THESIS DECLARATION

The undersigned

SURNAME	Catonini
NAME	Emiliano
PhD Registration Number	1370014

Thesis title:

Essays on hierarchies of beliefs and non-monotonic strategic reasoning

PhD in	Economics
Cycle	24
Candidate's tutor	Prof. Pierpaolo Battigalli
Year of discussion	2013

DECLARES

Under his responsibility:

1) that, according to the President's decree of 28.12.2000, No. 445, mendacious declarations, falsifying records and the use of false records are punishable under the penal code and special laws, should any of these hypotheses prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;

2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree of 30th April 1999 protocol no. 224/1999, to keep copy of the thesis on deposit at the Biblioteche Nazionali Centrali di Roma e Firenze, where consultation is permitted, unless there is a temporary embargo in order to protect the rights of external bodies and industrial/commercial exploitation of the thesis;

3) that the Servizio Biblioteca Bocconi will file the thesis in its 'Archivio istituzionale ad accesso aperto' and will permit on-line consultation of the complete text (except in cases of a temporary embargo);

4) that in order keep the thesis on file at Biblioteca Bocconi, the University requires

that the thesis be delivered by the candidate to Società NORMADEC (acting on behalf of the University) by online procedure the contents of which must be unalterable and that NORMADEC will indicate in each footnote the following information:

- thesis Essays on hierarchies of beliefs and non-monotonic strategic reasoning;
- by Catonini Emiliano;
- discussed at Università Commerciale Luigi Bocconi – Milano in 2013;
- the thesis is protected by the regulations governing copyright (law of 22 April 1941, no. 633 and successive modifications). The exception is the right of Università Commerciale Luigi Bocconi to reproduce the same for research and teaching purposes, quoting the source;
- **only in cases where another declaration has been undersigned requesting a temporary embargo:** the thesis is subject to a temporary embargo for (indicate duration of the embargo) …… months;

5) that the copy of the thesis deposited with NORMADEC by online procedure is identical to those handed in/sent to the Examiners and to any other copy deposited in the University offices on paper or electronic copy and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;

6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (law of 22 April 1941, no. 633 and successive integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal and shall be exempt from any requests or claims from third parties;

7) that the PhD thesis is not the result of work included in the regulations governing industrial property, it was not produced as part of projects financed by public or private bodies with restrictions on the diffusion of the results; it is not subject to patent or protection registrations, and therefore not subject to an embargo;

Date 10/31/2012

SURNAME	Catonini	NAME	Emiliano

Essays on hierarchies of beliefs and non-monotonic strategic reasoning

Emiliano Catonini

Contents

Acknowledgements

I want to thank heartily Pierpaolo Battigalli for introducing me to the magic and the mysteries of epistemic game theory and guiding me until here. Thank you also to all the PhD people who made this journey so satisfactory: classmates, professors and staff. Finally, a big big thank you to those who made my days out there, Elisabetta, Pantaleone, Federica, Franco, Alessandro, Nicola, Mattia, Veronica and Eleonora.

Introduction

Epistemically founded models of strategic reasoning derive predictions about each player's moves from the hierarchies of beliefs about opponents' moves that they may hold. Both in static and dynamic games, restrictions of the set of conceivable hierarchies can be used to refine the predictions of the models. Especially in dynamic games, the effects of such restrictions are highly non-trivial: the predictions may change in a non-monotonic way, i.e. they need not be a subset of the predictions in absence of restrictions. In epistemic game theory, this is a well known phenomenon. In a state space endowed with a type space for conditional probability systems [39], the strong belief operator [11] takes an event (a subset of the state space) and delivers another event whose epistemic types believe in the first event at all the information sets that are compatible with its realization. The strong-belief operator is non-monotonic: if the first event is restricted, the delivered event need not shrink, because the restricted event may be incompatible with more information sets[1] and there the player is now allowed to believe in any opponent type (thus to hold any hierarchy of beliefs whose first-order level is compatible with the information set).

Restrictions to hierarchies of beliefs can be used to model fundamental economic situations. Pre-play non-binding agreements among players are one of the most interesting cases. 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. Part I analyzes the effects of restrictions that may be induced by possibly incomplete non-binding agreements. In dynamic games, players may observe a deviation from the agreement before the game is over. The attempt to rationalize the deviation may lead players to revise their beliefs about opponents' behavior in the continuation of the game. This instance of forward induction reasoning is based not just on beliefs about rationality, but also on interactive beliefs about the compliance with the agreement itself, modeled precisely as restricted sets of hierachies of beliefs. Here I study the effects of such rationalization on the self-enforceability of an agreement, that is on the possibility that it is commonly believed and, once believed, that players comply with it. Accordingly, outcomes of the game are deemed to be enforceable by some

[1] This requires that the projection of the event on the strategy space is not full, so restrictions to hierarchies of beliefs alone are not sufficient. Rationality is the link between restrictions to hierarchies and projections that do not coincide with the whole strategy space.

kind of agreement or not. Conclusions can be very different from what the equilibrium refinement tradition suggests. For instance, a subgame imperfect Nash equilibrium may represent a self-enforcing agreement, while a subgame perfect equilibrium need not be self-enforcing. Incomplete agreements that do not attempt to restrict behaviour after a violation correspond to just agree on an outcome to achieve. Their self-enforceability is investigated in detail and conclusions are robust to an important "epistemic priority" issue: do players retain the beliefs about the agreement or the beliefs about rationality when, given the observed behaviour, the two are at odds? In the first case, strong-delta-rationalizability [7] is the appropriate tool to tackle the problem; for the second case, selective rationalizability is the right one. Both procedures can be epistemically chacterized through the strong belief operator: indeed, they both model a non-monotonic strategic reasoning process.

While strong-delta-rationalizability has already been epistemically characterized in [10], selective rationalizability is a novel solution concept and it is characterized in Part II. In a dynamic game, consider rational players holding common strong belief in rationality. That is, they are rational, they believe opponents are rational as long as not contradicted by observation, and so on. Very often, many different conjectures about opponents behavior are compatible with common strong belief in rationality. As a result, strong rationalizability [11] delivers more than one plan of actions to a player: which one will be implemented? The player may have exogenous arguments to form some particular conjecture; opponents may be aware that the player holds such arguments and may take it into account as long as her moves do not contradict this; and so on. For instance, arguments of this kind can arise precisely from non-binding agreements discussed above: then, both "lone" strategic reasoning and coordination with the opponents would be captured by this process. While usual rationalizability tools and equilibrium concepts are only able to capture one of the two, strong-delta-rationalizability already incorporates first-order-belief restrictions in a strategic reasoning process. But there, any-order belief in the restrictions is so strong to prevail over the same-order belief in rationality when (jointly) they are inconsistent with the observed moves. Here, instead, epistemic priority is inverted: players retain all the orders of belief in rationality that are per se consistent with the information set and drop the orders of belief in the restrictions that are at odds with them. The corresponding rationalizability tool is called selective rationalizability because it delivers a selection of strongly rationalizable strategies. The formal epistemic characterization clarifies all these issues.

Also in static games, restrictions to hierarchies of beliefs can explain particular strategic reasoning processes. Iterated admissibility is one of the most appealing solution concepts for complete-information strategic-form games. Yet, its epistemic analysis has turned out to be elusive. To understand when it is the appropriate solution concept, conditions under which players want to avoid strategies that are weakly dominated in some reduced game along the procedure (although possibly not in the final set!) must be provided. It is intuitive that these conditions have to incorporate some cautious attitude of the players. Yet, to what extent players are cautious and assume that opponents are must be carefully defined in order to provide a correct motivation for iterated admissibility. Brandenburger, Friedenberg and Keisler [18] define a notion of rationality, including an "open-mindedness" requirement for lexicographic beliefs, which delivers iterated admissibility when players adopt it, assume (to a defined extent) that opponents adopt it, and so on, up to some finite level. This notion of rationality cannot be commonly assumed by players unless heavy exogenous restrictions to beliefs apply. Here, in Part III, I provide a weaker notion of cautiousness that can be commonly assumed by players and still captures iterated admissibility, and which has a very clear and realistic interpretation. Players only form hierarchies of beliefs where, at each order, everyone gives positive probability to every opponents' strategy subprofile, at some level of the lexicographic conjecture. In order to avoid arbitrary restrictions and identify clearly the ones of interest, I carry on the analysis in a type space that encompasses all meaningful lexicographic hierarchies of beliefs, the canonical one, of which I show constructively the existence.

Part I
Non-binding agreements and forward induction reasoning

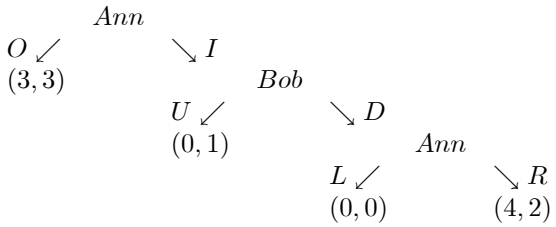
1 Introduction

When the players of a dynamic game are given the opportunity to communicate among themselves before the game starts, they are likely to exploit it to reach a possibly incomplete agreement about how to play. In most cases, the context allows them to reach only a non-binding agreement, which cannot be enforced by an external 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. The paper sheds light about what the viable alternatives are, that is, which agreements are able to induce the common belief that everyone will comply with them and, among them, which ones players will actually comply with. Moreover, in an implementation perspective, the paper aims to understand which outcomes of the game can be ensured by *some* agreement. The paper will not deal with the pre-play bargaining phase. But the evaluation of appealing agreements has a clear and strong feedback on which ones are likely to be selected.

Here I take the view that players will believe in the agreement only if this is compatible with reasonable assumptions about rationality, beliefs in rationality and their interaction with the beliefs in the agreement (of all orders). If compatibility holds, one must still check whether all possible behavioral implications are consistent with the agreement itself. Both issues can be elucidated with appropriate rationalizability concepts. Strong-delta-rationalizability [7] captures the hypothesis that beliefs about the agreement are given higher (or the same[2]) epistemic priority than beliefs about rationality; that is, in contingencies that some player would not have reached under the beliefs in the agreement and rationality up to some order, opponents abandon the belief in rationality of that order. When instead opponents want to retain all orders of belief in rationality that are per se compatible with the contingency and rather drop the beliefs in the

[2] The behavioural consequences are the same because the belief in the agreement of order n has no bite without the belief in rationality of order $n-1$. See [10] for details.

agreement that are at odds with them, selective rationalizability [20] is the right tool. Both strong-delta-rationalizability and selective rationalizability deliver either an empty set (if the first-order belief restrictions are not compatible with the strategic reasoning hypotheses) or the expected implications of the agreement (otherwise). The important difference between the two epistemic priority assumptions can be understood from the following perfect information game, taken from [20].



Bob states that he would play U if he is called to act. If Ann is rational[3] and believes that Bob would play U, she will play O. Then, if Bob observes Ann playing I, he cannot believe at the same time that Ann is rational and believes he will play U. If Bob puts priority on the belief of Ann that he will play U, he must drop the belief that Ann is rational (and then he could also expect Ann to play L after D, hence he could truly play U). If Bob puts priority on Ann being rational, he must drop the belief that Ann believes that he will play U (and expecting Ann to play R after D, he will play D, violating his statement).

Strong-delta-rationalizability and selective rationalizability allow to tackle the main issues for all dynamic games with complete information[4] and perfect recall. Yet, for notational simplicity, the focus is restricted to the wide class of finite[5] games with observable actions.[6] First: which agreements are credible and will be complied with? Which outcomes of the game can be achieved through some agreement? To answers these questions the concepts of, respectively, self-enforceability (of agreements)

[3] i.e. expected utility maximizer given the continuation conjecture at every information set.

[4] The tools employed in the paper apply also to dynamic games with incomplete information. The possibility to extend the analysis to such games is discussed in the conclusions.

[5] Finite sets of actions and finite horizon.

[6] Games where every player who is called to act knows exactly the current history of the game, i.e. information sets are singletons. All repeated games with perfect monitoring, for instance, are games with observable actions.

and enforceability (of outcomes) are defined. In particular, two relevant classes of agreements are deeply investigated: complete agreements and path agreements[7]. While a complete agreement specifies a single action for every player at every contingency, a path agreement does not attempt to restrict behavior at contingencies that follow a violation of the agreement itself. Thus, it corresponds to just agreeing on an outcome to achieve. For path agreements, self-enforceability under priority to the agreement and under priority to rationality are equivalent, hence the analysis is robust to the epistemic priority assumption. Only for a strongly rationalizable[8] [11] subgame perfect equilibrium (henceforth, SPE) outcome, the corresponding path agreement can be self-enforcing, but not even its credibility is guaranteed: the class of *equilibrium paths that can be upset by a convincing deviation* introduced by Osborne [36] is an example. When the path agreement is not self-enforcing, off-the-path restrictions are needed to enforce the corresponding outcome: leaving some mystery about on-the-path moves is of no help. Introducing off-the-path restrictions, also the outcome of a Nash, subgame imperfect equilibrium may be enforceable, even when epistemic priority falls on rationality. On the other hand, while under priority to rationality the self-enforceability of pure SPE is guaranteed, under priority to rationality there could not exist any agreement that enforces a given pure SPE outcome. At this point, one may wonder whether there always exist SPE outcomes that can be enforced by some agreement. This would be true only allowing players to agree on mixed strategies, a possibility that is not considered of particular interest here. However, its investigation brings to a broader result, which is interesting per se: in every game with observable actions, there is always the support of a SPE outcome distribution that is induced also by strongly rationalizable strategies. That is, backward induction and forward induction (as captured by strong rationalizability) never give disjoint predictions.

To introduce the main theoretical issues and their relevance to economics, Section 2 presents two applied examples. In the first one, players can profitably agree on a Nash, non subgame perfect equilibrium of the game, even when they commonly believe in rationality as long as it is not contradicted by observation. In the second one, players would like to achieve a desirable subgame perfect equilibrium outcome, but forward

[7] Path agreements have already been defined and analyzed by [27] through a solution concept that does not capture forward induction reasoning.

[8] Strong rationalizability is often referred to as extensive form rationalizability [37], meaning its correlated version.

induction reasoning makes the corresponding agreement not credible. Section 3 introduces the formal framework and the analytic tools. Section 4 presents the main results. Section 5 concludes. In the Appendix I apply the proposed methodology to a case in the literature and I present the formal solutions of the examples of Section 2 and the proofs of the theorems.

2 Motivating examples

Example 1 In a city, two parties can form a credible 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 8. If they both choose a *M*oderate positioning, their coalition will win and the surplus to split grows to 10. The problem is that party 2 may be tempted to take a radical positioning even if party 1 chooses a moderate one. (And then P_1 wouldn't be able to credibly switch to the radical positioning too.) Indeed, in this case, P_2's candidate would not win at the first round but would gain the ballot for sure and then possibly take all the surplus. At the last debate before the ballot, P_1 can declare whether *S*upporting P_2's candidate or the *A*lternative one and at the same time P_2 can make a political *O*ffer to P_1's voters or *N*ot. P_1 gets a payoff of 2 if supporting the winning candidate, unless P_2's candidate wins and does not make any offer. P_2's candidate wins for sure if P_1 supports, with probability 1/2 if not but P_2 makes the offer to P_1's voters.

$P_1 \backslash P_2$	M	R
M	$(5,5)$	$\cdot-$
R	$(0,0)$	$(4,4)$

$- - -$ $- \rightarrow$

$P_1 \backslash P_2$	N	O
A	$(2,0)$	$(1,4)$
S	$(0,8)$	$(2,6)$

The game has only one SPE, where P_1 plays R, A with probability 1/3 and S with probability 2/3 and P_2 plays R, N with probability 1/3 and O with probability 2/3. But the game features also a Nash equilibrium that Pareto-dominates the SPE, namely where P_1 plays M and A and P_2 plays M and N (or O).

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 still credible that P_1 will reply with A? After observing the deviation, P_1 must either believe that P_2 did not believe in the agreement or that it is irrational. If P_1 would rather drop the belief that P_2 is rational, 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 and so the agreement is believed and players comply with it. If P_1 is not willing to drop the belief that P_2 is rational, it can expect any rational move. But both $(R.N)$ and $(R.O)$ can be rationalized as best responses to some conjecture and A is a best reply to N in the subgame.[9] 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 and so the agreement is credible and players comply with it.

Notice that agreeing on the whole SPE requires instead to agree on mixed strategies, but this is unrealistic. Our formal definition of agreement will rule out this possibility. Agreeing on the whole support amounts instead to agreeing just on playing R in the first stage. This is enough here for the credibility and the compliance with the agreement: P_1 has no incentive to switch to M whatever it expects in the second stage. This type of agreement, which is silent about off-the-path behavior, has an intuitive appeal, as it will be pointed out. Yet, forward induction reasoning based on the belief in the path may rule out the off-the-path beliefs that are necessary to prevent a deviation.[10] This is the case in the next example.

Example 2 The duopolists of the Cola market, A and B, have to decide their marketing strategy before two big sport events, which will gather the population in front of the television. There are 10 million buyers: 2 of the 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 any,

[9] And O is the best reply to A, S is the best reply to O and N is a best reply to S, so the incentive to play A is compatible with the beliefs in rationality of all orders.

[10] 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 opponent 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. The risk of deviation is ruled out via forward induction.

otherwise 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. Also the loyal buyers switch to the brand making discounts, but the profit drops to 0.2 per million customers.

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 a best reply only to the other brand doing the same, and the profits of both brands 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 exists a SPE where the two firms collude in the first stage. Namely, (s_A^*, s_B^*) where:

$$s_i^*(h) = \begin{cases} C \; if \; h = h^0 \text{ (i.e. at the start of the game)} \\ D \; if \; h = (CC) \\ P \; else \end{cases} \qquad i = A, B.$$

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. There is common knowledge of the discount option, but it is not mentioned in the conversation. All this sums up to agree on the equilibrium path.

The agreement does not rule out punishment P after a deviation; therefore, it seems to be possible (although not guaranteed) that players fear it and comply with the path. Instead, once accounting for forward induction reasoning,[11] punishment is actually ruled out and thus the path agreement is not credible. B, if he is rational and believes that A will comply with the agreement, will defect in the first stage only if he expects no punishment in the second. A, if she believes this and observes D, will then (by

[11] This instance of forward induction reasoning differs from the most frequent ones in the literature, because it does not rely only on beliefs in rationality but also on beliefs about the respect of the agreement and their interaction.

forward induction) expect B to play D again. B, if he believes that A will interpret the deviation in this way, will actually play D in the first stage. A, if she believes that B expects such interpretation, will anticipate the deviation and will not believe in the agreement from the start.

A possible objection is that players, anticipating that the path agreement would not work, would explicitely threaten the punishment if collusion does not succeed in the first stage. First, this is not what happens in many real-life situations. Discussing what to do in case the partner defects is an inconvenient way to start a relationship. Second, agreeing on a sequence of moves in repeated games, implicitely assuming that the agreement is valid until not violated, or agreeing on an outcome to reach in an extensive-form game, is a simple, natural, and hence appealing agreement. Third, the belief in the remainder of an agreement that has already been violated is very likely to fall; in this case, it is as if players had just agreed on the path as above.[12]

In a nutshell, non subgame perfect equilibria can be credible agreements, even when players put epistemic priority on rationality, while it is very easy to construct examples where some SPE fails to do so. On the other hand, putting epistemic priority on the agreement, while any entire SPE is a credible agreement, this could not be the case when off-the-path restrictions fail to hold. For these reasons, the credibility of agreements can depart substantially from what the equilibrium refinement tradition would suggest. The formal tools to carry out a deep and well-founded analysis of the problem are presented in the next section. These tools are used in the Appendix to analyze formally the two examples.

3 Agreements and beliefs in dynamic games with complete information

3.1 Agreements, complete agreements and path agreements.

For any of the following player-specific sets X_i, $X := \prod_{i \in I} X_i$, $X_{-i} := \prod_{j \neq i} X_i$.

Define the following primitive sets:

[12] Players may instead give a lower priority to off-the-path beliefs but still keep them if they are compatible with the beliefs in the path (which is not the case in this example anyway). This situation is not formally modeled in the paper.

- I is the set of players;
- $(\overline{A}_i)_{i\in I}$ are the sets of actions potentially available to each player;
- $\overline{H} \subset \overline{A}^{<\mathbb{N}}$ is the set of histories of the game with the following properties:
  - $h^0 := \emptyset \in \overline{H}$;
  - for every $h = (\widetilde{a}^1, ..., \widetilde{a}^t) \in H$ and $l < t$, $(\widetilde{a}^1, ..., \widetilde{a}^l) \in H$;
  - for every $h \in \overline{H}$, if there exists $a \in \overline{A}$ such that $(h, a) \in \overline{H}$, then there exist non-empty $(\widetilde{A}_i \subseteq \overline{A}_i)_{i\in I}$ such that $\widetilde{a} \in \widetilde{A}$ if and only if $(h, \widetilde{a}) \in \overline{H}$.

For any two $h, h' \in \overline{H}$, define a precedence relation by letting $h \prec h'$ if and only if $h' = (h, x)$ for some non-empty $x \in \overline{A}^{<\mathbb{N}}$. I denote by $p(h)$ the predecessor of h, i.e. $\widetilde{h} \in \overline{H}$ such that $\widetilde{h} \prec h$ and for every $\widehat{h} \not\preceq \widetilde{h}$, $\widehat{h} \not\prec h$. The set of histories $\overline{H}$ endowed with the precedence relation $\prec$ is a tree with root h^0.

Denoting by $Z := \left\{z \in \overline{H} : \nexists h \in \overline{H}, z \prec h\right\}$, for every $i \in I$ let $u_i : Z \to \mathbb{R}$ be the payoff function. I denote an extensive form game with complete information and observable actions with $\Gamma = \langle I, X, (u_i)_{i\in I}\rangle$.

Now I can derive the following objects. Let $H := \overline{H}\backslash Z$ be the set of non-terminal histories and, for every $i \in I$ and $h \in H$, let $A_i(h) := proj_{A_i}\left\{a \in \overline{A} : (h, a) \in \overline{H}\right\}$ be the set of feasible actions of player i at history h. 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 set of histories $P \subset \overline{H}$ is defined as:

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

For any subset of strategies $\widehat{S}_i \subset S_i$, $\widehat{S}_i(h) := S_i(h) \bigcap \widehat{S}_i$.

On the other hand, the set of histories that are compatible with a set of strategy (sub-)profiles $\widehat{S} \subseteq \prod_{j\in J\subseteq I} S_j$ is defined as:

$$H(\widehat{S}) := \left\{(\widetilde{a}^1, ..., \widetilde{a}^t) \in H : \exists \widehat{s} \in \widehat{S}, \forall j \in J, \forall l < t, \right. \\ \left. \widehat{s}_j((\widetilde{a}^1, ..., \widetilde{a}^l)) = \widetilde{a}_j^{l+1}, \widehat{s}_j(h^0) = \widetilde{a}_j^1\right\}.$$

Analogously, the set of outcomes that are compatible with a set of strategy (sub-) profiles $\widehat{S} \subseteq \prod_{i \in J \subseteq I} S_j$ is denoted by

$$\zeta(\widehat{S}) := H(\widehat{S}) \bigcap Z.$$

In games with observable actions, a player who is called to act knows exactly the current history of the game. That is, information sets are singletons (histories). Notice moreover that to simplify notation every player is expected to play an action at every history: when a player is not truly active at a history, the set of feasible actions consists of just one "dummy" action. In this way, for every player the set of information sets coincides with the set of histories. In the following, I will refer to information sets rather than histories when it is conceptually appropriate and information sets are the objects to use in the analysis of a game without observable actions.

I will use a notion of SPE in mixed strategies instead of behavioral strategies. A profile of mixed strategies $\sigma \in \prod_{i \in I} \Delta(S_i)$ is a SPE in mixed strategies if there exists a SPE in behavioral strategies

$$((\beta_i^h)_{h \in H})_{i \in I} \in \prod_{i \in I} (\Delta(\overline{A}_i))^{|H|}$$

such that for every $i \in I$ and $s_i \in supp\sigma_i$, $\sigma_i(s_i) = \prod_{h \in H} \beta_i^h(s_i(h))$.

In this context, I consider players coming up with an *agreement* of the following form.

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

That is, an agreement specifies at every information set the pure actions among which players are expected to choose. If the agreement assigns to a player the whole set of available actions at an information set $h \in H$, $A_i(h)$, the agreement is silent about how the player should play at that information set. Notice that an agreement can restrict the behavior of players also at information sets that are precluded by the agreement itself.

When this is not the case and the agreement allows to reach only one outcome of the game, I will call it a *path agreement*. Formally:

Definition 2 (Path Agreement) *A path agreement is an agreement* $e = (e_i)_{i \in I}$ *such that there exists* $(\widetilde{a}^1, ..., \widetilde{a}^t) \in Z$ *such that for every* $i \in I$ *and* $l < t$, $e_i(h^0) = \widetilde{a}_i^1$, $e_i((\widetilde{a}^1, ..., \widetilde{a}^l)) = \widetilde{a}_i^{l+1}$ *and* $e_i(h) = A_i(h)$ *otherwise.*

Path agreements are particularly interesting for applied reasons. First, they correspond to a very natural kind of agreement: choosing an outcome as a goal. Second, they are likely to be taken when players either do not feel like discussing the possibility that someone could deviate from the agreement, or they just anticipate that they would not trust the agreement anymore once it has already been violated (see the second example). Moreover, as it will be proved later, path agreements have a desirable property: players do not have to bother about the epistemic priority issue to evaluate their credibility and self-enforceability. Consider furthermore the case in which off-the-path subgames do not feature pure strategies equilibria: players could find it difficult to reach an agreement for those contigencies (see the first example). This is one of the motivations for the study of incomplete codes in Gossner [26]. Gossner finds that incomplete codes (such as path agreements) can be more robust than complete ones to payoff perturbations. Interestingly, Gossner defines a class of "credible" codes. Once translated in the present framework, it is possible to apply the analytical tools developed here to prove that such class is actually credible, in the epistemically founded notion of this paper: see the Appendix for details. Path agreements are analyzed also in [27] through a solution concept that does not capture forward induction reasoning.

When instead players want to pinpoint how they should behave in every possible contingency of the game, they will come up with a *complete agreement* of the following form.

Definition 3 (Complete Agreement) *A complete agreement is an agreement* $e = (e_i)_{i \in I}$ *such that for every* $i \in I$ *and* $h \in H$, $|e_i(h)| = 1$.

Complete agreements are particularly interesting too because they can correspond to an entire equilibrium of the game.

However, before focusing on these two kinds of agreements, the analytical tools for the evaluation of a generic agreement will be provided.

3.2 Beliefs induced by an agreement

The elementary belief that an agreement may be able to induce is that opponents will comply with it. Such belief can be represented as a re-

stricted set of conjectures about what strategies opponents are going to play. In this extensive-form framework, conjectures are modeled as conditional probability systems [39] (henceforth, CPS). Here I define the concept of CPS directly for the problem at hand.

Definition 4 *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:*

1. *for every* $C \in (S_{-i}(h))_{h\in H}$, $\mu(C|C) = 1$;
2. *for every* $C \in (S_{-i}(h))_{h\in H}$, $\mu(\cdot|C)$ *is a probability measure on* S_{-i};
3. *for every* $E \in S_{-i}$ *and* $C, D \in (S_{-i}(h))_{h\in H}$, *if* $E \subseteq D \subseteq C$, *then* $\mu(E|D)\mu(D|C) = \mu(E|C)$.

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

The third axiom means that players update their conjectures via Bayes rule whenever possible as the game unfolds. For brevity, the conditioning events will be indicated with just the information set, which represents all the information acquired by players from observation.

Definition 5 *Consider an agreement* $e = (e_i)_{i\in I}$. *For every* $i \in I$, *and* $h \in H$ *define:*

$$S^e_{-i}(h) := \left\{ \widetilde{s}_{-i} \in S_{-i}(h) : \forall j \neq i,\ \forall \widetilde{h} \gtrsim h,\ \widetilde{s}_j(\widetilde{h}) \in e_j(\widetilde{h}) \right\}.$$

The set of first-order belief restrictions corresponding to the agreement is $\Delta^e_i \subseteq \Delta^H(S_{-i})$ *where for every* $\mu_i = (\mu_i(\cdot|h))_{h\in H} \in \Delta^e_i$ *and* $h \in H$, $supp\mu_i(\cdot|h) \subseteq S^e_{-i}(h)$.

The agreement is believed at every information set, for what concerns the continuation of the game.[13] Notice that any two players have the same restrictions about any third player.

[13] This simply amounts to assuming that players do not change their beliefs about opponents' behavior in the continuation of the game while the game unfolds. This rules out the attitude of believing in an off-the-path portion of the agreement until the path is followed and forming doubts about it once the path has actually been violated.

I will 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 opponents and themselves will play in the continuation of the game. This is equivalent [5] to choosing a *sequential best reply* to the CPS.

Definition 6 *A strategy $s_i \in S_i$ is a sequential best reply to a CPS $\mu_i \in \Delta^H(S_{-i})$ if for every $h \in H(s_i)$ and every $\widetilde{s}_i \in S_i(h)$,*

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

The set of sequential best replies to a CPS μ_i is denoted by $\rho_i(\mu_i)$.

Here I take the view that other than best replying to their conjectures, players try to refine them through strategic reasoning. As long as not contradicted by observation, players believe that opponents are rational and believe in the agreement, that opponents believe that everyone else is rational and believes in the agreement, and so on. Instead, at information sets where all these beliefs cannot hold together (in the sense that if they were all correct, the information set would not have been reached), players must make a choice. They can either retain all orders of belief in the agreement and drop the orders of belief in rationality that are at odds with them; or, they can retain all orders of beliefs 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. This is the epistemic priority issue: in the first case, I say that players give epistemic priority to the agreement; in the second case, that they give epistemic priority to rationality.

In case players give epistemic priority to the agreement, the appropriate solution concept is *strong-delta-rationalizability* [7], the augmentation with first-order beliefs restrictions of *strong rationalizability*, introduced by Battigalli and Siniscalchi [11]. Its ultimate definition [10] is translated here in the complete information framework of this paper.

Definition 7 (strong-delta-rationalizability) *Consider the following procedure.*

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

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

1. $s_i \in \rho_i(\mu_i)$
2. $\forall p = 0, ..., n-1$, $\forall h \in H$, $S^p_{-i,\Delta^e} \bigcap S_{-i}(h) \neq \emptyset \Rightarrow \mu_i(S^p_{-i,\Delta^e}|h) = 1$ *(i.e. μ_i strongly believes S^p_{-i,Δ^e});*

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

Strong rationalizability can be seen as a special case of strong-delta-rationalizability where no restriction applies and it will be denoted by dropping the subscript Δ^e. I will call *best rationalizable* strategies (see also [4]) of a player at an information set the ones that are compatible with the information set and survive more steps of strong rationalizability. Best rationalizable actions at an information set will be the ones prescribed by best rationalizable strategies.

Step by step, strong-delta-rationalizability captures the following assumptions:

1. players are rational and believe that opponents will comply with the agreement (and that opponents believe that everyone else will comply with the agreement, and so on);
2. 1 holds and players believe that 1 holds as long as not contradicted by observation;
3. 1 and 2 hold and players believe that 1 and 2 hold as long as not contradicted by observation;
4. ...

The sentence in brackets 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-delta-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 formal epistemic characterizations can be found in [10].

Strong-delta-rationalizability does two things. First, it constitutes a compatibility test for the belief in the agreement with the strategic reasoning assumptions. If the agreement does not pass the test, strong-delta-rationalizability delivers an empty set. This happens at a step where some information set is still compatible with the new assumptions, but the behavioral consequences contradict the first-order-belief restrictions at that information set (i.e., the remainder of the agreement). Second, if the agreement passes the test, the strong-delta-rationalizable strategy profiles coincide with the behavioral implications of the common belief in the agreement together with rationality and the beliefs in rationality of all orders, each of them holding as long as its conjunction with the opponents' belief in the agreement of the same order is not contradicted by observation.

In case players give epistemic priority to beliefs about rationality, the appropriate solution concept is *selective rationalizability* [20]:

Definition 8 (selective rationalizability) *Denote by* $(S^m)_{m\geq 0}$ *the strong rationalizability procedure. Consider the following procedure.*

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

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

1. $s_i \in \rho_i(\mu_i)$;
2. $\forall p = 0, ..., n-1, \forall h \in H, S^p_{-i,R\Delta^e} \bigcap S_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(S^p_{-i,R\Delta^e}|h) = 1$;
3. $\forall q = 0, ..., \forall h \in H, S^q_{-i} \bigcap S_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(S^q_{-i}|h) = 1$;

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

Selective rationalizability refines strong rationalizability by selecting those strategies that are compatible with the beliefs in the agreement. More precisely, step by step selective rationalizability captures the following assumptions:

1. players believe in the agreement and are rational, believe that opponents are rational (as long as not contradicted by observation), and so on;

2. 1 holds and players believe that 1 holds as long as not contradicted by observation;

3. 1 and 2 hold and players believe that 1 and 2 hold as long as not contradicted by observation;

4. ...

Selective rationalizability does the same two things of strong-delta-rationalizability. A non-empty set of selectively-rationalizable strategy profiles coincides with the behavioral implications of rationality, beliefs in rationality of all orders (each of them holding as long as not contradicted by observation) and beliefs in the agreement of all orders, each of them holding as long as its conjunction with the beliefs in rationality is not contradicted by observation.

Although similar in their structure, the two procedures can deliver sharply different results for the same restrictions. The perfect information game of the introduction is an example: strong-delta-rationalizability delivers a non-empty set of strategy profiles where Ann chooses O; selective rationalizability delivers the empty set.[14] In other cases, the two procedures can deliver two different non-empty sets of strategy profiles, inducing different sets of outcomes (see [20]).

To determine which one is the most appropriate one in applications is a subtle issue. Therefore, it will be useful to evaluate agreements with both tools.

[14] Bob claims that he would play U. If Ann's first order beliefs are actually restricted in this way, strong-delta-rationalizability delivers at the first step strategy O for Ann, while for Bob U and D are both rational. Hence, at the second step, Ann can still believe that Bob would play U and Bob can still play both U and D, being allowed not to believe that Ann is rational after observing I. Selective rationalizability, instead, requires Bob to believe that Ann is rational after observing I if there is a rational strategy of Ann that prescribes I. IR is the one, so Bob must choose D. But then, Ann cannot formulate a CPS that respects the restrictions and believes in Bob's rationality. Hence, selective rationalizability delivers an empty set.

4 Self-enforceability of agreements and enforceability of outcomes

4.1 Self-enforceability

Taking the perspective of evaluating 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 surely comply with the agreement not.

An agreement is credible when it passes the appropriate rationalizability test. For synthesis, 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 definitions under the two different epistemic priority assumptions.

Definition 9 (Credibility) *An agreement $e = (e_i)_{i \in I}$ is credible if $S_e^\infty \neq \emptyset$.*

Credibility does not necessarily imply that players will sure 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 10 (Self-enforceability) *An agreement $e = (e_i)_{i \in I}$ is self-enforcing if it is credible and for every $i \in I$, $h \in H(S_e^\infty)$ and $s_i \in S_{i,e}^\infty(h)$, $s_i(h) \in e_i(h)$.*

The definition requires explicitely that *every* rationalizable strategy complies with the agreement at the information sets that are compatible with itself and with the rationalizable strategy profiles. At an information set that is not compatible with the rationalizable profiles, instead, credibility only implies that among the best rationalizable strategies of each player, there is at least one that complies with the agreement. But this is enough to guarantee that the off-the-path beliefs of the players are in line with the agreement, so that the desired behavioral consequences apply.

One might imagine that a self-enforcing agreement under priority on rationality will be, a fortiori, self-enforcing under priority on the agreement. An example in the appendix of [20] shows that this is not the case.

Instead, there are interesting cases where the two things are equivalent,[15] so that the evaluation of the agreement is robust to the epistemic priority assumption, which may be difficult to identify in applications. The most interesting case is the path agreements one.

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-delta-rationalizability.

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

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

Proof. If $z \notin \zeta(S^\infty)$, $S^1_{R\Delta^e} = \emptyset$, because for some $i \in I$ there is no $\mu_i \in \Delta_i^e$ that strongly believes S^∞_{-i}. ■

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

Proposition 2 *Suppose that $\zeta(S^\infty)$ is a singleton. Then the corresponding path agreement is self-enforceable.*

Proof. For every $i \in I$, Δ_i^e is less restrictive than the third requirement of the selective rationalizability procedure. Hence selective rationalizability coincides with the steps of strong rationalizability after convergence and S^∞ is delivered.[16] ■

What about strongly rationalizable paths when there is more than one? Here there is no sharp answer. Looking back at the second example

[15] The simplest case is when the agreement actually selects among best-rationalizable actions. The third requirement for CPS in selective rationalizability becomes immaterial, because Δ_i^e is already not less restrictive, hence the procedure corresponds to strong-delta-rationalizability.

[16] Remark: the same proof shows that when the agreement allows at every information set all the best-rationalizable actions, it is self-enforcing under priority to rationality.

in Section 2 one can see that there can be both self-enforcing and non self-enforcing strongly rationalizable paths. As a self-enforcing path, one can consider $((D,D),(D,D))$: a deviation could be profitable only if it triggers cooperation in the second stage, but cooperating in the second stage is not rational. Instead, the path analyzed in Section 2 is an instance, in the same game, of a strongly rationalizable path that is not self-enforcing.

Similar paths have been already classified by Osborne [36] for 2-players, finitely-repeated, coordination games as *paths that can be upset by a convincing deviation.* Osborne formalizes the condition under which a path can be upset by a convincing deviation and asserts that such outcome path is not stable, in the sense of Kohlberg and Mertens [32]. Take the T-fold repetition G^T of an arbitrary two-players (i and j) strategic form game G. Let b^k and c^k be the first- and second-ranked stage-outcomes of G for player $k=i,j$ and $v_k : A_i \times A_j \to \mathbb{R}$, $k=i,j$ denote the stage payoffs.

Proposition 3 *[Osborne, [36]] Let $P=(\overline{a}^1,..,\overline{a}^T)$ be a pure Nash equilibrium outcome path of G^T. Suppose that there exist $\tau \in \{1,...,T-1\}$ and $\widetilde{a}_i \in A_i$ such that*

$$v_i((\widetilde{a}_i,\overline{a}_j^{\tau}))+v_i(c^i)+(T-\tau-1)v_i(b^i) < \sum_{t=\tau}^{T} v_i(\overline{a}^t) < v_i((\widetilde{a}_i,\overline{a}_j^{\tau}))+(T-\tau)v_i(b^i) \tag{1}$$

and

$$(T-\tau)v_j(b^i) > \max_{a_j} \left\{ v_j(b_i^i,a_j) : a_j \in A_j \text{ and } a_j \neq b_j^i \right\} + (T-\tau-1)v_j(b^j), \tag{2}$$

where $j \neq i$. Then the outcome path P is not stable.

In this framework, such paths can be characterized as non credible agreements.

Proposition 4 *Let $z=(\overline{a}^1,...,\overline{a}^T)$ be a path that can be upset by a convincing deviation. The corresponding path agreement is not credible.*

Proof: For $k=i,j$ and every $(a^1,...,a^T) \in Z$, define $u_k : Z \to \mathbb{R}$ as $u_k((a^1,...,a^T)) := \sum_{t=1}^{T} v_k(a^t)$. Set $h_0 := (\overline{a}^1,...,\overline{a}^{\tau-1})$, $h_1 := (\overline{a}^1,..,(\widetilde{a}_1,\overline{a}_j^{\tau}))$, $h_2 := (\overline{a}^1,..,(\widetilde{a}_1,\overline{a}_j^{\tau}),b^i,...,b^i)$.

For every $s_i \in S_i(h_1)$ such that $s_i \notin S_i(h_2)$ (i.e. all strategies of player i deviating in τ but not aiming to her best continuation afterwards),

$s_i \notin S^1_{i,\Delta^e}$, because for every $\mu_i \in \Delta^e_i$ and $s_j \in supp\mu_i(\cdot|h_0)$, there exists $s'_i \in S_i$ such that $\zeta(s'_i, s_j) = z$ and, by (1), $u_i(z) > u_i(\zeta(s_i, s_j))$, thus $s_i \notin \rho_i(\mu_i)$. Instead, for every $s_i \in S^1_{i,\Delta^e}(h_1)$ and $h_1 \prec h \prec h_2$, $s_i(h) = b^i_i$.

Hence, for every $s_j \in S_j(h_1)$ such that $s_j \notin S_j(h_2)$ (i.e. all strategies of player 2 following the path until τ and not replying with 1's best continuation after 1's deviation in τ), $s_j \notin S^2_{j,\Delta^e}$ because for every $\mu_j \in \Delta^e_j$ such that $supp\mu_j(\cdot|h_1) \subseteq S^1_{i,\Delta^e}$ and for every $s_i \in supp\mu_j(\cdot|h_1)$, there exists $s'_j \in S_j$ such that $\zeta(s_i, s'_j) = h_2$ and, by (2), $u_j(h_2) > u_j(\zeta(s_i, s_j))$, thus $s_j \notin \rho_j(\mu_j)$. Instead, for every $s_j \in S^2_{j,\Delta^e}(h_1)$ and $h_1 \prec h \prec h_2$, $s_j(h) = b^i_j$.

Hence, for every $s_i \in S_i(z)$, $s_i \notin S^3_{i,\Delta^e}$ because for every $\mu_i \in \Delta^e_i$ such that $supp\mu_i(\cdot|h_0) \subseteq S^2_{j,\Delta^e}$ and for every $s_j \in \text{supp}\mu_i(\cdot|h_0)$, there exists $s'_i \in S_i$ such that $\zeta(s'_i, s_j) = h_2$ and, by (1), $u_i(\zeta(s_i, s_j) = z) < u(h_2)$, thus $s_i \notin \rho_i(\mu_i)$.[17]

Hence, there does not exist $\mu_j \in \Delta^e_j$ such that $supp\mu_j(\cdot|h^0) \subseteq S^3_{i,\Delta^e}$, hence $S^4_{j,\Delta^e} = \emptyset$. ■

Other than delivering a class of non self-enforcing path agreements, the proposition above provides epistemic conditions under which the deviator can confidently upset the path.[18] They are the ones employed in the proof up to the footnote and give rise to the same instances of forward induction reasoning as in the second example of Section 2. That path would indeed fall in this class of paths that can be upset by a convincing deviation by extending its definition to all dynamic games with complete information in the natural way.

Analogously, while SPE paths can be self-enforcing agreements or not, non SPE paths are never self-enforcing.

Theorem 2 *Consider a path $z \in Z$. If there is no SPE $\sigma \in \Delta(S)$ such that $\zeta(supp\sigma) = \{z\}$, then the corresponding path agreement is not self-enforcing.*

Among SPE paths, it is possible that pure SPE paths are all self-enforcing or all not self-enforcing. Trivial examples of games where all

[17] Under the beliefs used so far (and notice that their specification at history h_0 suffices), player 1 is shown to be willing to deviate from the path, confident of having convinced player 2 to follow her preferred subpath after the deviation.

[18] This goes in the same spirit of the intuitive criterion [21] characterization provided by Battigalli and Siniscalchi [12]

paths are self-enforcing can be found in the class of repeated, pure coordination games. Instead, consider to repeat twice the following game. The two players must perform a task that gives a profit of 3 to each of them at the total effort cost of 2. If at least one player works, the task is performed; if only one player works, she pays the total cost of effort, if instead 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 pure SPE path is self-enforcing. 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.

For complete agreements, instead, the equivalence between self-enforceability under priority to the agreement and self-enforceability under priority to rationality does not hold.

As the first example of Section 2 shows, also a Nash, non subgame-perfect equilibrium can be a self-enforcing agreement, under both priority assumptions. One would hope that for a SPE this is always the case.

When the epistemic priority falls on the agreement, a SPE in pure strategies is actually self-enforcing.

Proposition 5 *Consider a SPE $s \in S$ of a game with observable actions and no relevant ties. The agreement $e = (e_i)_{i \in I}$ such that for every $i \in I$ and $h \in H$, $e_i(h) = s_i(h)$ is self-enforcing under priority to the agreement.*

Proof: By observable actions, for every history $h \in H$ I can define the subgame $\Gamma(h)$. Then, by subgame perfection and no relevant ties, for every $i \in I$, s_i is the sole strategy such that for every $h \in H$, $s_i|h$ is a best reply to $s_{-i}|h$. Hence, for every $i \in I$ and $\mu_i \in \Delta_i^e$, $\rho_i(\mu_i) = [s_i] := \{s_i' \in S_i : \forall h \in H(s_i),\ s_i'(h) = s_i(h)\}$. Thus, for every $i \in I$, $S^1_{i,\Delta^e} = [s_i]$ and the set of CPS that strongly believe S^1_{-i,Δ^e} is a superset of Δ_i^e, which implies $S^2_{i,\Delta^e} = [s_i]$. By induction, $S^\infty_{i,\Delta^e} = [s_i]$. I can conclude that the agreement corresponding to s is self-enforcing. ■

Battigalli and Friedenberg [9] provide a game without observable actions where a SPE outcome is not delivered by any extensive form best response set (see next section) and this implies that it cannot be delivered by strong-delta-rationalizability.

When the epistemic priority falls on rationality, instead, not all SPE in pure strategies are self-enforcing agreements. It is well known that some SPE outcomes are not strongly rationalizable;[19] hence there is no hope that the SPE passes the selective rationalizability test. Still, even for SPE whose outcomes are strongly rationalizable, the issue remains subtle. Taking exactly the equilibrium strategies may imply to deem an agreement not credible only because none of a player's best rationalizable actions at some off-the-path information set matches the equilibrium one. Even in a perfect information game like the centipede, the agreement corresponding to the sole SPE is not credible, because among the strongly rationalizable strategy profiles, despite inducing the SPE outcome, there is not the SPE one (see [38]). Hence it is more appropriate to tackle the problem through the alternative perspective: which SPE outcomes can be enforced by some agreement? This issue will be addressed in the next paragraph about enforceability.

4.2 Enforceability

Taking the opposite, implementation perspective, which outcomes of the game can the players hope to achieve through *some* agreement? Given a set of desirable outcomes, I say that it is enforceable if for some agreement the appriopriate rationalizability procedure delivers a non-empty subset of it.

Definition 11 (Enforceability) *Consider a set of outcomes $P \subset Z$. The set is enforceable if there exists an agreement $e = (e_i)_{i \in I}$ such that $\emptyset \neq \zeta(S_e^\infty) \subseteq P$.*

[19] Consider for instance the battle of sexes with an outside option: at the root of the game the man has the chance to take an outside option that gives him an intermediate payoff with respect to the coordination payoffs of the BoS, or he can go to play the BoS. The SPE where the man takes the option and the woman would play her preferred activity in the subgame is not strongly rationalizable. A rational man renouncing to the option will play his preferred activity in the subgame, a woman who believes in his rationality will then reply with his preferred activity too and a man who believes that the woman believes in his rationality will then renounce to the option.

When the epistemic priority falls on the agreement, the candidates for enforceability are the sets of outcomes for which there exists an extensive form best response set [9] (henceforth EFBRS) delivering a subset of them. EFBRS are defined by Battigalli and Friedenberg for the 2-players case and they are shown to be delivered by strong-delta-rationalizability for some restrictions, while strong-delta-rationalizability always delivers an EFBRS. The concept and the result are extended to the N-players, incomplete information case by Battigalli and Prestipino [10]. Here I present a simpler version of the two for the framework here.

Definition 12 (EFBRS) *An extensive form best response set is a cartesian subset of strategy profiles* $Q = \prod_{i \in I} Q_i \subset S$ *such that for every* $i \in I$ *and* $s_i \in Q_i$ *there exists* $\mu_i \in \Delta^H(S_{-i})$ *such that:*

1. $s_i \in \rho_i(\mu_i)$*;*
2. μ_i *strongly believes* Q_{-i} $(\forall h \in H,\, Q_{-i} \bigcap S_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(Q_{-i}|h) = 1)$*;*
3. $\rho_i(\mu_i) \subseteq Q_i$*.*

Proposition 6 *Consider a cartesian subset of strategy profiles* $Q = \prod_{i \in I} Q_i \subset S$*. The following are equivalent:*

1. Q *is an EFBRS;*
2. $Q = S^{\infty}_{\Delta}$ *for some first-order belief restrictions* $(\Delta_i)_{i \in I}$*.*

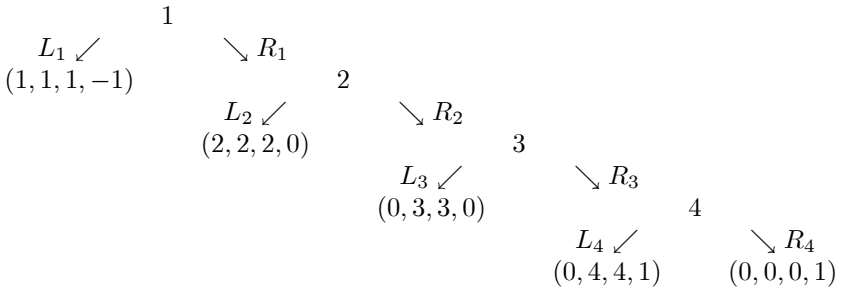
Now I can claim the following:

Proposition 7 *A set of outcomes* $P \subseteq Z$ *is enforceable only if there exists an EFBRS* Q *such that* $\zeta(Q) \subseteq P$*.*

Proof. Combine the previous proposition and definitions. ■

Though, the condition is not sufficient. The reason is that the first-order belief restrictions delivering the EFBRS could not derive from any agreement. First, it has to be ensured that the restrictions just exclude subsets of opponents pure strategies. Moreover, even when this is the case but players are more than 2, it has to be ensured that for any two players such restrictions exclude for both the same beliefs about a third player's

behavior. Unfortunately, this is not always true for a generic EFBRS. The following game is taken from Greenberg, Gupta and Luo [27].



$((R_1, L_2, R_3, L_4), (R_1, L_2, R_3, R_4))$ is an EFBRS. Indeed it is delivered by strong-delta-rationalizability where the restrictions impose to player 1 the belief that 2 will play L_2, to 2 the belief that 3 will play R_3 and 4 will play R_4 and to 3 the belief that 4 will play L_4. But these restrictions cannot come from an agreement because 2 and 3 have discordant ideas about how 4 will play. Indeed, according to the authors, the outcome of such EFBRS is reachable through an agreement that involves only the first two players.[20] For simplicity, the notion of agreement employed in this paper does not allow the exclusion of a subset of players from the agreement itself. Yet, also this case can be analyzed with the methodology proposed here.

What about enforceability when epistemic priority falls on rationality? Differently from self-enforceability, the possibility of enforcing a set of outcomes when the priority is on rationality is strictly limited by its enforceability in the priority to the agreement case.

Proposition 8 *If a set of outcomes is enforceable under priority to rationality, it is enforceable also under priority to the agreement.*

Proof. Consider a set $P \subset Z$ that is enforced by $e = (e_i)_{i \in I}$ under priority to rationality. Consider the reduction procedure $(\widetilde{S}^n)_{n=0}^{M+N} := (S^0, ..., S^M, S^1_{R\Delta^e}, ..., S^N_{R\Delta^e})$, where M is the smallest n such that $S^n = S^{n+1}$ and M is the smallest m such that $S^m_{R\Delta^e} = S^{m+1}_{R\Delta^e}$. For every $i \in$

[20] Also [41] deals with agreements among a subset of players in another non forward induction setting.

I and $h \in H$, define $m_i(h) := \max\left\{n = 0, ..., M+N : \widetilde{S}_i^n(h) \neq \emptyset\right\}$ and $e_i' : H \rightrightarrows \overline{A}_i$ in such a way that $a_i \in e_i'(h)$ if and only if $a_i \in e_i(h)$ and there exists $s_i \in \widetilde{S}_i^{m_i(h)}(h)$ such that $s_i(h) = a_i$. Then, for every $\mu_i \in \Delta_i^{e'}$, $\mu_i \in \Delta_i^e$ and strongly believes $(\widetilde{S}_{-i}^n)_{n=0}^{M+N}$, so that $S^1_{\Delta^{e'}} = S^\infty_{R\Delta^e}$ and $S^\infty_{\Delta^{e'}} = S^1_{\Delta^{e'}}$. Hence, $\zeta(S^\infty_{R\Delta^e})$ is enforced by $e' = (e_i')_{i\in I}$ under priority to the agreement, a fortiori for $P \supseteq \zeta(S^\infty_{R\Delta^e})$. ■

That is, if a set of outcomes is enforced by an agreement under priority to rationality, it is possible to enforce it also under priority to the agreement at least through a different agreement that imposes the conjectures allowed by selective rationalizability after convergence.

Speaking of single outcomes, one should first check if the corresponding path agreement is self-enforcing. If this is not the case, enforceability by loosening restrictions must be excluded.[21]

Theorem 3 *Consider an outcome $z \in Z$ and the corresponding path agreement $e = (e_i)_{i\in I}$. If z is enforced 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 realize a given outcome of the game, without willing or trusting the possibility to threaten in advance any sort of punishment in case of deviation, they cannot do any better than agreeing on the path. This is kind of small "revelation principle" for agreements design: leaving some mystery about on-the-path moves cannot be of any help for the goal. The negative side of it is that if the path is not self-enforcing, then off-the-path restrictions are necessary.

Allowing for off-the-path restrictions, which outcomes can be enforced by some agreement? An agreement can deliver a non subgame perfect Nash outcome, as the first example of Section 2 shows. Being a pure Nash outcome is a necessary condition under priority to the agreement, a fortiori under priority to rationality.

Proposition 9 *Consider an outcome $z \in Z$. If it is enforceable, then there exists a Nash equilibrium $s \in S$ such that $\zeta(s) = \{z\}$.*

Proof. Take an agreement $e = (e_i)_{i\in I}$ that enforces z. For every $i \in I$, take a $s_i \in S^\infty_{i,\Delta^e}$ such that for every $h \in H$, $s_i(h) \in e_i(h)$. Notice

[21] For agreements allowing multiple outcomes, this is not always true.

that s_i is a strategic-form best reply to s_{-i}, otherwise there would exist a $\mu_i \in \Delta_i^e$ that strongly believes $(S^n_{-i,\Delta^e})_{n\geq 0}$ such that $\mu_i(s_{-i}|h^0) = 1$ and $s_i' \in \rho_i(\mu_i)$ such that $\zeta((s_i', s_{-i})) \neq z$. ■

The condition is not sufficient: the game above provides an example of a Nash outcome that cannot be enforced by any agreement.

Moving to SPE outcomes, if priority is on the agreement by proposition 5 the corresponding agreement is self-enforcing, hence the outcome is enforceable. As already pointed out, this is not always the case when the priority falls on rationality. If the outcome is not delivered by strongly rationalizable strategies, there is no hope to enforce it. If it is, one must still check that some profile of strongly rationalizable strategies (as already argued, possibly different than the SPE itself) that delivers the outcome constitutes a Nash equilibrium of the game. Agreeing on the actions prescribed by those strategies will enforce the desired SPE outcome.

At this point, one may wonder whether enforceable SPE outcomes exist in every game when epistemic priority falls on rationality. This is clearly false even under priority on the agreement if one considers the fact that a game could feature only SPE in mixed strategies: the definition of agreement does not allow to agree on mixed actions and agreeing on the whole support does not guarantee self-enforceability. But allowing as an exercise to consider mixed strategies, the question acquires a wider interest that goes beyond the agreement problem: is there always the support of a SPE outcome distribution among strongly rationalizable outcomes? A positive answer would reconcile the two main approaches to solution concepts in dynamic games, backward induction and forward induction. The following theorem states that it is actually so[22] for the wide class of games with observable actions:[23] subgame perfection and strong rationalizability never give completely disjoint predictions.

Theorem 4 *Consider the set of strongly rationalizable strategy profiles S^∞. There exists a SPE $\sigma \in \Delta(S)$ and an equilibrium $\widetilde{\sigma} \in \Delta(S^\infty)$ such that for every $z \in Z$, $\sigma(S(z)) = \widetilde{\sigma}(S(z))$.*

[22] And these outcomes are all induced by reciprocal best replies, an important feature if one wants to go on proving enforceability with mixed actions.

[23] Probably it is possible to extend such class to all extensive form games with perfect recall.

5 Conclusions and further research

Pre-play non-binding agreements are widespread in the economic literature, but few papers addressed explicitely the issue of their credibility. In dynamic games, the evaluation of a pre-play non-binding agreement is much more controversial than in static games,[24] especially when the agreement is only partial. The reason is that forward induction reasoning, based not just on beliefs about rationality but also on beliefs about the compliance with the agreement itself and the interaction of the two, may allow a deviator to re-coordinate (without any explicit renegotiation!) with opponents on a more favorable subpath. With this motivation, Osborne [36] refines subgame perfection with forward induction reasoning considerations that can arise precisely from a pre-play non-binding agreement. Yet, the refinement applies only to a very small class of games (finitely repeated pure coordination games) and the justifying epistemic assumptions are not explicit (see Section 4). Greenberg, Gupta and Luo [27], instead, introduce the interesting concept of path agreement for a generic dynamic game. Yet, the credibility of such agreements is analyzed through a novel solution concept that does not capture forward induction reasoning and whose justifying epistemic assumptions are unknown.

Non-binding agreements can affect the behavior of players only through the beliefs they are able to induce in their minds. Therefore, the solution concepts used for their analysis must derive from clear assumptions about the initial beliefs that the agreement may induce, the way players update them as the game unfolds and their interaction with the beliefs in rationality of all orders. Whether players give priority to beliefs in the agreement or to beliefs in rationality when the two are at odds is a key issue. For the first case, the appropriate tool for the analysis, strong-delta-rationalizability ([7], [10]), already existed in the literature. For the second case, a novel solution concept, selective rationalizability (a refinement with first-order-belief restrictions of strong rationalizability [11]), has been developed and epistemically characterized in [20].

The paper uses these tools to develop a rigorous methodology for the evaluation of self-enforceability of agreements and enforceability of outcomes (through agreements). When players want to implement a certain outcome of the game, the simplest thing they can do is to reach the corresponding path agreement. Other than for their simplicity, path agreements

[24] Although also in static games the identification of self-enforcing agreements with Nash equilibria has been questioned by Aumann ([2]).

are particularly appealing for a more sophisticated reason: their credibility and self-enforceability are robust to the epistemic priority assumption, so that players (and the analyst) do not need to figure out what is the correct one to evaluate them. Unfortunately, limitations to the self-enforceability of path agreements apply. Only SPE, strongly rationalizable [11] paths can be self-enforcing, but some of them are not even credible: the class of *equilibrium paths that can be upset by a convincing deviation* introduced by Osborne [36] is an example. If the agreement corresponding to the desired outcome is not self-enforcing, players have to come up with a different agreement to enforce it. In the view that parsimonious agreements are preferrable and more likely to be believed in practise, one may wonder whether removing some on-the-path restrictions can be of any help for the goal. This is not the case: a sort of "revelation principle" for agreements says that if players are not willing to believe in off-the-path restriction, they cannot do any better than just declaring the outcome they want to implement. Putting off-the-path restrictions, instead, also non subgame perfect Nash outcomes could be enforced. The set of enforceable outcomes is bigger when the epistemic priority falls on the agreement and it is contained in the set of Nash outcomes. Pure SPE outcomes are all enforceable when the epistemic priority falls on the agreement: the corresponding complete agreement is self-enforcing. Under priority to rationality, being a pure SPE outcome is not sufficient for enforceability, although some non subgame perfect Nash outcomes still are. Thus, the conclusions can depart substantially from what the equilibrium refinement tradition suggests. The search for enforceable SPE outcomes under priority to rationality brings to a result of autonomous interest: in every game with observable actions, among strongly rationalizable outcomes there is always the support of a SPE outcome distribution. That is, backward induction and forward induction (as captured by strong rationalizability) never give disjoint predictions.

Although the tools of the analysis can be applied to dynamic games with incomplete information, the focus of the paper has been kept on complete information for interpretative and notational easiness. Yet, in an incomplete information environment, players reaching an agreement at the interim stage may discuss their types in the bargaining process or not. In the first case, first-order-beliefs restrictions could be extended to payoff-relevant types. In the second case, what the agreement suggests about opponents types would be embodied in the rationalizability procedure. However, in both cases, the analysis could be easily extended to games

with incomplete information.[25]

The methodology can instead be already applied to games with infinite horizon and I conjecture that the results of the paper would keep on holding. The paper has focused on games with finite horizon because the epistemic characterization of strong-delta-rationalizability in the infinite horizon case is still under development. Results would interestingly change, instead, by extending the analysis to psychological games [8]. Belief-dependent payoffs, like in the case of guilt-averse players, could sustain the self-enforceability a wider range of agreements. 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.

Moreover, 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. The paper proposes which the credible agreements and enforceable outcomes are. How will players ultimately choose among them? Welfare considerations[26] and a theory of bargaining could refine the answer. Dufwenberg, Servátka and Vadovic [24] propose an interesting approach to the pre-play bargaining issue. Miller and Watson [34], 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 the renegotiation proofness [25] one. Past moves and consequent forward induction considerations could influence the bargaining power of players in renegotiation.

Finally, the tools and results developed in the paper can be applied to a wide range of economic problems. For instance, the macroeconomic problem of commitment about the taxation of capital (see, for instance, [3]) poses the credibility issue investigated here: government promises about future taxation are not binding. To provide an example of application, the appendix exploits an existing work by Gossner [27] in the literature about incomplete codes.

[25] Strong-delta-rationalizability has already been defined for incomplete information games. Selective rationalizability can be extended in the same way.

[26] Welfare opportunities may be wider for less sophisticated players. This is another issue that is worth being explored.

6 Appendix

6.1 An application

Gossner [26] studies the credibility of incomplete codes. The most typical incomplete code coincides with a path agreement. Gossner provides a definition of "credible" incomplete code that for a path agreement case can be stated as follows: in every subgame that follows a unilateral deviation from the path every SPE does not provide an incentive to actually deviate. This ends up being a sufficient condition for the credibility of the path agreement in the epistemically founded sense of this work. This result cannot be directly applied to the case in Gossner because he deals with games with infinite horizon. However, it is reasonable to conjecture that an extension to games with infinite horizon of the analytic tools employed here would provide the same conclusion. Moreover, the result applies to the different but interesting case of codes in a game with finite horizon.

Theorem 5 *Consider a path $z \in Z$ such that for every h that follows a unilateral deviation by a player $l \in I$ and every SPE of $\Gamma(h)$ σ^h, $u_l(z) \geq \sum_{s^h \in supp\sigma^h} \sigma^h(s^h) u_l(\zeta(h, s^h(h), s^h(h, s^h(h)), ...)$. The corresponding path agreement is credible.*

Gossner treats such incomplete codes as self-enforcing. Yet, it has to be noticed that in the sense of this paper, self-enforceability is not guaranteed. Consider the following simple counterexample. Player 1 can *C*omply with the rules or *B*reak them. If she breaks them, her lawyer and the prosecutor play a game with two strategies each.

$1 \cdot -$ $\quad B- \rightarrow$

$C \downarrow$

$(2, \cdot, \cdot)$

2/3	L	R
U	$(3,3,0)$	$(0,0,3)$
D	$(0,0,3)$	$(3,3,0)$

In the only equilibrium of this subgame all strategies are equally likely and the SPE of the whole game induces player 1 to comply with the rules. Indeed, the code that prescribes to player 1 to comply is credible. Yet it is not self-enforcing. Player 1 could be convinced that the lawyer will play U and the prosecutor L. In this case, she would deviate from the path.

6.2 Formal analysis of Section 2

Formal analysis of example 1 in Section 2.

$P_1 \backslash P_2$	M	R
M	$(5,5)$	$\cdot-$
R	$(0,0)$	$(4,4)$

$- - - \quad - \rightarrow$

$P_1 \backslash P_2$	N	O
A	$(2,0)$	$(1,4)$
S	$(0,8)$	$(2,6)$

Agreement:
$e_1(h^0) = \{M\}$, $e_1((M,R)) = \{A\}$; $e_2(h^0) = \{M\}$, $e_2((M,R)) = \{N\}$.

First-order-belief restrictions:
$\Delta_1^e = \left\{\mu_1 \in \Delta^H(S_2) : \mu_1((M.N)|h^0) = 1, \mu_1(R.N|(M,R)) = 1\right\}$;
$\Delta_2^e = \left\{\mu_2 \in \Delta^H(S_1) : \mu_2((M.A)|h^0) = 1, \mu_2(M.A|(M,R)) = 1\right\}$.
Notice that the two sets are singletons.

Strong-delta-rationalizability:
$S^1_{1,\Delta^e} = \{M.A\}, S^1_{2,\Delta^e} = \{M.N, M.O\}$; $S^\infty_{1,\Delta^e} = \{M.A\}, S^\infty_{2,\Delta^e} = \{M.N, M.O\}$.

The sequential best replies of player 2 do not reach (M,R); hence, at the second step, player 1 is still allowed to believe in $R.N$ after (M,R) and the procedure already comes to convergence. Strong-delta-rationalizable strategies comply with the agreement at the reached information sets. The agreement is self-enforcing under priority to the agreement.

Strong rationalizability:
$S^1_1 = \{M.A, M.S, R.A, R.S\}, S^1_2 = \{M.N, M.O, R.N, R.O\}$;
$S^\infty_1 = \{M.A, M.S, R.A, R.S\}, S^\infty_2 = \{M.N, M.O, R.N, R.O\}$.

Selective rationalizability:
$S^1_{1,R\Delta^e} = \{M.A\}, S^1_{2,R\Delta^e} = \{M.N, M.O\}$; $S^\infty_{1,R\Delta^e} = \{M.A\}, S^\infty_{2,R\Delta^e} = \{M.N, M.O\}$.

The procedure is equivalent to strong-delta-rationalizability, because all strategies are strongly rationalizable, hence requirement 3 has no bite.

Formal analysis of example 2 in Section 2.

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

Agreement:
$e_i(h^0) = \{C\}$, $e_i((C,C)) = \{D\}$, $e_i(h) = \{C, D, P\}$ $\forall h \neq h^0, (C,C)$, $i = A, B$;

First-order-belief restrictions:
$\Delta_i^e = \left\{\mu_i \in \Delta^H(S_{-i}) : \mu_i(S_{-i}((C,C),(D,D)))|h^0) = 1\right\}$, $i = A, B$.
Conjectures at (C,C) are restricted in the right way by Bayes rule.

Strong-delta-rationalizability:

$$S^1_{i,\Delta^e} = \left\{s_i \in S_i : s_i((s_i(h^0), C)) = D, \right.$$
$$\left. s_i((s_i(h^0), D)) \neq C \neq s_i((s_i(h^0), P))\right\},$$

i.e. it is worth deviating in the first stage only if after the expected C by the opponent no P is expected, hence D is played and there is no incentive to cooperate in the second stage;

$$S^2_{i,\Delta^e} = \left\{s_i \in S^1_{i,\Delta^e} : s_i(h^0) \neq P, \right.$$
$$\left. s_i(h^0) = C \Rightarrow s_i((C,D)) = s_i((C,P)) = D\right\},$$

i.e. there is no incentive to punish in the first stage since this cannot induce the opponent to cooperate in the second and after cooperating there is no incentive to punish;
$S^3_{i,\Delta^e} = \left\{s_i \in S^2_{i,\Delta^e} : s_i(h^0) = D\right\}$, i.e. there is no incentive to cooperate in the first stage since defecting will not trigger the punishment;
$S^4_{i,\Delta^e} = \emptyset$, i.e. the agreement is not credible.

Selective rationalizability:
$S^1_{i,R\Delta^e} = \left\{s_i \in S^1_{i,\Delta^e} : s_i(h^0) \neq P\right\}$, i.e. the intersection of S^1_{i,Δ^e} and S_i^∞: since no rational strategy of the opponent prescribes to cooperate in

the second stage, punishing in the first stage cannot induce cooperation in the second;[27]

$S^2_{i,R\Delta^e} = \left\{s_i \in S^1_{i,R\Delta^e} : s_i(h^0) = C \Rightarrow s_i(C,D)) = D\right\}$, i.e. after cooperating against a defection, there is no incentive to punish (but notice that it is not a subset of S^2_{i,Δ^e} because having already excluded punishment in the first stage, after cooperation and punishment there is no constraint to expect defection);

$S^3_{i,R\Delta^e} = \left\{s_i \in S^2_{i,R\Delta^e} : s_i(h^0) = D\right\}$, i.e. there is no incentive to cooperate in the first stage since defecting will not trigger the punishment;

$S^4_{i,R\Delta^e} = \emptyset$, i.e. the agreement is not credible, as expected from theorem 1.

6.3 Proofs of the theorems

The proofs of the theorems are based on some common lemmata about strategic reasoning, in the way it is captured by strong-delta-rationalizability and selective rationalizability. Recall that strong rationalizability can be seen as a special case of strong-delta-rationalizability without actual restrictions.

Additional notation:

- μ always denotes a CPS and the letter h always denotes a history;
- the superscript h indicates that the object refers to the subgame $\Gamma(h)$;
- for $\widehat{h} \succsim h$ (and $\widehat{H}$ such that for every $\widetilde{h} \in \widehat{H}$, $\widetilde{h} \succsim \widehat{h}$):
  - $s_i^h =^{\widehat{h}} s_i^{\widehat{h}}$ ($s_i^h =^{\widehat{H}} s_i^{\widehat{h}}$) means that for every $\widetilde{h} \succsim \widehat{h}$ ($\widetilde{h} \in \widehat{H}$), $s_i^h(\widetilde{h}) = s_i^{\widehat{h}}(\widetilde{h})$;
  - $\mu_i^h =^{\widehat{h}} \mu_i^{\widehat{h}}$ ($\mu_i^h =^{\widehat{H}} \mu_i^{\widehat{h}}$) means that for every $\widetilde{h} \succsim \widehat{h}$, there exist a partition of $supp\mu_i^h(\cdot|\widetilde{h})$ $(\widetilde{S}^{h,k}_{-i})_{k=1,\ldots,N}$ and a partition

[27] The other strategies in S^1_{i,Δ^e} are instead strongly rationalizable. The plan of action of the pure SPE featuring coordination is sequential best reply to the conjecture that the opponent will do the same and, after observing a deviation, is instead playing a sequence of Nash actions (and sequences of Nash actions are obviously strongly rationalizable, which justifies also the strong rationalizability of the other strategies).

of $supp\mu_i^{\widehat{h}}(\cdot|\widetilde{h})$ $(\widetilde{S}_{-i}^{\widehat{h},k})_{k=1,...,N}$ such that for every $k = 1, ..., N$, $s_{-i}^h \in \widetilde{S}_{-i}^{h,k}$ and $s_{-i}^{\widehat{h}} \in \widetilde{S}_{-i}^{\widehat{h},k}$, $s_{-i}^h =^{\widehat{h}} s_{-i}^{\widehat{h}}$ $(s_{-i}^h =^{\widehat{H}} s_{-i}^{\widehat{h}})$ and $\mu_i^h(\widetilde{S}_{-i}^{h,k}|\widetilde{h}) = \mu_i^{\widehat{h}}(\widetilde{S}_{-i}^{\widehat{h},k}|\widetilde{h})$;

– $\mu_i^h =^{\widehat{h}} \widetilde{\sigma}_{-i}^{\widehat{h}}$ $(\mu_i^h =^{\widehat{H}} \widetilde{\sigma}_{-i}^{\widehat{h}})$ means that there exist a partition of $supp\mu_i^h(\cdot|\widehat{h})$ $(\widetilde{S}_{-i}^{h,k})_{k=1,...,N}$ and a partition of $supp\widetilde{\sigma}_{-i}^{\widehat{h}}$ $(\widetilde{S}_{-i}^{\widehat{h},k})_{k=1,...,N}$ such that for every $k = 1, ..., N$, $s_{-i}^h \in \widetilde{S}_{-i}^{h,k}$ and $s_{-i}^{\widehat{h}} \in \widetilde{S}_{-i}^{\widehat{h},k}$, $s_{-i}^h =^{\widehat{h}} s_{-i}^{\widehat{h}}$ $(s_{-i}^h =^{\widehat{H}} s_{-i}^{\widehat{h}})$ and $\mu_i^h(\widetilde{S}_{-i}^{h,k}|\widehat{h}) = \widetilde{\sigma}_{-i}^{\widehat{h}}(\widetilde{S}_{-i}^{\widehat{h},k}|\widehat{h})$; for $\sigma^{\widehat{h}} =^{\widehat{H}} \widetilde{\sigma}^{\widehat{h}}$ see $\mu_i^h =^{\widehat{H}} \widetilde{\sigma}_{-i}^{\widehat{h}}$ substituting $\mu_i^h(\cdot|\widehat{h})$ with $\sigma^{\widehat{h}}(\cdot)$;

- I denote by $\sigma_i|\widetilde{S}_i$ the conditional on $\widetilde{S}_i$ of the distribution σ_i over S_i; by $\sigma_i|h$ the $\sigma_i^h \in \Delta(S_i^h)$ such that for every $s_i^h \in S_i^h$, if $h \in H(supp\sigma_i^h)$, $\sigma_i^h(s_i^h) = \sigma_i(\widetilde{S}_i)/\sigma_i(S_i(h))$, where $\widetilde{S}_i$ is the set of $s_i \in S_i(h)$ such that $s_i =^h s_i^h$, otherwise $\sigma_i^h(s_i^h) = \sigma_i(\widetilde{S}_i)$ where $\widetilde{S}_i$ is the set of $s_i \in S_i$ such that $s_i =^h s_i^h$;

- $r_i(\sigma_{-i})$ is the set of strategic-form best replies to $\sigma_{-i} \in \Delta(S_{-i})$;

- $D_l(\overline{S}^h)$ is the set of histories that follow a unilateral deviation by player l from $H(\overline{S}^h)$, i.e. the set of histories $(h, \widetilde{a}^1, ..., \widetilde{a}^t) \succ h$ such that there exists $a_l \neq \widetilde{a}_l$ such that $(h, \widetilde{a}^1, ..., (a_l, \widetilde{a}_{-l}^t)) \in H(\overline{S}^h)$;

- the depth of $\Gamma(h)$ is the lenght T of the longest $(a^1, ..., a^T) \in Z^h$.

- $\zeta^{\emptyset h} : S^h \to Z$ is the function that associates to each $s^h \in S^h$ the path $z = (h, s^h(h), s^h(h, s^h(h)), ...) \in Z$.

Fix a $h \in H$. For every $i \in I$, consider a reduction procedure $(S_{i,n}^h)_{n\geq 0}$ such that $S_{i,0}^h = S_i^h$ and for every $n \geq 1$, $s_i^h \in S_{i,n}^h$ only if there exists μ_i^h that strongly believes $S_{-i,n-1}^h, ..., S_{-i,0}^h$ such that $s_i^h \in \rho_i(\mu_i^h)$. This encompasses not just strong-delta-rationalizability but also selective rationalizability, setting $(S_{i,n}^{h^0})_{n\geq 0} := ((S_i^1, ..., S_i^M), (S_{i,R\Delta^e}^m)_{m\geq 1})$, where $(S_i^m)_{m\geq 0}$ is strong rationalizability and M is the smallest step such that $S^{M+1} = S^M$.

The first lemma claims that a player who is surprised by history $\widehat{h}$ can reach it whatever she conjectures about opponents' behavior after $\widehat{h}$

and, as a consequence, can play any sequential best reply after history $\widehat{h}$ whatever she planned to play otherwise.

Lemma 1 *Consider a player i, a history $\widehat{h}$, a step n and a μ_i that strongly believes $(S^h_{-i,q})^{n-1}_{q=0}$ such that $\mu^h_i(S^h_{-i}(\widehat{h})|p(\widehat{h})) = 0$ and $\rho_i(\mu^h_i) \bigcap S^h_i(\widehat{h}) \neq \emptyset$. Then:*

1. *for every $\mu^{\widehat{h}}_i$ that strongly believes $(S^h_{-i,q}(\widehat{h})|\widehat{h})^{n-1}_{q=0}$, there exists $\widetilde{\mu}^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$ that strongly believes $(S^h_{-i,q})^{n-1}_{q=0}$ such that for every $\widetilde{h} \not\succsim \widehat{h}$, $\widetilde{\mu}^h_i(\cdot|\widetilde{h}) = \mu^h_i(\cdot|\widetilde{h})$;*

2. *for every $s^{\widehat{h}}_i \in \rho_i(\mu^{\widehat{h}}_i)$ and $s^h_i \in \rho_i(\mu^h_i) \bigcap S^h_i(\widehat{h})$, there exists $\widetilde{s}^h_i \in \rho_i(\widetilde{\mu}^h_i)$ such that $\widetilde{s}^h_i =^{\widehat{h}} s^{\widehat{h}}_i$ and for every $\widetilde{h} \not\succsim \widehat{h}$, $\widetilde{s}^h_i(\widetilde{h}) = s^h_i(\widetilde{h})$.*

Proof.

1. For every $s^{\widehat{h}}_{-i} \in S^{\widehat{h}}_{-i}$, call $m(s^{\widehat{h}}_{-i})$ the biggest $m < n$ such that $s^{\widehat{h}}_{-i} \in S^h_{-i,m}(\widehat{h})|\widehat{h}$ and take a $s^h_{-i}(s^{\widehat{h}}_{-i}) \in S^h_{-i,m(s^{\widehat{h}}_{-i})}(\widehat{h})$ such that $s^h_{-i}(s^{\widehat{h}}_{-i}) =^{\widehat{h}} s^{\widehat{h}}_{-i}$. Take a candidate CPS $\widetilde{\mu}^h_i$ such that for every $\widetilde{h} \not\succsim \widehat{h}$, $\widetilde{\mu}^h_i(\cdot|\widetilde{h}) = \mu^h_i(\cdot|\widetilde{h})$ and for every $\widetilde{h} \succsim \widehat{h}$ and $s^{\widehat{h}}_{-i} \in S^{\widehat{h}}_{-i}$, $\widetilde{\mu}^h_i(s^h_{-i}(s^{\widehat{h}}_{-i})|\widetilde{h}) = \mu^{\widehat{h}}_i(s^{\widehat{h}}_{-i}|\widetilde{h})$. It is a CPS because from $p(\widehat{h})$ to $\widehat{h}$ Bayes rule cannot be applied and it is true that $\widetilde{\mu}^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$. Moreover, $\widetilde{\mu}^h_i$ strongly believes $(S^h_{-i,q})^{n-1}_{q=0}$ by construction.

2. Take any $s^{\widehat{h}}_i \in \rho_i(\mu^{\widehat{h}}_i)$, $s^h_i \in \rho_i(\mu^h_i)$ and the $\widetilde{s}^h_i =^{\widehat{h}} s^{\widehat{h}}_i$ such that for every $\widetilde{h} \not\succsim \widehat{h}$, $\widetilde{s}^h_i(\widetilde{h}) = s^h_i(\widetilde{h})$. I have to check that $\widetilde{s}^h_i \in \rho_i(\widetilde{\mu}^h_i)$. For every $\widetilde{h} \not\succsim \widehat{h}$, $\widetilde{h} \in H(\widetilde{s}^h_i)$ and $s^h_{-i} \notin S^h_{-i}(\widehat{h})$, $\zeta(s^h_i, s^h_{-i}) = \zeta(\widetilde{s}^h_i, s^h_{-i})$; moreover $\mu^h_i(S^h_{-i}(\widehat{h})|\widetilde{h}) = 0$, so $\widetilde{s}^h_i$ respects the definition of sequential best reply to μ^h_i at $\widetilde{h}$, but then also to $\widetilde{\mu}^h_i$ because $\widetilde{\mu}^h_i(\cdot|\widetilde{h}) = \mu^h_i(\cdot|\widetilde{h})$. For every $s^h_{-i} \in S^h_{-i}(\widehat{h})$, $\zeta^{\emptyset\widehat{h}}(s^{\widehat{h}}_i, (s^h_{-i}|\widehat{h})) = \zeta^{\emptyset h}(\widetilde{s}^h_i, s^h_{-i})$, hence $\widetilde{s}^h_i|\widehat{h}$ respects the definition of sequential best reply to $\mu^{\widehat{h}}_i$ at every $\widetilde{h} \succsim \widehat{h}$, but then also $\widetilde{s}^h_i$ to $\widetilde{\mu}^h_i$ because $\widetilde{\mu}^h_i =^{\widehat{h}} \mu^{\widehat{h}}_i$. ■

The second lemma claims that if a player abandons the paths of an EFBRS (also amending the maximality requirement 3) whenever she has

a given conjecture that strongly believes those paths, then there is a unilateral deviation from those paths that she takes whatever she conjectures thereafter. The reason is that she can expect the opponents to be surprised by the deviation, hence they could react in any way (see previous lemma).

Lemma 2 *Take a history h, a cartesian set $\overline{S}^h \subset S^h$, a player l, a step n and a (uncorrelated)[28] $\overline{\mu}_l^h$ that strongly believes $\overline{S}_{-l}^h$ such that:*

1. *for every $i \neq l$ and (uncorrelated) μ_i^h that strongly believes $(S_{-i,q}^h)_{q=0}^{n-1}$ and $\widetilde{S}_{-i}^h := \left\{\widetilde{s}_{-i}^h \in S_{-i}^h : \exists \overline{s}_{-i}^h \in \overline{S}_{-i}^h,\ \widetilde{s}_{-i}^h =^{H(\overline{S}^h)} \overline{s}_{-i}^h\right\}$, $\rho_i(\mu_i^h) \subseteq S_{i,n}^h$;*

2. *for every $i \in I$ and $\overline{s}_i^h \in \overline{S}_i^h$ there exist a $s_i^h(\overline{s}_i^h) =^{H(\overline{S}^h)} \overline{s}_i^h$ and a (uncorrelated) $\mu_i^h(\overline{s}_i^h)$ that strongly believes $(S_{-i,q}^h)_{q=0}^{n-1}$ and $\widetilde{S}_{-i}^h$ such that $s_i^h(\overline{s}_i^h) \in \rho_i(\mu_i^h(\overline{s}_i^h))$;*

3. *for every (uncorrelated) $\mu_l^h =^{H(\overline{S}^h)} \overline{\mu}_l^h$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$, $\rho_l(\mu_l^h) \bigcap S_l^h \backslash S_l^h(D_l(\overline{S}^h)) = \emptyset$;*

Then, there exists $\widehat{h} \in D_l(\overline{S}^h)$ such that for every (uncorrelated) $\mu_l^{\widehat{h}}$ that strongly believes $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$, there exists (uncorrelated) $\mu_l^h =^{H(\overline{S}^h)} \overline{\mu}_l^h$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$ such that $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$ and $\rho_l(\mu_l^h) \bigcap S_l^h(\widehat{h}) \neq \emptyset$.

Proof.

The proof is the same for both the correlated and uncorrelated case.

Suppose by contraposition that such $\widehat{h}$ does not exist. I show that under assumptions 1 and 2, assumption 3 is violated. For every $\widehat{h} \in D_l(\overline{S}^h)$ take a $\mu_l^{\widehat{h}}$ that strongly believes $(S_{-l,q}^h(\widehat{h})|\widehat{h})_{q=0}^n$ such that for every $\mu_l^h =^{H(\overline{S}^h)} \overline{\mu}_l^h$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$ such that $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$, $\rho_l(\mu_l^h) \bigcap S_l^h(\widehat{h}) = \emptyset$. For every $i \neq l$, let F_i be the set of all functions $\varphi : \widehat{h} \in D_l(\overline{S}^h) \mapsto s_i^{\widehat{h}} \in proj_i supp \mu_l^{\widehat{h}}(\cdot|\widehat{h})$. For every $\overline{s}_i^h \in \overline{S}_i^h$ and

[28] i.e., for every $\widetilde{h} \succsim h$, $\mu_i^h(\cdot|\widetilde{h}) = \prod_{j \neq i} marg_{S_j^h} \mu_i^h(\cdot|\widetilde{h})$.

$\varphi \in F_i$, let $s_i^h(\overline{s}_i^h, \varphi)$ be the strategy such that for every $\widehat{h} \in D_l(\overline{S}^h)$, $s_i^h(\overline{s}_i^h, \varphi) =^{\widehat{h}} \varphi(\widehat{h})$ and $s_i^h(\overline{s}_i^h, \varphi)(\cdot) = s_i^h(\overline{s}_i^h)(\cdot)$ elsewhere. Order arbitrarily the elements of $D_l(\overline{S}^h) \bigcap H(\overline{s}_i^h)$ as $\widehat{h}^1, ..., \widehat{h}^N$. Set $k = 1$, $\mu_i^{h,1} := \mu_i^h(\overline{s}_i^h)$ and $s_i^{h,1} := s_i^h(\overline{s}_i^h)$.

Recursive step (k).

Take a $\mu_i^{\widehat{h}^k}$ *that strongly believes* $(S_{-i,q}^h(\widehat{h}^k)|\widehat{h}^k)_{q=0}^{n-1}$ *such that* $\varphi(\widehat{h}^k) \in \rho_i(\mu_i^{\widehat{h}^k})$. *Since* $\mu_i^{h,k}$ *strongly believes* $\widetilde{S}_{-i}^h$, $\mu_i^{h,k}(S_{-i}^h(\widehat{h}^k)|p(\widehat{h}^k)) = 0$. *Then, by lemma 1, there exist* $\widetilde{\mu}_i^h =^{\widehat{h}} \mu_i^{\widehat{h}^k}$ *that strongly believes* $(S_{-i,q}^h)_{q=0}^{n-1}$ *such that for every* $\widetilde{h} \not\succeq \widehat{h}$, $\widetilde{\mu}_i^h(\cdot|\widetilde{h}) = \mu_i^{h,k}(\cdot|\widetilde{h})$, *and* $\widetilde{s}_i^h \in \rho_i(\widetilde{\mu}_i^h)$ *such that* $\widetilde{s}_i^h =^{\widehat{h}} \varphi(\widehat{h}^k)$ *and for every* $\widetilde{h} \not\succeq \widehat{h}^k$, $\widetilde{s}_i^h(\widetilde{h}) = s_i^{h,k}(\widetilde{h})$. *Set* $\mu_i^{h,k+1} := \widetilde{\mu}_i^h$ *and* $s_i^{h,k+1} := \widetilde{s}_i^h$; *if* $k < N$, *run the recursive step again increasing* k *by* 1, *if* $k = N$, *observe that by assumption 1* $s_i(\overline{s}_i^h, \varphi) = s_i^{N+1} \in S_{i,h}^n$.

Clearly, using all the $s_i^h(\overline{s}_i^h, \varphi)$, one can build a $\mu_l^h =^{H(\overline{S}^h)} \overline{\mu}_l^h$ that strongly believes $(S_{-l,q}^h)_{q=0}^n$ such that for every $\widehat{h} \in D_l(\overline{S}^h)$, $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$. By the contrapositive hypothesis, $\rho_l(\mu_l^h) \bigcap S_l^h(D_l(\overline{S}^h)) = \emptyset$, hence assumption 3 is violated. ■

Proofs of theorems 1 and 3 are applications of the following lemma, where $(S_e^n)_{n \geq 0}$ can represent either strong-delta-rationalizability or selective rationalizability (in the shape above). The lemma is based on the idea that if in a procedure a player wants to deviate from the paths of an EFBRS, in another procedure where the same paths survive also the deviation should have survived (when in both cases the restrictions allow to believe in those paths and off-the-paths there are no restrictions).

Lemma 3 *Consider a path* $z \in Z$, *the corresponding path agreement* $e^1 = (e_i^1)_{i \in I}$ *and an agreement* $e^2 = (e_i^2)_{i \in I}$ *such that for every* $i \in I$ *and* $h \in H$, $e_i^2(h) \supseteq e_i^1(h)$. *Then* $\zeta(S_{e^2}^\infty) \supseteq \zeta(S_{e^1}^\infty)$ *and if* $\zeta(S_{e^2}^\infty) = z$, $\zeta(S_{e^1}^\infty) = z$.

Proof.

I show that $\zeta(S_{e^2}^\infty) \supseteq \zeta(S_{e^1}^\infty)$. When $\zeta(S_{e^2}^\infty) = z$, the same proof can be employed to show that $\zeta(S_{e^2}^\infty) \subseteq \zeta(S_{e^1}^\infty)$ and so $\zeta(S_{e^1}^\infty) = z$.

The proof is recursive. Suppose that $\zeta(S_{e^1}^\infty) \neq \emptyset$, otherwise it is automatically true.

Initialization. Set $k=1$, $h^1=h^2=h^0$, $(S^1_n)_{n\geq 0}$ and S^1_∞ as $(S^n_{e^2})_{n\geq 0}$ and $S^\infty_{e^2}$, and $(S^2_n)_{n\geq 0}$ and S^2_∞ as $(S^n_{e^1})_{n\geq 0}$ and $S^\infty_{e^1}$ (notice the inversion of 1 and 2!).

Recursive step (k): $\zeta^{\emptyset h^{k+1}}(S^{k+1}_\infty)\subseteq \zeta^{\emptyset h^k}(S^k_\infty)$.
The proof is inductive. For every $i\in I$, call

$$\widetilde{S}^{h^{k+1}}_{-i} := \left\{ \widetilde{s}^{h^{k+1}}_{-i} \in S^{h^{k+1}}_{-i} : \exists s^{h^{k+1}}_{-i} \in S^{h^{k+1}}_{-i,\infty}, \widetilde{s}^{h^{k+1}}_{-i} =^{H(S^{k+1}_\infty)} s^{h^{k+1}}_{-i} \right\}$$

and $\widetilde{\Delta}^{k+1}_{i,n-1}$ the set of $\mu^{h^{k+1}}_i$ that strongly believe $(S^k_{-i,q}(h^{k+1})|h^{k+1}))^{n-1}_{q=0}$ and $\widetilde{S}^{h^{k+1}}_{-i}$. Notice that:

$$\forall i\in I, \forall n>0, \forall \mu^{h^{k+1}}_i \in \widetilde{\Delta}^{k+1}_{i,n-1}, \rho_i(\mu^{h^{k+1}}_i)\subseteq S^k_{i,n}(h^{k+1})|h^{k+1}, \qquad \text{(C1)}$$

by the absence of off-the-path restrictions and, for $k>1$, by construction (consider C2 and C3 for $k-1$). So, it is enough to prove the following.

Inductive hypothesis (n).

$$\forall i\in I, \forall \overline{s}^{h^{k+1}}_i \in S^{k+1}_{i,\infty}, \exists \mu^{h^{k+1}}_i \in \widetilde{\Delta}^{k+1}_{i,n-1}, \rho_i(\mu^{h^{k+1}}_i) \ni s^{h^{k+1}}_i =^{H(S^{k+1}_\infty)} \overline{s}^{h^{k+1}}_i.$$

Basis step (1): $S^k_{-i,0}(h^{k+1})|h^{k+1}=S^{h^{k+1}}_{-i}$ and S^{k+1}_∞ is an EFBRS.

Inductive step (n+1).

Setting h^{k+1} as h and S^{k+1}_∞ as $\overline{S}^h$, assumption 2 of lemma 2 holds by the inductive hypothesis. Assumption 1 holds by C1. Suppose by contradiction that the inductive hypothesis fails at $n+1$ for some $l\in I$ and $\overline{s}^{k+1}_l\in S^{k+1}_{l,\infty}$. Then also assumption 3 must hold for $\overline{\mu}^{h^{k+1}}_l=\mu^{h^{k+1}}_l(\overline{s}^{k+1}_l)$. Hence there exists $h^{k+2}\in D_l(S^{k+1}_\infty)$ such that, calling $\Delta^{h^{k+2}}_{l,n}$ the set of $\mu^{h^{k+2}}_l$ that strongly believe $(S^k_{-l,q}(h^{k+2})|h^{k+2})^n_{q=0}$,

$$\forall \mu^{h^{k+2}}_l \in \Delta^{h^{k+2}}_{l,n}, \exists \mu^{h^{k+1}}_l \in \widetilde{\Delta}^{k+1}_{l,n}, \quad \mu^{h^{k+1}}_l =^{h^{k+2}} \mu^{h^{k+2}}_l, \rho_l(\mu^{h^{k+1}}_l)\bigcap S^{h^k}_l(h^{k+2})\neq\emptyset. \quad \text{(C2)}$$

Moreover, for every $i\neq l$, by the inductive hypothesis there exists $\mu^{h^{k+1}}_i\in\widetilde{\Delta}^{k+1}_{i,n-1}$ such that $\rho_i(\mu^{h^{k+1}}_i)\bigcap S^{h^{k+1}}_i(h^{k+2})\neq\emptyset$. Hence, by lemma

1,

$$\forall i \neq l, \forall \mu_i^{h^{k+2}} \in \Delta_{i,n-1}^{h^{k+2}}, \exists \mu_i^{h^{k+1}} \in \widetilde{\Delta}_{i,n-1}^{k+1},$$
$$\mu_i^{h^{k+1}} =^{h^{k+2}} \mu_i^{h^{k+2}}, \rho_i(\mu_i^{h^{k+1}}) \bigcap S_i^{h^k}(h^{k+2}) \neq \emptyset. \quad \text{(C3)}$$

Define the sequence

$$(S_{i,m}^{k+2})_{m\geq 0} := (S_{i,0}^k(h^{k+2})|h^{k+2}, ..., S_{i,n}^k(h^{k+2})|h^{k+2}), (S_{i,q}^{k+2})_{q>n})$$

such that for every $q > n$, $s_i^{h^{k+2}} \in S_{i,q}^{k+2}$ if and only if there exists a $\mu_i^{h^{k+2}}$ that strongly believes $(S_{-i,m}^{k+2})_{m=0}^{q-1}$ such that $s_i^{h^{k+2}} \in \rho_i(\mu_i^{h^{k+2}})$. For every $q \in \mathbb{N}$, $S_{i,q}^{k+2} \neq \emptyset$, by the absence of off-the-path restrictions and because by C2 and C3 $S_{i,n}^{k+2} \supseteq S_{i,n+1}^{k+2}$, so that $(S_{i,m}^{k+2})_{m\geq 0}$ is a reduction procedure. Set $S_\infty^{k+2} := \bigcap_{n\in\mathbb{N}} S_n^{k+2}$: my claim is that then $\zeta^{\emptyset h^{k+2}}(S_\infty^{k+2}) \subseteq \zeta^{\emptyset h^{k+1}}(S_\infty^{k+1})$, a contradiction. Run the recursive step increasing k by 1 to show it.

The iterations must stop at some point because $\Gamma(h^{k+1})$ becomes less and less deep, until it becomes of depth 1, hence h^{k+2} cannot exist and the contradiction cannot hold. ■

Proof of theorem 1.
The lemma can be applied for $e^1 = e^2 = e$, so that both inclusions hold. ■

Proof of theorem 3.
The lemma can be applied for $e^2 := e'$ and $e^1 := e$. ■

The lemma has also an interesting corollary about the relationship between strong-rationalizabillity and strong-delta rationalizability. The latter does not in general deliver a subset of the former, in terms of strategy profiles. Yet, it delivers a subset of outcomes when there are no off-the-path restrictions with respect to a chosen path. Thus there exists a monotonicity with respect to path restrictions in terms of outcomes.

Corollary 6 *Consider a path agreement $e = (e_i)_{ieI}$. It holds $S_{\Delta^e}^\infty \subseteq S^\infty$.*

Proof. Set $e^2 := e$; seeing strong rationalizability as strong-delta-rationalizability when delta restrictions arise from a silent agreement (e^1), I can apply the lemma. ■

Proofs of theorems 2 and 4 are applications of the following lemma.

Lemma 4 *Take a history h, an agreement $e = (e_i)_{i \in I}$, a player $j \in I$, a step m, a set of unordered non-terminal histories $\widehat{H}$ following h, a set of SPE $(\sigma^{\widetilde{h}})_{\widetilde{h} \in \widehat{H}}$ such that:*

1. *for every $i \in I$ and $\widetilde{h} \succsim h$, $e_i(\widetilde{h}) = A_i(\widetilde{h})$;*

2. *for every $\widetilde{h} \in \widehat{H}$, there exist $l \in I$, an equilibrium $\widetilde{\sigma}^{h,\widetilde{h}} \in \Delta(S^m_{\Delta^e}(h)|h)$ and an equilibrium $\widetilde{\sigma}^{\widetilde{h}} \in \Delta(S^m_{\Delta^e}(\widetilde{h})|\widetilde{h})$ such that $\widetilde{h} \in D_l(supp\widetilde{\sigma}^{h,\widetilde{h}})$ and $\widetilde{\sigma}^{\widetilde{h}} =^{H(supp\sigma^{\widetilde{h}})} \sigma^{\widetilde{h}}$;*

3. *for every $i \neq j$, there exists μ_i that strongly believes $(S^q_{-i,\Delta^e})^{m-1}_{q=0}$ such that $\mu_i(S_{-i}(h)|p(h)) = 0$ and $\rho_i(\mu_i) \bigcap S_i(h) \neq \emptyset$;*

4. *for every $n \leq m$, SPE of $\Gamma(h)$ σ^h such that for every $\widetilde{h} \in \widehat{H}$, $\sigma^h|\widetilde{h} = \sigma^{\widetilde{h}}$, and $\mu^h_j =^{H(supp\sigma^h)} \sigma^h_{-j}$ that strongly believes $(S^q_{-j,\Delta^e}(h)|h)^n_{q=0}$, there exists $\mu_j =^h \mu^h_j$ that strongly believes $(S^q_{-j,\Delta^e})^n_{q=0}$ such that $\rho_j(\mu_j) \bigcap S_j(h) \neq \emptyset$.*

Then, there exist a SPE σ^h such that for every $\widetilde{h} \in \widehat{H}$, $\sigma^h|\widetilde{h} = \sigma^{\widetilde{h}}$, and an equilibrium $\widetilde{\sigma}^h =^{H(supp\sigma^h)} \sigma^h$ such that $supp\widetilde{\sigma}^h \subseteq S^m_{\Delta^e}(h)|h$.

Proof.

The proof is inductive. Throughout the proof, notice that all CPS are valid by assumption 1.

Inductive hypothesis (d).

The lemma holds for every $\widetilde{h} \succ h$ satisfying the hypothesis of the lemma such that $\Gamma(\widetilde{h})$ has depth not bigger than d.

Basis step (1).

For every $n = 0, ..., m-1$ and equilibrium of $\Gamma(h)$ σ^h such that $supp\sigma^h \in S^n_{\Delta^e}(h)|h$, by assumption 4 $r_j(\sigma^h_{-j}) \subseteq S^{n+1}_{j,\Delta^e}(h)|h$. Moreover, for every $i \neq j$, since assumption 3 allows to apply lemma 1, $r_i(\sigma^h_{-i}) \subseteq S^{n+1}_{i,\Delta^e}(h)|h$. Inductively, it holds that $r_i(\sigma^h_{-i}) \subseteq S^m_{i,\Delta^e}(h)|h$.

INDUCTIVE STEP (d+1).

Additional notation. For any set of unordered non-terminal histories $\widetilde{H}$ and a corresponding set of SPE $(\sigma^{\widetilde{h}})_{\widetilde{h}\in\widetilde{H}}$, call $E^{\overline{h}}(\widetilde{H})$ the set of SPE of $\Gamma(\overline{h})$ $\sigma^{\overline{h}}$ such that for every $\widetilde{h} \in \widehat{H}$ such that $\widetilde{h} \succsim \overline{h}$, $\sigma^h|\widetilde{h} = \sigma^{\widetilde{h}}$.

Main contradicting hypothesis. Suppose that for every $\sigma^h \in E^h(\widehat{H})$ there does not exist an equilibrium $\widetilde{\sigma}^h =^{H(supp\sigma^h)} \sigma^h$ such that $supp\widetilde{\sigma}^h \subseteq S^m_{\Delta^e}(h)|h$.

In the rest of the proof, all CPS are meant to be uncorrelated, as well as the mixed profiles (i.e. they are cartesian products of mixed strategies).

Part 1 of the proof shows that the non-existence of such equilibrium requires that there exists a unilateral deviation from the SPE paths such that the deviator deviates whatever she conjectures about opponents moves after the deviation. This amounts to showing that under the assumptions of the lemma and the main contradicting hypothesis, the assumptions of lemma 2 hold.

For every $\sigma^h \in E^h(\widehat{H})$, take the biggest $n^* \leq m+1$ such that:

$$\forall n < n^*,\ \forall i \in I,\ \forall s_i^h \in supp\sigma_i^h,\ \exists \mu_i(s_i^h) =^{H(supp\sigma^h)} \sigma^h_{-i} \text{ that str. bel.}$$
$$(S^q_{-i,\Delta^e})_{q=0}^{n-1},\ \rho_i(\mu_i(s_i^h)) \bigcap S_i(h) \ni s_i(s_i^h) =^{H(supp\sigma^h)} s_i^h. \quad \text{(D1)}$$

By assumption 4 and the fact that σ^h is an equilibrium, $n^* \neq 0$.

Now I build an equilibrium $\widetilde{\sigma}^h =^{H(supp\sigma^h)} \sigma^h$ such that $supp\widetilde{\sigma}^h \subseteq S^n_{\Delta^e}(h)|h$. For every $p \in I$ and $\widehat{h} \in D_p(supp\sigma^h)$, take a $s_p^h \in supp\sigma_p^h$ such that $p(\widehat{h}) \in H(s_p^h)$ and set $\widetilde{S}^{\widehat{h}}_{-p} := (supp\mu_p(s_p^h)(\cdot|\widehat{h}))|\widehat{h}$. For every $i \in I$, let F_i be the set of all functions $\varphi : \widehat{h} \in (D_p(supp\sigma^h))_{p\neq i} \longmapsto s_i^{\widehat{h}} \in \widetilde{S}_i^{\widehat{h}}$. For every $s_i^h \in supp\sigma_i^h$ and $\varphi \in F_i$, let $s_i(s_i^h, \varphi)$ be the strategy such that for every $\widehat{h} \in (D_p(supp\sigma^h))_{p\neq i}$, $s_i(s_i^h,\varphi) =^{\widehat{h}} \varphi(\widehat{h})$ and $s_i(s_i^h,\varphi)(\cdot) = s_i(s_i^h)(\cdot)$ elsewhere. Clearly, using all the $s_i(s_i^h,\varphi)|h$ it is possible to construct the desired equilibrium, but I have to show that $s_i(s_i^h,\varphi) \in S^n_{i,\Delta^e}(h)$. Now I construct recursively a justifying CPS. Order arbitrarily the elements of $(D_p(supp\sigma^h))_{p\neq i} \bigcap H(s_i^h)$ as $\widehat{h}^1, ..., \widehat{h}^N$. Set $k = 1$, $\mu_i^1 := \mu_i(s_i^h)$ and $s_i^1 := s_i(s_i^h)$.

Recursive step (k). *Take a* $\mu_i^{\widehat{h}^k}$ *that strongly believes* $(S^q_{-i,\Delta^e}(\widehat{h}^k)|\widehat{h}^k)_{q=0}^{n-1}$ *such that* $\varphi(\widehat{h}^k) \in \rho_i(\mu_i^{\widehat{h}^k})$. *Observe that* μ_i^k *is such that* $\mu_i^k(S_{-i}(\widehat{h}^k)|p(\widehat{h}^k)) =$

0, *because* μ_i^k *strongly believes, from* h *on,* $S_{-i}(\zeta^{\emptyset h}(supp\sigma^h_{-i}))$ *and* $s_i^k \in \rho_i(\mu_i^k) \bigcap S_i(\widehat{h}^k)$. *Then, by lemma 1, there exist* $\widetilde{\mu}_i =^{\widehat{h}} \mu_i^{\widehat{h}^k}$ *that strongly believes* $S^{n-1}_{-i,\Delta}, ..., S^0_{-i,\Delta}$ *and for every* $\widetilde{h} \not\succeq \widehat{h}$, $\widetilde{\mu}_i(\cdot|\widetilde{h}) = \mu_i^k(\cdot|\widetilde{h})$, *and* $\widetilde{s}_i \in \rho_i(\widetilde{\mu}_i)$ *such that* $\widetilde{s}_i =^{\widehat{h}} \varphi(\widehat{h}^k)$ *and for every* $\widetilde{h} \not\succeq \widehat{h}^k$, $\widetilde{s}_i(\widetilde{h}) = s_i^k(\widetilde{h})$. *Set* $\mu_i^{k+1} := \widetilde{\mu}_i$ *and* $s_i^{k+1} := \widetilde{s}_i$; *if* $k < N$, *run the recursive step again increasing* k *by* 1, *if* $k = N$, $s_i(s_i^h, \varphi) = s_i^{N+1} \in S^n_{i,\Delta^e}(h)$.

Then, by the main contradicting hypothesis, $n^* \leq m$, so that for the reduction procedure $S^0_{\Delta^e}(h)|h, ..., S^{n^*}_{\Delta^e}(h)|h$, for $\overline{S}^h := supp\sigma^h$ and $n := n^*$, assumption 1 of lemma 2 holds by assumptions 1, 3 and 4. Assumption 2 holds by D1 and assumption 3 holds by the negation of D1 at n^* and assumption 4 of this lemma. (Notice that for every $\overline{\mu}_i^h =^h \sigma^h_{-i}$, $\overline{\mu}_i^h$ strongly believes $supp\sigma^h_{-i}$ and $\mu_i =^{H(supp\sigma^h)} \sigma^h_{-i}$ implies $\mu_i =^{H(supp\sigma^h)} \overline{\mu}_i^h$). Hence I can find a $\widehat{h} \in D_l(supp\sigma^h)$ such that for every $\mu_l^{\widehat{h}}$ that strongly believes $(S^q_{-l,\Delta^e}(\widehat{h})|\widehat{h})^{n^*}_{q=0}$, there exists $\mu_l^h =^{H(supp\sigma^h)} \sigma^h_{-l}$ that strongly believes $(S^q_{-l,\Delta^e}(h)|h)^{n^*}_{q=o}$ such that $\mu_l^h =^{\widehat{h}} \mu_l^{\widehat{h}}$ and $\rho_l(\mu_l^h) \bigcap S_l(\widehat{h}) \neq \emptyset$. Moreover, for every $i \neq l$, there exists μ_i^h that strongly believes $(S^q_{-i,\Delta^e}(h)|h)^{n^*-1}_{q=0}$ such that $\mu_i^h(S_{-i}(\widehat{h})|p(\widehat{h})) = 0$ and $\rho_i(\mu_i^h) \bigcap S_i(\widehat{h}) \neq \emptyset$. Hence,

$$\forall i \in I,\ \forall \mu_i^{\widehat{h}} \text{ that strongly believes } (S_{-i,\Delta^e}(\widehat{h})|\widehat{h})^{n^*}_{q=0},\ \rho_i(\mu_i^{\widehat{h}}) \subseteq S^{n^*}_{i,\Delta^e}(\widehat{h})|\widehat{h}. \tag{D2}$$

Part 2 shows that since everyone may play any sequential best reply after the deviation, the reduced subgame $S^{n^*}_{\Delta^e}(\widehat{h})|\widehat{h}$ features an equilibrium corresponding to a SPE complying with appropriate backward induction choices like the ones in $\widehat{H}$.

Consider now a set of unordered non-terminal histories $\widehat{H}^1$ following $\widehat{h}$ and a set of SPE $(\sigma^{\widetilde{h}})_{\widetilde{h} \in \widehat{H}^1}$ such that assumption 2 holds for step $m := n^*$ and $h := \widehat{h}$. Put in $\Sigma^{\widehat{h},1}$ the profiles $\widehat{\sigma}^{\widehat{h}}$ of the mixed extension of $S^{n^*}_{\Delta^e}(\widehat{h})|\widehat{h}$ such that for every $\widetilde{h} \in \widehat{H}^1 \bigcap H(supp\widehat{\sigma}^{\widehat{h}})$, $\widehat{\sigma}^{\widehat{h}}|\widetilde{h} = \widetilde{\sigma}^{\widetilde{h}}$.

Here I show that $\Sigma^{\widehat{h},1}$ is non-empty. For every $i \in I$, for every subset of $\widehat{H}^1$ $\widetilde{H}$ such that $\bigcap_{\widetilde{h} \in \widetilde{H}} S_i^{\widehat{h}}(\widetilde{h}) \neq \emptyset$ and for every pair $\widetilde{h}, \overline{h}$ in $\widetilde{H}$, there exists a $p \neq i$ such that $S_p^{\widehat{h}}(p(\widetilde{h})) \bigcap S_p^{\widehat{h}}(p(\overline{h})) = \emptyset$, otherwise $\widetilde{h}$ and $\overline{h}$ would not be unordered. Hence, there exists $\mu_i^{\widehat{h}}$ that strongly believes $(S^q_{-i,\Delta^e}(\widehat{h})|\widehat{h}))^{n^*}_{q=o}$ such that for every $\widetilde{h} \in \widetilde{H}$, $\mu_i^{\widehat{h}} =^{\widetilde{h}} \widetilde{\sigma}^{\widetilde{h}}_{-i}$, and by D2 $\rho_i(\mu_i^{\widehat{h}}) \subseteq S^{n^*}_{i,\Delta^e}(\widehat{h})|\widehat{h}$.

Here I show that every equilibrium of $\Sigma^{\widehat{h},1}$ $\widetilde{\sigma}^{\widehat{h}}$ is an equilibrium of the whole $\Gamma(\widehat{h})$. For every $i \in I$ and $\mu_i^{\widehat{h}} =^{\widehat{h}} \widetilde{\sigma}_{-i}^{\widehat{h}}$ that strongly believes $(S^q_{-i,\Delta^e}(\widehat{h})|\widehat{h})_{q=0}^{n^*-1}$, there exists $s_i^{\widehat{h}} \in \rho_i(\mu_i^{\widehat{h}}) \subseteq r_i(\widetilde{\sigma}_{-i}^{\widehat{h}})$ such that for every $\widetilde{h} \in \widehat{H}^1 \bigcap H(supp\widetilde{\sigma}^{\widehat{h}})$, $s_i^{\widehat{h}}|\widetilde{h} \in supp\widetilde{\sigma}_i^{\widetilde{h}}$, and by D2 $\rho_i(\mu_i^{\widehat{h}}) \subseteq S^{n^*}_{i,\Delta^e}(\widehat{h})|\widehat{h}$.

Now I show with a recursive procedure that there exists a $\sigma^{\widehat{h}} \in E^{\widehat{h}}(\widehat{H}^1)$ and an equilibrium $\widetilde{\sigma}^{\widehat{h}} \in \Sigma^{\widehat{h},1}$ such that $\widetilde{\sigma}^{\widehat{h}} =^{H(supp\sigma^h)} \sigma^h$. Suppose by contradiction that such SPE does not exist, take any equilibrium of $\Sigma^{\widehat{h},1}$ and call it $\widetilde{\sigma}^{\widehat{h},1}$. Set $k = 1$.

Recursive step (k). *By the contradicting hypothesis, there exists* $p \in I$ *and* $\widehat{h}^k \in D_p(supp\widetilde{\sigma}^{\widehat{h},k})$ *such that for every* $\sigma^{\widehat{h}^k} \in E^{\widehat{h}^k}(\widehat{H}^k)$,

$$\sum_{s^{\widehat{h}^k} \in supp\sigma^{\widehat{h}^k}} u_p(\sigma^{\widehat{h}^k}) > \sum_{s^{p(\widehat{h}^k)} \in supp\widetilde{\sigma}^{\widehat{h},k}|p(\widehat{h}^k)} u_p(\widetilde{\sigma}^{\widehat{h},k}|p(\widehat{h}^k)). \qquad \text{(D3)}$$

Notice that for every $\widehat{h}^w \in \widehat{H}^k$, $\widehat{h}^k \not\succeq \widehat{h}^w$, *because for every* $\widehat{\sigma}^{\widehat{h}} \in \Sigma^{\widehat{h},k}$ *and* $\widehat{h}^w \in \widehat{H}^k$, $\widehat{\sigma}^{\widehat{h}}|\widehat{h}^w = \widetilde{\sigma}^{\widehat{h}^w} =^{H(supp\sigma^{\widehat{h}^w})} \sigma^{\widehat{h}^w} \in E^{\widehat{h}^w}(\widehat{H}^k)$. *Moreover, for every* $i \neq p$ *and* $s_i^{\widehat{h}} \in supp\widetilde{\sigma}_i^{\widehat{h},k} \bigcap S_i^{\widehat{h}}(\widehat{h}^k)$, *there exists* μ_i *that strongly believes* $(S^q_{-i,\Delta^e})_{q=0}^{n^*-1}, ..., S^0_{-i,\Delta^e}$ *such that* $\mu_i(S_{-i}(\widehat{h}^k)|p(\widehat{h}^k)) = 0$ *and* $s_i^{\widehat{h}} \in (\rho_i(\mu_i) \bigcap S_i(\widehat{h}^k))|\widehat{h}$, *hence for* $m := n^*$ *and* $h := \widehat{h}^k$ *assumption 3 holds. Then, by lemma 1, for every* $n \leq n^*$ *and* $s_i^{\widehat{h}^k} \in S^{n-1}_{i,\Delta^e}(\widehat{h}^k)|\widehat{h}^k$, *there exists* $s_i \in S^n_{i,\Delta^e}(h)$ *such that* $s_i =^{\widehat{h}^k} s_i^{\widehat{h}^k}$ *and* $(s_i|\widehat{h})(\cdot) = s_i^{\widehat{h}}(\cdot)$ *elsewhere. Hence, for every* $\sigma^{\widehat{h}^k} \in E^{\widehat{h}^k}(\widehat{H}^k)$ *there exists* $\mu_p =^{H(supp\sigma^{\widehat{h}^k})} \sigma^{\widehat{h}^k}$ *that strongly believes* $(S^q_{-p,\Delta^e})_{q=0}^n$ *such that, by D3,* $\rho_p(\mu_p) \bigcap S_p(\widehat{h}^k) \neq \emptyset$, *thus assumption 4 holds. Assumptions 1 and 2 hold by construction. So, by the inductive hypothesis there exist* $\sigma^{\widehat{h}^k} \in E^{\widehat{h}^k}(\widehat{H}^k)$ *and an equilibrium of* $\Gamma(\widehat{h}^k)$ $\widetilde{\sigma}^{\widehat{h}^k} =^{H(supp\sigma^{\widehat{h}^k})} \sigma^{\widehat{h}^k}$ *such that* $supp\widetilde{\sigma}^{\widehat{h}^k} \subseteq S^m_{\Delta^e}(\widehat{h}^k)|\widehat{h}^k$. *For every* $w \leq k$, *put* $\widehat{h}^w$ *in* $\widehat{H}^{k+1}$ *if for every* $q \leq k$, $\widehat{h}^w \not\succ \widehat{h}^q$. *Denote by* $\Sigma^{\widehat{h},k+1}$ *the set of profiles* $\widetilde{\sigma}^{\widehat{h}}$ *of the mixed extension of* $S^{n^*}_{\Delta^e}(\widehat{h})|\widehat{h}$ *such that for every* $\widehat{h}^w \in \widehat{H}^{k+1}$, $\widetilde{\sigma}^{\widehat{h}}|\widehat{h}^w = \widetilde{\sigma}^{\widehat{h}^w}$. *(It has the same properties of* $\Sigma^{\widehat{h},1}$*.) Take any equilibrium* $\widetilde{\sigma}^{h,k+1}$ *in* $\Sigma^{\widehat{h},k+1}$ *and run the recursive step increasing* k *by* 1. *At some step the histories are exhausted, so I have a contradiction.*

Part 3 uses the post-deviation surviving SPE to construct always a new SPE of the whole $\Gamma(h)$ from which players could deviate only at

strictly earlier histories, until histories are exhausted and I have the desired contradiction.

Set $k = 1$, $\widetilde{H}^0 := \emptyset$, $\sigma^{h,1}$ as one of the $\sigma^h \in E^h(\widehat{H})$ with the biggest n^* and n^1 as such n^*.

Recursive step (k). *Following the procedure above for* $\sigma^{h,k}$ *and* n^k*, take the corresponding* $\widetilde{\sigma}^h$ *and* $\widehat{h}$ *(call them* $\widetilde{\sigma}^{h,k}$ *and* $\widehat{h}^k$*), set* $\widehat{H}^1 := \widehat{H} \bigcup \widetilde{H}^{k-1}$ *(it satisfies assumption 2 because* $n^k \leq n^{k-1} \leq ... \leq n^1$*) and derive the corresponding* $\sigma^{\widehat{h}}$ *and* $\widetilde{\sigma}^{\widehat{h}}$ *(call them* $\sigma^{\widehat{h}^k}$ *and* $\widetilde{\sigma}^{\widehat{h}^k}$*). Notice that for every* $w < k$, $\widehat{h}^k \not\succsim \widehat{h}^w$*, because for every* $\widetilde{h} \succsim \widehat{h}^w$ *such that there exists* $p \in I$ *such that* $\widetilde{h} \in D_p(supp\sigma^{h,k})$*, and for every* $\mu_p = {}^{H(supp\sigma^{h,k})}\sigma^{h,k}_{-p}$ *such that* $\mu_p(\cdot|\widetilde{h}) = \widetilde{\sigma}^{\widehat{h}^w}_{-p}|\widetilde{h} \subseteq \Delta(S^{n^k-1}_{-p,\Delta^e}(\widetilde{h})|\widetilde{h})$ *(since* $n^k \leq n^w$*),* $\rho_p(\mu_p) \bigcap S_p(\widetilde{h}) = \emptyset$*. Let*

$$\widetilde{H}^k := \left\{\widetilde{h} \in \widetilde{H}^{k-1} : \forall \overline{h} \in \widetilde{H}^{k-1} \bigcup \left\{\widehat{h}^k\right\}, \widetilde{h} \not\succ \overline{h}\right\} \bigcup \left\{\widehat{h}^k\right\}.$$

Notice that by construction, for every $w < k$ *such that* $\widehat{h}^w \succ \widehat{h}^k$, $\sigma^{\widehat{h}^k}|\widehat{h}^w = \sigma^{\widehat{h}^w}$*, so that* $E^h(\widehat{H} \bigcup \widetilde{H}^k) \subseteq E^h(\widehat{H} \bigcup \widetilde{H}^{k-1})$*. Set* $\sigma^{h,k+1}$ *as one of the* $\sigma^h \in E^h(\widehat{H} \bigcup \widetilde{H}^k)$ *with the biggest* n^**, call it* n^{k+1}*, and clearly* $n^{k+1} \leq n^k$*. Run the recursive step increasing* k *by* 1 *until the histories are exhausted, so I have a contradiction.* ■

Proof of theorem 2.

For a step after convergence, the contrapositive of the lemma can be applied for all the subgames that follow a unilateral deviation from the self-enforcing path (with empty $\widehat{H}$), showing that there are SPE of these subgames that support it as a SPE of the whole game. ■

Proof of theorem 4.

By seeing strong rationalizability as strong delta rationalizability when the agreement is silent, the lemma can be applied with $h := h^0$ for a step after convergence (with empty $\widehat{H}$). ■

Theorem 5 can be proved with a few arguments already employed.

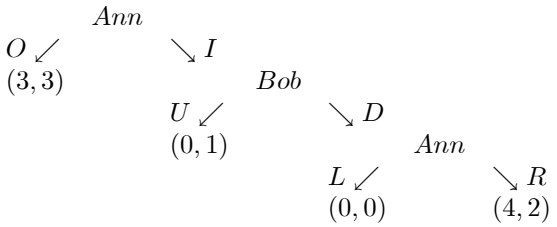
Proof of theorem 5.

Take a SPE σ inducing z. Suppose by contraposition that the path agreement corresponding to z is not credible. Then there exists a smallest $n > 1$ such that for some $l \in I$ and every μ_l that strongly believes $S^n_{-l,\Delta^e}, ..., S^0_{-l,\Delta^e}$ and $S_{-l}(z)$, $\rho_l(\mu_l) \bigcap S_l(z) = \emptyset$. This allows to apply lemma 2 (for correlated conjectures) and find a history $\widehat{h}$ following a unilateral deviation by l from z. Follow part 2 of the proof of the last lemma (with empty $\widehat{H}^1$). I find a SPE $\sigma^{\widehat{h}}$ and an equilibrium $\widetilde{\sigma}^{\widehat{h}} =^{H(supp\sigma^{\widehat{h}})} \sigma^{\widehat{h}}$ of $\Gamma(\widehat{h})$ such that $\widetilde{\sigma}^{\widehat{h}} \in \Delta(S^n_{\Delta^e}(\widehat{h})|\widehat{h})$, but then $u_l(z) < \sum_{s^h \in supp\sigma^h} \sigma^{\widehat{h}}(s^{\widehat{h}}) u_l(\zeta(s^{\widehat{h}}))$. ■

Part II
Selecting strongly rationalizable strategies

7 Introduction

Consider the following game with perfect information.



Suppose that Ann is rational[29] and, at the beginning of the game, believes with probability 1 that Bob would play U after I. Then she would clearly play O. Suppose that Bob is rational and believes with probability 1 that Ann is rational and holds such first-order-beliefs restriction. Then he would expect Ann to play O. So what would he do after observing I? He cannot believe at the same time that Ann is rational and believes that he would play U after I: the two things are at odds given I. Bob has to keep only one of the two beliefs. This is the epistemic priority issue. Suppose he keeps the belief that Ann believes that he would play U after I. Then, after observing I, he has to drop the belief that Ann is rational. Thus he could also expect Ann to play L after (I, D) and so play U. Then, if Ann believes that Bob is rational and holds such beliefs, she can keep her first-order-belief restriction. That is, Ann's belief that Bob would play U after I is consistent with the common belief in this first-order-belief restriction and the highest order of belief in rationality that is compatible with such common belief, according to the reached information set.

[29] i.e. expected utility maximizer given the continuation conjecture at every information set.

This strategic reasoning process is captured by strong-delta-rationalizability ([7], see next section), as shown by its epistemic characterization in [10]. In this process, the faith in the restrictions is so strong that Bob is ready to deem Ann as irrational after I. This could be the case if, for instance, Bob playing U is suggested by some convention that always holds in context of the game ([9]). Suppose instead that restrictions arise from a pre-play non-binding talk between the players. Bob declares he would play U after I. If Bob observes that Ann plays I anyway, he might think that Ann has not taken his words seriously, rather than thinking that Ann is irrational. Then, Bob would expect Ann to play R after D, hence he would play D instead of U. If Ann believes that Bob keeps on believing that she is rational after I, she must believe that Bob will play D, differently than what the first-order-belief restriction suggests. Hence, under this reasoning scheme, such first-order-beliefs restriction cannot hold.

In Section 2 I construct a rationalizability procedure, selective rationalizability, that captures this strategic reasoning process for complete information dynamic games with perfect recall. Just like strong-delta-rationalizability, selective rationalizability could be easily extended to games with incomplete information. However, for notational simplicity and since the pre-play non-binding agreement interpretation does not address beliefs about types, the focus is kept on complete information. In Section 3, the procedure will be epistemically characterized to explicit the epistemic hypotheses that motivate it. Section 4 concludes. The Appendix presents the formal analysis of the previous example, the proof of the characterization theorem and a counterexample for a naively intuitive but wrong result from section 2.

8 Selective rationalizability

Consider an extensive form game with complete information

$$\Gamma = \left\langle I, X, (\overline{A}_i, H_i, u_i)_{i \in I} \right\rangle$$

where:

- I is the set of players;
- $\overline{A}_i$ is the set of actions potentially available to the player, and I write $\overline{A}_J := \prod_{j \in J \subseteq I} \overline{A}_j$;

- $X \subseteq \bigcup_{t\in\{1,...,T\}} \left(\bigcup_{\emptyset\neq J\in 2^I} \overline{A}_J \right)^t \bigcup \left\{ h^0 := \emptyset \right\}$ [30] is a set of histories[31] such that:

  1. $h^0 \in X$;
  2. for every $(\widetilde{a}^1, ..., \widetilde{a}^l) \in X$ and every $t < l$, $(\widetilde{a}^1, ..., \widetilde{a}^t) \in X$, and I write $(\widetilde{a}^1, ..., \widetilde{a}^t) \prec (\widetilde{a}^1, ..., \widetilde{a}^l)$ (and $h^0 \prec x$ for every $x \neq h^0$);
  3. letting $Z := \left\{ z \in X : \forall a \in \bigcup_{\emptyset\neq J\in 2^I} \overline{A}_J,\ (z, a) \notin X \right\}$ denote the set of terminal histories, for every $(\widetilde{a}^1, ..., \widetilde{a}^l) \in X \backslash Z$ there exist $J \in 2^I \backslash \emptyset$ and $(A_j \in 2^{\overline{A}_j} \backslash \emptyset)_{j\in J}$ such that $(\widetilde{a}^1, ..., \widetilde{a}^l, a) \in X$ if and only if $a \in \prod_{j\in J} A_j$;[32]

- $H_i \subset 2^X$ is a set of information sets such that:

  1. it partitions $\left\{ x \in X \backslash Z : \exists a \in \bigcup_{\emptyset\neq J\in 2^I\backslash\{i\}} (\overline{A}_J \times \overline{A}_i),\ (x, a) \in X \right\}$;[33]
  2. for every $(x_1, ..., x_k) \in H_i$, $n, m \leq k$ and $(x_n, \widetilde{a}) \in X$, there exists $(x_m, (\widetilde{a}_i, ...)) \in X$;[34]
  3. for every $h = (x_1, ..., x_k) \in H_i$ and $n, m \leq k$, $x_n \not\prec x_m$; moreover, for every $(\widetilde{x}, \widetilde{a}) \precsim x_n$ such that $\widetilde{x} \in \widetilde{h} \in H_i$, there exists $(\widehat{x}, \widehat{a}) \precsim x_m$ such that $\widehat{x} \in \widetilde{h} \in H_i$ and $\widehat{a}_i = \widetilde{a}_i$;[35]

- $u_i : Z \to \mathbb{R}$ is the payoff function.

Given these primitives, one can retrieve the correspondence that assigns to the player the actions available at a given information set, $A_i(\cdot)$:

[30] h^0 is the empty, initial history, or root of the game.

[31] T is the finite maximum lenght of a history: the game has finite horizon.

[32] At every non-terminal history, the possible new action subprofiles are a cartesian set.

[33] The set of histories where the player is active.

[34] Players have the same feasible actions at every history in an information set. Players indeed cannot distinguish histories in the same information set from observation.

[35] Perfect recall. The first statement means that players remember whether they had been called to act at earlier stages. The second statement means that players can distinguish two histories if they were able to distinguish two predecessors or if they follow two different own moves.

$H_i \rightrightarrows \overline{A}_i$, as

$$A_i(h) = \left\{ a_i \in \overline{A}_i : \forall x \in h,\ \exists (x, (a_i, ...)) \in X \right\}.$$

Then, a strategy can be defined as a function $s_i : H_i \to \overline{A}_i$ such that for every $h \in H_i$, $s_i(h) \in A_i(h)$. The set of all feasible strategies is denoted by S_i. Calling $p(x)$ the predecessor of history x, i.e. $\widetilde{x} \in X$ such that $\widetilde{x} \prec x$ and for every $\widehat{x} \not\preceq \widetilde{x}$, $\widehat{x} \not\prec x$, the set of strategies that are compatible with an information set h (not necessarily of the same player!) is defined as:

$$S_i(h) := \left\{ s_i \in S_i : \exists (\widetilde{a}^1, ..., \widetilde{a}^l) \in h,\ \forall p(\widetilde{a}^1, ..., \widetilde{a}^{t \leq l}) \in \widetilde{h} \in H_i,\ s_i(\widetilde{h}) = \widetilde{a}_i^{t+1} \right\}.$$

For any subset of strategies $\widehat{S}_i \subset S_i$, $\widehat{S}_i(h) := S_i(h) \bigcap \widehat{S}_i$.

On the other hand, the set of information sets that are compatible with a set of strategy (sub-)profiles $\widehat{S} \subseteq \prod_{j \in J \subseteq I} S_j$ is defined as:

$$H_i(\widehat{S}) := \Big\{ h \in H_i : \exists \widehat{s} \in \widehat{S}, \exists (\widetilde{a}^1, ..., \widetilde{a}^l) \in h, \forall j \in J, \\ \forall p((\widetilde{a}^1, ..., \widetilde{a}^{t \leq l})) \in \widetilde{h} \in H_j,\ \widehat{s}_j(\widetilde{h}) = \widetilde{a}_j^t \Big\}.$$

Analogously, the set of outcomes that are compatible with a set of strategy (sub-)profiles $\widehat{S} \subseteq \prod_{i \in J \subseteq I} S_j$ is denoted by

$$\zeta(\widehat{S}) := \Big\{ (\widetilde{a}^1, ..., \widetilde{a}^l) \in Z : \exists \widehat{s} \in \widehat{S}, \forall j \in J, \\ \forall p((\widetilde{a}^1, ...,\ \widetilde{a}^{t \leq l})) \in \widetilde{h} \in H_j, \widehat{s}_j(\widetilde{h}) = \widetilde{a}_j^t \Big\}.$$

For all the player-specific objects, profiles or sets of profiles, i.e. cartesian products across all players, will be denoted by removing the subscript. The subscript $-i$ will instead denote the cartesian product across all players but i.

Players update their conjectures about the state of the world as the game unfolds. Information sets represent all the information acquired by players from observation. This process is captured by modeling conjectures as conditional probability systems [39] (henceforth CPS). Here I define directly CPS over the whole state space $\Omega := S \times T$, where type sets $(T_i)_{i \in I}$ will be defined in the next section.

Definition 13 *A conditional probability system (or CPS) on* $(\Omega_{-i}, (T_{-i} \times S_{-i}(h))_{h \in H_i})$, *with Borel sigma algebra* $\boldsymbol{B}(\Omega_{-i})$, *is a mapping* $\mu(\cdot|\cdot) : \boldsymbol{B}(\Omega_{-i}) \times (T_{-i} \times S_{-i}(h))_{h \in H_i} \to [0,1]$ *satisfying the following axioms:*

1. *for every* $C \in (T_{-i} \times S_{-i}(h))_{h \in H_i}$, $\mu(C|C) = 1$;
2. *for every* $C \in (T_{-i} \times S_{-i}(h))_{h \in H_i}$, $\mu(\cdot|C)$ *is a probability measure on* Ω_{-i};
3. *for every* $E \in \boldsymbol{B}(\Omega_{-i})$, $B, C \in (T_{-i} \times S_{-i}(h))_{h \in H_i}$, *if* $E \subseteq B \subseteq C$ *then* $\mu(E|B)\mu(B|C) = \mu(E|C)$.[36]

The set of all CPS is denoted by $\Delta^{H_i}(\Omega_{-i})$. Endowing with the set of probability measure on Ω_{-i} ($\Delta(\Omega_{-i})$) with the topology of weak convergence, $\Delta^{H_i}(\Omega_{-i})$ can be seen as a compact metrizable of $(\Delta(\Omega_{-i}))^{H_i}$, endowed with the product topology In this section, CPS on just strategies will be used. They can be obtained by replacing $\Omega_{-i}$with S_{-i} and $(T_{-i} \times S_{-i}(h))_{h \in H_i}$ with $(S_{-i}(h))_{h \in H_i}$. For brevity, conditioning events will be indicated with just the information set. For every $i \in I$, I take a compact subset of $\Delta^{H_i}(\Omega_{-i})$, Δ_i (the "first-order-belief restrictions"). It collects all the conjectures that player i could possibly hold, given the arguments that she has to exclude the others.

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 opponents and themselves will play in the continuation of the game. This is equivalent [5] to choosing a *sequential best reply* to the CPS.

Definition 14 *A strategy* $s_i \in S_i$ *is a sequential best reply to a CPS* $\mu_i \in \Delta^{H_i}(S_{-i})$ *if for every* $h \in H_i(s_i)$ *and every* $\widetilde{s}_i \in S_i(h)$,

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

The set of sequential best replies to a conjecture μ_i is denoted by $\rho_i(\mu_i)$.

[36] This means applying Bayes rule whenever possible.

In order to introduce selective rationalizability, I first have to introduce strong rationalizability [11].[37] Strong rationalizability can be seen as a special case of strong-delta-rationalizability and strong-delta-rationalizability can be usefully compared to selective rationalizability to understand the epistemic priority issue. Hence, I first translate the ultimate definition of strong-delta-rationalizability [10] in the complete information framework of this paper.

Definition 15 (strong-delta-rationalizability) *Fix a collection* $\Delta = (\Delta_i)_{i \in I}$ *of compact subsets of CPS. Consider the following procedure.*

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

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

1. $s_i \in \rho_i(\mu_i)$
2. $\forall p = 0, ..., n-1$, $\forall h \in H_i$, $S^p_{-i,\Delta} \bigcap S_{-i}(h) \neq \emptyset \Rightarrow \mu_i(S^p_{-i,\Delta}|h) = 1$;

Finally let $S^\infty_{i,\Delta} = \bigcap_{n \geq 0} S^n_{i,\Delta}$*. The profiles in* S^∞_Δ *are called strongly-delta-rationalizable.*

When for every $i \in I$, $\Delta_i = \Delta^{H_i}(S_{-i})$, i.e. first-order beliefs are not restricted, this is strong rationalizability. Strong rationalizability delivers the behavioral implications of common strong belief in rationality [11]. More generally, the steps of strong rationalizability deliver the *best rationalizable* [6] strategies of each player at each information set, i.e. player's strategies that are compatible with the information set and with the highest possible order of belief in rationality.

Otherwise, strong-delta-rationalizability can also deliver an empty set. This means that for some player at some information set the first-order-belief restrictions are at odds with the behavioral consequences of this reasoning process.

Now I can define selective rationalizability.

[37] The precursor of strong rationalizability is extensive-form-rationalizability from [37].

Definition 16 (selective rationalizability) *Fix a collection* $\Delta = (\Delta_i)_{i \in I}$ *of compact subsets of CPS. Denote by* $(S^m)_{m=0}^{\infty}$ *the strong rationalizability procedure. Consider the following procedure.*

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

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

1. $s_i \in \rho_i(\mu_i)$*;*
2. $\forall p = 0, ..., n-1, \forall h \in H_i, S^p_{-i,R\Delta} \bigcap S_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(S^p_{-i,R\Delta} | S_{-i}(h)) = 1$*;*
3. $\forall q = 0, ..., \forall h \in H_i, S^q_{-i} \bigcap S_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(S^q_{-i} | S_{-i}(h)) = 1$*;*

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

Notice two important facts. First, selective rationalizability can be seen as a refinement of strong rationalizability, since requirement 3 imposes the same conjectures as the strong rationalizability procedure once arrived to convergence. Second, selective rationalizability can be seen as a special case of strong-delta-rationalizability, since requirement 3 is a constant restriction on CPS that could be incorporated in Δ_i, while requirements 1 and 2 are the same.

Requirement 3 ensures that epistemic priority falls on rationality. At every step, at every information set, players are obliged to put probability 1 on the best rationalizable opponents' strategies at the information set. Only among them, players select the ones that are allowed by the first-order-belief restrictions, are compatible with opponents using their own restrictions (if any), and so on (requirement 2). Players choose sequential best replies to such conjectures (requirement 1).

When the procedure delivers an empty set, it is possible that shifting the epistemic priority from rationality to the restrictions, the latter become credible and strong-delta-rationalizability delivers a non-empty set. This is what happens in the example of the introduction. One could conjecture that the opposite is not possible: if strong-delta-rationalizability delivers an empty set, so, a fortiori, will selective rationalizability. This is not the case: see the appendix for a counterexample.

9 Epistemic framework and characterization theorem

For the formalization of the beliefs that motivate the use of selective rationalizability, I present a simplification of the framework in [10] where the incompleteness of information dimension is dropped. However, the choice of this framework will allow to easily extend the analysis to dynamic games with incomplete information in future work.

For every $i \in I$, take space of epistemic types T_i and a belief map $g_i = (g_{i,h})_{h \in H_i} : T_i \to \Delta^{H_i}(\Omega_{-i})$ such that T_i is compact metrizable and g_i is continuous and onto[38]; $(\Omega_i, T_i, g_i)_{i \in I}$ is a belief-complete type structure. In order to find the subset of the state space where the restrictions hold, I must first retrieve the marginal CPS on opponents strategies. For every $i \in I$, define $f_i = (f_{i,h})_{h \in H_i} : T_i \to \Delta^{H_i}(S_{-i})$ as $f_{i,h}(t_i) = marg_{S_{-i}} g_{i,h}(t_i)$; the subset of the state space where the restrictions hold for player i is $[\Delta_i] := \{(s_i, t_i, \omega_{-i}) \in \Omega : f_i(t_i) \in \Delta_i\}$, and for all players $[\Delta] := \bigcap_{i \in I} [\Delta_i]$.[39]. $[\Delta]$ is compact because for every $i \in I$, Δ_i is compact and f_i is continuous.

In order to define subsets of the state space where beliefs over events in the state space itself hold, it will be useful to extend belief mappings on the player herself. For every $i \in I$, define $g_i^* := (g_{i,h}^*)_{h \in H_i} : \Omega_i \to \Delta^{H_i}(\Omega)$ from g_i through the following formula: for every $(s_i, t_i) \in \Omega_i$, $h \in H_i$ and $E \in \boldsymbol{B}(\Omega)$,

$$g_{i,h}^*((s_i, t_i))(E) = g_{i,h}(t_i)(\{\omega_{-i} \in \Omega_{-i} : ((s_i^h, t_i), \omega_{-i}) \in E\}),$$

where s_i^h is the unique strategy in $S_i(h)$ that coincides with s_i at each information set that does not strictly precede h (thus, $s_i^h = s_i$ if and only if $s_i \in S_i(h)$).

The closed subset of the state space where player $i \in I$ believes in an event $E \subset \Omega$ at an information set $h \in H_i$ is defined as

$$B_{i,h}(E) := \left\{(s, t) \in \Omega : g_{i,h}^*((s_i, t_i))(E) = 1\right\}.$$

The subset of the state space where i believes in E at every information set is then $B_i(E) := \bigcap_{h \in H_i} B_{i,h}(E)$. When E is not a purely epistemic event,

[38] This imposes to choose sets of types with the cardinality of the continuum.

[39] Notice that the correct operation is the intersection and not the cartesian product because even when they refer to a single players events in the state space already span all the space across players.

i.e. $proj_S E \neq S$, it could be impossible to believe in it at some information set, because it may be contradicted by observation.[40] However, to perform forward induction it is necessary to believe in events of this kind as long as observation does not contradict them: this allows to rationalize opponents' behaviour and forecast future moves based on past ones. The subset of the state space where this persistency of a belief holds is represented by the *strong belief* [11] operator:

$$SB_i(E) := \bigcap_{h \in H_i : proj_S E \bigcap S(h) \neq \emptyset} B_{i,h}(E).$$

For belief operators, the absence of the subscript (or the subscript $-i$) will denote their intersection across all players (but i).[41] The correct strong belief operator is defined directly for the conjunction of players: $CSB(E)) := E \bigcap SB(E)$.

First-order beliefs may not be enough to explain the strategic reasoning of the players. Higher order beliefs are defined as powers of the belief operators: for any operator $O : \Omega \to \Omega$, $O^{n+1}(E) := O(O^n(E))$. When an epistemic event is believed by all players at every order, I say it is *transparent* and I write $B^*(E) := \bigcap_{n \in \mathbb{N}} B^n(E)$. For the correct strong belief operators, which will work with non-epistemic events, I define common strong belief as $CSB^\infty(E) := \bigcap_{n \geq 0} CSB^n(E)$.

Since $[\Delta]$ is an epistemic event, it cannot have any behavioral implication per se. Moreover, the belief in it of any order does not say anything about opponents' behavior either: in order to have bite, they have to accompanied by the corresponding beliefs in rationality. The elementary brick for this scope is the rationality event. The subset of the state space where i plays a sequential best reply to her conjecture is denoted by

$$R_i := \{(s_i, t_i, \omega_{-i}) \in \Omega : s_i \in \rho_i(f_i(t_i))\}.$$

R_i is closed because $\rho_i \circ f_i$ is upper-hemicontinuous. The rationality event is $R := \bigcap_{i \in I} R_i$.

Here I want to represent the situation in which players, along the game, hold all orders of beliefs in rationality that are consistent with the observed behavior and, within this event, form their conjectures according to their

[40] This cannot happen for epistemic events because beliefs are not observable.

[41] Though, if E is a cartesian product, $B(E)$, $SB(E)$, $CSB(E)$ are cartesian sets.

first-order-beliefs restrictions. Moreover, I suppose that players believe as long as not contradicted by observation that the opponents form conjectures in this way, and so on. The first sentence is captured by the event $[\Delta] \bigcap CSB^{\infty}(R)$. $CSB^{\infty}(R)$ is the common strong belief in rationality event, which characterizes strong rationalizability [11]. It is always non-empty. Its conjunction with $[\Delta]$ can instead be empty, because at some information set a player may find no allowed conjecture over the best rationalizable strategies of the opponents. Capturing also the second sentence delivers a subset of this event that is going to be the event of interest here: $CSB^{\infty}([\Delta] \bigcap CSB^{\infty}(R))$.

Now the following characterization theorem can be stated:

Theorem 7 *Fix a collection $\Delta = (\Delta_i)_{i \in I}$ of compact subsets of CPS. Then, for every $n \geq 0$,*

$$S^{n+1}_{R\Delta} = \text{proj}_S CSB^n([\Delta] \bigcap CSB^{\infty}(R)),$$

and

$$S^{\infty}_{R\Delta} = \text{proj}_S CSB^{\infty}([\Delta] \bigcap CSB^{\infty}(R)).$$

The comparison of this characterization of selective rationalizability with the characterization of strong-delta-rationalizability proposed by [10] clarifies the epistemic priority difference behind the two solution concepts. In the event $CSB^{\infty}(R \bigcap B^*([\Delta])) \subset B^*([\Delta])$ that characterizes strong-delta-rationalizability,[42] players hold at every information set all orders of beliefs in the restrictions and only the orders of belief in rationality whose pairwise conjunctions with the belief in the restriction of the same order deliver opponents strategies that are consistent with the information set. In the event $CSB^{\infty}([\Delta] \bigcap CSB^{\infty}(R)) \subset CSB^{\infty}(R)$, instead, the opposite holds: players hold at every information set all the orders of beliefs in rationality that are per se consistent with the information set and only the orders of beliefs in the restrictions whose pairwise conjunctions with the belief in rationality of the same order deliver opponents strategies that are consistent with the information set. Both events are empty when at some information set the restrictions themselves are at odds with the behavioral

[42] Also the event $CSB^{\infty}(R \bigcap [\Delta])$ characterizes strong-delta-rationalizability (see [10]) because dropping also the belief in the restriction of some order when it is at odds with the belief in rationality of the same order does not enlarge the set of conjectures over opponents' strategies, since the former has no bite without the latter.

consequences of the previously described beliefs that are consistent with the information set.

Notice instead that the event $CSB^{\infty}(R) \bigcap B^{*}([\Delta])$[43] is not just wrong to characterize selective rationalizability, but it is also not very compelling per se: it is likely to be empty for intuitively credible restrictions. Indeed, it is empty even for restrictions corresponding to a strongly rationalizable SPE: off-the-path, it is impossible to believe at the same time that the deviator is rational and believes in the equilibrium.[44] In the correct event, instead, the strong belief in the conjunction of restrictions and common strong belief in rationality allows to drop the belief that the deviator believes in the restrictions.

10 Conclusions and further research

In most dynamic games, common strong belief in rationality is not sufficient to derive sharp predictions on how opponents will behave. In this case, players will try to refine further their conjectures using exogenous information or seeking some kind of coordination with the opponents before starting to play. Especially in the second case, players will tentatively believe in the coordination; will believe that opponents believe in the coordination, and so on. Yet, all these beliefs, together with beliefs in rationality, can be at odds with observed behavior. In this case, players holding common strong belief in rationality are forced to drop the beliefs in the restriction of conjectures that give rise to the incompatibility.

This process is captured here by selective rationalizability. Differently than strong-delta-rationalizability [7], which captures players holding all order of beliefs in the restrictions at every information set and rather dropping the incompatible beliefs in rationality, selective rationalizability delivers a subset of strongly rationalizable strategy profiles. This subset can be the empty set when the restrictions themselves are not compatible with the strategic reasoning depicted above. Selective rationalizability is then epistemically characterized to explicit all the assumptions on which the reasoning process is based.

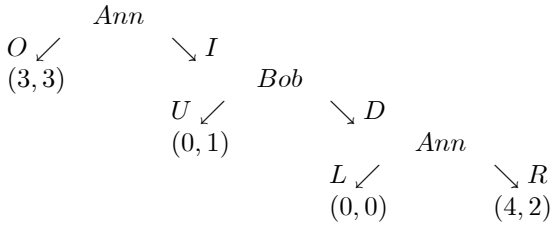
[43] The same event had been conjectured to characterize strong-delta-rationalizability in [12].

[44] Unless some tie occurs.

The interpretation of the first-order-belief restriction as arising from a pre-play non-binding agreement is of particular interest. Catonini [19] uses selective rationalizability (and strong-delta-rationalizability) to study the self-enforceability of appealing classes of agreements. However, first-order-beliefs restrictions can find a different interpretation in many fields of economics. In financial markets, for instance, they can correspond to information arriving to the market, and be used by agents who hold common strong belief in rationality to refine their conjectures. A context of this kind is likely to feature incompleteness of information. The restictions can concern precisely the information that is unknown to some player. Therefore, it would be easy and profitable to extend selective rationalizability to an incomplete information framework.

11 Appendix

Formal analysis of the example in the introduction



First-order-beliefs restrictions:
$\Delta_1 := \left\{\mu_1 \in \Delta^{H_1}(S_2) : \ \mu_1(D|h^0) = 0\right\}$; $\Delta_2 := \Delta_2^{H_2}(S_1)$.
(player 2 threatens to play U)

Strong rationalizability:
$S_1^1 = \{O.L, O.R, I.R\}$, $S_2^1 = \{U, D\}$;
$S_1^2 = \{O.L, O.R, I.R\}$, $S_2^2 = \{D\}$;
$S_1^3 = \{I.R\} = S_1^\infty$, $S_2^3 = \{D\} = S_2^\infty$.

Strong-delta-rationalizability:
$S_{1,\Delta}^1 = \{O.L, O.R\} = S_1^\infty$, $S_{2,\Delta}^1 = \{U, D\} = S_{2,\Delta}^\infty$.

Selective rationalizability:
$S^1_{1,R\Delta} = \emptyset$.

PROOF OF THEOREM 7.

First, I prove a generalized version of the theorem. Applying this generalized version to strong rationalizability yields the hypotheses to run the same proof for selective rationalizability and prove the theorem.

Consider this generalized rationalizability procedure:

Definition 17 *Fix two collections,* $\Delta = (\Delta_i)_{i\in I}$ *and* $\Delta^G = (\Delta^G_i)_{i\in I}$*, of compact subsets of CPS. Consider the following procedure.*

(Step 0) *For every* $i \in I$*, let* $S^0_{i,G} = S_i$*. Moreover let* $S^0_{-i,G} = \prod_{j\neq i} S^0_{j,G}$ *and* $S^0_G = \prod_{i\in I} S^0_{i,G}$*.*

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

1. $s_i \in \rho_i(\mu_i)$*;*
2. $\forall p = 1, ..., n-1, \forall H_i, S^p_{-i,G} \bigcap S_{-i}(h) \neq \emptyset \Longrightarrow \mu_i(S^p_{-i,G}|S_{-i}(h)) = 1$*;*
3. $\mu_i \in \Delta^G_i$*.*

Moreover, let $S^n_{-i,G} = \prod_{j\neq i} S^n_{j,G}$ *and* $S^n_G = \prod_{i\in I} S^n_{i,G}$*.*
Finally, let $S^\infty_G = \bigcap_{n\geq 0} S^n_G$*.*

Consider the following property for a cartesian event $E \subseteq \Omega$.

Definition 18 *A cartesian event* $E \subseteq \Omega$ *satisfies the "completeness" property if for every*

- $i \in I$*,*
- $(s_i, t_i) \in E_i$*,*
- $\tau_{-i} : s'_{-i} \in \text{proj}_{S_{-i}} E \longmapsto (s'_{-i}, t_{-i}) \in E_{-i}$*,*

there exists $t_i' \in \text{proj}_{T_i} E$ *such that:*

- $(s_i, t_i') \in E_i$,
- $f_i(t_i') = f_i(t_i)$,
- $g_{i,\cdot}(t_i')\left[\tau_{-i}(s_{-i})\right] = f_i(t_i)\left[s_{-i}|S_{-i}(\cdot)\right]$ *for every* $s_{-i} \in \text{proj}_{S_{-i}} E$.

Now I can state the generalized version of the theorem.

Lemma 5 *Suppose that there exists a compact, cartesian event* $E \subseteq R$ *with the completeness property such that:*

1. *for every* $i \in I$ *and for every* $\mu_i \in \Delta_i \bigcap \Delta_i^G$, *there exists a* $\omega = (s,t) \in E$ *such that* $f_i(t_i) = \mu_i$ $(\Longrightarrow S_G^1 \subseteq \text{proj}_S E)$;
2. *for every* $\omega = (s,t) \in E$ *and for every* $i \in I$, $f_i(t_i) \in \Delta_i \bigcap \Delta_i^G$ $(\Longrightarrow S_G^1 \supseteq \text{proj}_S E)$.

Then, for every $n \geq 1$, $CSB^{n-1}(E)$ *has the completeness property and:*

1. *for every* $i \in I$ *and for every* $\mu_i \in \Delta_i \bigcap \Delta_i^G$ *satisfying requirements 2 (up to* $n-1$*), there exists a* $\omega = (s,t) \in CSB^{n-1}(E)$ *such that* $f_i(t_i) = \mu_i$ $(\Longrightarrow S_G^n \subseteq \text{proj}_S CSB^{n-1}(E))$;
2. *for every* $\omega = (s,t) \in CSB^{n-1}(E)$ *and for every* $i \in I$, $f_i(t_i) \in \Delta_i \bigcap \Delta_i^G$ *and satisfies requirements 2 (up to* $n-1$*)* $(\Longrightarrow S_G^n \supseteq \text{proj}_S CSB^{n-1}(E))$.

Moreover, the same holds replacing n *with* ∞.

Proof of the lemma.

First I show that for every $n \geq 1$, $CSB^{n-1}(E)$ has the completeness property. Take any $i \in I$, $\omega_i = (s_i, t_i) \in proj_{\Omega_i} CSB^{n-1}(E)$ and $\tau_{-i} : s_{-i}' \in proj_{S_{-i}} CSB^{n-1}(E) \longmapsto (s_{-i}', t_{-i}) \in proj_{\Omega_{-i}} CSB^{n-1}(E)$. Extend τ_{-i} to $\tau_{-i}' : s_{-i}' \in proj_{S_{-i}} E \longmapsto (s_{-i}', t_{-i}) \in E_{-i}$ in such a way that for every $s_{-i}' \in proj_{S_{-i}} CSB^{n-1}(E)$, $\tau_{-i}'(s_{-i}') = \tau_{-i}(s_{-i}')$ and for every $m < n-1$ and $s_{-i}' \in proj_{S_{-i}} CSB^{m}(E)$, $\tau_{-i}'(s_{-i}') \in proj_{\Omega_{-i}} CSB^{m}(E)$. By the completeness property of E, there exists a $\omega_i' = (s_i, t_i') \in E_i$ such that $f_i(t_i') = f_i(t_i)$ and for every $s_{-i} \in proj_{S_{-i}} E$, $g_{i,\cdot}(t_i')\left[\tau_{-i}'(s_{-i})\right] = f_i(t_i)\left[s_{-i}|S_{-i}(\cdot)\right]$. Notice that by construction, for every $m < n-1$, t_i' strongly believes $CSB^m(E)$. Therefore, for every $m < n-1$, $\omega_i' \in$

$proj_{\Omega_i} SB_i(CSB^m(E))$. Hence $\omega_i' \in proj_{\Omega_i} CSB^{n-1}(E)$,[45] $f_i(t_i') = f_i(t_i)$ and for every $s_{-i} \in proj_{S_{-i}} CSB^{n-1}(E)$, $g_{i,\cdot}(t_i')\,[\tau_{-i}(s_{-i})] = f_i(t_i)\,[s_{-i}|S_{-i}(\cdot)]$.

To show that $CSB^{\infty}(E)$ has the completeness property too, the same procedure can be applied with ∞ in place of n.

Now I can prove the lemma by induction.

Inductive hypothesis: the lemma holds for step n.

Basis step: the lemma holds for step 1 by hypothesis.

Inductive step:

1. Take any $i \in I$ and any $\mu_i \in \Delta_i \bigcap \Delta_i^G$ satisfying requirements 2 (up to n). By the inductive hypothesis, there exists a $\omega_i = (s_i, t_i) \in CSB_i^{n-1}(E)$ such that $f_i(t_i) = \mu_i$. Take any $\tau_{-i} : s_{-i}' \in proj_{S_{-i}} CSB^{n-1}(E) \longmapsto (s_{-i}', t_{-i}) \in proj_{\Omega_{-i}} CSB^{n-1}(E)$. As shown before, $CSB^{n-1}(E)$ has the completeness property, hence there also exists a $\omega_i' = (s_i, t_i') \in proj_{\Omega_i} CSB^{n-1}(E)$ such that $f_i(t_i') = f_i(t_i)$ and for every $s_{-i} \in proj_{S_{-i}} CSB^{n-1}(E)$, $g_{i,\cdot}(t_i')\left[\tau_{-i}'(s_{-i})\right] = f_i(t_i)\,[s_{-i}|S_{-i}(\cdot)]$. Moreover, notice that $g_i(t_i')$ strongly believes $CSB^{n-1}(E)$, so that $\omega_i' \in proj_{\Omega_i} SB(CSB^{n-1}(E))$. Hence $\omega_i' \in proj_{\Omega_i} CSB^n(E)$.

2. Take any $\omega \in CSB^n(E)$. Since $\omega \in CSB^{n-1}(E)$, for every $i \in I$, $f_i(t_i) \in \Delta_i \bigcap \Delta_i^G$ and satisfies requirements 2 (up to n). Moreover, since $\omega \in SB(CSB^{n-1}(E))$, $g_i(t_i)$ strongly believes$CSB^{n-1}(E)$, hence $f_i(t_i)$ strongly believes $proj_{S_{-i}} CSB^{n-1}(E)$. Hence $f_i(t_i)$ satisfies also requirement 2 up to $n+1$.

Finally I prove that the lemma holds with ∞ in place of n.

1. First, observe that by the first part of the lemma, if $S_G^{\infty} \neq \emptyset$ then $CSB^n(E) \neq \emptyset$ for every $n \geq 0$, so that the family of nested, non-empty closed sets $\{CSB^n(E)\}_{n \geq 0}$ has the finite intersection property. Hence, being $E = CSB^0(E)$ a compact set, $CSB^{\infty}(E) \neq \emptyset$. Take any $i \in I$ and any $\mu_i \in \Delta_i \bigcap \Delta_i^G$ satisfying requirements 2

[45] When E is a cartesian product, also $SB(E)$ is a cartesian product. Then, if I can pick a ω in their intersection, meaning that it is non-empty, and ω_i' belongs to both projections, there is a ω' in the intersection whose projection is ω_i'.

(for every $m \geq 0$). Take a $\omega_i = (s_i, t_i) \in E$ such that $f_i(t_i) = \mu_i$. Thanks to the non-emptiness of $CSB^{\infty}(E)$, I can take a $\tau'_{-i} : s'_{-i} \in proj_{S_{-i}}E \longmapsto (s'_{-i}, t_{-i}) \in E_{-i}$ such that for every $m \geq 0$ and $s'_{-i} \in proj_{S_{-i}}CSB^m(E)$, $\tau'_{-i}(s'_{-i}) \in proj_{\Omega_{-i}}CSB^m(E)$. By the completeness property of E, there exists a $\omega'_i = (s_i, t'_i) \in E_i$ such that $f_i(t'_i) = f_i(t_i)$ and for every $s_{-i} \in proj_{S_{-i}}E$, $g_{i,\cdot}(t'_i)\left[\tau'_{-i}(s_{-i})\right] = f_i(t_i)\left[s_{-i}|S_{-i}(\cdot)\right]$. Notice that by construction and by the lemma holding for every $m \geq 0$, t'_i strongly believes $CSB^m(E)$. Hence, for every $m \geq 0$, $\omega'_i \in proj_{\Omega_i}SB_i(CSB^m(E))$, so $\omega'_i \in proj_{\Omega_i}CSB^m(E)$. Hence $\omega'_i \in proj_{\Omega_i}CSB^{\infty}(E)$ and $f_i(t'_i) = \mu_i$.

2. Take any $\omega \in CSB^{\infty}(E)$. Since for every $m \geq 0$, $\omega \in CSB^m(E)$, for every $i \in I$, by the lemma holding for every $m \geq 0$, $f_i(t_i) \in \Delta_i \bigcap \Delta_i^G$ and satisfies requirements 2 (for every $m \geq 0$). ■

Proof of theorem 7.

For every $i \in I$, set as Δ_i^G the set of CPS satisfying requirement 3 of the selective rationalizability procedure. Δ_i^G is compact. To apply the lemma for $E = [\Delta] \bigcap CSB^{\infty}(R)$, I need to show that such E satisfies the hypotheses of the lemma.

To show that the compact set $[\Delta] \bigcap CSB^{\infty}(R)$ has the completeness property, it is enough to show that $CSB^{\infty}(R)$ has the completeness property, because the intesection with $[\Delta]$ selects the types only according to first-order-beliefs about strategies. Therefore, for every $(s_i, t_i) \in proj_{\Omega_i}([\Delta] \bigcap CSB^{\infty}(R))$, all the (s_i, t'_i) verifying the definition of completeness for $CSB^{\infty}(R)$ also belong to $proj_{\Omega_i}([\Delta] \bigcap CSB^{\infty}(R))$, because $f_i(t_i) = f_i(t'_i)$.

In turn, to apply the lemma for $E = R$ and prove that $CSB^{\infty}(R)$ has the completeness property, I need to show that the compact set R has the completeness property.

Take any $i \in I$, $(s_i, t_i) \in R_i$ and $\tau_{-i} : s'_{-i} \in proj_{S_{-i}}R \longmapsto (s'_{-i}, t_{-i}) \in R_{-i}$. If there exists a CPS $\nu_i \in \Delta^{H_i}(S_{-i} \times T_{-i})$ such that for every $s_{-i} \in proj_{S_{-i}}R$, $\nu_{i,\cdot}\left[\tau_{-i}(s_{-i})\right] = f_i(t_i)\left[s_{-i}|S_{-i}(\cdot)\right]$ and $marg_{S_{-i}}\nu_i = f_i(t_i)$, I am done, because by completeness of the type space g_i is onto, hence there exists a type t'_i such that $g_i(t'_i) = \nu_i$, and $s_i \in \rho_i(marg_{S_{-i}}\nu_i)$, hence $(s_i, t'_i) \in R_i$. Define the candidate CPS by extending arbitrarily τ_{-i} to $\tau'_{-i} : s'_{-i} \in S_{-i} \longmapsto (s'_{-i}, t_{-i}) \in \Omega_{-i}$ in such a way that for every $s'_{-i} \in proj_{S_{-i}}R$, $\tau'_{-i}(s'_{-i}) = \tau_{-i}(s_{-i})$ and setting for every $s_{-i} \in S_{-i}$, $\nu_{i,\cdot}\left[\tau'_{-i}(s_{-i})\right] =$

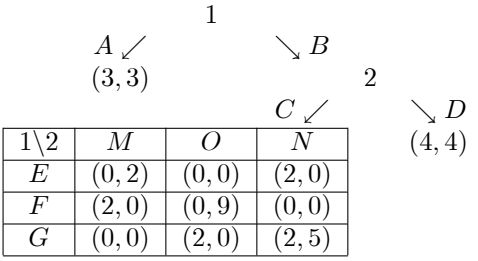
$f_i(t_i)\,[s_{-i}|S_{-i}(\cdot)]$ and extending the assignments by additivity. Now notice that τ'_{-i} yields an embedding of $\bigcup_{h\in H_i} supp f_i(t_i)[\cdot|S_{-i}(h)]$ in $S_{-i}\times T_{-i}$, so that f_i being a CPS implies that ν_i is a CPS too.

To show that for every $i \in I$ and $\mu_i \in \Delta_i \bigcap \Delta_i^G$, there exists a $\omega = (s,t) \in [\Delta] \bigcap CSB^\infty(R)$ such that $f_i(t_i) = \mu_i$, it is enough to show that the same holds for a $\omega \in CSB^\infty(R)$, because any such ω also belongs to $[\Delta]$ by $f_i(t_i) = \mu_i \in \Delta_i$.

To show that for every $\omega = (s,t) \in [\Delta] \bigcap CSB^\infty(R)$ and $i \in I$, $f_i(t_i) \in \Delta_i \bigcap \Delta_i^G$, it is enough to show that for every $\omega = (s,t) \in CSB^\infty(R)$ and $i \in I$, $f_i(t_i) \in \Delta_i^G$.

Notice that applying the lemma for $E = R$ yields both results, because Δ_i^G is the set of CPS satisfying requirement 2 (for every $m \geq 0$) of the strong-rationalizability procedure. ■

Counterexample for the wrong conjecture that emptiness of strong-delta-rationalizability implies emptiness of selective rationalizability (for the same restrictions)



$1\backslash 2$	M	O	N
E	$(0,2)$	$(0,0)$	$(2,0)$
F	$(2,0)$	$(0,9)$	$(0,0)$
G	$(0,0)$	$(2,0)$	$(2,5)$

First-order-beliefs restrictions:
$\Delta_1 := \left\{\mu_1 \in \Delta^{H_1}(S_2) : \mu_1(C \cdot |h^0) = 0\right\}$.
(player 2 promises to play D)
$\Delta_2 := \left\{\mu_2 \in \Delta^{H_2}(S_1) : \forall h \in H_2,\ \mu_2(B.G|h) = 0\right\}$;
(player 1 threatens not to play the pure Nash in the subgame if player 2 breaks the promise)

Strong rationalizability:
$S_1^1 = S_1$, $S_2^1 = \{D.\cdot, C.O, C.N\}$;
$S_1^2 = \{A.\cdot, B.E, B.G\}$, $S_2^2 = S_2^1$;
$S_1^3 = S_1^2$, $S_2^3 = \{D.\cdot, C.N\}$;

$S_1^4 = S_1^3 = S_1^\infty$, $S_2^4 = S_2^3 = S_2^\infty$.

Strong-delta-rationalizability:
$S_{1,\Delta}^1 = \{B.\cdot\}$, $S_{2,\Delta}^1 = \{D.\cdot, C.O\}$;
$S_{1,\Delta}^2 = \{B.G\}$, $S_{2,\Delta}^2 = S_{2,\Delta}^1$.
$S_{1,\Delta}^3 = S_{1,\Delta}^2$, $S_{2,\Delta}^3 = \emptyset$.

Selective rationalizability:
$S_{1,R\Delta}^1 = \{B.E, B.G\}$, $S_{2,R\Delta}^1 =\{D.\cdot\}$;
$S_{1,R\Delta}^2 = S_{1,R\Delta}^1 = S_{1,R\Delta}^\infty$, $S_{2,R\Delta}^2 = S_{2,R\Delta}^1 = S_{2,R\Delta}^\infty$.

Part III
Common assumption of cautious rationality and iterated admissibility

12 Introduction

In the huge variety of solution concepts for complete-information strategic-form games, iterated admissibility, i.e. iterated deletion of weakly dominated strategies, is surely one of the most appealing. First, it is a decision criterion that does not rely on any pre-existing equilibrium motivation: players can perform it from scratch through nothing else than their strategic reasoning. Second, it reflects an intuitevely reasonable way to behave: to the minimum, it avoids choosing a strategy when there is another one that, when it makes a difference, can only do better.[46] Still, it has to be identified more precisely when iterated admissibility is actually the appropriate solution concept and why, more generally, it is a sound way for players to choose their strategies.

The first step to this end is detecting which kind of conjectures and optimality concept motivate players to avoid strategies that are weakly dominated in some reduced game along the procedure. The following game,[47] where strategy L is eliminated in the first round and strategy B

[46] Moreover, in case the strategic form is derived from an extensive form game without relevant ties among payoffs, iterated admissibility operationalizes extensive form rationalizability ([37] and [11]). Yet, the analysis of the extensive form solution concept is required to understand the epistemic motivations: see [11].

[47] This example (with one strategy added) is due to Pierpaolo Battigalli.

is eliminated in the second, is helpful to follow the next few arguments.

$1\backslash 2$	L	C	R
U	$(4,1)$	$(4,1)$	$(0,1)$
M	$(0,1)$	$(0,1)$	$(4,1)$
D	$(3,1)$	$(2,1)$	$(2,1)$
B	$(9,0)$	$(0,1)$	$(0,1)$

It is known from Pearce [37] that a strategy is not weakly dominated if and only if it is a best reply to some fully mixed conjecture over opponents' strategies. But notice that, differently than the iterated deletion of strongly dominated strategies, iterated admissibility can exclude strategies that are not weakly dominated in the final reduced game (L). To justify this, a player must still consider the possibility that some opponent might play some previously deleted strategy. But to what extent? If previously deleted strategies could be given a positive probability, a player would clearly run the opposite risk of rescuing strategies that are weakly dominated in the final reduced game (B). This tension is solved by lexicographic conjectures and lexicographic best replies.[48] A lexicographic conjecture is a finite list of simple conjectures in a priority ordering. They allow to take into consideration previously deleted strategies and yet, pushing them farther in the list, to deem them as infinitely less likely than strategies that survive more steps of the procedure. A lexicographic best reply is a strategy that, for any other strategy, does not worse than the latter against the conjectures of the list up to the end or up to one against which the former does strictly better. Formal definitions will be provided in section 2. Notice that to justify strategy D player 1 must be allowed to hold a lexicographic conjecture with overlapping supports. In the final set, D is a best reply to the simple conjecture that considers C and R equally likely, but it is not a strict best reply. Hence player 1 may wonder about a secondary conjecture to check the desirability of D. If the secondary conjecture were obliged to put probability 1 on the deleted strategy L, strategy D would not be a lexicographic best reply. Instead, considering L and R equally likely as secondary hypothesis makes strategy D a lexicographic best reply.

The second step consists of finding the epistemic hypotheses that identify and conceptually motivate the right lexicographic conjectures, whose

[48] See also Stahl ([40])

lexicographic best replies correspond to the iteratevely admissible strategies. These hypotheses will be defined as notions of cautious rationality, assumption of opponents' cautious rationality, and so on, which characterize players who perform iterated admissibility.

Here comes the contribution of the paper. In section 3, a canonical type space for lexicographic hierarchies of beliefs is constructed. The canonical type space allows players to conceive any meaningful lexicographic hierarchy of beliefs about strategies, so that no exogenous restriction is super-imposed and the states of interest will be entirely identified by the conceptually relevant events. In section 4, compelling notions of assumption, cautiousness and rationality are defined and put at work in this epistemic environment. These notions allow to construct events that not only capture any step of iterated admissibility but can also hold together, defining a cautious rationality and common assumption of cautious rationality non-empty event in which players "share" their being cautious and rational in the sense of this paper.

Brandenburger, Friedenberg and Keisler [18] (henceforth BFK), to whom this work is much indebted, define notions of rationality and assumption that, opportunely combined, deliver the iteratevely admissible strategies. They obtain this result by incorporating in rationality a very strong open-mindedness requirement: players put every state of world in the support of their beliefs, at some level of their lexicographic probability system over the state space. This means that players conceive at the same time every lexicographic hierarchy of beliefs allowed by the type space and consider it possible to some extent. Then, the authors prove the impossibility result that for a rich enough type structure (complete and continuous), players are unable to commonly assume this notion of rationality: the corresponding event is empty. The impossibility ceases to hold for poorer type structures, but this means imposing exogenous restrictions to the hierarchies of beliefs, which could find no justification in the context at hand. Now, suppose that players were able to prove to each other that they are rational in this sense. Then players should assume that everyone is rational; assume that everyone is rational and assumes that everyone is rational; and so on. But if common assumption of rationality is impossible, at some point players must start forming doubts. Why should they? This puzzling result has inspired different papers other than this. Keisler and Lee [31] show that relaxing the continuity hypothesis in the type space, the impossibility may cease to hold. Heifetz, Meier and Schipper [28] take a more radical way out by changing

the solution concept.[49] The aim of this paper, instead, is to epistemically characterize precisely iterated admissibility, for its intuitive appeal, but obtaining a non-empty "cautious rationality and common assumption of cautious rationality" event through interpretationally clear innovations. Switching from open-mindedness to a milder cautiousness requirement allows to preserve the characterization and let players commonly believe in their cautiousness and rationality. The idea is simple and realistic: players cannot or are not interested in conceiving and weighing all possible hierarchies of beliefs at the same time.[50] Cautious players just conceive all possible opponents' strategies, the payoff relevant objects. Then, they make a minimal use of higher-order beliefs[51] to put those strategies in a likelihood order, according to hypotheses about opponents' strategic reasoning. For instance, opponents' strategies that are best replies to some cautious conjecture (i.e. cautiously rational ones) are given priority with respect to the ones that are not. Such reduction of the computational burden for players is strictly connected with their ability to commonly assume this notion of cautious rationality.

13 Iterated admissibility and lexicographic beliefs

For all the following player-specific sets X_i, let $X := \prod_{j \in I} X_j$ and $X_{-i} := \prod_{j \neq i} X_j$.

Consider a finite strategic form game $\langle I, (S_i, u_i)_{i \in I} \rangle$, where I is the set of players and for every $i \in I$, S_i is the set of strategies and $u_i : S \to \mathbb{R}$ is the payoff function. For any finite set X, let $\Delta(X)$ be the set of probability measures on it. Define the expected payoff function π_i on $\Delta(S_i) \times \Delta(S_{-i})$

[49] Also Asheim and Dufwemberg [1] defined a solution concept (fully admissible sets) that captures a form of cautiousness and full belief in rationality and that does not refine, nor is refined, by iterated admissibility.

[50] This is different than impoverishing the type structure: players can still conceive all the meaningful hierarchies of beliefs, simply they will not be obliged to.

[51] I do not rule out in any way that players can put more than necessary or even all hierarchies of beliefs in their conjectures. But the possibility to make a parsimonious use of them is enough to allow common assumption of cautious rationality.

by setting for every $(\sigma_i, \sigma_{-i}) \in \Delta(S_i) \times \Delta(S_{-i})$,

$$\pi_i(\sigma_i, \sigma_{-i}) := \sum_{s_{-i} \in supp\sigma_{-i}} \sum_{s_i \in supp\sigma_i} u_i(s_i, s_{-i})\sigma_i(s_i)\sigma_{-i}(s_{-i}).$$

A pure strategy or a pure opponents' subprofile of strategies as argument of π_i will indicate the probability distribution putting probability 1 on it.

Iterated admissibility is a reduction procedure of the set of strategy profiles that relies on a weak dominance criterion.

Definition 19 *For every player $i \in I$, take a set $\widehat{S}_i \subseteq S_i$. For every strategy $s_i \in \widehat{S}_i$, s_i is weakly dominated over $\widehat{S}$ if there exists $\sigma_i \in \Delta(\widehat{S}_i)$ such that for every $s_{-i} \in \widehat{S}_{-i}$, $\pi_i(s_i, s_{-i}) \leq \pi_i(\sigma_i, s_{-i})$ and there exists $\widehat{s}_{-i} \in \widehat{S}_{-i}$ such that $\pi_i(s_i, \widehat{s}_{-i}) < \pi_i(\sigma_i, \widehat{s}_{-i})$.*

Now iterated admissibility can be defined formally.

Definition 20 *The iterated admissibility procedure is a finite chain of cartesian sets of strategy profiles $S^0 := \prod_{i \in I} S_i^0 \supset ... \supset S^M := \prod_{i \in I} S_i^M$ such that for every $i \in I$ and $s_i \in S_i$:*

1. *$S_i^0 = S_i$;*
2. *for every $n < M$, $s_i \in S_i^{n+1}$ if and only if $s_i \in S_i^n$ and s_i is not weakly dominated over S^n;*
3. *$s_i \in S_i^M$ if and only if s_i is not weakly dominated over S^M.*

Notice that inclusions are strict: then, the chain is finite because the sets of strategies are finite. Moreover, S^M is non-empty because for a player there is always at least one strategy that is not weakly dominated.

When a strategy is not weakly dominated over a set, there exists a fully mixed conjecture over opponents' subprofiles in the set against which the strategy is a best reply.

Proposition 10 *Consider a cartesian set of strategy profiles $\widehat{S} \subseteq S$. For every $i \in I$ and $s_i \in \widehat{S}_i$, if s_i is not weakly dominated over $\widehat{S}$, then there exists $\sigma_{-i} \in \Delta(\widehat{S}_{-i})$ such that for every $s_{-i} \in \widehat{S}_{-i}$, $\sigma_{-i}(s_{-i}) > 0$ and for every $\widehat{s}_i \in \widehat{S}_i$, $\pi_i(s_i, \sigma_{-i}) \geq \pi_i(\widehat{s}_i, \sigma_{-i})$.*

As already argued, looking only at simple fully mixed conjectures may wrongly justify the choice of an iteratevely inadmissible strategy: for a player $i \in I$ there may be strategies that are not weakly dominated over S^M and yet do not belong to S_i^M. The reason is that a player who performs iterated admissibility wants to avoid also strategies that are weakly dominated over some previous set of the chain. Thus, she considers every opponents' subprofile in that set still possible to some extent, but the ones that do not survive the following step are not considered nearly as likely as the ones that do. Therefore, the epistemic characterization will need lists of conjectures that allow to put the states of the world at uncomparable levels of likelihood. These lists are defined here as *lexicographic beliefs.*

Definition 21 *Consider a measurable space X and let $\Delta(X)$ denote the space of probability measures on its Borel field. A lexicographic belief is a finite list $\lambda = (\lambda_1, ..., \lambda_k) \in (\Delta(X))^k$ of such probability measures.*

I will denote by $\Delta^{LEX}(X) := \bigcup_{k \in \mathbb{N}} (\Delta(X))^k$ the set of all lexicographic beliefs over X.

When X is the space of opponents' strategy subprofiles, I will call the lexicographic beliefs *lexicographic conjectures.* As argued in the introduction, I am interested in lexicographic conjectures with possibly overlapping supports, i.e. where there can exist $n \neq m$ such that $supp\lambda_n \bigcap supp\lambda_m \neq \emptyset$. With respect to lexicographic conjectures, I take the standard definition of *lexicographic best reply.*

Definition 22 *Consider a player $i \in I$ and a lexicographic conjecture $\lambda \in \Delta^{LEX}(S_{-i})$. A strategy $s_i \in S_i$ is a lexicographic best reply to $\lambda = (\lambda_1, ..., \lambda_k)$ if for every $s_i' \neq s_i$, there exists $j \leq k$ such that for every $h \leq j$, $\pi_i(s_i, \lambda_h) \geq \pi_i(s_i', \lambda_h)$ and, if $j < k$, $\pi_i(s_i, \lambda_j) > \pi_i(s_i', \lambda_j)$.*

Instead, when X will be the section with respect to opponents of the state space (cross product of the strategy space and of the type space I will construct), I will be interested in lexicographic beliefs with nonoverlapping measures.[52] Such lexicographic beliefs represent a list of mutually exclusive hypotheses about the state of the world: the primary hypothesis, the secondary hypothesis, and so on. This will not prevent marginal lexicographic beliefs on strategies to have overlapping supports; it will just

[52] This is actually a slightly weaker requirement than nonoverlapping supports when the underlying space is infinite.

require the belief in the same opponents' strategy subprofile in two different hypotheses to be motivated by two different states of the world. As in BFK, this property is called *mutual singularity* and the lexicographic beliefs that satisfy it are called *lexicographic probability systems.*[53]

Definition 23 *Consider a measurable space X. A lexicographic belief $\lambda = (\lambda_1, ..., \lambda_k) \in \Delta^{LEX}(X)$ is mutually singular if there are measurable sets $E_1, ..., E_k$ in X such that for every $j \leq k$ and $h \neq j$, $\lambda_j(E_j) = 1$ and $\lambda_j(E_h) = 0$. A mutually singular lexicographic belief is called lexicographic probability system.*

I will denote by $\Delta^{LPS}(X) \subset \Delta^{LEX}(X)$ the set of all lexicographic probability systems (henceforth, LPS) over X.

Lexicographic hierachies of beliefs about strategies will be defined in the next section, where they are used to construct the type space that captures them.

14 A canonical type space for lexicographic hierarchies of beliefs

Here I construct a canonical type space for lexicographic hierachies of beliefs.

I will metrize spaces as follows:

- S_i with the discrete metric;
- $\Delta(X)$, where X is a separable complete metric space (Polish), with the Prohorov metric;
- $\Delta^{LEX}(X)$, where X is a Polish space, by setting the distance between two elements of the same lenght $\lambda = (\lambda_1, ..., \lambda_k)$ and $\widehat{\lambda} = (\widehat{\lambda}_1, ..., \widehat{\lambda}_k)$ as the maximum over $h \leq k$ of the Prohorov distances between λ_h and $\widehat{\lambda}_h$, and the distance between two elements of different lenghts to 1;[54]

[53] The term was coined by Blume, Brandenburger and Dekel [14] with reference also to lists of overlapping measures.

[54] The Prohorov distance between two elements is at most 1, so triangular inequality is respected.

- the product of Polish spaces with the product metric.

With these choices, all the spaces are Polish themselves (see [22]).
For every $i \in I$ and $n \in \mathbb{N}$, define inductively the following sets:

$$\begin{aligned} X_i^1 &: = S_{-i}; \\ X_i^{n+1} &: = X_i^n \times \Delta^{LEX}(X_{-i}^n). \end{aligned}$$

Moreover, define

$$Z_i^1 := X_i^1,\ Z_i^{n+1} := \Delta^{LEX}(X_{-i}^n);$$

then $X_i^n = \prod_{m=1}^{n} Z_i^m$.

Now I can define a *(coherent) lexicographic hierarchy of beliefs about strategies.*

Definition 24 *A (coherent) lexicographic hierarchy of beliefs about strategies is a finite list* $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)$ *such that for every* $h \leq k$, $\delta_h = (\delta^1, \delta^2, ...) \in \prod_{n\in\mathbb{N}} \Delta(X_i^n)$ *(and for every* $n \in \mathbb{N}$, $\text{marg}_{X_i^n}\delta^{n+1} = \delta^n$*)*

Since all the sets previously defined are Polish spaces, the following version of lemma 1 in Brandenburger and Dekel [16] holds.[55]

Lemma 6 *Let* $D_i := \left\{\delta = (\delta^1, \delta^2, ...) \in \prod_{n\in\mathbb{N}} \Delta(X_i^n) : \text{marg}_{X_i^n}\delta^{n+1} = \delta^n\right\}$. *There exists a unique function* $f_i : D_i \to \Delta(\prod_{n\in\mathbb{N}} Z_i^n)$ *such that for every* $\delta \in D_i$ *and* $h \in \mathbb{N}$, $\text{marg}_{X_i^h} f_i(\delta) = \delta^h$. *Moreover,* f_i *is a homeomorphism.*

Proof. See [16].

Define the set of coherent lexicographic hierarchies of beliefs $C_i := \bigcup_{k\in\mathbb{N}} (D_i)^k$ and metrize it by setting the distance between two elements of the same lenght $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)$ and $\boldsymbol{\delta'} = (\delta'_1, ..., \delta'_k)$ as the maximum over

[55] In [16] the lemma only claims the existence of the homeomorphism because it suffices for the purposes of the paper. However, their proof constructs exactly the homeomorphism specified here through a version of Kolmogorov Existence Theorem (from [22]), which also claims the uniqueness of the images with respect to the marginals requirement (hence the uniqueness of the function with this feature).

$h \leq k$ of the distances between δ_h and δ'_h, and the distance between two elements of different lenghts to 1. Then the function

$$g_i : C_i \to \Delta^{LEX}(\prod_{n\in\mathbb{N}} Z_i^n) \text{ such that } g_i(\boldsymbol{\delta} = (\delta_1, ..., \delta_k)) := (f_i(\delta_1), ..., f_i(\delta_k)).$$

is a homeomorphism.[56]

Clearly $\prod_{n\in\mathbb{N}} Z_i^n$ is a strict superset of $S_{-i} \times C_{-i}$ because $\prod_{n>1} Z_i^n$ contains also non coherent hierarchies. Moreover I want to achieve mutual singularity in the final type space. The following inductive procedure allows to restrict the sets in the desired way and close the final type space.

Define:

- $\Lambda_i^0 := \left\{ \boldsymbol{\delta} \in C_i : g_i(\boldsymbol{\delta}) \in \Delta^{LPS}(\prod_{n\in\mathbb{N}} Z_i^n) \right\}$;
- $\Lambda_i^n := \left\{ \boldsymbol{\delta} = (\delta_1, ..., \delta_k) \in \Lambda_i^{n-1} : \forall h = 1, ..., k,\ f_i(\delta_h)[S_{-i} \times \Lambda_{-i}^{n-1}] = 1 \right\}$;
- $\Lambda_i := \bigcap_{n\in\mathbb{N}} \Lambda_i^n$.

I have to show that for every $n \in \mathbb{N}$, Λ_i^n is well defined, that is, $S_{-i} \times \Lambda_{-i}^{n-1}$ is measurable.

By corollary C.1 in BFK, for every Polish space X, $\Delta^{LPS}(X)$ is a Borel set in $\Delta^{LEX}(X)$. The function g_i is measurable. Hence, Λ_i^0 is a Borel set in C_i and Λ_{-i}^0 is a Borel set in C_{-i}.

By theorem 17.24 in [30], for every Polish space X, the Borel sigma-algebra on $\Delta(X)$ generated by the Prohorov metric is generated also by the family of maps $\mu \mapsto \mu(A)$ with $\mu \in \Delta(X)$ and Borel set $A \subseteq X$. This requires that for every Borel set $W \subseteq X$, the set $\left\{ \delta \in \Delta(X) : \delta(W^C) > 0 \right\}$ is Borel, hence its complement $\left\{ \delta \in \Delta(X) : \delta(W) = 1 \right\}$ is Borel too. For every lenght $k \in \mathbb{N}$ and $h \leq k$, the projection function $\lambda = (\lambda_1, ..., \lambda_k) \mapsto \lambda_h$ with $\lambda \in (\Delta(X))^k$ is continuous, hence since $\left\{ \delta \in \Delta(X) : \delta(W) = 1 \right\}$ is Borel,

$$L_h^k := \left\{ \lambda = (\lambda_1, ..., \lambda_k) \ \in (\Delta(X))^k : \lambda_h(W) = 1 \right\}$$

[56] For any $\boldsymbol{\delta} = (\delta_1, ..., \delta_k) \in C_i$, take a ball around $g_i(\boldsymbol{\delta}) = (f_i(\delta_1), ..., f_i(\delta_k))$ of radius ρ. Since f_i is a homeomorphism, for every $h \leq k$ and for the ball around $f_i(\delta_h)$ of radius ρ, there is a ball around δ_h whose image is contained in the previous ball. Take the smallest radius ε among those balls around δ_h over $h \leq k$. The image of the ball around $\boldsymbol{\delta}$ of radius ε is contained in the ball of $g_i(\boldsymbol{\delta})$ of radius ρ. The same reasoning can be applied inverting g_i.

is Borel too.

$$L^k := \left\{\lambda = (\lambda_1, ..., \lambda_k) \in (\Delta(X))^k : \forall h = 1, ..., k,\ \lambda_h(W) = 1\right\} = \bigcap_{h=1,...,k} L_h^k,$$

so it is Borel too.

$$L := \left\{\lambda = (\lambda_1, ..., \lambda_l) \in \Delta^{LEX}(X) : \forall h = 1, ..., l,\ \lambda_h(W) = 1\right\} = \bigcup_{k\in\mathbb{N}} L^k,$$

so it is Borel too. Setting $X := \prod_{n\in\mathbb{N}} Z_i^n$, and $W := S_{-i} \times \Lambda_{-i}^{n-1}$, $\Lambda_i^n = g_i^{-1}(L) \bigcap \Lambda_i^{n-1}$, so it is Borel. Hence, Λ_{-i}^n is a Borel set in C_{-i}.

Now consider that:

- Λ_i is homeomorphic to $g_i(\Lambda_i)$;
- $g_i(\Lambda_i) = \left\{(\lambda_1, ..., \lambda_k) \in \Delta^{LPS}(\prod_{n\in\mathbb{N}} Z_i^n) : \forall h = 1, ..., k, \lambda_k[S_{-i} \times \Lambda_{-i}] = 1\right\}$, because g_i is onto;
- the latter is homeomorphic to $\Delta^{LPS}(S_{-i} \times \Lambda_{-i})$.

The last homeomorphism is the function that preserves the measures of all sets. Redefine g_i as the composition of itself with this last homeomorphism. So g_i is now a homeomorphism between Λ_i and $\Delta^{LPS}(S_{-i} \times \Lambda_{-i})$ such that for every $\delta \in \Lambda_i$ and for every $h \in \mathbb{N}$, $marg_{X_i^h} g_i(\delta) = \delta^h$. This closes the canonical type space for LPS ($(T_i, g_i)_{i\in I}$ from now on).

All hierarchies in Λ_i are *collectively coherent* (they are coherent, believe that opponents are coherent, and so on); moreover, they display common certainty in mutual singularity. The type space is canonical in the sense that it represents all hierarchies of this kind. Notice that the common certainty in mutual singularity does not mean that the lexicographic hierarchies are composed by mutually singular beliefs of all orders Indeed, beliefs of all first n orders could be even identical at different likelihood levels.

Heifetz, Meier and Schipper [28] construct a canonical type space for LPS with a bottom-up procedure, i.e. building directly only the desired hierarchies and putting them together in the type space. Since they introduce an epistemic hypothesis of mutual singularity of conjectures over opponents' strategies, they obtain mutual singularity in the final type space automatically. The top-down procedure here, instead, allows to

throw away only those hierarchies whose representation as lexicographic beliefs over the state space is not mutually singular. Our construction is therefore bigger and the represented hierarchies can be composed by overlapping beliefs for any finite number of orders.

15 Common assumption of cautious rationality and the characterization theorem

In the canonical type space just constructed, the goal is now to identify the conceptually meaningful events that imply iteratevely admissible strategies as behavioral projections. These events will be the result of clear and realistic hypotheses about players' strategic reasoning, which allow to estabilish under which conditions iterated admissibility is the appropriate solution concept.

The first event of interest is the rationality one and it is based on the hypothesis that players play lexicographic best replies to their lexicographic conjectures.

Definition 25 *Rationality is the event* $R := \prod_{i \in I} R_i \subset S \times T$ *such that for every* $i \in I$ *and* $\omega_i = (s_i, t_i) \in R_i$, s_i *is a lexicographic best reply to* $\text{marg}_{S_{-i}} g_i(t_i) = (\lambda_1, ..., \lambda_k)$.

The second event of interest is the cautiousness one and it is based on the hypothesis that players' lexicographic conjectures deem all the opponents' strategy subprofiles as possible to some extent.

Definition 26 *Cautiousness is the event* $C := \prod_{i \in I} C_i \subset S \times T$ *such that for every* $i \in I$ *and* $\omega_i = (s_i, t_i) \in C_i$, $\text{marg}_{S_{-i}} g_i(t_i) = (\lambda_1, ..., \lambda_k)$ *has the following property: for every* $s_{-i} \in S_{-i}$ *there exists* $j \leq k$ *such that* $\lambda_k(s_{-i}) > 0$.

The conjunction of the two is the cautious rationality event. It is the one that translates into the use of a weak dominance criterion.

Definition 27 *Cautious rationality is the event* $R^1 := \prod_{i \in I} (R_i \bigcap C_i) = R \bigcap C$.

The projection on the strategy space of the event cautious rationality will coincide with the first iteration of the iterated admissibility procedure, i.e. with non weakly dominated strategies. To capture the further iterations, I need to identify the events where conjectures give the right priority to the iteratevely admissible strategies, in terms of their likelihood. These events are based on the hypothesis that players hold a kind of belief in opponents' cautious rationality up to some order. This kind of belief in an event (such as cautious rationality) shall not necessarily rule out completely that the event does not occur. This concept is defined here as *assumption*.

Definition 28 *A LPS* $\lambda_i = (\lambda_1, ..., \lambda_k) \in (\Delta(S_{-i} \times T_{-i}))^k$ *assumes* $B \subset S_{-i} \times T_{-i}$ *(at level* h*) if there exists* $h \leq k$ *such that:*

1. *there are measurable sets* $E_1, ..., E_k$ *in* $S_{-i} \times T_{-i}$ *such that for every* $j \leq h$, $E_j \subseteq B$ *and* $\lambda_j(E_j) = 1$ *and for every* $j > h$, $E_j \bigcap B = \emptyset$ *and* $\lambda_j(E_j) = 1$.

2. *for every* $s_{-i} \in \text{proj}_{S_{-i}} B$, *there exists* $j \leq h$ *such that* $\text{marg}_{S_{-i}} \lambda_j[s_{-i}] > 0$.

The first requirement has the interpretation that players deem the event *infinitely more likely* than its complementary. The definition of this concept in BFK is different because it applies again to open-minded (i.e. full support) LPS only, and has been given a preference-based representation.[57] The second requirement means that players consider every possible behavioral implication of the event infinitely more likely than all other strategy subprofiles. This reflects the view that players, as argued in the introduction, are concerned about conceving all possible moves by the opponents, while using higher-order beliefs only to rank them in a likelihood order. However, point 2 holds also in BFK, although in their model there is no need to specify it in the definition of assumption, because it is already a consequence of open-mindedness.

With this notion of assumption, the corresponding operator that maps subsets of the state space into subsets of the state space can be defined as follows:

$$A_i(B) := \{\omega_i = (s_i, t_i) \in (S_i \times T_i) : g_i(t_i) \text{ assumes } B\}.$$

[57] It would be interesting to check how different an axiomatic treatment of the definition here should be.

Using the cautious rationality event and the last operator, the right *cautious rationality and m–th order assumption of rationality* events and the *cautious rationality and common assumption of rationality* event can be defined inductively as follows:

$$\begin{aligned} \forall m \quad &\geq \quad 1,\ R^{m+1} := R^m \bigcap (\prod_{i \in I} A_i(R^m_{-i})); \\ R^{\infty} \quad &: \quad = \bigcap_{m \in \mathbb{N}} R^m. \end{aligned}$$

The behavioral implications of the first events correspond step-by-step to the iteratevely admissible strategy profiles. The second event is non-empty too and its behavioral implications coincide with the final set of the iterated admissibility procedure. These facts are summarized in the following characterization theorem.

Theorem 8 *For every* $n \geq 0$, $S^n = \text{proj}_S R^n$. *Moreover,* $S^M = \text{proj}_S R^{\infty}$.

Proof.

For every $n \leq M$, $i \in I$ and $s_i \in S_i^n$, take a $\mu_i^n(s_i) \in \Delta(S_{-i}^{n-1})$ such that $supp\mu_i^n(s_i) = S_{-i}^{n-1}$ and for every $s_i' \in S_i$, $\pi_i(s_i, \mu_i^n(s_i)) \geq \pi_i(s_i', \mu_i^n(s_i))$ (it exists by proposition 3). Moreover, for every $i \in I$ and $s_i \in S_i^M$, take a $\mu_i^{M+1}(s_i) \in \Delta(S_{-i}^M)$ such that $supp\mu_i^{M+1}(s_i) = S_{-i}^M$ and for every $s_i' \in S_i$, $\pi_i(s_i, \mu_i^{M+1}(s_i)) \geq \pi_i(s_i', \mu_i^{M+1}(s_i))$.

For every $k \leq M$, define the types $U_i^k := \bigcup_{s_i \in S_i^k} (s_i \times k)$ and set $U_i := \bigcup_{0 \leq k \leq M} U_i^k$.

For every $i \in I$, define $h_i : U_i \to \Delta^{LPS}(S_{-i} \times U_{-i})$ with the following procedure:

- for every $s_i \in S_i$, take a $\lambda \in \Delta(S_{-i} \times U_{-i})$ such that $supp(marg_{S_{-i}}\lambda) \neq S_{-i}$ and let $h_i((s_i, 0)) := \lambda$;
- for every $0 < k < M$ and $s_i \in S_i^k$, take the $\lambda \in \Delta(S_{-i} \times U_{-i}^{k-1})$ such that for every $s_{-i} \in S_{-i}^{k-1}$, $\lambda[(s_{-i}, (s_{-i}, k-1))] = \mu_i^k(s_i)[s_{-i}]$ and let $h_i((s_i, k)) := (\lambda, h_i((s_i, k-1)))$;

- for every $s_i \in S_i^M$, take the $\lambda_2 \in \Delta(S_{-i} \times U_{-i}^{M-1})$ such that for every $s_{-i} \in S_{-i}^{M-1}$, $\lambda_2[(s_{-i},(s_{-i},M-1))] = \mu_i^M(s_i)[s_{-i}]$ and take the $\lambda_1 \in \Delta(S_{-i} \times U_{-i}^M)$ such that for every $s_{-i} \in S_{-i}^M$, $\lambda_1[(s_{-i},(s_{-i},M))] = \mu_i^{M+1}(s_i)[s_{-i}]$ and let $h_i((s_i,M)) := (\lambda_1, \lambda_2, h_i((s_i,M-1)))$.

For every $j \in I$ and $u_j \in U_j$, take the lexicographic hierarchy of beliefs $\boldsymbol{\delta}_j(u_j) = (\delta_1, ..., \delta_k)$ induced by $h_j(u_j)$ in the finite type space $(U_i, h_i)_{i \in I}$ and rename $\boldsymbol{\delta}_j(u_j)$ as u_j in C_j (see section 3). Now by the definition of g_j, it must be $g_j(u_j) = h_j(u_j)$ because in such case it is true that for every $l \leq k$ and for every $h \in \mathbb{N}$, $marg_{X_l^h} f_j(\delta_l) = \delta_l^h$ and by lemma 6 there is only one function satisfying this property. Moreover, $g_j(u_j)$ is mutually singular and $\boldsymbol{\delta}_j(u_j)$ is collectively coherent, so it survives all steps of the reduction of the type space and finally $u_j \in T_j$.

Define $m(s_i) := \max\{n \in \mathbb{N} : s_i \in S_i^n\}$. Clearly, for every $i \in I$ and every $s_i \in S_i^1$, $(s_i,(s_i,m(s_i))) \in R_i^1$. By induction, it is immediate to show that $g_i((s_i,m(s_i)))$ assumes $R_{-i}^{m(s_i)-1}, ..., R_{-i}^1$. So it holds that for every $n \leq M$, $S^n \subset proj_S R^n$.

Moreover, notice that for every $i \in I$ and $s_i \in S_i^M$, $g_i((s_i,M))$ assumes also R_{-i}^M, so that $(s_i,(s_i,M)) \in R_i^{M+1}$. But then by induction it is immediate to show that for every $n \in \mathbb{N}$, $g_i((s_i,M))$ assumes R_{-i}^n. So, $S^M \subseteq proj_S R^\infty$.

For the opposite inclusion, take as inductive hypothesis that $S^n \supseteq proj_S R^n$.

Setting $R^0 := S \times T$, it is trivially verified for $n = 0$.

Take any $\omega = (s,t) \in R^{n+1} \subseteq R^n$. Notice that s_i is a lexicographic best reply to the lexicographic conjecture $marg_{S_{-i}} g_i(t_i) = (\nu_1, ..., \nu_k)$, where $g_i(t_i)$ assumes R^n at some level $l \leq k$. By the inductive hypothesis, for every $i \in I$, $s_i \in S_i^n$. Hence it is enough to show that there exists a measure $\sigma_{-i} \in \Delta(S_{-i}^n)$, with $supp\sigma_{-i} = S_{-i}^n$, such that $\pi_i(s_i,\sigma_{-i}) \geq \pi_i(s_i',\sigma_{-i})$ for every $s_i' \in S_i^n$. Any measure $\nu := \alpha_1\nu_1 + ... + \alpha_l\nu_l$ such that the sum-1 weights $\alpha_1, ..., \alpha_l$ satisfy $\alpha_{n+1} \cdot l \cdot (\max_{s \in S} \pi_i(s) - \min_{s \in S} \pi_i(s)) < \alpha_n$ works.

Moreover, since $S^M \supseteq proj_S R^M \supseteq proj_S R^\infty$, it holds $S^M \supseteq proj_S R^\infty$.

■

16 Conclusions and further research

Players are expected to play iteratevely admissible strategies when they are cautiously rational, assume opponents are cautiously rational, and so on, where being cautious, rational and assuming an event like opponents' cautious rationality must be carefully defined. A player is rational when she plays a lexicographic best reply to her lexicographic conjecture about opponents' strategies. A player is cautious when she forms the lexicographic conjecture by taking into consideration every opponents' strategy subprofile as possible to some extent. The definition of assumption allows players to consider in their conjectures also the possibility that the event does not occur, but assigning likelihood priority to the event and to all its possible behavioral implications.

BFK characterize iterated admissibility with rationality and assumption of rationality events, but incorporating in rationality an open-mindedness requirement which is stronger than the cautiousness requirement here: players must form conjectures that assign a priority level and a probability weight to (a neighborhood of) every state of world. As a consequence, in every complete and continuous type space, players cannot commonly assume this notion of open-minded rationality since the corrisponding event is empty. This reflects the computational burden required to players.

This impossibility has been eliminated here by weakening the requirement on players' conjectures. Players are allowed to form parsimonious conjectures, whose supports can also be constituted by a finite number of states of the world. But players always care to order and weigh all possible opponents' moves, the payoff-relevant objects. Assuming opponents' rationality, and so on, and associating to strategies higher-order beliefs allows players to put them in a meaningful likelihood order.

As a result, players are able to form conjectures that commonly assume this notion of cautious rationality, also in a rich type space that does not prevent them from coming up with any meaningful lexicographic hierarchy of beliefs about strategies. The existence of such canonical type space has been shown constructively.

The passage from open-mindedness to cautiousness has the further advantage of reducing the complexity of the analysis. The characterization does not depend on the topology of the type space,[58] which could then be

[58] The topological construction of the type space allows to claim the continuity of

costructed as a simple measure-theoretical object, like in [29] for the non-lexicographic case. The analysis could be further simplified by removing the mutual singularity requirement. It has to be noticed that players' lexicographic beliefs of any order are not required to be mutually singular. Hence, removing mutual singularity over states of the world as a whole would not change the interpretation of the epistemic characterization.

Whether the constructed type space is universal or not has not been investigated yet. In the proof of the characterization theorem, a finite type space is mapped into the canonical type space and thus shown to be a belief-closed subset of the latter. If this could be done for any type space, the canonical type space would also be terminal. Hence, a universal type space would exist. On the other hand, it would be interesting to check the existence of a type space for LPS with finite joint support. It is reasonable to think that players will not introduce an infinite dimension to justify the likelihood order they want to give to a finite number of opponents' strategy subprofiles.

However, the epistemic model set up here for the characterization of iterated admissibility can be used for different scopes. For instance, I conjecture that by simply removing the marginal support requirement from the definition of assumption, the same events would characterize the elimination of weakly dominated strategies followed by many rounds of elimination of strongly dominated strategies (which is the appropriate solution concept also under the hypotheses of [23] and [15]).

Finally, the characterization (as in BFK) relies also on strategy-type pairs that are not cautiously rational not because the strategy is not a lexicographic best reply to the conjecture, but because the conjecture does not respect the cautiousness requirement (and the same applies in BFK with the open-mindedness one). This is necessary to assume cautious rationality and form at some level a fully mixed conjecture whenever an opponent has a dominant strategy, which is always rational, and a dominated strategy, which is always irrational. If types without full marginal support on strategies could be put out of the picture, LPS would coincide with Conditional Probability Systems a la Myerson [35] and a unified framework for the epistemic analysis of solution concepts in static and dynamic games could be developed. To avoid the use of such incautious conjectures it is enough to allow that an assumed event, rationality in this case, be given a nonnull probability after the level at which it stops having probability

belief maps and compare the results with BFK's ones. However, the topology is a dispensable object for the characterization result.

one. This requires to rethink the interpretation of assumption as deeming an event infinitely more likely than the complementary.

References

[1] Asheim, G., Dufwenberg, M., "Admissibility and common belief", *Games and Economic Behavior*, **42**, 2003, 208-234.

[2] Aumann, R., "Nash equilibria are not self-enforcing", in: J. J. Gabszewicz, J. F. Richard and L. A. Wolsey, "Economic Decision-Making: Games, Econometrics and Optimization", Elsevier, Amsterdam, 1990.

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

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

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

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

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

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

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

[10] 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*, forthcoming.

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

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

[13] Ben-Porath, E., and E. Dekel, "Signaling Future Actions and the Potential for Sacrifice", *Journal of Economic Theory*, **57**, 1992, 36-51.

[14] Blume, L., Brandenburger, A., Dekel, E., "Lexicographic probabilities and equilibrium refinements", *Econometrica*, **59**, 1991, 81-98.

[15] Borgers, T., "Weak dominance and approximate common knowledge", *Journal of Economic Theory*, **64**, 1994, 265-276.

[16] Brandenburger, A. and E. Dekel, "Hierarchies of beliefs and common knowledge", *Journal of Economic Theory*, **59**, 1993, 189-198.

[17] Brandenburger, A. and A. Friedenberg, "Self-Admissible Sets", *Journal of Economic Theory*, **145**, 2010, 785-811.

[18] Brandenburger, A., Friedenberg, A., Keisler, J., "Admissibility in games", *Econometrica*, **76**, 2008, 307-352.

[19] Catonini, E., "Non-binding agreements and forward induction reasoning", job market paper, 2012.

[20] Catonini, E., "Selecting strongly rationalizable strategies", working paper, 2012.

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

[22] Dellacherie, C., Meyer, P., "Probabilities and potential", *Math. Stud.,* **29**, 1978.

[23] Dekel, E. and Fudenberg, D., "Rational behavior with payoff uncertainty", *Journal of Economic Theory*, **52**, 1990, 243-267.

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

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

[26] Gossner, O., "The robustness of incomplete penal codes in repeated interactions", working paper, 2012.

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

[28] Heifetz, A., Meier, M., Schipper, B., "Comprehensive rationalizability", working paper.

[29] Heifetz, A., Samet, D., "Topology-free typology of beliefs", *Journal of Economic Theory*, **82**, 1998, 324-341.

[30] Kechris, A., "Classical descriptive set theory", Springer, 1995.

[31] Keisler, J., Lee, B., "Common assumption of rationality", working paper.

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

[33] Mertens, J.F., "Stable Equilibria - A Reformulation", *Mathematics of Operations Research*, **14**, 1989, 575-625.

[34] Miller, D., Watson, J., "A Theory of Disagreement in Repeated Games with Bargaining", working paper, 2012.

[35] Myerson, R., "Multistage games with communication", *Econometrica*, **54**, 1986, 323-358.

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

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

[38] Reny, P.J., "Common Belief and the Theory of Games with Perfect Information", *Journal of Economic Theory*, **59**, 1993, 257-274.

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

[40] Stahl, D., "Lexicographic rationalizability and iterated admissibility", *Economic Letters*, **47**, 1995, 155-159.

[41] Tebaldi, P., "Non-Binding Agreements in Dynamic games", MSc Thesis, 2011, Bocconi University.