilibrium of $\Gamma(\widetilde{h})$;
- for any set of unordered[58] non-terminal histories $\widetilde{H} \subseteq H$ and any set of SPE $\Sigma^{\widetilde{H}} = (\sigma^{\widetilde{h}})_{\widetilde{h} \in \widetilde{H}}$ of the corresponding subgames, $E^h(\Sigma^{\widetilde{H}})$ is the set of SPE of $\Gamma(h)$ σ^h such that for every $\widetilde{h} \in \widetilde{H} \cap H^h$, $\sigma^h|\widetilde{h} = \sigma^{\widetilde{h}}$.

I will often use the fact that $=^*$ and $=^{\widehat{h}}$ are transitive and that $=^{\widehat{h}}$ implies $=^*$.[59] Moreover note that when $\widetilde{\sigma}_i^h =^* \sigma_i^{\widehat{h}}$:

♡ for every $\widetilde{h} \succ \widehat{h}$ with $p(\widetilde{h}) \in H(\sigma^{\widehat{h}})$ and $\widehat{h} \preceq \overline{h} \prec \widetilde{h}$, $(\widetilde{\sigma}_i^h|\overline{h})(S_i^{\overline{h}}(\widetilde{h})) = (\sigma_i^{\widehat{h}}|\overline{h})(S_i^{\overline{h}}(\widetilde{h}))$;[60]

♠ if for every $j \neq i$ and $\widetilde{h} \in D_j(\sigma^{\widehat{h}})$, $\widetilde{\sigma}_i^h|\widetilde{h} = \sigma_i^{\widehat{h}}|\widetilde{h}$, then $\widetilde{\sigma}_i^h =^{\widehat{h}} \sigma_i^{\widehat{h}}$.[61]

When $\sigma^{\widehat{h}}$ is an equilibrium and $\widetilde{\sigma}^h =^{\widehat{h}} \sigma^{\widehat{h}}$, I will often use the fact that for every $\widetilde{h} \in H(\sigma^{\widehat{h}})$, $\widetilde{\sigma}^h|\widetilde{h}$ is an equilibrium. Moreover:

♦ if $\widetilde{\sigma}_{-i}^{\widehat{h}} =^* \sigma_{-i}^{\widehat{h}}$, $\pi(\widetilde{\sigma}_{-i}^{\widehat{h}}) \geq \pi(\sigma_{-i}^{\widehat{h}}) = \pi_i(\sigma^{\widehat{h}})$ and if $\widetilde{\sigma}^{\widehat{h}}$ is an equil., $\pi(\widetilde{\sigma}_{-i}^{\widehat{h}}) = \pi(\sigma_{-i}^{\widehat{h}})$;[62]

♣ if $\mu_i^{\widehat{h}} =^{\widehat{h}} \sigma_{-i}^{\widehat{h}}$, for every $s_i^{\widehat{h}} \in \text{Supp}\sigma_i^{\widehat{h}}$, there is $\widetilde{s}_i^{\widehat{h}} \in \rho(\mu_i^{\widehat{h}})$ such that $\widetilde{s}_i^{\widehat{h}} =^{H(\sigma^{\widehat{h}})} s_i^{\widehat{h}}$.[63]

For the next five Lemmata, fix $n \in \mathbb{N}$, $h \in H$, a red. procedure $(S_q^h)_{q \geq 0}$ and $\widetilde{\sigma}^h \in \Delta(S_{n-1}^h)$. Let $H^\sigma := H(\widetilde{\sigma}^h)$. For every $i \in I$, let $D_i := D_i(\widetilde{\sigma}^h)$ and $D_{-i} := \cup_{j \neq i} D_j$.

The first Lemma, like in the proof of Lemma 3, combines different reactions of player i to unexpected deviations from H^σ. Its proof is in the Online Appendix too.

Lemma 5 *Fix $v \leq n$ and $i \in I$ such that $\widetilde{\sigma}_i^h \in \Delta(r(\widetilde{\sigma}_{-i}^h))$ and for every $\mu_i^h =^* \widetilde{\sigma}_{-i}^h$ t.s.b. $(S_{-i,q}^h)_{q=0}^{v-1}$, $\rho(\mu_i^h) \subseteq S_{i,v}^h$. For every $\widetilde{h} \in D_{-i}$, fix $\widehat{\sigma}_i^{\widetilde{h}} \in \Delta(S_{i,v}^h(\widetilde{h})|\widetilde{h})$. There exists $\widehat{\sigma}_i^h \in \Delta(S_{i,v}^h)$ such that $\widehat{\sigma}_i^h =^* \widetilde{\sigma}_i^h$ and for every $\widetilde{h} \in D_{-i}$, $\widehat{\sigma}_i^h|\widetilde{h} = \widehat{\sigma}_i^{\widetilde{h}}$.*

[58] For every two histories h, h' in the set, $h \not\succeq h'$ and $h' \not\succeq h$.

[59] This fact resembles the set monotonicity of $=^{\widetilde{H}}$ for CPS's.

[60] The equivalent condition for CPS's is incorporated in the definition of $=^{\widehat{H}}$. Here it can be derived from the conjectures at $\widehat{h}$ because $\overline{h}$ is reached with positive probability.

[61] This is the analogous of ♠ for CPS's. Since $\widetilde{\sigma}_i^h =^* \sigma_i^{\widehat{h}}$, $(\widetilde{\sigma}_i^h|\widehat{h})(S_i^{\widehat{h}}(\widetilde{h})) = \sigma_i^{\widehat{h}}(S_i^{\widehat{h}}(\widetilde{h}))$. For every $z \succ \widetilde{h}$, since $\widetilde{\sigma}_i^h|\widetilde{h} = \sigma_i^{\widehat{h}}|\widetilde{h}$, $(\widetilde{\sigma}_i^h|\widetilde{h})(S_i^{\widetilde{h}}(z)) = (\sigma_i^{\widehat{h}}|\widetilde{h})(S_i^{\widetilde{h}}(z))$. Together, $(\widetilde{\sigma}_i^h|\widehat{h})(S_i^{\widehat{h}}(z)) = \sigma_i^{\widehat{h}}(S_i^{\widehat{h}}(z))$.

[62] The equivalent payoff relation for CPS's was implied by $\mu_i^{\widehat{h}} =^{H_{\widehat{h}}} \widehat{\mu}_i^{\widehat{h}}$ where $\widehat{\mu}_i^{\widehat{h}}$ and $H_{\widehat{h}}$ satified A0, but the latter condition instead of its implication for payoffs was employed via Lemma 2.

[63] From the observation in Section 3 about the relationship between continuation and sequential best replies. It is the analogous of Lemma 2 for equilibria, net of indifferences: if the CPS has an initial equilibrium conjecture, there are sequential best replies that mimic the equilibrium strategies within the histories that precede the equilibrium outcomes.

For the next four lemmata suppose that $\widetilde{\sigma}^h$ is an equilibrium and:

A0 for every $v \leq n$, $i \in I$ and $\mu_i^h =^* \widetilde{\sigma}_{-i}^h$ t.s.b. $(S_{-i,q}^h)_{q=0}^{v-1}$, $\rho(\mu_i^h) \subseteq S_{i,v}^h$,

so that every $i \in I$ satisfies the hypotheses of Lemma 5.

Lemma 6 is a characterization of equilibrium which will turn out to be useful. Since the arguments for it are standard, the proof is omitted.

Lemma 6 *Fix $(\widehat{\sigma}_i^h)_{i\in I} \in \times_{i\in I}\Delta(S_i^h)$: $\widehat{\sigma}^h$ is an equilibrium if and only if for every $\overline{h} \in H(\widehat{\sigma}^h)$, $i \in I$ and $a_i \in A_i(\overline{h})\backslash\widehat{\sigma}_i^h[\overline{h}]$, calling $H_{a_i}^{\overline{h}} := (\overline{h}, (a_i, a_{-i}))_{a_{-i}\in\widehat{\sigma}_{-i}^h[\overline{h}]}$,*

$$\sum_{\widetilde{h}\in H_{a_i}^{\overline{h}}\backslash Z} \pi(\widehat{\sigma}_{-i}^h|\widetilde{h}) \cdot (\widehat{\sigma}_{-i}^h|\overline{h})(S_{-i}^{\overline{h}}(\widetilde{h})) + \sum_{z\in H_{a_i}^{\overline{h}}\cap Z} u_i(z) \cdot (\widehat{\sigma}_{-i}^h|\overline{h})(S_{-i}^{\overline{h}}(z)) \leq \pi_i(\widehat{\sigma}^h|\overline{h}). \quad (\bigstar)$$

Lemma 7 converts a condition on CPS's into $\bigstar$ for some related conjectures.

Lemma 7 *Fix $\widehat{\sigma}^h =^* \widetilde{\sigma}^h$, $\overline{h} \in H^\sigma$, $i \in I$, $a_i \in A_i(\overline{h})\backslash\widehat{\sigma}_i^h[\overline{h}]$, $\widehat{h} \in H_{a_i}^{\overline{h}}\backslash Z$, $v \leq n$ and $\widetilde{\mu}_i^{\widehat{h}}$ t.s.b. $(S_{-i,q}^h(\widehat{h})|\widehat{h})_{q=0}^v$ such that (i) $\widehat{\sigma}_{-i}^h|\widehat{h} \leq \widetilde{\mu}_i^{\widehat{h}}$, (ii) for every $\widetilde{h} \in H_{a_i}^{\overline{h}}\backslash(Z \cup \{\widehat{h}\})$, $\widehat{\sigma}_{-i}^h|\widetilde{h} \leq \overline{\sigma}_{-i}^{\widetilde{h}}$ for some $\overline{\sigma}_{-i}^{\widetilde{h}} \in \Delta(S_{-i,v}^h(\widetilde{h})|\widetilde{h})$, and (iii) for every $\mu_i^h =^* \widetilde{\sigma}_{-i}^h$ t.s.b. $(S_{-i,q}^h)_{q=0}^{v-1}$, if $\mu_i^h =^{\widehat{h}} \widetilde{\mu}_i^{\widehat{h}}$, then $\rho(\mu_i^h)(\widehat{h}) = \emptyset$. Then $\bigstar$ holds.*

Proof. Let $\overline{\sigma}_{-i}^{\widehat{h}} := \widetilde{\mu}_i^{\widehat{h}}(\cdot|\widehat{h})$. By Lemma 5 there exists $\overline{\sigma}_{-i}^h \in \Delta(S_v^h)$ such that $\overline{\sigma}_{-i}^h =^* \widetilde{\sigma}_{-i}^h$, for every $\widetilde{h} \in H_{a_i}^{\overline{h}}\backslash Z$, $\overline{\sigma}_{-i}^h|\widetilde{h} = \overline{\sigma}_{-i}^{\widetilde{h}}$ and for every $\widetilde{h} \in D_i\backslash H_{a_i}^{\overline{h}}$, $\overline{\sigma}_{-i}^h|\widetilde{h} = \widetilde{\sigma}_{-i}^h|\widetilde{h}$. Fix $\mu_i^h =^h \overline{\sigma}_{-i}^h$ t.s.b. $(S_{-i,q}^h)_{q=0}^v$ such that $\mu_i^h =^{\widehat{h}} \widetilde{\mu}_i^{\widehat{h}}$ (one exists because $\mu_i^h(\cdot|h) = \overline{\sigma}_{-i}^h$ implies $\mu_i^h(\cdot|\widehat{h})|\widehat{h} = \overline{\sigma}_{-i}^h|\widehat{h} = \widetilde{\mu}_i^{\widehat{h}}(\cdot|\widehat{h})$).

For every $\widetilde{h} \in D_i\backslash H_{a_i}^{\overline{h}}$ and $z \in Z$ such that $z \succ \widetilde{h}$, by the argument used for $\spadesuit$, $\overline{\sigma}_{-i}^h(S_{-i}^h(z)) = \widetilde{\sigma}_{-i}^h(S_{-i}^h(z))$. For every $z \in \varsigma(\mathrm{Supp}\widetilde{\sigma}^h)$, by $\overline{\sigma}^h =^* \widetilde{\sigma}^h$, $\overline{\sigma}_{-i}^h(S_{-i}^h(z)) = \widetilde{\sigma}_{-i}^h(S_{-i}^h(z))$. Hence every $s_i^h \notin S_i^h(\widehat{h}) = \cup_{\widetilde{h}\in H_{a_i}^{\overline{h}}} S_i^h(\widetilde{h})$ induces with $\overline{\sigma}_{-i}^h$ and $\widetilde{\sigma}_{-i}^h$ (1) and with $\overline{\sigma}_{-i}^h|\overline{h}$ and $\widetilde{\sigma}_{-i}^h|\overline{h}$ (2) the same distribution over outcomes. By 1 and $r(\widetilde{\sigma}_{-i}^h)(\overline{h}) \neq \emptyset$, $r(\overline{\sigma}_{-i}^h)(\overline{h}) \cup r(\overline{\sigma}_{-i}^h)(\widehat{h}) \neq \emptyset$; by $\overline{h} \prec \widehat{h}$, $r(\overline{\sigma}_{-i}^h)(\overline{h}) \neq \emptyset$; by $\overline{h} \in H(\overline{\sigma}_{-i}^h)$, $\rho(\mu_i^h)(\overline{h}) \neq \emptyset$;[64] by $\rho(\mu_i^h)(\widehat{h}) = \emptyset$, $\widehat{r}(\mu_i^h, \overline{h})(\widehat{h}) = \emptyset$; by $\mu_i^h(\cdot|\overline{h})|\overline{h} = \overline{\sigma}_{-i}^h|\overline{h}$, $r(\overline{\sigma}_{-i}^h|\overline{h})(\widehat{h}) = \emptyset$; by 2, $\pi(\overline{\sigma}_{-i}^h|\overline{h}) = \pi(\widetilde{\sigma}_{-i}^h|\overline{h}) = \pi_i(\widetilde{\sigma}^h|\overline{h})$; thus,

$$\sum_{\widetilde{h}\in H_{a_i}^{\overline{h}}\backslash Z} \pi(\overline{\sigma}_{-i}^{\widetilde{h}}) \cdot (\overline{\sigma}_{-i}^h|\overline{h})(S_{-i}^{\overline{h}}(\widetilde{h})) + \sum_{z\in H_{a_i}^{\overline{h}}\cap Z} u_i(z) \cdot (\overline{\sigma}_{-i}^h|\overline{h})(S_{-i}^{\overline{h}}(z)) \leq \pi_i(\widetilde{\sigma}^h|\overline{h}).$$

[64] See the relationship between continuation and sequential best replies in Section 3.

By $\widetilde{\sigma}^h =^* \widehat{\sigma}^h$, $\pi_i(\widetilde{\sigma}^h|\overline{h}) = \pi_i(\widehat{\sigma}^h|\overline{h})$. By $\overline{\sigma}^h_{-i} =^* \widetilde{\sigma}^h_{-i} =^* \widehat{\sigma}^h_{-i}$ and $\heartsuit$, for every $\widetilde{h} \in H^{\overline{h}}_{a_i}$, $(\overline{\sigma}^h_{-i}|\overline{h})(S^{\overline{h}}_{-i}(\widetilde{h})) = (\widehat{\sigma}^h_{-i}|\overline{h})(S^{\overline{h}}_{-i}(\widetilde{h}))$. For each $\widetilde{h} \in H^{\overline{h}}_{a_i} \backslash Z$, $\overline{\sigma}^{\widetilde{h}}_{-i} \geq \widehat{\sigma}^h_{-i}|\widetilde{h}$. So $\bigstar$ holds. ■

For the next two Lemmata, fix a set of unordered histories $\widehat{H} \subseteq H^h$ and a set of SPE $\Sigma^{\widehat{H}} = (\sigma^{\widetilde{h}})_{\widetilde{h} \in \widehat{H}}$ such that:

A1 for every $\widetilde{h} \in \widehat{H}$, there exists an equilibrium $\widetilde{\sigma}^{\widetilde{h}} =^* \sigma^{\widetilde{h}}$ such that $\widetilde{\sigma}^{\widetilde{h}} \in \Delta(S^h_n(\widetilde{h})|\widetilde{h})$.

Lemma 8 augments Lemma 4 for the simpler case in which $\overline{S}^h$ is the support of a SPE. If only until step $n-1$ an equilibrium that mimics the SPE survives, at step n one deviation is always *strictly* preferred to continuing as the equilibrium prescribes.

Lemma 8 *Suppose that there exists $\sigma^h \in E^h(\Sigma^{\widehat{H}})$ such that $\widetilde{\sigma}^h =^* \sigma^h$ but there is no equilibrium $\widehat{\sigma}^h \in \Delta(S^h_n)$ such that $\widehat{\sigma}^h =^* \sigma^h$. Then there exist $l \in I$ and $\widehat{h} \in D_l \backslash (\cup_{\widetilde{h} \in \widehat{H}} H^{\widetilde{h}})$ such that for every $\widehat{\sigma}^{\widehat{h}}_{-l} \in \Delta(S^h_{-l,n}(\widehat{h})|\widehat{h})$, $v \leq n$ and $\widetilde{\mu}^{\widehat{h}}_l \geq \widehat{\sigma}^{\widehat{h}}_{-l}$ t.s.b. $(S^h_{-l,q}(\widehat{h})|\widehat{h})^v_{q=0}$, there exists $\widetilde{\mu}^h_l =^* \widetilde{\sigma}^h_{-l}$ t.s.b. $(S^h_{-l,q})^{v-1}_{q=0}$ such that $\widetilde{\mu}^h_l =^{\widehat{h}} \widetilde{\mu}^{\widehat{h}}_l$ and $\rho(\widetilde{\mu}^h_l)(\widehat{h}) \neq \emptyset$ (so by A0 $\rho(\widetilde{\mu}^h_l) \subseteq S^h_{l,v}$).*

Proof. Suppose not. For every $i \in I$ and $\widehat{h} \in D_i \backslash (\cup_{\widetilde{h} \in \widehat{H}} H^{\widetilde{h}}) =: \overline{D}_i$, fix $\widehat{\sigma}^{\widehat{h}}_{-i}$, $v(\widehat{h})$ and $\widetilde{\mu}^{\widehat{h}}_i$ that violate the statement and let $\widetilde{\sigma}^{\widehat{h}}_{-i} := \widehat{\sigma}^{\widehat{h}}_{-i}$.

By Lemma 5 there exists $\widehat{\sigma}^h \in \Delta(S^h_n)$ such that $\widehat{\sigma}^h =^* \widetilde{\sigma}^h =^* \sigma^h$ and for every $i \in I$, $\widetilde{h} \in \widehat{H} \cup \overline{D}_i$ and $\widehat{h} \in D_i \cap H^{\widetilde{h}}$, $\widehat{\sigma}^h_{-i}|\widehat{h} = \widetilde{\sigma}^{\widetilde{h}}_{-i}|\widehat{h}$. I show that $\widehat{\sigma}^h$ is an equilibrium, a contradiction.

Fix $\overline{h} \in H^\sigma$, $i \in I$ and $a_i \in A_i(\overline{h}) \backslash \widetilde{\sigma}^h_i[\overline{h}]$. If there exists $\widetilde{h} \preceq \overline{h}$ such that $\widetilde{h} \in \widehat{H}$, $\widehat{\sigma}^h =^* \sigma^h =^{\widetilde{h}} \sigma^{\widetilde{h}} =^* \widetilde{\sigma}^{\widetilde{h}}$, so by $\spadesuit$ $\widehat{\sigma}^h =^{\widetilde{h}} \widetilde{\sigma}^{\widetilde{h}}$; then $\widehat{\sigma}^h|\overline{h}$ is an equilibrium, so by Lemma 6 ("only if") $\bigstar$ holds. If $H^{\overline{h}}_{a_i} \backslash Z \subseteq \widehat{H}$, for every $\widehat{h} \in H^{\overline{h}}_{a_i} \backslash Z$, $\widehat{\sigma}^h_{-i}|\widehat{h} = \widetilde{\sigma}^{\widehat{h}}_{-i}$ and by $\blacklozenge$, $\pi(\widetilde{\sigma}^{\widehat{h}}_{-i}) = \pi(\sigma^h_{-i}|\widehat{h})$; by $\widehat{\sigma}^h =^* \sigma^h$, $\pi_i(\widehat{\sigma}^h|\overline{h}) = \pi_i(\sigma^h|\overline{h})$ and by $\heartsuit$, $(\widehat{\sigma}^h_{-i}|\overline{h})(S^{\overline{h}}_{-i}(\widehat{h})) = (\sigma^h_{-i}|\overline{h})(S^{\overline{h}}_{-i}(\widehat{h}))$; so since $\sigma^h|\overline{h}$ is an equilibrium by Lemma 6 ("only if") $\bigstar$ holds. If $H^{\overline{h}}_{a_i} \cap \overline{D}_i \neq \emptyset$, fix $v := \min_{\widetilde{h} \in H^{\overline{h}}_{a_i} \cap \overline{D}_i} v(\widetilde{h})$ and $\widehat{h} := \arg\min_{\widetilde{h} \in H^{\overline{h}}_{a_i} \cap \overline{D}_i} v(\widetilde{h})$: for every $\widetilde{h} \in H^{\overline{h}}_{a_i} \cap \overline{D}_i$, $\widehat{\sigma}^h_{-i}|\widetilde{h} \leq \widetilde{\mu}^{\widetilde{h}}_i(\cdot|\widetilde{h}) \in \Delta(S^h_{-i,v}(\widetilde{h})|\widetilde{h}) \neq \emptyset$[65] and for every $\widetilde{h} \in H^{\overline{h}}_{a_i} \backslash (Z \cup \overline{D}_i)$, $\widehat{\sigma}^h_{-i}|\widetilde{h} \in \Delta(S^h_{-i,v}(\widetilde{h})|\widetilde{h})$; therefore by Lemma 7 $\bigstar$ holds. Thus by Lemma 6 ("if"), $\widehat{\sigma}^h$ is an equilibrium. ■

Lemma 9 is the "dual" of Lemma 8: if an equilibrium has survived n steps but it does not mimic a SPE (within a subset), then there is a deviation from one of the

[65] For every $j \neq i$, there is $\mu^h_j =^h \widetilde{\sigma}^h_{-j}$ t.s.b. $(S^h_{-l,q})^{n-1}_{q=0}$: by $\clubsuit$, $\rho(\mu^h_j)(\widetilde{h}) \neq \emptyset$; by A0 $\rho(\mu^h_j) \subseteq S^h_{j,n}$.

equilibrium paths that the deviator could take whenever thereafter she expects at least the payoff of a SPE of the subgame (within a subset).

Lemma 9 *Suppose that* $\widetilde{\sigma}^h \in \Delta(S^h_n)$ *and for every* $\widetilde{h} \in \widehat{H}$, $\widetilde{\sigma}^h =^{\widetilde{h}} \widetilde{\sigma}^{\widetilde{h}}$, *but there is no* $\sigma^h \in E^h(\Sigma^{\widehat{H}})$ *with* $\widetilde{\sigma}^h =^* \sigma^h$. *Then there exist* $p \in I$ *and* $\overline{h} \in D_p(\widetilde{\sigma}^h) \backslash (\cup_{\widetilde{h} \in \widehat{H}} H^{\widetilde{h}})$ *such that for every* $\sigma^{\overline{h}} \in E^{\overline{h}}(\Sigma^{\widehat{H}})$, $v \leq n$ *and* $\widetilde{\mu}^{\overline{h}}_p \geq \sigma^{\overline{h}}_{-p}$ *t.s.b.* $(S^h_{-p,q}(\overline{h})|\overline{h})^v_{q=0}$, *there exists* $\widetilde{\mu}^h_p =^* \widetilde{\sigma}^h_{-p}$ *t.s.b.* $(S^h_{-p,q})^{v-1}_{q=0}$ *such that* $\widetilde{\mu}^h_p =^{\overline{h}} \widetilde{\mu}^{\overline{h}}_p$ *and* $\rho(\widetilde{\mu}^h_p)(\overline{h}) \neq \emptyset$ *(so by A0* $\rho(\widetilde{\mu}^h_p) \subseteq S^h_{p,v}$*).*

Proof. Suppose not. For every $i \in I$ and $\widehat{h} \in D_i \backslash (\cup_{\widetilde{h} \in \widehat{H}} H^{\widetilde{h}}) =: \overline{D}_i$, fix $\sigma^{\widehat{h}}$, $v(\widehat{h})$ and $\widetilde{\mu}^{\widehat{h}}_i$ that violate the statement.

Construct $\widehat{\sigma}^h =^* \widetilde{\sigma}^h$ such that for every $\widetilde{h} \in \cup_{i \in I} \overline{D}_i \cup \widehat{H}$ and $\widehat{h} \in \cup_{i \in I} D_i \cap H^{\widetilde{h}}$, $\widehat{\sigma}^h|\widehat{h} = \sigma^{\widetilde{h}}|\widehat{h} \in E^{\widehat{h}}(\Sigma^{\widehat{H}})$, and for every $\widetilde{h} \notin \cup_{i \in I} D_i$ such that $p(\widetilde{h}) \in H^\sigma$, $\widehat{\sigma}^h|\widetilde{h} \in E^{\widetilde{h}}(\Sigma^{\widehat{H}})$. I show that $\widehat{\sigma}^h$ is an equilibrium such that for every $\widetilde{h} \in \widehat{H} \cap H^\sigma$, $\widehat{\sigma} =^{\widetilde{h}} \sigma^{\widetilde{h}} \in E^{\widetilde{h}}(\Sigma^{\widehat{H}})$. Then, by Lemma 6 ("only if") for every $\widetilde{h} \in H^\sigma$, $\widehat{\sigma}^h|\widetilde{h}$ is an equilibrium, and so $\widehat{\sigma}^h \in E^h(\Sigma^{\widehat{H}})$, a contradiction.

Fix $\overline{h} \in H^\sigma$, $i \in I$ and $a_i \in A_i(\overline{h}) \backslash \widetilde{\sigma}^h_i[\overline{h}]$. If there exists $\widetilde{h} \preceq \overline{h}$ such that $\widetilde{h} \in \widehat{H}$, $\widehat{\sigma}^h =^* \widetilde{\sigma}^h =^{\widetilde{h}} \widetilde{\sigma}^{\widetilde{h}} =^* \sigma^{\widetilde{h}}$, so by ♠ $\widehat{\sigma} =^{\widetilde{h}} \sigma^{\widetilde{h}}$; then $\widehat{\sigma}^h|\overline{h}$ is an equilibrium, so by Lemma 6 ("only if") ★ holds. If $H^{\overline{h}}_{a_i} \backslash Z \subseteq \widehat{H}$, for every $\widehat{h} \in H^{\overline{h}}_{a_i} \backslash Z$, $\widehat{\sigma}^h_{-i}|\widehat{h} = \sigma^{\widehat{h}}_{-i}$ and by ♦, $\pi(\sigma^{\widehat{h}}_{-i}) = \pi(\widetilde{\sigma}^h_{-i}|\widehat{h})$; by $\widehat{\sigma}^h =^* \widetilde{\sigma}^h$, $\pi_i(\widehat{\sigma}^h|\overline{h}) = \pi_i(\widetilde{\sigma}^h|\overline{h})$ and by ♡, $(\widehat{\sigma}^h_{-i}|\overline{h})(S^{\overline{h}}_{-i}(\widehat{h})) = (\widetilde{\sigma}^h_{-i}|\overline{h})(S^{\overline{h}}_{-i}(\widehat{h}))$; so since $\widetilde{\sigma}^h|\overline{h}$ is an equilibrium by Lemma 6 ("only if") ★ holds. If $H^{\overline{h}}_{a_i} \cap \overline{D}_i \neq \emptyset$, fix $v := \min_{\widetilde{h} \in H^{\overline{h}}_{a_i} \cap \overline{D}_i} v(\widetilde{h})$ and $\widehat{h} := \arg\min_{\widetilde{h} \in H^{\overline{h}}_{a_i} \cap \overline{D}_i} v(\widetilde{h})$: for every $\widetilde{h} \in H^{\overline{h}}_{a_i} \cap \overline{D}_i$, $\widehat{\sigma}^h_{-i}|\widetilde{h} \leq \widetilde{\mu}^{\widetilde{h}}_i(\cdot|\widetilde{h}) \in \Delta(S^h_{-i,v}(\widetilde{h})|\widetilde{h}) \neq \emptyset$[66] and for every $\widetilde{h} \in H^{\overline{h}}_{a_i} \backslash (Z \cup \overline{D}_i)$, $\widehat{\sigma}^h_{-i}|\widetilde{h} \leq \widetilde{\sigma}^{\widetilde{h}} \in \Delta(S^h_{-i,v}(\widetilde{h})|\widetilde{h})$; hence by Lemma 7 ★ holds. Thus by Lemma 6 ("if") $\widehat{\sigma}^h$ is an equilibrium. ■

Now I can prove the main Lemma.

Lemma 10 *Fix* $h \in H$, $m \in \mathbb{N}$, *a red. proced.* $(S^h_q)_{q \geq 0}$, *a set of unordered histories* $\widehat{H} = \{h^1, ..., h^w\} \subseteq H^h$ *and a set of SPE* $\Sigma^{\widehat{H}} = (\sigma^{\widetilde{h}})_{\widetilde{h} \in \widehat{H}}$ *s.t. A1 holds for* $n = m$ *and:*

A2 for every $v \leq w$, *there exists an equilibrium* $\widehat{\sigma}^{h,v} \in \Delta(S^h_{m-1})$ *such that* $h^v \in \cup_{i \in I} D_i(\widehat{\sigma}^{h,v})$ *and for every* $q < v$, *if* $h^q \in H(\widehat{\sigma}^{h,v})$, $\widehat{\sigma}^{h,v} =^* \widetilde{\sigma}^{h^q}$;

A3 for each $i \in I$, $n \leq m$, $\sigma^h \in E^h(\Sigma^{\widehat{H}})$ *and* $\mu^h_i \geq \sigma^h_{-i}$ *t.s.b.* $(S^h_{-i})^{n-1}_{q=0}$, $\rho(\mu^h_i) \subseteq S^h_{i,n}$.

[66] See the previous footnote.

Then there exist $\sigma^h \in E^h(\Sigma^{\widehat{H}})$ *and an equilibrium* $\widetilde{\sigma}^h \in \Delta(S^h_m)$ *such that* $\widetilde{\sigma}^h =^* \sigma^h$.

Proof. The proof is by induction on the depth of $\Gamma(h)$.

Inductive hypothesis (d)

The Lemma holds for every $h \in H$ such that $\Gamma(h)$ has depth not bigger than d.

Basis step (1) For every $i \in I$, $n \leq m$ and equilibrium of $\Gamma(h)$ σ^h such that Supp$\sigma^h \subseteq S^h_{n-1}$, by A3 $r(\sigma^h_{-i}) \subseteq S^h_{i,n}$. Inductively, Supp$\sigma^h \subseteq S^h_m$.

INDUCTIVE STEP (d+1) Suppose not. I will find a contradiction through a recursive procedure. Set $k = 0$ and $\overline{H}^0 := \widehat{H}$.

Recursive step (k)

If $k > 0$, $\overline{H}^k$ and $\Sigma^{\overline{H}^k}$ are defined in step $k-1$. Let $n \leq m$ be the greatest $q \in \mathbb{N}$ such that there exist $\sigma^{h,k} \in E^h(\Sigma^{\overline{H}^k})$ and an equilibrium $\widetilde{\sigma}^{h,k} \in \Delta(S^h_{q-1})$ with $\widetilde{\sigma}^{h,k} =^* \sigma^{h,k}$. If $k > 0$, by the last remark of the previous steps, $E^h(\Sigma^{\overline{H}^k}) \subseteq ... \subseteq E^h(\Sigma^{\overline{H}^0})$. Then by ♦ A3 implies A0. Moreover, n weakly decreases with k. Then $\Sigma^{\overline{H}^k}$ satisfies A1 and A2 with n in place of m.[67] Lemma 8 yields $l \in I$ and $\widehat{h} \in D_l(\widetilde{\sigma}^{h,k}) \setminus \cup_{\overline{h} \in \overline{H}^k} H^{\overline{h}}$.

Define the reduction procedure $((S^{\widehat{h}}_{i,q})_{i\in I})^{\infty}_{q=0} := ((S^h_{i,q}(\widehat{h})|\widehat{h}))_{i\in I})^{\infty}_{q=0}$. Fix $i \neq l$ and $v \leq n$. For every $\widetilde{\mu}^{\widehat{h}}_i$ t.s.b. $(S^{\widehat{h}}_{-i,q})^{v-1}_{q=0}$, since $\widehat{h} \in D_l(\widetilde{\sigma}^{h,k})$, by Lemma 1 there exists $\widetilde{\mu}^h_i =^h \widetilde{\sigma}^{h,k}_{-i}$ t.s.b. $(S^h_{-i,q})^{v-1}_{q=0}$ such that $\widetilde{\mu}^h_i =^{\widehat{h}} \widetilde{\mu}^{\widehat{h}}_i$. By A0, $\rho(\widetilde{\mu}^h_i) \subseteq S^h_{i,v}$; by ♣, $\rho(\widetilde{\mu}^h_i)(\widehat{h}) \neq \emptyset$. Hence, together with Lemma 8, for every $i \in I$, $v \leq n$, $\widehat{\sigma}^{\widehat{h}}_{-i} \in \Delta(S^{\widehat{h}}_{-i,n})$ and $\widetilde{\mu}^{\widehat{h}}_i \geq \widehat{\sigma}^{\widehat{h}}_{-i}$ t.s.b. $(S^{\widehat{h}}_{-i,q})^v_{q=0}$, $\rho(\widetilde{\mu}^{\widehat{h}}_i) \subseteq S^{\widehat{h}}_{i,v} \neq \emptyset$ (**F**).

Let $\widetilde{H}^0 := \overline{H}^k \cap H^{\widehat{h}}$: I show that there exist $\sigma^{\widehat{h}} \in E^{\widehat{h}}(\Sigma^{\widetilde{H}^0})$ and an equilibrium $\widetilde{\sigma}^{\widehat{h}} =^* \sigma^{\widehat{h}}$ such that Supp$\widetilde{\sigma}^{\widehat{h}} \subseteq S^{\widehat{h}}_n$. Suppose not (**G**). I will find a contradiction through a recursive procedure. For every $q \leq w + k$ such that $h^q \in \widetilde{H}^0$, let $\overline{h}^q := h^q$ and $\widehat{\sigma}^{\widehat{h},q} := \widehat{\sigma}^{h,q}|\widehat{h}$, which is an equilibrium because $\cup_{\overline{h}\in\overline{H}^k} H^{\overline{h}} \not\ni \widehat{h} \prec \overline{h}^q \in \cup_{i\in I} D_i(\widehat{\sigma}^{h,q})$, so $\widehat{h} \in H(\widehat{\sigma}^{h,q})$. Set $t = 0$.

Recursive step (t) *If* $t > 0$, $\widetilde{H}^t$ *and* $\Sigma^{\widetilde{H}^t}$ *are defined in step* $t-1$, *and satisfy A1 and A2 with* n *in place of* m *and* $\widehat{h}$ *in place of* h.[68] *For every* $i \in I$, *let* $\Sigma^{\widehat{h},t}_i$ *be the set of* $\widehat{\sigma}^{\widehat{h}}_i \in \Delta(S^{\widehat{h}}_{i,n})$ *such that for every* $\widetilde{h} \in \widetilde{H}^t \cap H(\widehat{\sigma}^{\widehat{h}}_i)$, $\widehat{\sigma}^{\widehat{h}}_i =^{\widetilde{h}} \widetilde{\sigma}^{\widetilde{h}}_i$.

First I show that $\Sigma^{\widehat{h},t}$ *is non-empty and features an equilibrium of* $\Gamma(\widehat{h})$. *Let*

[67] For every $h^q \in \overline{H}^k$, $\widetilde{\sigma}^{h^q} \in \Delta(S^h_n(h^q)|h^q)$ and $\widehat{\sigma}^{h,q} \in \Delta(S^h_{n-1})$ come from A1 and A2 if $h^q \in \widehat{H}$, from some previous step if $h^q \notin \widehat{H}$.

[68] For every $\overline{h}^q \in \widetilde{H}^t$, $\widetilde{\sigma}^{\overline{h}^q} \in \Delta(S^h_n(\overline{h}^q)|\overline{h}^q)$ and $\widehat{\sigma}^{\widehat{h},q} \in \Delta(S^{\widehat{h}}_{n-1})$ come from the outer recursive step if $\overline{h}^q \in \overline{H}^k$, from some previous step if $\overline{h}^q \notin \overline{H}^k$.

$\tau := w + k + t$. *Note that for every* $i \in I$,

$$\Sigma_i^{\widehat{h},t} = \cap_{\widetilde{h} \in \widetilde{H}^t} \cap_{z \in Z^{\widetilde{h}} \cup H^{\widetilde{h}}} \{\widehat{\sigma}_i^{\widehat{h}} \in \Delta(S_i^{\widehat{h}}) : \widehat{\sigma}_i^{\widehat{h}}(S_i^{\widehat{h}}(z)) = \widehat{\sigma}_i^{\widehat{h}}(S_i^{\widehat{h}}(\widetilde{h})) \cdot \widetilde{\sigma}_i^{\widetilde{h}}(S_i^{\widetilde{h}}(z))\} \cap \Delta(S_{i,n}^{\widehat{h}}),$$

an intersection of convex and compact sets.[69] *Hence* $\Sigma_i^{\widehat{h},t}$ *is convex and compact. Then, since expected utility is linear, the reduced game with strategy sets* $(\Sigma_i^{\widehat{h},t})_{i \in I}$*, if non-empty, features an equilibrium* $\widetilde{\sigma}^{\widehat{h},t}$*. Fix* $i \in I$ *and* $\mu_i^{\widehat{h}} =^{\widehat{h}} \widetilde{\sigma}_{-i}^{\widehat{h},t}$ *t.s.b.* $(S_{-i,q}^{\widehat{h}})_{q=0}^n$*. There exists* $\widehat{\sigma}_i^{\widehat{h},\tau+1} \in \Delta(\rho(\mu_i^{\widehat{h}})) \subseteq \Delta(r(\widetilde{\sigma}_{-i}^{\widehat{h},t}))$ *such that for every* $\widetilde{h} \in \widetilde{H}^t \cap H(\widetilde{\sigma}_{-i}^{\widehat{h},t})$*, since* $\widetilde{\sigma}_{-i}^{\widehat{h},t} =^{\widetilde{h}} \widetilde{\sigma}_{-i}^{\widetilde{h}}$*, by* ♣ $\widehat{\sigma}_i^{\widehat{h},\tau+1} =^* \widetilde{\sigma}_i^{\widetilde{h}}$*. By F,* $\widehat{\sigma}_i^{\widehat{h},\tau+1} \in \Delta(S_{i,n}^{\widehat{h}})$*. If* $t = 0$ *and* $\widetilde{H}^0 = \emptyset$*,* $\Sigma^{\widehat{h},0} = \Delta(S_n^{\widehat{h}}) \neq \emptyset$ *(by* F*) and* $\widehat{\sigma}_i^{\widehat{h},\tau+1} \in \Sigma_i^{\widehat{h},0}$*, so that* $\widetilde{\sigma}_i^{\widehat{h},t} \in \Delta(r(\widetilde{\sigma}_{-i}^{\widehat{h},t}))$ *too. Else, for notational convenience let* $\widehat{\sigma}_{-i}^{\widehat{h},\tau+1} := \widetilde{\sigma}_{-i}^{\widehat{h},t}$ *and proceed as follows.*

For every $\widetilde{h} \in H^{\widehat{h}}$ *and* $q \leq \tau$*, let* $Q_q^{\widetilde{h}} := \{g \leq q : \overline{h}^g \in \widetilde{H}^t \cap H^{\widetilde{h}}\}$ *and* $Q_{\tau+1}^{\widetilde{h}} := Q_\tau^{\widetilde{h}}$*. I show that for every* $q \in Q_\tau^{\widehat{h}} \cup \{\tau+1\}$*, there exists* $\overline{\sigma}_i^{\widehat{h},q} \in \Delta(S_{i,n}^{\widehat{h}})$ *such that* $\overline{\sigma}_i^{\widehat{h},q} =^* \widehat{\sigma}_i^{\widehat{h},q}$ *and for every* $g \in Q_q^{\widehat{h}}$*,* $\overline{\sigma}_i^{\widehat{h},q} =^{\overline{h}^g} \widetilde{\sigma}_i^{\overline{h}^g}$*. Fix* $q \in Q_\tau^{\widehat{h}} \cup \{\tau+1\}$ *and suppose to have shown it already for every* $g \in Q_q^{\widehat{h}} \backslash \{q\}$*. Since* $\widehat{\sigma}_i^{\widehat{h},q} \in \Delta(r(\widehat{\sigma}_{-i}^{\widehat{h},q}))$*, by* $\widehat{\sigma}^{\widehat{h},q} \in \Delta(S_n^{\widehat{h}})$ *and F, Lemma 5 yields* $\overline{\sigma}_i^{\widehat{h},q} \in \Delta(S_{i,n}^{\widehat{h}})$ *such that* $\overline{\sigma}_i^{\widehat{h},q} =^* \widehat{\sigma}_i^{\widehat{h},q}$ *and for every* $\widetilde{h} \in D_{-i}(\widehat{\sigma}^{h,q})$*: if* $\widetilde{h} \in \widetilde{H}^t$*,* $\overline{\sigma}_i^{\widehat{h},q}|\widetilde{h} = \widetilde{\sigma}_i^{\widetilde{h}}$*; if* $\widetilde{h} \notin \widetilde{H}^t$ *but* $Q_q^{\widetilde{h}} \neq \emptyset$*,* $\overline{\sigma}_i^{\widehat{h},q}|\widetilde{h} = \overline{\sigma}_i^{\widehat{h},\max Q_q^{\widetilde{h}}}|\widetilde{h}$ *(where* $\max Q_q^{\widetilde{h}} < q$ *because* $\overline{h}^q \in \widetilde{H}^t \cap (\cup_{j \in I} D_j(\widehat{\sigma}^{h,q}))$ *and* $\widetilde{H}^t \cap (\cup_{j \in I} D_j(\widehat{\sigma}^{h,q})) \cap H^{\widetilde{h}} = \emptyset$*); if* $\widetilde{h} \succ \overline{h}$ *for some* $\overline{h} \in \widetilde{H}^t$*,* $\overline{\sigma}_i^{\widehat{h},q}|\widetilde{h} = \widetilde{\sigma}_i^{\overline{h}}|\widetilde{h}$*, so that by* $\overline{\sigma}_i^{\widehat{h},q} =^* \widehat{\sigma}_i^{\widehat{h},q} =^* \widetilde{\sigma}_i^{\overline{h}}$ *and* ♠*,* $\overline{\sigma}_i^{\widehat{h},q} =^{\overline{h}} \widetilde{\sigma}_i^{\overline{h}}$*. Then* $\overline{\sigma}_i^{\widehat{h},\tau} \in \Sigma_i^{\widehat{h},t} \neq \emptyset$*.*[70] *So* $\widetilde{\sigma}^{\widehat{h},t}$ *and* $\widehat{\sigma}^{\widehat{h},\tau+1}$ *exist and* $\overline{\sigma}_i^{\widehat{h},\tau+1} \in \Sigma_i^{\widehat{h},t}$*. Since* $\overline{\sigma}_i^{\widehat{h},\tau+1} =^* \widehat{\sigma}_i^{\widehat{h},\tau+1}$*,* $\overline{\sigma}_i^{\widehat{h},\tau+1} \in \Delta(r(\widetilde{\sigma}_{-i}^{\widehat{h},t}))$*. Then* $\widetilde{\sigma}_i^{\widehat{h},t} \in \Delta(r(\widetilde{\sigma}_{-i}^{\widehat{h},t}))$ *too.*

By ♦*,* F *implies A0 with* $\widehat{h}$ *in place of* h*;* $\widetilde{H}^t$ *satifies A1 with* n *in place of* m*. By the last remark of the previous steps* $E^h(\Sigma^{\widetilde{H}^t}) \subseteq E^h(\Sigma^{\widetilde{H}^0})$*, so, by* G*,* $\widetilde{\sigma}^{\widehat{h},t}$ *satisfies the hypotheses of Lemma 9.*[71] *Lemma 9 yields* $p \in I$ *and* $\overline{h} \in D_p(\widetilde{\sigma}^{\widehat{h},t}) \backslash \cup_{\widetilde{h} \in \widetilde{H}^t} H^{\widetilde{h}}$*.*

Define the reduction procedure $((S_{i,q}^{\overline{h}})_{i \in I})_{q=0}^\infty := ((S_{i,q}^{\widehat{h}}(\overline{h})|\overline{h}))_{i \in I})_{q=0}^\infty$*. Fix* $i \neq p$ *and* $v \leq n$*. For every* $\widetilde{\mu}_i^{\overline{h}}$ *t.s.b.* $(S_{-i,q}^{\overline{h}})_{q=0}^v$*, since* $\overline{h} \in D_p(\widetilde{\sigma}^{\widehat{h},t})$*, by Lemma 1 there exists* $\widetilde{\mu}_i^{\widehat{h}} =^{\widehat{h}} \widetilde{\sigma}_{-i}^{\widehat{h},t}$ *t.s.b.* $(S_{-i,q}^{\widehat{h}})_{q=0}^{v-1}$ *such that* $\widetilde{\mu}_i^{\widehat{h}} =^{\overline{h}} \widetilde{\mu}_i^{\overline{h}}$*. By A0,* $\rho(\widetilde{\mu}_i^{\widehat{h}}) \subseteq S_{i,v}^{\widehat{h}}$*; by* ♣*,* $\rho(\widetilde{\mu}_i^{\widehat{h}})(\overline{h}) \neq \emptyset$*. Hence, together with Lemma 9, for every* $i \in I$*,* $v \leq n$*,* $\sigma^{\overline{h}} \in E^{\overline{h}}(\Sigma^{\widetilde{H}^t})$ *and* $\widetilde{\mu}_i^{\overline{h}} \geq \sigma_{-i}^{\overline{h}}$ *t.s.b.* $(S_{-i,q}^{\overline{h}})_{q=0}^v$*,* $\rho(\mu_i^{\overline{h}}) \subseteq S_{i,v}^{\overline{h}}$*. Moreover, for every* $q \in Q_\tau^{\overline{h}}$*, since*

[69] Clearly, $\Delta(S_{i,n}^{\widehat{h}})$ is convex and compact. Each set of the kind $\{\widehat{\sigma}_i^{\widehat{h}} \in \Delta(S_i^{\widehat{h}}) : (\widehat{\sigma}_i^{\widehat{h}})(S_i^{\widehat{h}}(z)) = (\widehat{\sigma}_i^{\widehat{h}})(S_i^{\widehat{h}}(\widetilde{h})) \cdot c\}$, where c is a constant, is clearly convex and compact too. Notice that if $\widetilde{h} \notin H(\widehat{\sigma}_i^{\widehat{h}})$, $\widehat{\sigma}_i^{\widehat{h}}$ satisfies $(\widehat{\sigma}_i^{\widehat{h}})(S_i^{\widehat{h}}(z)) = (\widehat{\sigma}_i^{\widehat{h}})(S_i^{\widehat{h}}(\widetilde{h})) \cdot c$ as $0 = 0$.

[70] By recursive step $t-1$, $\max Q_\tau^{\widehat{h}} = \tau$.

[71] Without loss of generality assume that for every $i \in I$ and $\widetilde{h} \in \widetilde{H}^t \backslash H(\widetilde{\sigma}_i^{\widehat{h},t})$, $\widetilde{\sigma}_i^{\widehat{h},t} =^{\widetilde{h}} \widetilde{\sigma}_i^{\widetilde{h}}$.

$\overline{h} \notin \widetilde{H}^t$, $\overline{h}^q \succ \overline{h} \in H(\widehat{\sigma}^{\widehat{h},q})$, *so set* $\widehat{\sigma}^{\overline{h},q} := \widehat{\sigma}^{\widehat{h},q}|\overline{h}$. *Then A3, A2 and A1 are satisfied with* $\widetilde{H}^t \cap H^{\overline{h}}$ *in place of* $\Sigma^{\widehat{H}}$, n *in place of* m *and* $\overline{h}$ *in place of* h. *So by the I.H. there exist* $\sigma^{\overline{h}} \in E^{\overline{h}}(\Sigma^{\widetilde{H}^t})$ *and an equilibrium* $\widetilde{\sigma}^{\overline{h}} =^* \sigma^{\overline{h}}$ *such that* $\text{Supp}\widetilde{\sigma}^{\overline{h}} \subseteq S_n^{\overline{h}} = S_n^{\widehat{h}}(\overline{h})|\overline{h}$.

Since $\overline{h} \notin \cup_{\widetilde{h} \in \widetilde{H}^t} H^{\widetilde{h}}$ *and* $H^{\widehat{h}}$ *is finite,* $\widetilde{H}^{t+1} := \{\widetilde{h} \in \widetilde{H}^t : \widetilde{h} \not\succeq \overline{h}\} \cup \{\overline{h}\}$ *is a set of unordered histories that keep shortening with* t, *until a contradiction is obtained. Before, let* $\Sigma^{\widetilde{H}^{t+1}} := \Sigma^{\widetilde{H}^t} \cup \{\sigma^{\overline{h}}\} \backslash (\sigma^{\widetilde{h}})_{\widetilde{h} \in \widetilde{H}^t \backslash \widetilde{H}^{t+1}}$, $\overline{h}^{\tau+1} := \overline{h}$, $\widehat{\sigma}^{\widehat{h},\tau+1} := \widetilde{\sigma}^{\widehat{h},t}$.[72] *Then, increase* t *by* 1 *and run again noting what follows: for every* $\widetilde{h} \in \widetilde{H}^t$ *such that* $\widetilde{h} \succ \overline{h}$, $\sigma^{\overline{h}}|\widetilde{h} = \sigma^{\widetilde{h}}$, *so that* $E^h(\Sigma^{\widetilde{H}^{t+1}}) \subseteq E^h(\Sigma^{\widetilde{H}^t})$.

Since $\widehat{h} \notin \cup_{\overline{h} \in \overline{H}^k} H^{\overline{h}}$ and H^h is finite, $\overline{H}^{k+1} := \{\widetilde{h} \in \overline{H}^k : \widetilde{h} \not\succeq \widehat{h}\} \cup \{\widehat{h}\}$ is a set of unordered histories that keep shortening with k, until a contradiction is obtained. Before, let $\Sigma^{\overline{H}^{k+1}} := \Sigma^{\overline{H}^k} \cup \{\sigma^{\widehat{h}}\} \backslash (\sigma^{\overline{h}})_{\overline{h} \in \overline{H}^k \backslash \overline{H}^{k+1}}$, $h^{w+k+1} := \widehat{h}$, $\widehat{\sigma}^{h,w+k+1} := \widetilde{\sigma}^{h,k}$.[73] Increase k by 1 and run again noting what follows: for every $\overline{h} \in \overline{H}^k$ such that $\overline{h} \succ \widehat{h}$, $\sigma^{\widehat{h}}|\overline{h} = \sigma^{\overline{h}}$, so that $E^h(\Sigma^{\overline{H}^{k+1}}) \subseteq E^h(\Sigma^{\overline{H}^k})$. ■

Proof of theorem 2. Fix $j \in I$, $h \in D_j(S^\infty_{\Delta^e})$ and $(S^h_n)^\infty_{n=0} = (S^n_{\Delta^e}(h)|h)^\infty_{n=0}$. Fix $m \in \mathbb{N}$, $i \neq j$ and μ_i^h t.s.b. $(S^h_{-i,n})^{m-1}_{n=0}$. By self-enforceability ($\zeta(S^\infty_{\Delta^e}) = \{z\}$) there exists μ_i t.s.b. $(S^n_{-i,\Delta^e})^{m-1}_{n=0}$ such that $\mu_i(S_{-i}(h)|p(h)) = 0$ and $\rho(\mu_i)(h) \neq \emptyset$. Hence, by Lemma 1 there exists $\widetilde{\mu}_i =^h \mu_i^h$ t.s.b. $(S^n_{-i,\Delta^e})^{m-1}_{n=0}$ such that $\rho(\widetilde{\mu}_i)(h) \neq \emptyset$. So A3 holds for every $i \neq j$; A1 and A2 hold with $\widehat{H} = \emptyset$. By $\zeta(S^\infty_{\Delta^e}) = \{z\}$ there is $m \in \mathbb{N}$ such that $S^h_m = \emptyset$, so Lemma 10 cannot hold. Thus A3 must be violated for j and some $v \leq m$, SPE σ^h of $\Gamma(h)$ and $\mu_j^h \geq \sigma^h_{-j}$ t.s.b $(S^h_{-j,v})^v_{q=0}$. Fix $s_{-j} \in S^\infty_{-j,\Delta^e} \subseteq S_{-j}(z)$: for every $s^h_{-j} \in \text{Supp}\mu_j^h(\cdot|h)$, by Lemma 1 I can construct $\widetilde{s}_{-j} \in S^v_{-j,\Delta^e}(z)$ such that $\widetilde{s}_{-j} =^h s^h_{-j}$ and $\widetilde{s}_{-j} =^{H\backslash\{h\}} s_{-j}$. Using all such $\widetilde{s}_{-j}$'s I can construct $\mu_j =^h \mu_j^h$ t.s.b $(S^h_{-j,v})^v_{q=0}$ and $S_{-j}(z)$ so that $\rho(\mu_j)(h) = \emptyset$ and $\rho(\mu_j)(z) \neq \emptyset$. Then j prefers z to the distribution over outcomes induced by σ^h. Hence, z is a SPE path. ■

Proof of theorem 6. Lemma 10 can be applied with strong rationalizability as reduction procedure, $h := h^0$, empty $\widehat{H}$, and m after convergence. ■

References

[1] Bassetto, M., "Equilibrium and government commitment", *Journal of Economic Theory*, **124(1)**, 2005, 79-105.

[72] Overwrite the previous, temporary choice.

[73] Note that for each $\overline{h} \in \overline{H}^k \cap H(\widetilde{\sigma}^{h,k})$, $\widetilde{\sigma}^{h,k} =^* \sigma^{h,k} =^* \sigma^{\overline{h}} =^* \widetilde{\sigma}^{\overline{h}}$, so $\widetilde{\sigma}^{h,k}$ satisfies A2 with $\overline{H}^{k+1}$.

[2] Battigalli, P., "Comportamento razionale ed equilibrio nei giochi e nelle situazioni sociali", 1987, undergraduate dissertation, Universita' Bocconi, Milano.

[3] Battigalli, P., "Dynamic Consistency and Imperfect Recall", *Games and Economic Behavior*, **20(1)**, 1997, 31-50.

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

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

[6] Battigalli, P. and M. Dufwenberg, "Dynamic psychological games", *Journal of Economic Theory*, **144(1)**, 2009, 1-35.

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

[8] 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.

[9] Battigalli, P. and M. Siniscalchi, "Interactive beliefs, epistemic independence and strong rationalizability", *Research in economics*, **53,** 1999, 247-273.

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

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

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

[13] Catonini, E., "Selecting strongly rationalizable strategies", working paper, 2015.

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

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

[16] Dufwenberg, M., Servátka, M., Vadovic, R., "ABC on Deals", Working Papers in Economics 12/14, University of Canterbury, Dpt. of Economics and Finance.

[17] Farrell, J. P., and Maskin, E., "Renegotiation in repeated games", *Games and Economics Behavior*, **1(4)**, 1989, 327-360.

[18] Fudenberg, D., and D. Levine, "Self-confirming equilibrium", *Econometrica,* **61**, 1993, 523–546.

[19] Gossner, O., "The robustness of incomplete penal codes in repeated interactions", working paper, 2014.

[20] Govindan, S., Wilson, R., "On forward induction," *Econometrica*, **77**, 2009, 1-28.

[21] Greenberg, J., Gupta, S., Luo, X., "Mutually acceptable courses of action", *Economic Theory*, **40**, 2009, 91-112.

[22] Harrington, J. "A Theory of Collusion with Partial Mutual Understanding", working paper, 2015.

[23] 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.

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

[25] Miller, D., and J. Watson, "A Theory of Disagreement in Repeated Games with Bargaining", *Econometrica*, **81(6)**, 2013, 2303–2350.

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

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

[28] Reny, P., "Backward Induction, Normal Form Perfection and Explicable Equilibria", *Econometrica*, **60(3)**, 1992, 627-49.

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