Self-enforcing agreements and forward induction reasoning*

Emiliano Catonini†

November 2016

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

Keywords: Agreements, Self-Enforceability, Forward Induction, Extensive-Form Rationalizability, Strategic Stability.

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

*A special thanks goes to Pierpaolo Battigalli, this paper would not exist without his mentoring. Thank you also to Adam Brandenburger, Shurojit Chatterji, Yi-Chun Chen, Alfredo Di Tillio, Amanda Friedenberg, Amanda Jakobsson, Atsushi Kajii, Mattia Landoni, Xiao Luo, Elena Manzoni, Andres Perea, Burkhard Schipper, Madhav Shrihari Aney, Satoru Takahashi, Elias Tsakas, and Dimitrios Tsomocos for precious suggestions.

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

1 Introduction

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

I take the view that players will believe in the agreement only if this is compatible with beliefs in rationality[2] and their interaction with the beliefs in the agreement of all orders. Ann will believe in the agreement only if Bob may comply with it in case he is rational, he believes in the agreement, he believes that Ann is rational and believes in the agreement (which may add non-agreed upon restrictions on what Bob expects Ann to do), and so on. Moreover, I take the view that deviations, or more generally past actions, are not interpreted as mistakes but as intentional choices. Suppose that for Bob, in case he is rational and believes in the agreement, some move makes sense only if he plans to play a certain action thereafter. Ann, upon observing such move, will believe that Bob will play that action (and Bob may use the move to signal this). This instance of forward induction reasoning is based not just on the belief in Bob's rationality, but also on its interaction with the belief that Bob believes in the agreement. Example 3 in Section 2 is a case in point. Consider now a move that Bob, if he is rational and believes in the agreement, cannot find profitable whatever he plays thereafter. Example 1 in Section 2 illustrates a situation of this kind. Then Ann cannot keep believing that Bob is rational and, at the same time, that he believes in the agreement. Which belief will she maintain? Given the cheap talk nature of the agreement, I take the view that Ann will keep believing that Bob is rational (if this

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

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

is per se compatible with Bob's behavior). However, in Section 5 I argue that the main insights of the paper go through under the opposite assumption. In addition, if compatible with Bob's rationality, Ann may maintain the belief that Bob believed that she would have not violated the agreement *before him*. In Section 6 I show that the main insights of the paper go through under this additional assumption.

For notational simplicity, the focus is restricted to the class of finite games with complete information, observable actions,[3] and no chance moves. However, the methodology can be applied to all dynamic games with perfect recall and countably many information sets,[4] hence possibly infinite horizon. Which agreements will be believed and complied with? Which outcomes of the game can be achieved through some agreement? To answer these questions, the concepts of *credibility*, *self-enforceability* (of agreements) and *implementability* (of outcomes) are introduced. An agreement is credible if believing in it is compatible with the strategic reasoning hypotheses. A credible agreement is self-enforcing if it induces *only* paths of play which are allowed by the agreement itself. An outcome is implementable if it is the *only* outcome induced by some self-enforcing agreement.

In two-players games, an outcome is implementable if and only if it is induced by a "strict"[5] Nash equilibrium in extensive-form rationalizable strategies (Pearce [26]; Battigalli and Siniscalchi [9]). Thus, standard elimination procedure and fixed point condition provide to the analyst (or to a mediator) the set of outcomes that can be achieved through pre-play coordination (and for each outcome, an agreement that implements it). Subgame perfection is not a necessary condition for implementability. This result may be surprising for two reasons. First, it is obtained under all the possible orders of belief in rationality, also after deviations from the agreed upon path. Second, the literature has always assigned to subgame perfection a dominating role. At the end of Section 6 I will elaborate further on why I find this emphasis misplaced.[6]

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

[4] This limitation allows to use Conditional Probability Systems (see Section 3), which require a countable set of conditioning events.

[5] i.e. without best replies to the equilibrium conjecture which would induce a different outcome: see Section 3.1 for a formal definition.

[6] The relationship between subgame perfection and strategic reasoning in absence of agreements has already been extensively studied for perfect information games (i.e. without simultaneous moves)

In games with more than two players, not all strict Nash equilibria in extensive-form rationalizable strategies induce an implementable outcome: the threats of two players towards a third player may be mutually incompatible. Thus, conditions on the off-the-path subgames are required. To accomplish this task, I define a new, set-valued solution concept in reduced strategies: *Self-Enforcing Set* (SES). SES's can be seen as the counterpart of subgame perfect equilibrium (henceforth, SPE), where the plans of deviators are not exogenously given, but are determined by forward induction. To implement a SES outcome, players can agree on the SES itself. Thus, they do not need to promise (and co-players trust) what they would do after an own violation of the agreement. That SES's are set-valued reflects the incompleteness of the agreement, which may be crucial for the implementation of an outcome: see Example 2 in Section 2.

Sometimes, the implementation of an outcome is possible only if players declare in advance what they would do after a deviation from the path. To fully characterize implementable outcomes, SES's are enriched through the notion of *tight agreement*. Tight agreements only require to verify one-step conditions instead of many steps of reasoning. Moreover, they implement exactly the outcomes they allow. In this sense, tight agreements are *truthful*. Hence, the characterization of implementable outcomes with tight agreements provides a revelation principle for agreements design: players need not be vague about the outcome they want to achieve.

In many contexts, there are limitations to which agreements players can actually reach. On the one hand, players may be unable (or unwilling) to coordinate on a precise outcome.[7] On the other hand, in some contexts it may be natural to agree simply on an outcome to reach, without discussing what to do in case of a deviation. The methodology developed in the paper allows to evaluate agreements with any kind of incompleteness.

---

with no relevant ties. Reny [27] shows that backward and forward induction strategies do not coincide. Nonetheless, Battigalli [4] proves that backward and forward induction yield the same unique outcome. This result is proved also by Heifetz and Perea [19] and by Chen and Micali [13]. The latter show that in all games with perfect recall, forward induction refines backward induction without equilibrium reasoning in terms of outcomes. In a previous work I find an overlapping between forward induction and SPE outcomes in games with observable actions.

[7] For instance, Harrington [18] documents instances of "mutual partial understanding" among firms which leaves the exact path of price increase undetermined to escape antitrust sanctions. Such mutual understanding can be modeled as an incomplete agreement, whose consequences can be studied with the methodology developed in this paper.

This work is greatly indebted to the literature on rationalizability in dynamic games. In this literature, restrictions to first-order beliefs are usually accounted for through *Strong-Δ-Rationalizability* (Battigalli, [6]; Battigalli and Siniscalchi, [10]). Strong-Δ-Rationalizability is based on the hypothesis that players *do not* maintain the belief in the rationality of the co-players when they display behavior which cannot be optimal under their first-order belief restrictions. Battigalli and Prestipino [8] show that Strong-Δ-Rationalizability actually captures transparency of the first-order belief restrictions, i.e. the assumption that all orders of belief in the restrictions always hold in the game. Battigalli and Friedenberg [7] interpret the restrictions as the context in which the game takes place; for instance, a well-established convention.

To characterize the different hypotheses of this paper, another rationalizability procedure with first-order belief restrictions, *Selective Rationalizability*, is constructed and characterized epistemically in [12]. Selective rationalizability captures *common strong belief in rationality* (Battigalli and Siniscalchi [9]), i.e. the assumption that any order of belief in rationality holds as long as not contradicted by the observed behavior. Thus, it combines unconstrained (i.e. based only on beliefs in rationality) and constrained (i.e. based also on first-order belief restrictions) strategic reasoning. In Section 5, I show how the assumptions and the notions adopted in this paper explain the differences in the results with respect to this literature.

Kohlberg and Mertens [20] were the first to introduce forward induction considerations into equilibrium reasoning, through the set-valued notion of *strategically stable* equilibria. Govindan and Wilson [16] refine sequential equilibrium with a notion of forward induction. However, these two prominent works and the related literature share the two same shortcomings. First, they never question subgame perfection as a must-have for a "strategically stable" solution. Second, the strategic reasoning that leads to play such equilibria is unclear or limited.[8] The rationalizability approach adopted in this paper, which is backed by epistemic foundations, allows to eliminate both shortcomings. First, there is no constraint about how precisely and on which kind of equilibrium behavior players agree. Second, there is transparency about which particular agreements, beliefs, and epistemic assumptions induce different lines of reasoning, with a clear demarcation between unconstrained and constrained forward induction reasoning (missing in this literature).

[8] A similar critique to strategic stability has been put forward also by Van Damme [30].

In this sense, this work can also be interpreted as the axiomatic realization of a program akin to Kohlberg and Mertens' (see [20], p. 1020).[9] Full-fledged forward induction reasoning is captured and clarified. Agreements provide clear motivation and intuitive implementation, whereas strategic stability requires to retrieve hard-to-guess mixed strategies for the verification of the most intuitive outcomes. Implementable outcomes are proved and not assumed to be *strict* Nash, but not necessarily subgame perfect. In Section 6, I take a class of strategically unstable equilibria and show precisely which kind of forward induction reasoning is able to rulethem out. It turns out that the idea behind subgame perfection is at deep contradiction precisely with this kind of forward induction reasoning.

To introduce intuitively these ideas, Section 2 discusses three simple examples. Section 3 presents the theoretical framework and the analytic tools for the formal treatment of Section 4. Sections 5 and 6 discuss the relationship with the literature on rationalizability and on equilibrium in dynamic games, and the robustness of the analysis to different kinds of forward induction reasoning. Section 7 illustrates an applied example. The Appendix collects the proofs of theorems and propositions, the formalization of Example 3 and of the applied example, and two additional examples recalled in the paper. The proof of remarks is left to the reader.

2 Examples

Example 1 Consider the following game.

$A\backslash B$	W	E
N	$3,3$	$\cdot-$
S	$0,0$	$2,2$

$\longrightarrow$

$A\backslash B$	L	R
U	$1,1$	$2,2$
D	$0,6$	$3,5$

The subgame has only one equilibrium, where all actions are played with probability $1/2$. Hence, the unique SPE of the game induces outcome (S,E), which is Pareto dominated by (N,W). Suppose, Ann and Bob agree to play (N,W) and that Ann

[9] Kohlberg and Mertens [20] write: "We agree that an ideal way to discuss which equilibria are stable, and to delineate this common feeling, would be to proceed axiomatically. However, we do not yet feel ready for such an approach; we think the discussion in this section will abundantly illustrate the difficulties involved." Nowadays, the achievements of epistemic game theory allow to overcome many of these difficulties.

should play U in case of deviation of Bob. Is the agreement credible? If Bob is rational, he may deviate only if he does not believe in N, or does not believe in U, or both. Then, after the deviation, Ann cannot believe at the same time that Bob is rational and believes in the agreement. If she drops the belief that Bob believes in the agreement and maintains the belief that Bob is rational, she *can* believe that Bob does not believe in U and that he will play L. Thus, she can react with U. Anticipating this, Bob can expect N and U, and refrain from deviating. Further steps of reasoning do not modify the conclusion: the agreement is credible and, once believed, players will comply with it.

Note that U is played with positive probability also in the SPE. Example 4 (in the Appendix) displays instead a credible threat which differs from the unique equilibrium action of the subgame.

Example 2. In this 3-players game, in the subgame, Cleo chooses the matrix, Ann the row, and Bob the column.

$Ann \longrightarrow$	$M1$	L	C	R	$M2$	L	C	R
$\downarrow O \quad I$	U	$8,5,0$	$9,0,0$	$1,4,1$	U	$8,5,0$	$9,5,0$	$1,4,0$
$4,4,4$	D	$9,5,0$	$8,5,0$	$0,4,0$	D	$9,0,0$	$8,5,0$	$0,4,1$

In any equilibrium of the subgame, R cannot be played with probability higher than $1/2$, otherwise Ann would choose U and then Bob would switch to L. Hence, O is not a SPE outcome. Suppose that Bob and Cleo want to induce Ann to choose O. If they try to coordinate on a joint threat, they fail: if Bob knows the matrix, he prefers L or C to R. So, suppose that Bob threatens Ann with R and Cleo remains silent. If Ann plays I and is rational, she does not believe in R. Thus, she may play U or D, and Cleo may react with $M1$ or $M2$. Then, it is credible that Bob will react with R.

Example 3. Consider now the twofold repetition of the following game.

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

Ann and Bob agree that only Ann will work in the first period and, if this happens, only Bob will work in the second period. They do not agree on what to do if the

agreement is violated in the first period. Suppose that Bob deviates to Work in the first period. Ann can still believe that Bob is rational and believed in the agreement. But then, she must believe that Bob will not work in the second period, otherwise his deviation cannot be profitable. So she reacts to the deviation by working also in the second period. If Bob believes that Ann believes that he is rational and believes in the agreement, he anticipates this reaction and chooses to deviate. Anticipating this, Ann cannot believe in the agreement. The agreement is not credible.

So, Ann and Bob agree that only Bob will work in both periods. But then, Bob can signal with a deviation his intention to free ride also in the second period, so Ann works in the second period and Bob benefits from the deviation.

Two objections may be raised at this point. First, Ann could interpret the deviation as follows: "Bob believed that I would have not complied with the agreement, and best replied by not complying himself." But then, if the beliefs of Ann are Bayes-consistent, she must believe that Bob does not trust her from the start: the deviation of Bob is not at odds with the belief that Ann complies with the agreement. Second, Ann and Bob could agree beforehand on what to do in case of deviation. For social convenience, they may not be willing to do so. Or, when Bob displays disbelief in the agreement, Ann may still believe that he believed that she would have not violated the agreement before him. This belief gives rise to the rationalization of deviations depicted above (and further discussed in Section 6).

3 Agreements, beliefs and strategic reasoning

3.1 Preliminaries

Primitives of the game.[10] Let I be the finite set of *players*. For any profile $(X_i)_{i\in I}$ and any $\emptyset \neq J \subseteq I$, I write $X_J := \times_{j\in J} X_j$, $X := X_I$, $X_{-i} := X_{I\setminus\{i\}}$. Let $(\overline{A}_i)_{i\in I}$ be the finite sets of *actions* potentially available to each player. Let $\overline{H} \subseteq \cup_{t=1,...,T} \overline{A}^t \cup \{\emptyset\}$ be the set of histories, where $h^0 := \{\emptyset\} \in \overline{H}$ is the root of the game and T is the finite horizon. For any $h = (a^1, ..., a^t) \in \overline{H}$ and $l < t$, it holds $h' = (a^1, ..., a^l) \in \overline{H}$, and I write $h' \prec h$.[11] Let $Z := \{z \in \overline{H} : \forall h \in \overline{H}, z \not\prec h\}$ be the set of terminal histories

[10] The basic notation for games is mostly taken from Osborne and Rubinstein [25].

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

(henceforth, *outcomes* or *paths*)[12], and $H := \overline{H} \backslash Z$ the set of non-terminal histories (henceforth, just *histories*). For each $i \in I$, let $A_i : H \rightrightarrows \overline{A}_i$ be the correspondence that assigns to each history h, always observed by player i, the set of actions $A_i(h) \neq \emptyset$[13] available at h. I impose on $\overline{H}$ the following property: For every $h \in H$, $(h, a) \in \overline{H}$ if and only if $a \in A(h)$. For each $i \in I$, let $u_i : Z \to \mathbb{R}$ be the *payoff function*. The list $\Gamma = \left\langle I, \overline{H}, (u_i)_{i \in I} \right\rangle$ is a *finite game with complete information and observable actions*.

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

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

For any $(\overline{S}_j)_{j \in I} \subset S$, let $\overline{S}_i(h) := S_i(h) \cap \overline{S}_i$. For any $J \subseteq I$, let $H(\overline{S}_J) := \left\{ h \in H : \overline{S}_J(h) \neq \emptyset \right\}$ denote the set of histories compatible with $\overline{S}_J$. For any $h = (h', a) \in \overline{H}$, let $p(h)$ denote the immediate predecessor h' of h.

Throughout the paper, what a strategy prescribes at histories that are precluded by the strategy itself will be completely immaterial. Thus, the domain of each strategy s_i is restricted to $H(s_i)$; however, the term *strategy* rather than *reduced strategy* or *plan of actions* will be kept for brevity. At times, the domain of strategies will be further restricted to the histories that follow a given one. The restriction of a strategy $s_i \in S_i(h)$ to the histories following h is denoted by $s_i|h$ and is called *continuation plan*. A continuation plan can also be seen as a strategy of the subgame with root h, denoted by $\Gamma(h)$. Let S_i^h be the set of continuation plans from h on (or, equivalently, the strategies of $\Gamma(h)$) of player i. For any $\overline{S}_J \subset S_J$, let

$$\overline{S}_J|h := \left\{ s_J^h \in S_J^h : \exists s_J \in \overline{S}_J(h), s_J|h = s_J^h \right\}.$$

Histories and outcomes of $\Gamma(h)$ will be identified by the histories and outcomes of the whole game which follow h, and not redefined as shorter lists of action profiles.

[12] In many papers, paths and outcomes are different mathematical objects and a map from paths to outcomes is assumed. Since this distinction is immaterial for this paper, outcomes will be identified with paths, and the term "path" will be used with emphasis on the sequence of moves, and "outcome" with emphasis on the conclusion of the game.

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

Equilibria. A strategy profile $s = (s_i)_{i \in I} \in S$ is a *strict Nash equilibrium* if, for all $i \in I$ and $s'_i \notin S_i(\zeta(s))$, $u_i(\zeta(s)) > u_i(\zeta(s'_i, s_{-i}))$. A SPE is a profile of *non-reduced* strategies that, for each history $h \in H$, prescribes a profile of continuation plans $s^h = (s^h_i)_{i \in I} \in S^h$ which is a Nash equilibrium (not necessarily strict) of $\Gamma(h)$.

3.2 Agreements

Players discuss publicly how to play before the game starts. I assume that:

- Players do not coordinate explicitly as the game unfolds: all the opportunities for coordination are discussed beforehand.
- No subset of players can reach a private agreement, secret to co-players.
- Players do not agree on the use of randomization devices. Players would lack the incentive to (set the agreed-upon odds and) stick to the output of a (artificial) randomization device over the own actions.[14] Players also lack the ability to commit, otherwise it would not make sense to talk of non-binding agreements. Agreeing on the use of joint randomization devices, instead, would expand the set of outcomes players can achieve,[15] and could be analyzed with the methodology developed here.

Players can leave two kinds of strategic uncertainty, i.e. *agreement incompleteness*. First, and more importantly, players can be vague about which action they intend to play at some history. Second, players can promise to play an action at only one of two histories, without revealing at which one. This second kind of vagueness can naturally arise from strategic reasoning (also in absence of an agreement) and can be profitably exploited in agreements: see Example 5 (in the Appendix). A player can also declare what she plans to do in case she fails to implement her primary plans. And so on. Also the trust in a player who has already violated the agreement can be strategically exploited:[16] see again Example 5. Thus, agreements are formally modeled as follows.

[14] For this reason, I will talk of outcome sets instead of outcome distributions. As Pearce [26] puts it, "this indeterminacy is an accurate reflection of the difficult situation faced by players in a game." In games like matching pennies, an agreement is hardly conceivable.

[15] Similarly to how correlated equilibrium expands the set of Nash equilibrium outcome distributions.

[16] However, differently than in a SPE, this trust will be challenged with strategic reasoning.

Definition 1 *An Agreement is a profile of correspondences* $e = (e_i)_{i \in I}$ *with* $e_i : h \in H \mapsto e_i^h \subseteq S_i^h$ *such that for all* $i \in I$, $e_i^0 := e_i^{h^0} \neq \emptyset$, *and for all* $h \neq h^0$,

$$e_i^h \neq \emptyset \Rightarrow \cup_{h' \prec h} e_i^{h'}(h) = \emptyset \neq \cup_{h' \prec h} e_i^{h'}(p(h)).$$

Starting from the root of the game, an agreement can assign to a player a non-empty set of continuation plans only at histories that immediately follow a deviation by the player from the plans already assigned.[17] However, (i) the agreement may be empty at all such histories. Moreover, (ii) it may be *de facto* silent about a player's behavior also at histories that follow a deviation by anyone else. Agreements are particularly simple when (iii) players declare which actions they may play at each history, independently of what they plan to do at other histories.

Definition 2 *An agreement* $e = (e_i)_{i \in I}$ *is:*

i) *reduced if for every* $i \in I$ *and* $h \neq h^0$, $e_i^h = \emptyset$;

ii) *a path agreement on* $z \in Z$ *if it is reduced and for every* $i \in I$, $e_i^0 = S_i(z)$;[18]

iii) *on actions if for all* $i \in I$ *and* $h \in H$, $e_i^h = S_i^h \setminus \cup_{z \in V_i^h} S_i^h(z)$ *for some* $V_i^h \subseteq Z$.

A reduced agreement corresponds to a profile of strategy sets.[19] A path agreement corresponds to just agreeing on an outcome to achieve. An agreement on actions can be expressed not just through *vetos* V_i^h that players cast on outcomes, but also through actions instead of continuation plans assigned by the agreement at *each* history. Most agreements discussed in the paper are reduced and on action. Non-reduced agreements can be found in Example 4 and 5. An agreement not on actions is discussed in Example 5. Path agreements can be found in Example 3 and 4.

For any agreement $e = (e_i)_{i \in I}$, I refer to $\zeta(e^0)$ as the outcome set that the agreement *prescribes*.

[17] This is reminiscent of the notion of *basis* of a CPS introduced by Siniscalchi [29]: new theories are introduced only at histories that are not deemed as plausible as the previous ones under the theories already introduced.

[18] The term path agreement was first used by Greenberg et al. [17]: see also footnote 30.

[19] Recall that all strategies are reduced.

3.3 Belief in the agreement

Players' beliefs are modeled as Conditional Probability Systems (Renyi, [28]; henceforth, CPS). Here I define CPS's directly for the problem at hand.

Definition 3 *Fix $i \in I$ and let $\mathcal{C} := \left\{C \in 2^{S_{-i}} : \exists h \in H, C = S_{-i}(h)\right\}$. A Conditional Probability System on $(S_{-i}, \mathcal{C})$ is a mapping $\mu(\cdot|\cdot) : 2^{S_{-i}} \times \mathcal{C} \to [0,1]$ satisfying the following axioms:*

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

CPS-2 for every $C \in \mathcal{C}$, $\mu(C|C) = 1$;

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

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

A CPS is an array of probability measures, one for each history, which assign probability 1 to co-players' strategies compatible with the history. The array satisfies the chain rule (CPS-3). For brevity, the conditioning events will be indicated with just the history.

For any player i and any set of co-players $J \subseteq I \setminus \{i\}$, I say that a CPS μ_i *strongly believes* $\overline{S}_J \subseteq S_J$ if for every $h \in H(\overline{S}_J)$, $\mu_i(\overline{S}_J \times S_{I\setminus(J\cup\{i\})}|h) = 1$.[20] In formulae and proofs I will use the acronym "t.s.b." for "that strongly believes".

Note that a player can have correlated beliefs about the strategies of different co-players. This is not in contradiction with the absence of joint randomization devices in the agreement: players can believe in spurious correlations among co-players' strategies (see, for instance, Aumann [1]).[21] However, *strategic independence* (Battigalli, [3])[22] could be assumed throughout the paper and the results would not change.

I say that players believe in the agreement if, at each history, they believe in strategies of co-players which comply with the agreement from each co-player's last violation of the agreement onwards.

[20] In the original meaning of Strong Belief, due to Battigalli and Siniscalchi [9], $\overline{S}_J \times S_{I\setminus(J\cup\{i\})}$ and not $\overline{S}_J$ is "strongly believed". The slight difference in the use of the term is only for later notational convenience.

[21] For instance, a player can believe that a sunny day will induce more optimistic beliefs in two co-players.

[22] Roughly speaking, the assumption that a player has a separate CPS about the behavior of each co-player.

Definition 4 *Fix an agreement $e = (e_i)_{i \in I}$ and $\mu_i \in \Delta^H(S_{-i})$. I say that player i believes in the agreement when, for every $h \in H$, $s_{-i} = (s_j)_{j \neq i}$ with $\mu_i(s_{-i}|h) > 0$, $j \neq i$, and $\overline{h} \preceq h$,*

$$e_j^{\overline{h}}(h) \neq \emptyset \Rightarrow s_j|\overline{h} \in e_j^{\overline{h}}.$$

Let Δ_i^e be the set of all $\mu_i \in \Delta^H(S_{-i})$ where player i believes in the agreement.

Note that every $\mu_i \in \Delta_i^e$ strongly believes $(e_j^0)_{j \neq i}$.

3.4 Rationality and Rationalizability

I consider players who reply rationally to their conjectures. By rationality I mean that players, at every history, choose an action that maximizes expected utility given their belief about how co-players will play, and the expectation to reply rationally again in the continuation of the game. This is equivalent (Battigalli, [5]) to playing a *sequential best reply* to the CPS.

Definition 5 *Fix $\mu_i \in \Delta^H(S_{-i})$. A strategy $s_i \in S_i$ is a sequential best reply to μ_i if for each $h \in H(s_i)$, s_i is a continuation best reply to $\mu_i(\cdot|h)$, i.e. for each $\widetilde{s}_i \in S_i(h)$,*

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

The set of sequential best replies to μ_i (resp., to some $\mu_i \in \Delta_i^e$) is denoted by $\rho(\mu_i)$ (resp., by $\rho(\Delta_i^e)$). The set of normal-form best replies to a probability measure ν_i on S_{-i} is denoted by $r_i(\nu_i)$.

I say that a strategy s_i is *rational* if it is a sequential best reply to some $\mu_i \in \Delta^H(S_{-i})$. An important remark: Even when no rational strategy prescribes action a at two unordered histories h and h', there might be other two rational strategies, both compatible with h and h', which prescribe a only at, respectively, h and h'.

Here I take the view that players refine their first-order beliefs through strategic reasoning based on beliefs in rationality and beliefs in the belief in the agreement. In particular, I assume that every player, as long as not contradicted by observation, believes that each co-player is rational and believes in the agreement; that each co-player believes that each other player is rational and believes in the agreement; and

so on. At histories where common belief in, jointly, rationality and the belief in the agreement is contradicted by observation, I assume that players maintain all orders of belief in rationality that are per se compatible with the observed behavior, and drop the incompatible orders of belief in the agreement. I will call *independent rationalization* the hypothesis that players maintain a order of belief in rationality or in the agreement about a co-player when her *individual* behavior allows, as opposed to the hypothesis that players maintain such order of belief about all co-players only until none of them contradicts it.[23] The adoption of independent rationalization shows better the robustness of the main insights. After a deviation that displays the disbelief of the deviator in the agreement, without independent rationalization co-players' threats would not be demanded any degree of coordination, making departures from subgame perfection more likely. In Example 5, independent rationalization makes it much more challenging for players to find an effective agreement.

As shown in [12], the behavioral consequences of this kind of strategic reasoning are captured by Selective Rationalizability. Selective Rationalizability refines the following version of Extensive Form Rationalizability[24] (henceforth just **Rationalizability**).

Definition 6 *Let $S^0 := S$. Fix $n > 0$ and suppose to have defined $((S_j^q)_{j\neq i})_{q=0}^{n-1}$. For each $i \in I$ and $s_i \in S_i$, let $s_i \in S_i^n$ if and only if $s_i \in \rho(\mu_i)$ for some $\mu_i \in \Delta^H(S_{-i})$ that strongly believes $((S_j^q)_{j\neq i})_{q=0}^{n-1}$.*

Finally, let $S_i^\infty = \cap_{n\geq 0} S_i^n$. The profiles in S^∞ are called rationalizable.

It will be useful to introduce the following class of "realization equivalent" rationalizable continuation plans, under the hypothesis that the opponents play rationalizable plans. For any $h \in H(S^\infty)$ and $\overline{s}_i^h \in S_i^\infty|h$, let $\left[\overline{s}_i^h\right]^\infty$ be the set of all $s_i^h \in S_i^\infty|h$ such that $\zeta(s_i^h, s_{-i}^h) = \zeta(\overline{s}_i^h, s_{-i}^h)$ for all $s_{-i}^h \in S_{-i}^\infty|h$. For any $\overline{S}_i^h \subseteq S_i^\infty|h$, let $[\overline{S}_i^h]^\infty := \cup_{\overline{s}_i^h \in \overline{S}_i^h} \left[\overline{s}_i^h\right]^\infty$.

[23] This is not in contradiction with the absence of strategic independence: players can believe in spurious correlations among co-players' strategies, although they are ready to believe that different co-players have different orders of belief in rationality or in the agreement. For instance, the beliefs of a more and a less sophisticated players can be affected by weather in the same way.

[24] This notion of Extensive-Form-Rationalizability is the adaptation of Strong Rationalizability (Battigalli and Siniscalchi, [9]) to independent rationalization. Independent rationalization is also a feature of Independent Rationality Orderings (Battigalli [3]), where strategic independence is adopted. The original notion of Extensive-Form-Rationalizability, due to Pearce [26], adopts instead structural consistency (Kreps and Wilson [21]).

Selective Rationalizability can now be defined as follows.

Definition 7 *Let* $(S^0_{i,e})_{i\in I} := (S^\infty_i)_{i\in I}$. *Fix* $n > 0$ *and suppose to have defined* $((S^q_{j,e})_{j\neq i})^{n-1}_{q=0}$. *For each* $i \in I$ *and* $s_i \in S_i$, *let* $s_i \in S^n_{i,e}$ *if and only if there is* $\mu_i \in \Delta^e_i$ *that strongly believes* $((S^q_{j,e})_{j\neq i})^{n-1}_{q=0}$ *such that* $s_i \in \rho(\mu_i)$ *and:*

S3: μ_i *strongly believes* $((S^q_j)_{j\neq i})^\infty_{q=0}$.

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

S3 guarantees that a player always believes in co-players' strategies which are compatible with the highest possible order of belief in rationality. On top of this, at every step n and history h, a player believes in co-players' strategies which are compatible with the agreement and with the highest possile order $m \leq n-1$ of belief in the agreement. Note that the first-order belief in the agreement is mandatory. Then, the empty set is obtained when at some step some co-player can reach a history only with strategies that do not comply with the agreement from the history on. In this way, the compatibility of the belief in the agreement with the strategic reasoning hypotheses is tested.

S3 can be substituted by $s_i \in S^\infty_i$ for all the agreements $e = (e_i)_{i\in I}$ such that $e^h_i = \left[e^h_i\right]^\infty$ for all $i \in I$ and $h \in H$: see Lemma 3 in the Appendix. By Definition 12 and by Theorem 1, this class of agreements suffices to induce all the implementable outcome sets (and also the agreements that correspond to a Self-Enforcing Set fall in this class, see Definition 13). However, for any agreement, Rationalizability and Selective Rationalizability can be merged into one elimination procedure, where the belief in the agreement kicks in once the rationalizable profiles are obtained (see footnote 50). Finally, strong belief in $((S^q_{j,e})_{j\neq i})^{n-2}_{q=0}$ can be replaced by $s_i \in S^{n-1}_{i,e}$ only in 2-players games or dropping independent rationalization: see [12] for details.

Only in the Applied Example of Section 7, the game features non-rationalizable strategies. To see Selective Rationalizability at work, check the formalization of Example 3 in the Appendix. I will refer to $\zeta(S^\infty_e)$ as the set of outcomes *induced* by e.

4 Self-enforceability and implementability

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

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

Credibility does not imply that players will comply with the agreement, but only that they may do so *everywhere in the game*. Strategic reasoning on a credible agreement induces each player i to strongly believe in a subset of co-players' agreed-upon plans, namely $S_{-i,e}^\infty \cap e_{-i}^0$. I say that an agreement is self-enforcing if this belief will not be contradicted by the actual play.

Definition 9 *A credible agreement is self-enforcing if* $\zeta(S_e^\infty) = \zeta(S_e^\infty \cap e^0)$.

Self-enforceability implies that players will certainly comply with the agreement *on the agreed-upon paths*, so that no violation of the agreement will actually occur. That is, $\zeta(S_e^\infty) \subseteq \zeta(e^0)$. This condition is also sufficient for self-enforceability of a credible agreement on actions.

Proposition 1 *An agreement on actions is self-enforcing if and only if*

$$\emptyset \neq \zeta(S_e^\infty) \subseteq \zeta(e^0).$$

In Examples 1 and 2, the reduced agreements with, respectively, $e_A^0 = \{N.U\}, e_B^0 = \{W\}$, and $e_A^0 = S_A, e_B^0 = \{R\}, e_C^0 = S_C$ are self-enforcing. All strategies are rationalizable. At the first step of Selective Rationalizability, Ann eliminates S in Example 1 and selects O in Example 2, while Bob selects W in Example 1 and, like Cleo, does not eliminate any strategy in Example 2. In both cases, Selective Rationalizability is over at step 1. Example 3, formalized in the Appendix, provides two non-credible agreements.

A merely credible agreement fails to secure outcomes that players agreed upon and believed in. Moreover, only self-enforcing agreements are able to secure a specific outcome.

Proposition 2 *If $\zeta(S_e^\infty)$ is a singleton, then e is self-enforcing.*

For these reasons, in the remainder of the paper, the focus will be on self-enforcing agreements.

Which outcomes of the game can be achieved through self-enforcing agreements?

Definition 10 *A set of outcomes $P \subseteq Z$ is implementable if there exists a self-enforcing agreement such that $\zeta(S_e^\infty) = P$ (and I say the agreement* implements *P).*

With "implementable outcomes" I will refer specifically to implementable singletons. The set of outcomes prescribed by a self-enforcing agreement may be larger than the outcome set it induces. So, a natural question arises: for each implementable outcome set, is there an implementing agreement that prescribes precisely that set of outcomes? The answer is not obvious because simply restricting the initial plans of some self-enforcing agreement to those that induce the implemented outcome set may not work: see Example 4. Thus, consider the following classes of agreements.

Definition 11 *A self-enforcing agreement is truthful if $\zeta(S_e^\infty) = \zeta(e^0)$.*

Definition 12 *An agreement $e = (e_i)_{i \in I}$ is tight if for each $i \in I$,*

T1 For all $h \in H(S^\infty)$, $\cup_{\overline{h} \preceq h} e_i^{\overline{h}}(h) \neq \emptyset$ and $e_i^h = [e_i^h]^\infty$; else, $e_i^h = \emptyset$;

T2 For each $h \in H(\rho(\Delta_i^e) \cap S_i^\infty))$, $e_i^h \subseteq (\rho(\Delta_i^e) \cap S_i^\infty)|h$;

T3 For each μ_i that strongly believes e_{-i}^0, $\zeta(\rho(\mu_i) \times e_{-i}^0) \subseteq \zeta(e^0)$.

T3 says that players who believe in the agreement have no incentive to leave the paths it prescribes. Thus, the following holds.

Remark 1 *An agreement $e = (e_i)_{i \in I}$ where $\zeta(e^0)$ is a singleton satisfies T3 if and only if e^0 is a set of strict Nash equilibria.*

T1 says that a tight agreement reaches all the rationalizable histories with rationalizable continuation plans of all players; moreover, such plans do not restrict behavior at other histories, and no further plans are made. By T2, the prescribed plans must also be rational for a player who believes in the agreement and reaches the history. This guarantees that the agreed-upon plans never fall below other plans in the "likelihood order" of co-players who reason by forward induction about her. Thus, the following holds.

Proposition 3 *A tight agreement is truthful.*

On the other hand, for every implementable outcome set, there is always a tight agreement that prescribes it.

Theorem 1 *An outcome set is implementable if and only if there exists a tight agreement that prescribes it.*

Then, by Remark 1, the following holds.

Corollary 1 *Every implementable outcome is induced by a strict Nash equilibrium in rationalizable strategies.*[25]

By Theorem 1 and Proposition 3, the answer to the original question is affirmative.

Corollary 2 *Every implementable outcome set is implemented by a truthful agreement.*

Corollary 2 constitutes a *revelation principle* for agreements design: players need not be vague about the outcomes they want to achieve.

Corollary 1 restricts the search for implementable outcomes to the fixed points of the normal-form, best response correspondence, in the reduced game of rationalizable strategies.

Theorem 1 provides a *full* characterization of implementable outcome sets. Tight agreements simplify the (already finitely dimensional) search for implementable outcome sets and implementing agreements. First, Rationalizability is performed, without keeping memory of its steps afterwards. Once a candidate outcome (set) is fixed, Corollary 2, allows to restrict the search to agreements that prescribe it. Moreover, one can focus on initial plans that are rational under strong belief in the ones of co-players (by T2), and directly provide the incentive not to deviate from the desired paths (by T3). Then, the behavior of deviators must be specified as to satisfy T1 and T2 off-path. Note that T2 only requires to compute the sequential best replies to the belief in the agreement itself, as opposed to the multiple steps required by Selective Rationalizability, and without memory of the steps of Rationalizability.

[25] It is straightforward to prove this result directly by observing that if z is implemented by e, then any $s \in S_e^\infty$ is a strict Nash equilibrium in rationalizable strategies.

Example 5 illustrates an interesting tight agreement, which prescribes an outcome that cannot be implemented by an agreement on actions or without off-the-path restrictions.[26] However, tight agreements may be more complex than needed for the implementation of an outcome set. For a single outcome, the simplest and more natural agreement is the corresponding path agreement. Yet, very few path agreements are self-enforcing. In Example 4, not even the path agreement on the unique SPE outcome is self-enforcing. Thus, one may wonder which outcome sets can be implemented with reduced agreements and agreements on actions.

First, let us consider reduced agreements. A reduced agreement corresponds to a Cartesian set of strategy profiles. Recall that, throughout the paper, only reduced strategies are considered. This implies that, differently than in a SPE or than in a tight agreement, a reduced agreement remains silent about the behavior of deviators. However, the behavior of deviators is better predicted by forward induction. Thus, consider the following, set-valued solution concept.

Definition 13 *Fix* $S^* = \times_{i \in I} S_i^* \subseteq S$. *I say that* S^* *is a Self-Enforcing Set if for each* $i \in I$*:*

♠ *Rationalizability*: $S_i^* = [S_i^*]^\infty$.

♣ *Self-Justifiability:* $S_i^* \subseteq \left\{ s_i : \exists \mu_i \text{ t.s.b. } (S_j^*, S_j^\infty)_{j \neq i},\ s_i \in \rho(\mu_i) \right\} =: \overline{S}_i$;

♡ *Forward Induction:* $\overline{S}_i \subseteq \left\{ s_i : \exists \mu_i \text{ t.s.b. } (S_j^*, \overline{S}_j, S_j^\infty)_{j \neq i},\ s_i \in \rho(\mu_i) \right\}$;

♢ *Self-Enforceability*: *For each* μ_i t.s.b. S_{-i}^*, $\zeta(\rho(\mu_i) \times S_{-i}^*) \subseteq \zeta(S^*)$.

Rationalizability says that the SES prescribes rationalizable plans without restricting behavior at the non-rationalizable histories. Consider now players who strongly believe that *each* co-player will play as the SES prescribes and, alternatively, as rationalizability prescribes. Self-Justifiability says that they may play any strategy prescribed by the SES. This yields truthfulness when the SES induces multiple outcomes. Forward Induction says that all the strategies such players may play, thus including the SES strategies, are compatible with strong belief that co-players form beliefs in the same way. At each history h off-path, beside set-valuedness, the logics of Forward Induction differ from the logics of subgame perfection in two ways. On the one hand, Forward Induction "completes" the SES by determining the continuation plans of a deviator (j) from the SES with forward induction reasoning, based on her

[26] Implying by no reduced agreement under a discussed modification of the game.

belief in the SES if possible ($\overline{S}_j$) or just the beliefs in rationality otherwise (S_j^∞), as opposed to the exogenous prescriptions of a SPE. On the other hand, only the players who do not display disbelief in the continuation plans of-co-players, determined by the SES or by forward induction, are expected to keep best replying to them. This best response condition, imposed after one step of reasoning instead of just at the start, suffices to guarantee credibility after all steps of reasoning, which players do not actually need to perform when they agree on a SES.

On top of this, Self-Enforceability[27] guarantees that players will not leave the paths induced by the SES if they strongly believe that *all* co-players will play as the SES prescribes. Deviations from the SES threats can still occur and are interpreted with forward induction reasoning based on the belief in the SES. Still, Self-Enforceability suffices to yield self-enforceability of the corresponding agreement.

Theorem 2 *Fix a SES S^*. The reduced agreement e with $e^0 = S^*$ is truthful.*

For any self-enforcing agreement e, $S_e^\infty \cap e^0$ satisfies Self-Enforceability and Self-Justifiability, while restrictions to behavior at non-rationalizable histories can always be eliminated as to satisfy Rationalizability. Thus, these three conditions per se do not restrict the implementable outcome sets induced by a SES. Yet, $S_e^\infty \cap e^0$ may not satisfy Forward Induction. The sequential best replies of player i under strong belief in $(S_{j,e}^\infty \cap e_j^0)_{j \neq i}$ may not be, at some history, what co-players expect after all steps of reasoning under e. Such refinement of beliefs may be crucial to sustain the threats. For this reason, not every implementable outcome set, not even if implemented by a reduced agreement e, is induced by some SES. Example 5 provides a case in point. However, a SES always exists.

Remark 2 *S^∞ is a SES.*

The search for candidate SES's conveniently coincides with the search of the initial plans of a tight agreement. Then, Forward Induction must be checked. If no candidate SES for the implementation of an outcome set satisfies Forward Induction (like in Example 5), then one can try to transform a candidate SES into a tight agreement, by prescribing the behavior of deviators as to satisfy T1 and T2 off-path.

[27] I will write Self-Enforceability with capital letters to distinguish it from the self-enforceability of agreements.

Let us look for SES's in Example 2. All strategies are rationalizable, so Rationalizability is always satisfied. If $S_A^* = \{O\}$, by Self-Enforceability of Ann $S_B^* = \{R\}$, and by Self-Justifiability of Bob $S_C^* = \{M1, M2\}$. Let $S^* = \{(O, R, M1), (O, R, M2)\}$. Forward Induction holds because for each $i \in I$ and μ_i t.s.b. $(S_j^*, S_j^\infty)_{j \neq i}$, by $H(S_j^*) = H(\overline{S}_j)$ for all $j \neq i$, μ_i strongly believes $(S_j^*, \overline{S}_j, S_j^\infty)_{j \neq i}$. The construction of S^* and the verification of Forward Induction correspond to the informal arguments of Section 2. Note that any SES inducing O needs to be set-valued, albeit inducing a unique outcome.

If $\{I.U, I.D\} \subseteq S_A^*$, by Self-Enforceability $S_B^* = \{L.C.R\}$ and $S_C^* = \{M1, M2\}$, but then $O \in S_A^*$, i.e. $S^* = S^\infty = S$. If $\emptyset \neq \{I.U, I.D\} \cap S_A^* \neq \{I.U, I.D\}$, by Self-Justifiability $R \notin S_B^*$; then, by Self-Justifiability $O \notin S_A^*$, and by Self-Enforceability $S_C^* = \{M1, M2\}$. Let $S^* = \{I.x\} \times \{L, C\} \times \{M1, M2\}$ for $x = U, D$: Forward Induction is satisfied because $H(S^*) = H$.

Can the SES be implemented by a reduced agreement *on actions*? The answer is yes if the SES can be expressed through vetos cast by each player on rationalizable outcomes.

Proposition 4 *Fix $S^* = \times_{i \in I} S_i^* \subseteq S$ that satisfies ♣, ♡, ♢, and, for each $i \in I$:*

♠ *Rationalizable Vetos: $S_i^* = S_i^\infty \setminus \cup_{z \in W_i} S_i(z)$ for some $W_i \subseteq \zeta(S^\infty)$.*

Then, S^* is SES and $\zeta(S^*)$ is implemented by the reduced agreement on actions with vetos $V_i^0 := Z \setminus \zeta(S_i^* \times S_{-i})$ for all $i \in I$.

Casting unilateral vetos on outcomes is equivalent to exclude actions instead of strategies. The candidate SES is then the set of rationalizable strategies that do not prescribe the excluded actions. The implementing reduced agreement on actions is the set of *all* strategies that allow the SES outcomes.[28] Since the game in Example 2 has no pair of unordered histories, its SES's all satisfy Rationalizable Vetos (and S^∞ always does).

Focus now on implementable outcomes. By Rationalizability and Self-Enforceability, every SES that induces a unique outcome is a set of strict Nash equilibria in rationalizable strategies. Does the opposite hold? The answer is no: the threats of two different

[28] With $V_i^{h^0} = W_i$, the agreement may be not credible: at some $h \in H(S_i^\infty) \cap H(S_i \setminus \cup_{z \in W_i} S_i(z))$, there may not be any $s_i \in S_i^\infty(h) \setminus \cup_{z \in W_i} S_i(z)$, so strong belief in both S_i^∞ and $S_i \setminus \cup_{z \in W_i} S_i(z)$ is impossible.

players may be incompatible with each other. However, this cannot happen in a two-players game: each strict Nash in rationalizable strategies satisfies Self-Justifiability and Forward Induction.

Proposition 5 *Fix a two-players game and $z \in Z$. The set S^* of all strict Nash equilibria $s \in S^\infty(z)$, if non-empty, is a SES that satisfies Rationalizable Vetos.*

Moreover, for each $s \in S^*$, the reduced agreement e with $e^0 = \{s\}$ implements z.

Together with Corollary 1, the following holds.

Theorem 3 *In a two-players game, an outcome is implementable if and only if there exists a strict Nash equilibrium in rationalizable strategies that induces it.*

Together with Proposition 5, the following holds.

Corollary 3 *In a two-players game, every implementable outcome is implemented by a truthful, reduced agreement on actions.*

Thus, in two-players games, standard elimination procedure and fixed point condition suffice to find *all* implementable outcomes and, for each of them, a truthful, reduced agreement on actions that implements it.

5 Comparison with the rationalizability literature

The literature on strategic reasoning with first-order belief restrictions is mostly based on the use of Strong-Δ-Rationalizability ([6], [10]). The definition of Strong-Δ-Rationalizability with independent rationalization coincides with Definition 7 without S3 and with $S^0 = S$. The differences between the results of this paper and the results in this literature are due to (i) the adoption of Selective Rationalizability in place of Strong-Δ-Rationalizability, (ii) the structure on the first-order belief restrictions imposed by the notion of agreement, and (iii) the focus on self-enforceability rather than just credibility.

Differences and similarities between Selective Rationalizability and Strong-Δ-Rationalizability are deeply analyzed in [12]. Here I only recall the main conceptual difference behind the two solution concepts. Fix a move that a player would not

rationally pick were she to believe in the agreement. Contrary to Selective Rationalizability, Strong-Δ-Rationalizability captures the hypothesis that, upon observing such move, co-players *drop* the belief that the player is rational. This hypothesis is called "*(epistemic) priority to the agreement*" (as opposed to *rationality*). So, the question is: how would the adoption of Strong-Δ-Rationalizability instead of Selective Rationalizability affect the results?

In every example except the Applied Example of Section 7, all strategies are rationalizable; thus, Selective Rationalizability and Strong-Δ-Rationalizability coincide. Hence, the insights from the examples are robust to a shift of epistemic priority from rationality to the agreement.

What happens in games where not all strategies are rationalizable? Let $(S^q_{\Delta^e})^\infty_{q=0}$ be Strong-Δ-Rationalizability with independent rationalization.

Remark 3 *All the results of Section 4 hold through verbatim after substituting:*

1. *selectively-rationalizable strategies (S^∞_e) with strong-Δ-rationalizable strategies ($S^\infty_{\Delta^e}$) everywhere;*

2. *rationalizable strategies (S^∞) with all strategies (S) in the definitions of $[\cdot]^\infty$,*[29] *tight agreement and Self-Enforcing Set, and with rational strategies (S^1) in the statements of Corollary 1, Proposition 5, and Theorem 3.*

Remark 3 can be verified by operating the same substitutions in the proofs of the results, and skipping some passages as highlighted in footnotes. A credible agreement under priority to rationality needs not be credible under priority to the agreement: as shown in [12], Selective Rationalizability is not a refinement of Strong-Δ-Rationalizability for the same first-order belief restrictions. Across all agreements, instead, under priority to the agreement more outcome sets can be implemented.

Proposition 6 *If an outcome set is implementable under priority to rationality, then it is implementable under priority to the agreement.*

However, since agreements originate from mere, pre-play cheap talk, epistemic priority to rationality appears in my view as a more considerate hypothesis. Else, for instance, any Nash equilibrium in rational strategies of a two-players game would

[29] This is just to adapt to the formalism of Section 4: the equivalence classes become singletons.

correspond to a self-enforcing agreement, also when incompatible with just strong belief in rationality.

Battigalli and Friedenberg [7] capture the implications of Strong-Δ-Rationalizability without independent rationalization *across all* first-order belief restrictions with the notion of Extensive Form Best Response Set. An EFBRS is a Cartesian set of strategy profiles $\overline{S} = \times_{i \in I} \overline{S}_i$ satsfying the following condition:

EFBRS: for all $i \in I$ and $s_i \in \overline{S}_i$, $s_i \in \rho(\mu_i)$ for some μ_i t.s.b. $\overline{S}_{-i}$ with $\rho(\mu_i) \subseteq \overline{S}_i$.

The EFBRS Condition is the analogue of Self-Justifiability in absence of priority to rationality and independent rationalization, but with an additional "maximality" requirement: all the sequential best replies to some justifying beliefs must be in the EFBRS. These beliefs are not expressed by the EFBRS itself, whereas a SES directly provides the first-order belief restrictions that yield the SES outcomes. The restrictions that yield the EFBRS may impose the belief in specific randomizations, or, more fundamentally, differ across two players regarding the moves of a third player.[30] An agreement, instead, aligns any two player's beliefs about a third player's moves. For this reason, even with randomizations in agreements and without independent rationalization, EFBRS's would still be insufficient for implementability of the induced outcomes under priority to the agreement, calling for Self-Enforceability in place of maximality.

Battigalli and Siniscalchi [10] find out that, for first-order belief restrictions which correspond to the belief in an outcome, Strong-Δ-Rationalizability yields a non-empty set only if there exists a self-confirming equilibrium (Fudenberg and Levine [15], Battigalli [2]) inducing that outcome. Regardless of the epistemic priority choice, implementable outcomes are instead all Nash by Corollary 1 and Remark 3. Why is it the case? The reason lies in the difference between credibility and self-enforceability. Under a self-enforcing agreement, players have the incentive to stay on path for *all* their refined beliefs. This allows to find *independent* strategies of co-players against

[30] Greenberg et al. [17] define a (non-forward induction) solution concept, called "mutually acceptable courses of action". Their leading example represents an EFBRS outcome z. Strong-Δ-Rationalizability yields z for first-order belief restrictions that could be derived from an agreement for each player, but not from the same agreement for all players. Indeed, z is not implementable under priority to the agreement. Also allowing subsets of players to reach private agreements, z would still not be implementable, because the first-order belief restrictions of each player need instead to be transparent to all players (as they are under Strong-Δ-Rationalizability).

which there is no incentive to deviate. Credibility, instead, may be granted just by correlated beliefs about the reactions of co-players to the deviation.

Conversely, in signaling games, Battigalli and Siniscalchi [10] show that when an equilibrium outcome satisfies the Iterated Intuitive Criterion (Cho and Kreps [14]), Strong-Δ-Rationalizability yields a non-empty set for the corresponding first-order belief restrictions. Yet, even in the simplest examples of this paper, off-the-path restrictions are usually needed for self-enforceability. So, what does credibility under the path restrictions actually test when the agreement is reacher than the path agreement? The next section sheds light on this point.

6 Comparison with equilibrium literature

Kohlberg and Mertens [20] motivate their equilibrium analysis in a similar way to this paper: "*A noncooperative game is played without any possibility of communication between the players. However, we may think of the actual play as being preceded by a more or less explicit process of preplay communication (the course of which has to be common knowledge to all players), which gives rise to a particular choice of strategies.*" ([20], page 1004) Then, they introduce forward induction as implicit communication *during* the game, based on actual moves: "*Essentially what is involved here is an argument of "forward induction": a subgame should not be treated as a separate game, because it was preceded by a very specific form of preplay communication — the play leading to the subgame.* ([20], page 1013) Finally, they claim that the "forward induction" property of their notion of strategic stability, "*captures the "forward-induction" logic of our basic example.*" ([20], page 1029) The two examples of forward induction in the paper refer to a player who gives up an outside option. The consequent reasoning is not based on pre-play communication: unconstrained forward induction reasoning suffices for players to coordinate on the strategically stable solutions of two examples.

Govindan and Wilson [16], instead, use the Beer-Quiche game (Cho and Kreps, [14]) to show a different kind of forward induction reasoning. In Beer-Quiche, one of the two pure equilibria can be disregarded with a story of interactive beliefs in its outcome distribution. That is, constrained forward induction reasoning. However, both kinds of reasoning are hard to detect in the formal definition of forward

induction of [16], while depth of reasoning and scope of the analysis remain limited. As acknowledged by the authors themselves, their notion of forward induction only captures rationality and strong belief in rationality in two-players games ([16], page 11),[31] and fails in games with more than two players ([16], page 21). Moreover, it applies only to sequential equilibrium.

Osborne [24] identifies a class of non strategically stable SPE in two-players, finitely repeated, coordination games: those with an *equilibrium path that can be upset by a convincing deviation*. Differently than for the general definition of strategic stability, it is easy to match these equilibria with a precise line of forward induction reasoning: the one triggered by a path agreement. Indeed, equilibrium paths that can be upset by a convincing deviation can be characterized as non-credible path agreements, although the agreement on the whole SPE may well be self-enforcing. This confirms that also strategic stability captures (at least to some extent) constrained forward induction reasoning about the beliefs in an outcome (distribution). Thus, after the aforementioned characterization, I will analyze path-based forward induction reasoning, yet in presence of a richer-than-path agreement, such as a whole equilibrium profile. With this, I will show the robustness of the insights of the paper to this kind of strategic reasoning, and provide a general and transparent approach to the forward induction stories in the background of the equilibrium literature.

Fix a two-players (i and j) static game G with action sets A_i and A_j and payoff function $v_k : A_i \times A_j \to \mathbb{R}$, $k = i, j$. Let b^k and c^k be the first- and second-ranked stage-outcomes of G for player $k = i, j$. A path $(\overline{a}^1, .., \overline{a}^T)$ of pure Nash equilibria of the T-fold repetition of G *can be upset by a convincing deviation* ([24]) if there exist $\tau \in \{1, ..., T-1\}$ and $\widehat{a}_i \neq \overline{a}_i^\tau$ such that, letting $\overline{T} := T - \tau$,

$$v_i(\widehat{a}_i, \overline{a}_j^\tau) + v_i(c^i) + (\overline{T}-1)v_i(b^i) \;<\; \sum_{t=\tau}^{T} v_i(\overline{a}^t) < v_i(\widehat{a}_i, \overline{a}_j^\tau) + \overline{T}v_i(b^i); \qquad \text{(I)}$$

$$\overline{T}v_j(b^i) \;>\; \max_{a_j \in A_j \setminus \{b_j^i\}} v_j(b_i^i, a_j) + (\overline{T}-1)v_j(b^j). \qquad \text{(J)}$$

Condition I says that player i benefits from a unilateral deviation at τ only if

[31] I suggest that the two steps limitation (rationality and strong belief in rationality) on unconstrained reasoning extends to the constrained reasoning captured by forward induction. Moreover, I suggest that, once forward induction is immerged in sequential equilibrium, a further step of reasoning is captured *at the beginning of the game*. Indeed, the equilibrium selection in Beer-Quiche also requires a further step of reasoning at the beginning of the game.

followed by her preferred subpath. Condition J says that player j cannot benefit from a unilateral deviation from that subpath even if followed by her preferred subpath (which also shows that i's preferred stage-outcome is Nash, hence the restriction to coordination games).

Proposition 7 *Let $\overline{z} = (\overline{a}^1, ..., \overline{a}^T)$ be a path that can be upset by a convincing deviation. The path agreement on $\overline{z}$ is not credible.*

Example 3 provides two paths that can be upset by a convincing deviation,[32] although the agreements on the SPE that induce them are self-enforcing.

What does the non-credibility of the path agreement suggest when off-the-path threats are actually in place? It suggests that, under a particular way to interpret deviations (transparent to players), there is no credible threat that prevents some deviation. This interpretation of deviations relies on the belief that the deviator believes that no deviation by a co-player would have occured had she stayed on path. So, it stems from the common belief that everyone *trusts* that no-one is not going to violate the agreement unless someone else does first. The deviation proves that this trust towards the deviator was misplaced, but does not contradict the common belief in it. Thus, co-players, instead of dropping the belief that the deviator believes in the *whole* agreement, drop the belief that the deviator believes in the post-deviation threats, and save the belief that the deviator believed in the agreement on-path. In other words, the beliefs in the compliance with the agreement on-path have higher epistemic priority than the beliefs in the compliance with the agreement off-path. Assigning the highest epistemic priority to the beliefs in rationality, I call this finer epistemic priority order "*(epistemic) priority to the path*". Its behavioral consequences are captured by an extension of Selective Rationalizability, epistemically characterized in [12].[33] Fix $z \in Z$. Let $((S^q_{j,z})_{j \neq i})^{\infty}_{q=0}$ denote Selective Rationalizability under the path agreement on z, and call $(S^{\infty}_{j,z})_{j \neq i}$ *z-rationalizable*. Fix an agreement $e = (e_i)_{i \in I}$ with $\zeta(e^0) = \{z\}$.

[32] Formally, the paths do not satisfy the first strict inequality in (I), but this is immaterial because c^i $((W, W))$ and b^i $((W, FR))$ entail the same action for player i (Bob). This would not happen in pure coordination games that are in the focus of [24].

[33] The epistemic characterization in [12] works for general restrictions and would require $\Delta^e_i \subseteq \Delta^z_i$, which is typically false. Yet, I show in [11] that in this specific case, the use of Δ^e_i or $\Delta^e_i \cap \Delta^z_i$ is equivalent, and trivially $\Delta^e_i \cap \Delta^z_i \subseteq \Delta^z_i$.

Definition 14 *Let* $(S^0_{i,e^z})_{i\in I} = (S^\infty_{i,z})_{i\in I}$. *Fix* $n > 0$ *and suppose to have already defined* $((S^q_{j,e^z})_{j\neq i})^{n-1}_{q=0}$. *For each* $i \in I$ *and* $s_i \in S_i$, *let* $s_i \in S^n_{i,e^z}$ *if and only if there is* $\mu_i \in \Delta^e_i$ *that strongly believes* $((S^q_{j,e^z})_{j\neq i})^{n-1}_{q=0}$ *such that* $s_i \in \rho(\mu_i)$ *and:*

E3: μ_i *strongly believes* $((S^q_{j,z})_{j\neq i})^\infty_{q=0}$ *and* $((S^q_j)_{j\neq i})^\infty_{q=0}$.

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

E3 captures the interpretation of deviations depicted above. On top of this, players refine their beliefs according to the whole agreement. Thus, the credibility of the path agreement constitute a preliminary test for the implementability of z under the hypotheses of this section. This answers the question at the end of Section 5. If the outcome passes the test, there exist off-the-path beliefs, compatible with the interpretation of deviations depicted above, which induce players to stay on path. However, no agreement may be able to narrow down players' beliefs to those, like for the beliefs that sustain an EFBRS. An example of this is provided in [12], and it motivates the adoption of different belief restrictions in an epistemic priority order, instead of just turning to path restrictions and using credibility in place of self-enforceability.

Analogously to Selective Rationalizability, E3 can be substituted by $s_i \in S^\infty_{i,z}$ for all the agreements $e = (e_i)_{i\in I}$ such that $e^h_i = \left[e^h_i\right]^\infty$ for all $i \in I$ and $h \in H$, where $[\cdot]^\infty$ is redefined with S^∞_z in place of S^∞.[34] And again, this class of agreements suffices to induce all the implementable outcome sets under priority to the path. Indeed, restricting the focus for simplicity to agreements and strategy sets which prescribe a unique outcome z, the analysis of Section 4 can be replicated under this finer epistemic priority order.

Remark 4 *All the results of Section 4 hold through verbatim after substituting everywhere:*

1. *selectively-rationalizable strategies* (S_e) *with* z*-selectively-rationalizable strategies* (S_{e^z});[35]

2. *rationalizable strategies* (S^∞) *with* z*-rationalizable strategies* (S^∞_z).

[34] This is not proved formally in [12]. However, both S3 and E3 are maintained in the proofs.

[35] Self-Enforcing and truthful agreements which prescribe a unique outcome coincide. Then, restricting the attention to these agreement, all implementable outcomes are trivially implemented by a truthful agreement.

Remark 4 can be verified by operating the same substitutions in the proofs of the results. Although z-Selective Rationalizability does not refine Selective Rationalizability under the same agreement, the following holds.

Proposition 8 *If an outcome is implementable under priority to the path, then it is implementable under priority to rationality.*

In all the examples, the self-enforcing agreements remain self-enforcing under priority to the path. Thus, the insights are robust to the finer epistemic priority order adopted in this section. Strategic stability does not eliminate every non subgame perfect equilibrium;[36] yet, in the attempt to do so, valuable equilibria are disregarded.[37]

The final question is: does subgame perfection perform a meaningful further refinement under these strategic reasoning hypotheses? My answer is no. The idea behind subgame perfection is at deep contradiction with the interpretation of deviations behind this kind of forward induction reasoning. Fix a strict SPE. After any deviation from the SPE path, co-players will believe that the deviator believed in the path but does not believe in the threat. Then, they will not expect the deviator to best reply to the threat. But then, that the threat is a best reply to a plan of the deviator which is a best reply to the threat itself is of no additional value.[38] This breaks down the logics of subgame perfection. Example 4 illustrates this intuition. Thus, the insistence on subgame perfection in the forward induction literature is, in my view, particularly misplaced.[39]

[36] Kohlberg and Mertens [20] regard the inability to imply subgame perfection as a weakness of stability, and "hope that in the future some appropriately modified definition of stability will, in addition, imply connectedness and backwards induction." This paper suggests the opposite direction.

[37] Consider the (non-SPE) outcome T in Figure 6 in [20]. Its instability is claimed at page 1030, based on the substitutability of the zero-sum subgame with its equilibrium payoffs. But this amounts to assume that player 1 has the most pessimistic expectation for that subgame. Allowing for more optimistic beliefs, player 2 can believe that player 1 will try to reach the subgame. Thus, player 2 can react with R, a threat which implements T under all epistemic priority hypotheses.

[38] Also under the more agnostic interpretation of deviations of Section 4, even for a rationalizable SPE outcome, there may not be any threats that are compatible with both forward induction and subgame perfection.

[39] Interestingly, Man [23] finds out that also the "invariance" argument, used to motivate the notions of forward induction of Kohlberg and Mertens [20] and Govindan and Wilson [16], does not imply sequential equilibrium.

7 An Applied Example

Consider a linear city model of monopolistic competition between two firms, $i = 1, 2$.[40] Each firm i sets price p_i and, up to some prices, faces demand function

$$D_i(p_i, p_{-i}) = \begin{cases} 0 & \text{if } p_i > p_{-i} + 16 \\ 16 - p_i + p_{-i} & \text{if } p_i \in [p_{-i} - 16, p_{-i} + 16] \\ 32 & \text{if } p_i < p_{-i} - 16. \end{cases}$$

At the same time, each firm can choose between two production technologies, $k = 1, 2$. Technology $k = 1$ entails fixed cost $F_1 = 160$ and marginal cost $c^1 = 64$. Technology $k = 2$ entails fixed cost $F_2 = 800$ and marginal cost $c^2 = 32$. Conditional on using technology $k = 1, 2$, the best response function of firm i reads:

$$\widehat{p}_i(p_{-i}) = 8 + \frac{1}{2}c^k + \frac{1}{2}p_{-i}.$$

Conditional on employing $k = 1$, the unique equilibrium price vector is $(80, 80)$. Yet, the best reply to $p_{-i} = 80$ is $p_i = 64$, with the use of $k = 2$. Conditional on employing $k = 2$, the unique equilibrium price vector is $(48, 48)$. Yet, the best reply to $p_{-i} = 48$ is $p_i = 64$, with the use of $k = 1$. Note that profit is much higher under $(80, 80)$ and $k = 1$ than under $(48, 48)$ and $k = 2$.

Suppose now that firms compete for two periods. The upgrade from $k = 1$ to $k = 2$ between the two periods has a switching cost $W > 128$.[41] Then, if firms employed $k = 1$ in the first period, $(80, 80)$ is the unique rationalizable price vector in the second period. Firms want to reach an agreement to employ $k = 1$ in both periods. Consider a unilateral deviation by firm $i = 2$ to $k = 2$ in the first period. In the second period, firm 2 is indifferent between the two technologies for $p_1 = 72$. Firm 1 would need to pay W to adopt $k = 2$. So, for $\overline{p}_1 = 72$ and some $\overline{p}_2 > 80$, for each $i = 1, 2$ the best reply correspondence reads:

$$\widehat{p}_i(p_{-i}) = \begin{cases} 40 + \frac{1}{2}p_{-i} & \text{if } p_{-i} < \overline{p}_{-i} \\ \left\{40 + \frac{1}{2}p_{-i}, 24 + \frac{1}{2}p_{-i}\right\} & \text{if } p_{-i} = \overline{p}_{-i} \\ 24 + \frac{1}{2}p_{-i} & \text{if } p_{-i} > \overline{p}_{-i} \end{cases}$$

[40] The microfoundation of the demand functions in this model is presented in Green, et. al. [22], pages 396-397.

[41] W can be thought of as a firing cost, interpreting $k = 1$ as the labour intensive technology.

The set of rationalizable price vectors is $[70,78]\times([60,63]\cup[75,76])$. Each $p_1 \in [70,78]$ is a best reply to a conjecture over 60 and 76. Each $p_2 \in [60,63]$ is a best reply to some $p_1 \in [72,78]$ and each $p_2 \in [75,76]$ is a best reply to some $p_1 \in [70,72]$. Each $p_1 > 78$ can be best reply only to $p_2 > 76$, which can be best reply only to $p_1 > 104$, until the floor of price at which consumers buy is hit. Analogous arguments prove that all other p_1, p_2 are not rationalizable. There is no pure equilibrium: conditional on employing different technologies, the equilibrium price vector induces the firm that employs $k = 2$ to switch to $k = 1$. There is one equilibrium in which firm 1 sets $p_1 = 72$ (so that firm 2 can randomize) and firm 2 sets $p_2 = 60$ with probability 3/4 and $p_2 = 76$ with probability 1/4. From now on, assume for simplicity that firms can pick only integer prices.

Fix the path $z := (((1,80),(1,80)),((1,80),(1,80)))$, which yields profit $u_i(z) = 2\cdot(16^2 - F_1) = 192$ to $i = 1,2$. The best unilateral deviation of firm $-i$ (to $p_{-i} = 64$ and $k = 2$) in the first period, followed by the equilibrium of the subgame, yields to $-i$ profit $32^2 + 28^2 - 2F_2 = 208 > 192$. Thus, z is not a SPE path. Suppose instead that firm i reacts to the deviation with price $p_i = 70$. Then, the deviation is not profitable: $32^2 - F_2 + 11^2 - F_1 = 185 < 192$. Can firm i credibly threaten to fix $p_i = 70$ after such deviation? The answer is yes. First, note that at every rationalizable, non-initial history, the rationalizable prices of the two firms must constitute a best response set. Then, after a rationalizable deviation of firm $-i$ to $k = 2$, some $p_{-i} \in [60,63]$ and some $p_{-i} \in [75,76]$ must both be possible. But then, firm i can react with $p_i = 72$. Second, if expecting $p_i = 72$ makes the deviation profitable, firm $-i$ can then fix $p_{-i} = 60$, and the best reply of firm i to $p_{-i} = 60$ is precisely $p_i = 70$. In the Appendix I exploit this intuition to show formally the existence of a strict Nash equilibrium in rationalizable strategies that induces z. By Proposition 5, the corresponding agreement implements z.

Is z implementable also under priority to the path? Yes: by displaying the intention to gain a higher profit than under the path, firm $-i$ is not able to re-coordinate on a more profitable subpath with firm $-i$, who may always react with a lower price than firm $-i$ hoped for. In particular, if the least optimistic belief of $-i$ that justifies the deviation is $\widetilde{p}_i > 72$, the best reply to the best reply to p_i is smaller than p_i itself ($\widehat{p}_i(\widehat{p}_{-i}(p_i)) < p_i$); if $70 < \widetilde{p}_i \leq 72$, $-i$ may fix $p_{-i} = 60$, and i can react with $p_i = 70$. The construction of a Nash inducing z in the Appendix is valid also under priority to the path. By Remark 4, the corresponding agreement implements z under priority to

the path.

8 Appendix

8.1 Games

Construction of Nash for the Applied Example. For simplicity, I will omit the technology choice in the description of strategies. Note preliminarly that a unilateral deviation to $(2, p_{-i})$ with $p_{-i} = 61, ..., 67$ is profitable for firm $-i$ if followed by $p_i = 72$.

Fix $n \geq 0$ and suppose to have shown the existence of a strict Nash equilibrium $(s_i^*)_{i=1,2} \in S^n(z)$, and, for each $i = 1, 2$ and $h = ((1, 80), (2, p_{-i}))$ with $p_{-i} = 61, ..., 67$, of $s_{i,h}^* \in S_i^n$ such that $s_{i,h}^*(h) = 72$ and $s_{i,h}^*(h') = s_i^*(h')$ for all $h' \neq h$. Then, there exist μ_{-i} that strongly believes $(S_i^q)_{q=0}^{n-1}$ such that $\mu_{-i}(s_{i,h}^*|h^0) = 1$, and $s'_{-i,h}, s''_{-i,h} \in \rho(\mu_{-i})(h) \subseteq S_{-i}^n$ such that $s'_{-i,h}(h) = 60$ and $s''_{-i,h}(h) = 76$. Fix $i = 1, 2$. For each $h = ((1, 80), (k, p_{-i})) \in H(S_{-i}^n)$, fix $s_{-i,h} \in \arg\min_{s_{-i} \in S_{-i}^n(h)} s_{-i}(h)$. Fix μ_i^* that strongly believes $(S_{-i}^q)_{q=0}^n$ such that $\mu_i^*(s_{-i}^*|h^0) = 1$, and, for each $h = ((1, 80), (k, p_{-i})) \in H(S_{-i}^n)$ with $(k, p_{-i}) \neq (1, 80)$, $\mu_i^*(s_{-i,h}|h^0) = 1$. Fix $\overline{s}_i^* \in \rho(\mu_i) \subseteq r_i(\mu_i^*(\cdot|h^0)) = S_i(z)$ and μ_{-i} that strongly believes $(S_i^q)_{q=0}^{n+1}$ such that $\mu_{-i}(\overline{s}_i^*|h^0) = 1$. Fix $h = ((1, 80), (k, p_{-i})) \in H(S_{-i}^n)$. If $k = 2$ and $\overline{s}_i^*(h) > 72$, or $k = 1$ and $\overline{s}_i^*(h) > 80$, then $\widehat{p}_{-i}(\overline{s}_i^*(h)) = \widehat{p}_{-i}(\widehat{p}_i(s_{-i,h}(h))) < s_{-i,h}(h)$. But then, $\rho(\mu_{-i})(h) = \emptyset$, otherwise $s_{-i,h} \neq \arg\min_{s_{-i} \in S_{-i}^n(h)} s_{-i}(h)$. If $k = 2$, $p_{-i} \neq 61, ..., 67$, and $\overline{s}_i^*(h) \leq 72$, or $k = 1$ and $\overline{s}_i^*(h) \leq 80$, then $\rho(\mu_{-i})(h) = \emptyset$, because $\mu_{-i}(S_i(z)|h^0) = 1$ and the deviation cannot be profitable for $-i$. If $k = 2$ and $p_{-i} = 61, ..., 67$, then $\overline{s}_i^*(h) \leq \widehat{p}_i(s'_{-i,h}(h)) = \widehat{p}_i(60) = 70$, thus $\rho(\mu_{-i})(h) = \emptyset$. Since $\rho(\mu_{-i})(h) = \emptyset$ for all $h \notin H(S_{-i}^n)$, $r_{-i}((\mu_{-i}|h^0)) \subseteq S_{-i}(z)$. Hence, $\overline{s}^* \in S^{n+1}(z)$ is a strict Nash equilibrium. For each $h = ((1, 80), (2, p_{-i}))$ with $p_{-i} = 61, ..., 67$, fix $\mu_{i,h}^*$ that strongly believes $(S_{-i}^q)_{q=0}^n$ such that $\mu_{i,h}^*(s'_{-i,h}|h) \cdot 60 + \mu_{i,h}^*(s''_{-i,h}|h) \cdot 78 = 64$, and $\mu_{i,h}^*(\cdot|h') = \mu_i^*(\cdot|h)$ for all $h' \neq h$. Thus, there exists $\overline{s}_{i,h}^* \in \rho(\mu_{i,h}^*) \subseteq S_i^{n+1}$ such that $\overline{s}_{i,h}^*(h) = 72$ and $\overline{s}_{i,h}^*(h') = s_i^*(h')$ for all $h' \neq h$. Inductively, I find a strict Nash equilibrium $s^* \in S^\infty(z)$.

All employed μ_i strongly believe $S_{-i}(z)$. Thus, the procedure can be prolonged to obtain a strict Nash equilibrium $s^* \in S_z^\infty(z)$.

Formalization of Example 3.

$2 \times$

$A \backslash B$	W	F
W	$2,2$	$1,3$
F	$3,1$	$0,0$

For $i = A, B$, I will write a strategy s_i as $x.y.w$, where $x = s_i(h^0)$, $y = s_i((s_i(h^0), W))$, and $z = s_i((s_i(h^0), F))$. For any $z \in Z$, consider the path agreement $e^0 = S_A(z) \times S_B(z) = S(z)$; then $\Delta_i^e = \left\{\mu_i \in \Delta^H(S_{-i}) : \mu_i(S_{-i}(z)|h^0) = 1\right\}$, for $i = A, B$. All strategies are rational, hence rationalizable.
Let $z = ((W, F), (F, W))$. Selective Rationalizability goes as follows.

$$\begin{aligned}
S^1_{A,e} &= S_A(z);\ S^1_{B,e} = S_B(z) \cup \{W.F.W, W.F.F\};\\
S^2_{A,e} &= \{W.W.F\};\ S^2_{B,e} = S^1_{B,e};\\
S^3_{A,e} &= S^2_{A,e};\ S^3_{B,e} = \{W.F.W, W.F.F\};\\
S^4_{A,e} &= \emptyset.
\end{aligned}$$

Let $z := ((F, W), (F, W))$. Selective Rationalizability goes as follows.

$$\begin{aligned}
S^1_{A,e} &= S_A(z),\ S^1_{B,e} = S_B(z) \cup \{F.F.F, F.W.F\};\\
S^2_{A,e} &= \{F.F.W\},\ S^2_{B,e} = S^1_{B,e};\\
S^3_{A,e} &= S^2_{A,e},\ S^3_{B,e} = \{F.F.F, F.W.F\};\\
S^4_{A,e} &= \emptyset.
\end{aligned}$$

Example 4. Consider the following game.

$A \backslash B$	W	E
N	$6,6$	$\cdot-$
S	$0,0$	$2,2$

$\longrightarrow$

$A \backslash B$	L	C	R
U	$9,0$	$0,5$	$0,3$
M	$0,5$	$9,0$	$0,3$
D	$0,7$	$0,7$	$1,8$

All strategies are rational, hence rationalizable. The subgame has one pure equilibrium, (D, R), and no mixed equilibrium: for Ann to be indifferent between U and M, Bob must randomize over L, C, but when he is indifferent between them,

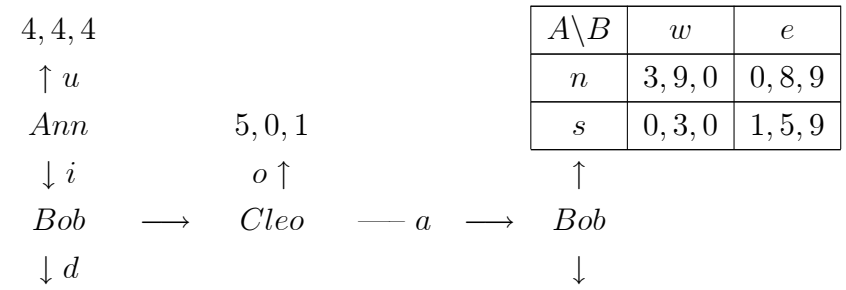
he prefers R; for Ann to be indifferent between U and D or M and D, Bob must randomize over, respectively, L, R and C, R, but R dominates L against U, D and C against M, D. So, the game has only one SPE, inducing path (S, E).

Players want instead to implement (N, W). Hence they reach the reduced agreement with $e_A^0 = \{N.U, N.M\}$ and $e_B^0 = \{W\}$. The agreement is self-enforcing: $S_e^1 = \{N.U, N.M, N.D\} \times \{W\}$, thus $S_e^\infty = S_e^1 = S((N, W))$. Also, the agreement is self-enforcing under priority to the path: each action of Bob in the subgame is a best reply to some belief over the actions of Ann which justifies the deviation. Formally, $S_z^\infty = S_z^1 = \{N.U, N.M, N.D\} \times S_B$, and $(S_{e^z}^q)_{q=0}^\infty = (S_e^q)_{q=0}^\infty$.

Note two things about the SPE. First, despite being unique, it requires off-the-path restrictions for its implementation. Under the path agreement on (S, E), Ann may deviate to N, hoping that Bob will reply with L or C, which are best replies against the potentially profitable deviations of Ann. Second, the SPE action D is not a potentially profitable deviation for Ann with respect to the path. Thus, if the deviation is interpreted as an attempt to improve the payoff with respect to the agreed-upon path, Bob cannot expect Ann to play D. Hence, the fact that R is best reply to D which is best reply to R itself is of no value.

Finally, consider the following, non-reduced, agreement: $e_A^0 = \{S\}, e_A^{(N,E)} = \{D\}, e_B^0 = S_B$. It implements (S, E): $S_e^1 = S_A \times \{E.R\}$; $S_e^2 = \{S\} \times \{E.R\}$; so $S_e^\infty = S_e^2 \subseteq S((S, E))$. Restrict now the plans of Bob to those compatible with (S, E), i.e. $e_B^0 = \{E.L, E.C, E.R\}$. Then, $S_{A,e}^1 = \{S, N.U, N.M\}$. But then, $S_{B,e}^2 = \emptyset$. Thus, a self-enforcing agreement cannot always be made truthful by excluding the plans that are not compatible with the path it implements.

Example 5. Consider the following game.



$A\backslash B$	w	e
n	$3, 9, 0$	$0, 8, 9$
s	$0, 3, 0$	$1, 5, 9$

$C\backslash B$	l	c	r
t	$5, 4, 1$	$5, 6, 0$	$5, 0, 0$
b	$5, 4, 0$	$5, 0, 1$	$5, 10, 1$

$A\backslash B$	w	e
n	$3, 9, 0$	$0, 8, 9$
s	$0, 3, 0$	$1, 5, 9$

Consider the following tight agreement, prescribing outcome (u):

$$\begin{aligned} e_A^0 &= \{u\}\,,\ e_B^0 = S_B \backslash \{d.l, d.c, d.r\}\,,\ e_C^0 = \{t.a\}\,; \\ e_A^{(i)} &= \{n.n, n.s, s.n\}\,,\ e_B^{(i,d)} = \{l, c, r\}\,. \end{aligned}$$

T3 holds, as $\rho(\Delta_A^e) = e_A^0 = \{u\}$. All strategies are rational, so $S^\infty = S$ and $H(S^\infty) = H$. Thus, for T1 to hold it is sufficient that all histories are reached by some plan: $H(e_A^0) = \{h^0\}$ and $H(e_A^{(i)}) = H \backslash \{h^0\}$; $H(e_B^0) = H \backslash \{i.d\}$ and $e_B^{(i,d)} \neq \emptyset$; $H(e_C^0) = H$. Finally, T2 holds. For Ann, $\rho(\Delta_A^e) = \{u\}$, so $e_A^0 \subseteq \rho(\Delta_A^e)$ and $(i) \notin H(\rho(\Delta_A^e))$. Bob expects Ann to play n with probability of at least $1/2$ in one of the two subgames, where his expected payoff is then at least 6.5. Moreover, he believes that Cleo will give him the opportunity to pick that subgame. After d, instead, he expects Cleo to play t, with a payoff of 6. Thus, $e_B^0 = \rho(\Delta_B^e)$, and $(i.d) \notin H(\rho(\Delta_B^e))$. For Cleo, $e_C^0 \subseteq \rho(\Delta_C^e) = S_C$. Since the agreement is tight, by Proposition 3 it implements (u).

Note that the agreement is not on actions: Ann promises to play n in one of the two subgames, but she does not say in which one. Is there an agreement on actions that implements u? No. For Ann to select u, Bob and Cleo must exclude from the agreement, or eliminate through strategic reasoning, d and o. If o is excluded or eliminated, Bob expects a payoff of at least 5 by not playing d. Thus, Bob will eliminate $d.l$. If Bob still considers $d.c$ or $d.r$ when $d.l$ is eliminated, Cleo will best reply with b. But then Bob will select $d.r$, and u cannot be implemented. So, the agreement must make sure that Bob eliminates $d.c$ and $d.r$ no later than $d.l$. For the elimination of $d.r$, it is necessary that Cleo excludes b from the agreement. Then Bob is confident that by playing $d.c$ he can get 6. So, for Bob to eliminate $d.c$, he must be confident of getting a higher payoff without playing d. So, he must be confident that in at least one of the two subgames, Ann will not play s. If this subgame was pinned down by the agreement or strategic reasoning, then Bob would play w in the subgame he moves to. Then, Cleo will select o, and u cannot be implemented. Hence, Ann, through the agreement or strategic reasoning, does have to exclude planning s in both subgames, but at the same time she must not reveal in which subgame she is not planning s. In this game, she can do this only through the agreement: if she rationally plays i, she hopes in d or o, and if d and o are not played, she could plan s in both subgames.

Thus, **agreements that are not on actions can be needed to implement an**

outcome. Then, the "only if" direction of Theorem 1 would fail if tight agreements were required to be on actions. This is true even if the focus is restricted to outcome sets implemented by agreements on actions: it is possible to complicate the game in such a way that any tight agreement prescribing (u) still requires the exclusion of $s.s$ and not of $n.s$ and $s.n$, but (u) is also implemented by an agreement on actions that leads to the elimination of $s.s$.[42]

Furthermore, Ann needs to restrict her agreed-upon plans at histories that follow her own deviation from the implemented path. Using Theorem 2 by contraposition, this shows that there is no SES that induces u. Yet, the tight agreement above is clearly equivalent to the following reduced agreement: $\overline{e}^0_A = \{u, i.n.n, i.n.s, i.s.n\}$, $\overline{e}^0_B = e^0_B$, $\overline{e}^0_C = e^0_C$. Thus, $\overline{e}$ is self-enforcing (but not truthful) and it implements (u). This shows that the reverse of Theorem 2 does not hold: SES's do not capture all the outcomes that can be implemented by a reduced agreement.

So, a final question arises: is the entire class of reduced agreements sufficient to implement all implementable outcomes? The answer is no. Imagine that at the initial history, Ann plays simultaneously with Bob, and needs to exclude i from the agreement to coordinate with Bob on an outcome equivalent to (u).[43] Then, Ann would need to exclude both i and the continuation plan $s.s$ after (i). Thus, **non-reduced agreements can be needed to implement an outcome.**

8.2 Proofs

Throughout, let $H^\infty := H(S^\infty)$ and $H_\infty := \{h \notin H^\infty : p(h) \in H^\infty\}$. For any $\mu_i \in \Delta^H_i(S_{-i})$, let $H^{\mu_i} := \{h^0\} \cup \{h \in H^\infty : \mu_i(S_{-i}(h)|p(h)) = 0\}$.

Proof of Proposition 1. "Only if": trivial. "If": e is credible by $\zeta(S^\infty_e) \neq \emptyset$, and $\zeta(S^\infty_e) \supseteq \zeta(S^\infty_e \cap e^0)$ is obvious; for the opposite inclusion I show that for every $s = (s_i)_{i\in I} \in S^\infty_e$, there exists $s^* \in S^\infty_e \cap e^0$ such that $\zeta(s^*) = \zeta(s)$. Fix $i \in I$ and $\mu_i \in \Delta^e_i$ t.s.b. $((S^q_{j,e})_{j\neq i})^\infty_{q=0}$ and $((S^q_j)_{j\neq i})^\infty_{q=0}$with $s_i \in \rho(\mu_i)$. By $\zeta(S^\infty_e) \subseteq \zeta(e^0)$, for each $h \in H(s_i) \cap H(S^\infty_e)$, $s_i(h) = \overline{s}_i(h)$ for some $\overline{s}_i \in e^0_i(h)$. Since the agreement is on actions, there exists $\overline{s}_i \in e^0_i$ such that $\overline{s}_i(h) = s_i(h)$ for all $h \in H(s_i) \cap H(S^\infty_e)$. Fix $h \in H' := \{h' \in H(s_i)\backslash H(S^\infty_e) : p(h') \in H(S^\infty_e)\}$. Since $p(h) \in H(s_i) \cap H(S^\infty_e)$,

[42] The modified game is available upon request.

[43] This also makes it plausible that Ann wants to contribute to the credibility of not playing i: in the example above, she just destroys any hope to get a higher payoff than her outside option.

$h \in H(s_i) \cap H(\overline{s}_i)$. Then $h \in H(S^\infty_{i,e}) \cap H(e^0_i)$. Thus, since e is credible, $e^0_i \cap S^\infty_{i,e}(h) \neq \emptyset$. Fix $s_{i,h} \in e^0_i(h) \cap S^\infty_{i,e}(h)$ and $\mu_{i,h} \in \Delta^e_i$ t.s.b. $((S^q_{j,e})_{j\neq i})^\infty_{q=0}$ and $((S^q_j)_{j\neq i})^\infty_{q=0}$ with $s_{i,h} \in \rho(\mu_{i,h})$. Since μ_i strongly believes $S^\infty_{-i,e}$, $\mu_i(S_{-i}(h)|p(h)) = 0$. Thus, there exists $\mu^*_i \in \Delta^e_i$ t.s.b. $((S^q_{j,e})_{j\neq i})^\infty_{q=0}$ and $((S^q_j)_{j\neq i})^\infty_{q=0}$ such that $\mu^*_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S^\infty_e)$, and $\mu^*_i(\cdot|h') = \mu_{i,h}(\cdot|h')$ for all $h \in H'$ and $h' \succeq h$. So, there is $s^*_i \in \rho(\mu^*_i) \subseteq S^\infty_{i,e}$ such that $s^*_i(h) = s_i(h) = \overline{s}_i(h)$ for all $h \in H(s_i) \cap H(S^\infty_e)$, and $s^*_i|h = s_{i,h}|h$ for all $h \in H'$. Since the agreement is on actions, $s^*_i \in e^0_i$, and by $H(s^*) \subseteq H(S^\infty_e)$, $\zeta(s^*) = \zeta(s)$. ■

Proof of Proposition 2. Since e is credible, $S^\infty_e \cap e^0 \neq \emptyset$. Since $\zeta(S^\infty_e)$ is a singleton and $\zeta(S^\infty_e) \supseteq \zeta(S^\infty_e \cap e^0)$, $\zeta(S^\infty_e) = \zeta(S^\infty_e \cap e^0)$. ■

Lemma 1 *Fix an agreement e. If e^0 satisfies T3 and $e^0 \subseteq S^\infty_e$, e is truthful.*

Proof. First, I show that $\zeta(S^\infty_e) \subseteq \zeta(e^0)$. Fix $s = (s_i)_{i\in I} \in S^\infty_e$ and $h \in H(s) \cap H(e^0)$. Since e^0 is Cartesian, so is $A^h_e := \left\{a \in A(h) : (h,a) \in \overline{H}(e^0)\right\}$. For each $i \in I$, since $s_i \in \rho(\Delta^e_i)(h)$ and $e^0_{-i}(h) \neq \emptyset$, by T3 $s_i(h) \in A^h_{i,e}$. Thus $(h, s(h)) \in \overline{H}(e^0)$. By induction, $\zeta(s) \in \zeta(e^0)$.

So, by $e^0 \subseteq S^\infty_e$, $\zeta(S^\infty_e \cap e^0) = \zeta(e^0) = \zeta(S^\infty_e)$. ■

Lemma 2 *Fix $i \in I$, $\overline{h} \in H^\infty$, $s^{\overline{h}}_i \in S^\infty_i|\overline{h}$, and $h \in H(s^{\overline{h}}_i) \cap H_\infty$. Thus, $[s^{\overline{h}}_i]^\infty|h = S^\infty_i|h$.*[44]

Proof. Fix $s_i, s'_i \in S^\infty_i(h)$ with $s_i|\overline{h} = s^{\overline{h}}_i$. Fix μ_i, μ'_i t.s.b. $((S^q_j)_{j\neq i})^\infty_{q=0}$ with $s_i \in \rho(\mu_i)$ and $s'_i \in \rho(\mu'_i)$. Since $h \in H(S^\infty_i) \backslash H^\infty$, $p(h) \in H(S^\infty_{-i})$, and μ_i strongly believes S^∞_{-i}, $\mu_i(S_{-i}(h)|p(h)) = 0$. Then, there exists μ^*_i t.s.b. $((S^q_j)_{j\neq i})^\infty_{q=0}$ such that $\mu^*_i(\cdot|h') = \mu_i(\cdot|h')$ for all $h' \not\succeq h$, and $\mu^*_i(\cdot|h') = \mu'_i(\cdot|h')$ for all $h' \succeq h$. Thus, there exists $s^*_i \in \rho(\mu^*_i) \subseteq S^\infty_i$ such that $s^*_i|h = s'_i|h$ and $s^*_i(h') = s_i(h')$ for all $h' \not\succeq h$ with $h' \in H(s_i)$. So $s^*_i|\overline{h} \in [s^{\overline{h}}_i]^\infty$. ■

Lemma 3 *Fix a rationalizable agreement $e = (e_i)_{i\in I}$. For each $i \in I$ and $\mu_i \in \Delta^e_i$ t.s.b. $(S^\infty_j)_{j\neq i}$, $[\rho(\mu_i)]^\infty \subseteq S^1_{i,e}$.*[45]

[44] This lemma is not needed under priority to the agreement.

[45] This means that for agreements in this class, such as tight agreements and agreements that correspond to a SES, S3 can be substituted by $s_i \in S^\infty_i$ at the first step. An easy induction argument extends this fact to all steps.

Proof. Fix $s_i \in [\rho(\mu_i)]^\infty \subseteq S_i^\infty$ and $\overline{s}_i \in \rho(\mu_i)$ with $\overline{s}_i(h) = s_i(h)$ for all $h \in H^\infty$. Fix μ_i' t.s.b. $((S_j^q)_{j\neq i})_{q=0}^\infty$ with $s_i \in \rho(\mu_i')$. For each $h \in H^\infty \cap H(\overline{s}_i) = H^\infty \cap H(s_i)$, by $\mu_i(S_{-i}^\infty|h) = 1$, $s_i \in S_i^\infty$, and $\overline{s}_i(h) = s_i(h)$ for all $h \in H^\infty$, also s_i is a continuation best reply to $\mu_i(\cdot|h)$. Fix $h \in H(s_i) \cap H_\infty$. Fix $s_{-i} = (s_j)_{j\neq i} \in S_{-i}(h)$. For each $j \neq i$, if $s_j \notin S_j^\infty$ or $\cup_{\overline{h}\prec h} e_j^{\overline{h}}(h) = \emptyset$, let $s_j' = s_j$. Else, fix $\overline{h} \prec h$ with $e_j^{\overline{h}}(h) \neq \emptyset$. By $e_j^{\overline{h}} = [e_j^{\overline{h}}]^\infty$, there exists $s_j' \in S_j^\infty$ such that $s_j'|\overline{h} \in e_j^{\overline{h}}$ and, by Lemma 2, $s_j'|h = s_j|h$. Let $\eta^h(s_{-i}) = (s_j')_{j\neq i}$. Since $h \in H(S_i^\infty)\backslash H^\infty$, $p(h) \in H^\infty$, and μ_i strongly believes S_{-i}^∞, $\mu_i(S_{-i}(h)|p(h)) = 0$. Then, there exists $\mu_i^* \in \Delta_i^e$ t.s.b. $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $\mu_i^*(\cdot|h') = \mu_i(\cdot|h')$ for all $h' \in H^\infty$, and $\mu_i^*(s_{-i}|h') = \mu_i'((\eta^h)^{-1}(s_{-i})|h')$ for all $h \in H(s_i) \cap H_\infty$, $h' \succeq h$, and $s_{-i} \in S_{-i}(h')$. Thus, $s_i \in \rho(\mu_i^*) \subseteq S_{i,e}^1$. ■

Proof of Proposition 3.

For each $i \in I$, let $\overline{S}_i := \rho(\Delta_i^e) \cap S_i^\infty$. I show that $e^0 \subseteq S_e^\infty$; then, by T3, the result follows from Lemma 1. By T2, $e_i^0 \subseteq \overline{S}_i$ for all $i \in I$. Now I show that $\overline{S}_i \subseteq S_{i,e}^1$.[46] Fix $s_i \in \overline{S}_i$, $\mu_i \in \Delta_i^e$, and μ_i' t.s.b. $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $s_i \in \rho(\mu_i) \cap \rho(\mu_i')$. Fix $h \in H^{\mu_i}$ and $s_{-i} = (s_j)_{j\neq i}$ with $\mu_i(s_{-i}|h) > 0$. For each $j \neq i$, by T1, there is $\overline{h} \preceq h$ such that $\emptyset \neq e_j^{\overline{h}}(h) \subseteq S_j^\infty|\overline{h}$ and, if $h \in H(\overline{S}_j)$, by T2, $e_j^{\overline{h}} \subseteq \overline{S}_j|\overline{h}$. By $\mu_i \in \Delta_i^e$, $s_j|\overline{h} \in e_j^{\overline{h}}$. Thus, there is $s_j' \in S_j^\infty$ such that $s_j'|\overline{h} = s_j|\overline{h} \in e_j^{\overline{h}}$ and, if $h \in H(\overline{S}_j)$, $s_j' \in \overline{S}_j$. Let $\eta^h(s_{-i}) := (s_j')_{j\neq i}$. Fix $h \in H(s_i) \cap H_\infty$ and $s_{-i} = (s_j)_{j\neq i} \in S_{-i}(h)$. Fix $j \neq i$. If $s_j \in S_j^\infty$ and $\cup_{\overline{h}\prec h} e_j^{\overline{h}}(h) \neq \emptyset$, by T1 there is $\overline{h} \prec h$ such that $\emptyset \neq e_j^{\overline{h}}(h) = [e_j^{\overline{h}}(h)]^\infty$ and, if $h \in H(\overline{S}_j)$, by T2, $e_j^{\overline{h}} \subseteq \overline{S}_j|\overline{h}$. By Lemma 2, $e_j^{\overline{h}}|h = S_j^\infty|h$. Thus, there is $s_j' \in S_j^\infty$ such that $s_j'|\overline{h} \in e_j^{\overline{h}}$, $s_j'|h = s_j|h$, and, if $h \in H(\overline{S}_j)$, $s_j' \in \overline{S}_j$. If $s_j \in S_j^\infty$, $\cup_{\overline{h}\prec h} e_j^{\overline{h}}(h) = \emptyset$, and (★) $s_j|h \in \overline{S}_j|h$, pick $s_j' \in \overline{S}_j$ such that $s_j'|h = s_j|h$. Else, let $s_j' := s_j$. Let $\eta^h(s_{-i}) := (s_j')_{j\neq i}$. Since $h \in H(S_i^\infty)\backslash H^\infty$, $p(h) \in H^\infty$, and, by $\mu_i \in \Delta_i^e$ and T1, $\mu_i(\{s_{-i} : s_{-i}|p(h) \in S_{-i}^\infty|p(h)\}|p(h)) = 1$, $\mu_i(S_{-i}(h)|p(h)) = 0$. Thus, there exists $\mu_i^* \in \Delta_i^e$ that strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $\mu_i^*(s_{-i}|h) = \mu_i((\eta^h)^{-1}(s_{-i})|h)$ for all $h \in H^{\mu_i}$ and s_{-i} with $\mu_i((\eta^h)^{-1}(s_{-i})|h) > 0$, and $\mu_i^*(s_{-i}|h') = \mu_i'((\eta^h)^{-1}(s_{-i})|h')$ for all $h \in H(s_i) \cap H_\infty$, $h' \succeq h$, and $s_{-i} \in S_{-i}(h')$. Clearly, $s_i \in \rho(\mu_i^*) \subseteq S_{i,e}^1$. Obviously, $\overline{S}_i \supseteq S_{i,e}^1$. So, $\overline{S} = S_e^1$.

For each $j \neq i$ and $h \in H_\infty \cap H(\overline{S}_j)$, by Lemma 2, $[\overline{S}_j]^\infty|h = S_j^\infty|h$. By T1 and Lemma 3, $[S_{j,e}^1]^\infty = S_{j,e}^1$. So, by $S_{j,e}^1 = \overline{S}_j$, $\overline{S}_j|h = S_j^\infty|h$. Then, $s_j \in S_j^\infty$ implies (★). So, μ_i^* strongly believes also $(\overline{S}_j)_{j\neq i} = (S_{j,e}^1)_{j\neq i}$; hence $s_i \in S_{i,e}^2$. Thus, $e^0 \subseteq \overline{S} = S_e^1 = S_e^2 = S_e^\infty$. ■

[46] Under priority to the agreement, the equality is obvious, but the construction that follows is still needed for the second step.

Proof of Theorem 2.

Define $\overline{S}$ like in Definition 13. I show that $e^0 = S^* \subseteq S_e^\infty$; then, since Self-Enforceability implies T3, the result follows from Lemma 1. By Self-Justifiability $S^* \subseteq \overline{S}$. By Lemma 3, $\overline{S} \subseteq S_e^1$.[47] Obviously, $\overline{S} \supseteq S_e^1$. So, $\overline{S} = S_e^1$. Now I show that $S_e^1 = S_e^\infty$

Fix $i \in I$ and $s_i \in \overline{S}_i \subseteq S_i^\infty$. Fix μ_i' t.s.b. $((S_j^q)_{j\neq i})_{q=0}^\infty$ and μ_i t.s.b. $(S_j^*, \overline{S}_j, S_j^\infty)_{j\neq i}$ such that $s_i \in \rho(\mu_i') \cap \rho(\mu_i)$ (μ_i exists by Forward Induction). Fix $h \in H(s_i) \cap H_\infty$ and $s_{-i} = (s_j)_{j\neq i} \in S_{-i}(h)$. Fix $j \neq i$. If $s_j \notin S_j^\infty$ or $h \notin H(\overline{S}_j)$, let $s_j' := s_j$. Else, by the argument in the proof of Proposition 3, $s_j \in S_j^\infty$ implies $s_j|h \in \overline{S}_j|h$. If $h \in H(S_j^*)$, by Rationalizability and Lemma 2, $s_j|h \in S_j^*|h$ too. So, there is $s_j' \in \overline{S}_j$ such that $s_j'|h = s_j|h$ and, if $h \in H(S_j^*)$, by $S^* \subseteq \overline{S}$, $s_j' \in S_j^*$. Let $\eta^h(s_{-i}) := (s_j')_{j\neq i}$. Since $h \in H(S_i^\infty)\backslash H^\infty$, $p(h) \in H^\infty$, and μ_i strongly believes S_{-i}^∞, $\mu_i(S_{-i}(h)|p(h)) = 0$. Thus, there exists μ_i^* t.s.b. $(S_j^*, \overline{S}_j)_{j\neq i}$.and $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $\mu_i^*(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H^\infty$, and $\mu_i^*(s_{-i}|h') = \mu_i'((\eta^h)^{-1}(s_{-i})|h')$ for all $h \in H(s_i) \cap H_\infty$, $h' \succeq h$, and $s_{-i} \in S_{-i}(h')$. Clearly, $s_i \in \rho(\mu_i^*) \subseteq S_{i,e}^2$. Thus, $S_e^1 = S_e^2 = S_e^\infty$. ■

Proof of Proposition 4.[48] First, I show that S^* is a SES, i.e. that Rationalizable Vetos implies Rationalizability. Fix $i \in I$, $s_i \in S_i^*$, and $s_i' \in [s_i]^\infty$. For each $z \in W_i$, by $z \in \zeta(S^\infty)$, $s_i(h) = s_i'(h)$ for all $h \prec z$. Thus, by $s_i \notin S_i(z)$, $s_i' \notin S_i(z)$. So, $s_i' \in S_i^*$.

Consider now the reduced agreements $e, \overline{e}$ with, for all $i \in I$, $\overline{e}_i^0 = S_i^\infty \backslash \cup_{z\in W_i} S_i(z)$ and $e_i^0 = S_i \backslash \cup_{z\in V_i} S_i(z)$ with $V_i := Z\backslash\zeta(\overline{e}_i^0 \times S_{-i})$. Fix $s_i \in \overline{e}_i^0$. Then, $\zeta(\{s_i\} \times S_{-i}) \cap V_i = \emptyset$. Thus, $s_i \in e_i^0$. So, $\overline{e}_i^0 \subseteq e_i^0$. Fix $z \in \zeta(e_i^0 \times S_{-i})$. Then, $z \notin V_i$. Thus, $z \in \zeta(\overline{e}_i^0 \times S_{-i})$. So, $\zeta(e_i^0 \times S_{-i}) \subseteq \zeta(\overline{e}_i^0 \times S_{-i})$. Then, by $\overline{e}_i^0 \subseteq e_i^0$, $H(e_i^0) = H(\overline{e}_i^0)$, and so $\Delta_j^{\overline{e}} \subseteq \Delta_j^e$ for all $j \in I$. Fix $s_i \in e_i^0 \cap S_i^\infty$. For every $z \in Z$ with $s_i \in S_i(z)$, $z \notin V_i \supseteq W_i$. Then, by $s_i \in S_i^\infty$, $s_i \in \overline{e}_i^0$. Thus, by $H(e_i^0) = H(\overline{e}_i^0) \subseteq H(S_i^\infty)$, for each $\mu_j \in \Delta_j^e$ t.s.b. $(S_i^\infty)_{i\neq j}$, $\mu_j \in \Delta_j^{\overline{e}}$. So, $\overline{e}$ and e are equivalent under S3. By Theorem 2, $\overline{e}$ implements $\zeta(S^*)$. So, e too. ■

Proof of Proposition 5.[49] Strict Nash obviously implies Self-Enforceability. Fix $i \in I$. For each $s_i \in S_i^\infty$ and $s_{-i} \in r_{-i}(s_i)$, $\zeta(s_i, s_{-i}) \in \zeta(S^\infty)$. Then, S_i^* is the set of all $s_i \in S_i^\infty(z)$ such that for each $\widehat{z} \in \zeta(S^\infty)$ with $u_{-i}(\widehat{z}) \geq u_{-i}(z)$, $s_i \notin S_i(\widehat{z})$. Thus,

[47] Under priority to the agreement, the equality is obvious.

[48] Under priority the agreement, just observe that (i) Rationalizability has no bite, so S^* is a SES, and (ii) the candidate implementing agreement on actions corresponds to the SES itself, so by Theorem 2 it does implement $\zeta(S^*)$.

[49] Under priority to the agreement, substitute S^∞ with S, and not with S^1.

Rationalizable Vetos holds. Define $\overline{S}$ like in Definition 13. Fix $s_i^* \in S_i^* \cup \overline{S}_i$ and μ_i t.s.b. S_{-i}^∞ such that $s_i^* \in \rho(\mu_i)$. Fix any μ_i^* t.s.b. S_{-i}^* and S_{-i}^∞ such that $\mu_i^*(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(s_i^*) \backslash H(S_{-i}^*)$, and $\mu_i^*(\overline{S}_{-i}|h) = 1$ for all $h \in H(\overline{S}_{-i}) \backslash (H(s_i^*) \cup H(S_{-i}^*))$. Clearly, $s_i^* \in \rho(\mu_i^*)$. So, $S^* \subseteq \overline{S}$, i.e. Self-Justifiability holds. Moreover, by Self-Enforceability, $H(\overline{S}) = H(S(z))$. Thus, $(H(\overline{S}_i) \backslash H(S_{-i}^*)) \cap H(\overline{S}_{-i}) = \emptyset$. Then, μ_i^* strongly believes also $\overline{S}_{-i}$. So, Forward Induction holds.

Fix $s^* \in S^*$ and let $e, \overline{e}$ be the reduced agreements with $e^0 = \{s^*\}$ and $\overline{e}^0 = \{S^*\}$. Fix $n \in \mathbb{N}$ and suppose to have shown that $S^* \subseteq S_{\overline{e}}^{n-1} = S_e^{n-1}$. Then, for each $i \in I$ and μ_i t.s.b. $\{s_{-i}^*\}$, $(S_{-i}^q)_{q=0}^\infty$, and $(S_{-i,e}^q)_{q=0}^{n-1}$, there is μ_i' t.s.b. S_{-i}^*, $(S_{-i}^q)_{q=0}^\infty$, and $(S_{-i,e}^q)_{q=0}^{n-1}$ with $\mu_i(\cdot|h) = \mu_i'(\cdot|h)$ for all $h \notin H(S_{-i}(z))$, and vice versa. For all $h \prec z$, $r_i(\mu_i(\cdot|h)) = S_i(z) = r_i(\mu_i'(\cdot|h))$, and for all $h \in H(S_i(z))$ with $h \not\prec z$, $h \notin H(S_{-i}(z))$. Thus, $\rho(\mu_i) = \rho(\mu_i')$. So, $S_e^n = S_{\overline{e}}^n$, and by Lemma **??**, $S^* \subseteq S_{\overline{e}}^n$. Hence, by Theorem 2, $\{z\} = \zeta(S_{\overline{e}}^\infty) = \zeta(S_e^\infty)$. Then, by Proposition 2, e is self-enforcing, and by $\{z\} = \zeta(S_e^\infty)$, truthful. ■

Lemma 4 *Fix an agreement e, a finite chain of Cartesian sets of strategy profiles $S = \overline{S}^0 \supset ... \supset \overline{S}^M \neq \emptyset$ and $L \leq M$ such that for all $i \in I$ and $s_i \in S_i$,*

1. *if $L \neq 0$, $s_i \in \overline{S}_i^L$ if and only if $s_i \in \rho(\mu_i)$ for some μ_i t.s.b. $((\overline{S}_j^q)_{j \neq i})_{q=0}^L$;*
2. *$s_i \in \overline{S}_i^M$ if and only if $s_i \in \rho(\mu_i)$ for some $\mu_i \in \Delta_i^e$ t.s.b. $((\overline{S}_j^q)_{j \neq i})_{q=0}^M$.*

Define $[\cdot]^L$ as $[\cdot]^\infty$ with $(\overline{S}_i^L)_{i \in I}$ in place of $(S_i^\infty)_{i \in I}$. Suppose that $\zeta(\overline{S}^M) = \zeta(\overline{S}^M \cap e^0)$. Then there exists an agreement $\overline{e}$ with $\zeta(\overline{e}^0) = \zeta(\overline{S}^M)$ which satisfies T1, T2, and T3 with L in place of ∞.

Proof. Let $H^L := H(\overline{S}^L)$ and $H_L := \{h \notin H^L : p(h) \in H^L\}$. Construct an agreement with the following inductive procedure. Let e^0 be the reduced agreement with $e_i^{0,0} := \overline{S}_i^M \cap e_i^0 \neq \emptyset$ for all $i \in I$. Fix $n > 0$ and suppose to have defined an agreement e^{n-1}. Fix $i \in I$ and let

$$H' := \left\{ h \in H^L : \cup_{\overline{h} \prec h} e_i^{\overline{h}, n-1}(p(h)) \neq \emptyset = \cup_{\overline{h} \preceq h} e_i^{\overline{h}, n-1}(h) \right\}.$$

For each $h \notin H'$, let $e_i^{n,h} := e_i^{n-1,h}$. Now fix $h \in H'$ and let $m := \max\{q \geq L : h \in H(\overline{S}_i^q)\}$. If there is $\overline{h} \preceq h$ with $e_i^{\overline{h}}(h) \neq \emptyset$, let $e_i^{n,h} := ((\overline{S}_i^m | \overline{h}) \cap e_i^{\overline{h}})|h$ (non-empty by $\overline{S}^M \neq \emptyset$

and 2.); else, let $e_i^{n,h} := \overline{S}_i^m|h$. Since histories in H' are unordered, $(e_j^n)_{j\in I}$ is an agreement. By finiteness of the game, the procedure stops at some e^K. Define $\overline{e}$ as, for each $i \in I$ and $h \in H$, $\overline{e}_i^h = [e_i^{h,K}]^L$ if $h \in H^L$ and $\overline{e}_i^h = \emptyset$ else. Then, by construction, $\overline{e}^h$ satisfies T1.

Fix $i \in I$ and $\mu_i \in \Delta_i^e$ t.s.b $((\overline{S}_j^q)_{j\neq i})_{q=0}^M$. Fix $h \in H^L$, $s_{-i} = (s_j)_{j\neq i}$ with $\mu_i(s_{-i}|h) > 0$, and $j \neq i$. By T1, there is h'' with $\overline{e}_j^{h''}(h) \neq \emptyset$. If there is $h' \preceq h$ with $e_i^{h'}(h) \neq \emptyset$, $s_j|h' \in e_j^{h'}$, and by construction of $\overline{e}$, $h'' \succeq h'$. So, since $s_j \in \overline{S}_j^m$ for all m with $\overline{S}_j^m(h) \neq \emptyset$, $s_j|h'' \in \overline{e}_j^{h''}$. Then, since by T1 $\overline{e}_j^{h''} = \emptyset$ for all $h'' \notin H^L$, $\mu_i \in \Delta_i^{\overline{e}}$ ($\bigstar$).

Fix $i \in I$ and $\mu_i \in \Delta_i^{\overline{e}}$. Fix $s_i \in \rho(\mu_i)$. Let

$$H^{L,\mu_i} := \left\{h^0\right\} \cup \left\{h \in H^L : \mu_i(S_{-i}(h)|p(h)) = 0\right\}.$$

For each $h \in H^{L,\mu_i}$ and $s_{-i} = (s_j)_{j\neq i}$ with $\mu_i(s_{-i}|h) > 0$, by construction of $\overline{e}$, there exists $\eta^h(s_{-i}) = (s'_j)_{j\neq i}$ such that, for all $j \neq i$: (i) $s'_j(h') = s_j(h')$ for all $h' \in H^L \cap H(s_j)$ with $h' \succeq h$; (ii) $s'_j \in \overline{S}_j^m$ for all $m \geq L$ with $\overline{S}_j^m(h) \neq \emptyset$; (iii) if there is $\overline{h} \preceq h$ with $e_j^{\overline{h}}(h) \neq \emptyset$, $s'_j|\overline{h} \in e_j^{\overline{h}}$. Construct any $\mu_i^* \in \Delta_i^e$ t.s.b. $((\overline{S}_j^q)_{j\neq i})_{q=0}^M$ such that $\mu_i^*(s_{-i}|h) = \mu_i((\eta^h)^{-1}(s_{-i})|h)$ for all $h \in H^{L,\mu_i}$ and s_{-i} with $\mu_i((\eta^h)^{-1}(s_{-i})|h) > 0$. By 1., for any $\widetilde{\mu}_i$ t.s.b. $\overline{S}_{-i}^L$ and $\widetilde{s}_i \in \rho(\widetilde{\mu}_i)$, $\zeta(\{\widetilde{s}_i\} \times \overline{S}_{-i}^L) \subseteq \zeta(\overline{S}^L)$. By (i), $\mu_i(S_{-i}(z)|h) = \mu_i^*(S_{-i}(z)|h)$ for all $h \in H^L$ and $z \in \zeta(\overline{S}^L)$ with $z \succeq h$. Thus, there exists $s_i^* \in \rho(\mu_i^*) \subseteq \overline{S}_i^M$ (by 2.) such that $s_i^*(h) = s_i(h)$ for all $h \in H^L \cap H(s_i)$ ($\blacktriangle$). So, $H(\rho(\Delta_i^{\overline{e}})) \cap H^L \subseteq H(\overline{S}_i^M) \cap H^L$ ($\blacktriangledown$).

Fix $\overline{\mu}_i$ t.s.b. $\overline{e}_{-i}^0$ and $\overline{s}_i \in \rho(\overline{\mu}_i)$. Clearly, there exist $\mu_i \in \Delta_i^{\overline{e}}$ with $\mu_i(\cdot|h) = \overline{\mu}_i(\cdot|h)$ for all $h \in H(\overline{e}_{-i}^0)$, and $s_i \in \rho(\mu_i)$ such that $s_i(h) = \overline{s}_i(h)$ for all $h \in H(\overline{e}_{-i}^0)$. By ($\blacktriangle$), there exists $s_i^* \in \overline{S}_i^M$ such that $s_i^*(h) = s_i(h) = \overline{s}_i(h)$ for all $h \in H^L \supseteq H(\overline{S}_i^M \times \overline{e}_{-i}^0)$. So,

$$\zeta(\{\overline{s}_i\} \times \overline{e}_{-i}^0) = \zeta(\{s_i^*\} \times \overline{e}_{-i}^0) = \zeta(\{s_i^*\} \times e_{-i}^{0,K}) \subseteq \zeta(\overline{S}^M) = \zeta(e^{0,K}) = \zeta(\overline{e}^0),$$

where the second equality holds by $s_i^* \in \overline{S}_i^L$ and $\zeta(\overline{S}_i^L \times \overline{e}_{-i}^0) = \zeta(\overline{S}_i^L \times e_{-i}^{0,K})$, the inclusion by $s_i^* \in \overline{S}_i^M$ and $e_{-i}^{0,K} \subseteq \overline{S}_{-i}^M$, the penultimate equality by $\zeta(\overline{S}^M) = \zeta(\overline{S}^M \cap e^0)$ and $\overline{S}^M \cap e^0 = e^{0,K}$, and the last equality by construction of $\overline{e}$. So, $\overline{e}^0$ satisfies T3.

Fix $\overline{h} \in H(\rho(\Delta_i^{\overline{e}}) \cap \overline{S}_i^L)$ with $\overline{e}_i^{\overline{h}} \neq \emptyset$. Then, by T1, $\overline{h} \in H^L$. Thus, by ($\blacktriangledown$), $\overline{h} \in H(\overline{S}_i^M)$. Fix $s_i^{\overline{h}} \in \overline{e}_i^{\overline{h}}$. By construction of $\overline{e}$, there is $s_i \in \overline{S}_i^M(\overline{h})$ such that $s_i(h) = s_i^{\overline{h}}(h)$ for all $h \in H^L \cap H(s_i)$ with $h \succeq \overline{h}$. By 2., there exists $\mu_i \in \Delta_i^e$ t.s.b

$((\overline{S}_j^q)_{j\neq i})_{q=0}^M$ such that $s_i \in \rho(\mu_i)$. By ($\bigstar$), $\mu_i \in \Delta_i^{\overline{e}}$. If $L=0$, let $\mu_i^* := \mu_i$. If $L \neq 0$, by T1 there is $s_i' \in \overline{S}_i^L$ with $s_i'|\overline{h} = s_i^{\overline{h}}$. By 1., there is μ_i' t.s.b. $((\overline{S}_j^q)_{j\neq i})_{q=0}^L$ such that $s_i' \in \rho(\mu_i')$. Fix $h \in H' := H(s_i^{\overline{h}}) \cap H_L$ and $s_{-i} = (s_j)_{j\neq i} \in S_{-i}(h)$. Fix $j \neq i$. If $s_j \notin \overline{S}_j^L$ or $\cup_{h' \prec h} \overline{e}_j^{h'}(h) = \emptyset$, let $s_j' := s_j$. Else, by T1 and Lemma 2 with L in place of ∞, there is $h' \prec h$ with $\overline{e}_j^{h'}|h = \overline{S}_j^L|h$. Thus, there exists $s_j' \in \overline{S}_j^L$ such that $s_j'|h = s_j|h$ and $s_j'|h' \in \overline{e}_i^{h'}$. Let $\eta^h(s_{-i}) := (s_j')_{j\neq i}$. Since μ_i strongly believes $\overline{S}_{-i}^L$, $\mu_i(S_{-i}(h)|p(h)) = 0$ for all $h \in H'$. Hence, there exists $\mu_i^* \in \Delta_i^{\overline{e}}$ t.s.b. $((\overline{S}_j^q)_{j\neq i})_{q=0}^L$ such that $\mu_i^*(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H^L$, and $\mu_i^*(s_{-i}|h') = \mu_i'((\eta^h)^{-1}(s_{-i})|h')$ for all $h \in H'$, $h' \succeq h$, and $s_{-i} \in S_{-i}(h')$. Thus, there exists $s_i^* \in \rho(\mu_i^*) \subseteq \rho(\Delta_i^{\overline{e}}) \cap \overline{S}_i^L$ such that $s_i^*|\overline{h} = s_i^{\overline{h}}$. Hence, $\overline{e}_i^{\overline{h}} \subseteq (\rho(\Delta_i^{\overline{e}}) \cap \overline{S}_i^L)|\overline{h}$. So, $\overline{e}$ satisfies T2. ■

Proof of Theorem 1. The "if" part coincides with Proposition 3. For the "only if" part, fix an implementable outcome set $P \subseteq Z$ and an agreement e with $\zeta(S_e^\infty) = \zeta(S_e^\infty \cap e^0) = P$. Apply Lemma 4 with[50] $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^L, (S_e^q)_{q=1}^K)$, where L and K are the smallest l and k such that $S^l = S^{l+1}$ and $S_e^k = S_e^{k+1}$.[51] ■

Proof of Proposition 6. Fix an implementable outcome set $P \subseteq Z$ under priority to rationality, and an implementing agreement e. Since e is self-enforcing under priority to rationality, I can apply Lemma 4 with $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^D, (S_e^q)_{q=1}^K)$, where D and K are the smallest d and k such that $S^d = S^{d+1}$ and $S_e^k = S_e^{k+1}$, and $L=0$. The obtained agreement $\overline{e}$ is tight under priority to the agreement. Thus, by Proposition 3 and Remark 3, $\overline{e}$ implements P under priority to the agreement. ■

Proof of Proposition 8. Fix an implementable outcome set $P \subseteq Z$ under priority to the path, and an implementing agreement e. Since e is self-enforcing under priority to the path, I can apply Lemma 4 with $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^L (S_z^q)_{q=1}^D, (S_{e^z}^q)_{q=1}^K)$, where L, D and K are the smallest l, d, and k such that $S^l = S^{l+1}$, $S_z^d = S_z^{d+1}$, and $S_{e^z}^k = S_{e^z}^{k+1}$. The obtained agreement $\overline{e}$ is tight under priority to rationality. Thus, by Proposition 3, $\overline{e}$ implements P under priority to rationality. ■

[50] Here Selective Rationalizability is merged with Rationalizability into a unique elimination procedure.

[51] Under priority to the agreement, let $(\overline{S}^q)_{q=0}^M = (S_{\Delta^e}^q)_{q=0}^M$ and $L=0$, where M is the smallest m such that $S_{\Delta^e}^m = S_{\Delta^e}^{m+1}$. Under priority to the path, let $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^B (S_z^q)_{q=1}^D, (S_{e^z}^q)_{q=1}^K)$ and $L = B + D$, where B, D and K are the smallest b, d, and k such that $S^b = S^{b+1}$, $S_z^d = S_z^{d+1}$, and $S_{e^z}^k = S_{e^z}^{k+1}$.

References

[1] Aumann, R. "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica*, **55**, 1987, 1–18.

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

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

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

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

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

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

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

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

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

[11] Catonini, E., "Rationalizability, order independence, and subgame perfection", working paper, 2016.

[12] Catonini, E., "Selecting strongly rationalizable strategies", working paper, 2016.

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

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

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

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

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

[18] Harrington, J. "A Theory of Collusion with Partial Mutual Understanding", working paper, 2016.

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

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

[21] Kreps, D.M. and R. Wilson, "Sequential equilibria", *Econometrica*, **50**, 1982, 863-94.

[22] Green, J. R., Mas-Colell, A., and Whinston, M., *Microeconomic Theory*, Oxford University Press, 2006.

[23] Man, P. "Forward Induction Equilibrium", *Games and Economic Behavior*, **75(1)**, 2012, 265-276.

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

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

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

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

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

[29] Siniscalchi, M., "Structural Rationality in Dynamic Games", working paper, 2016.

[30] Van Damme, E. "Stable Equilibria and Forward Induction", *Journal of Economic Theory*, **48**, 1989, 476–496.