# Self-enforcing agreements and forward induction reasoning*

**Emiliano Catonini**†

April 2019

**Abstract**

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 the deviator's 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 implementable by some agreement or not. Conclusions depart substantially from what the traditional equilibrium refinements suggest. A non subgame perfect equilibrium outcome may be induced by a self-enforcing agreement, while a SPE outcome may not. Agreement incompleteness can be crucial to implement an outcome.

**Keywords:** Agreements, Self-Enforceability, Forward Induction.

*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, Nicodemo De Vito, Alfredo Di Tillio, Amanda Friedenberg, Giacomo Lanzani, Andres Perea, Burkhard Schipper, Madhav Shrihari Aney, Satoru Takahashi, Elias Tsakas, two anonymous referees, and to all the attendants of the Stratemotions workshop (Università Bocconi, December 2017), of the ESEM-EEA conference 2017, of the Barcelona GSE Workshop (June 2018), and of the seminars at NUS, SMU, Oxford University and Maastricht University.

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

# 1 Introduction

In many economic situations, agents can communicate before they start to act. Players with strategic power may exploit this opportunity to coordinate on some desirable outcome, or to influence other players' behavior by announcing publicly how they plan to play. I will refer to the common, possibly partial understanding of how each player is expected to play as an *agreement*. In most cases, the context allows players 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. When the game is dynamic, even if players tentatively trust the agreement at the outset, they are likely to question this trust and revise their beliefs based on strategic reasoning and the observed behavior. The fact that an agreement is in place can modify the interpretation of unexpected behavior. All this can be decisive for the incentives to fight or accommodate a deviation from the agreed-upon play. Taking these forward induction considerations into account, 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 enforced by *some* agreement. The paper will not deal with the pre-play communication phase. Yet, assessing their credibility has a clear feedback on which agreements are likely to be reached.

In static games, it is well-known that Nash equilibrium characterizes the set of action profiles that can be played as the result of a non-binding agreement, reached at a pre-play round of cheap talk communication.[1] In dynamic games, this role is usually assigned to Subgame Perfect Equilibrium (henceforth, SPE). Because SPE induces a Nash equilibrium in every subgame, this seems *prima facie* a sensible choice. But does SPE truly characterize self-enforcing agreements in dynamic games?

Relevant economic decisions can seldom be interpreted as unintentional mistakes. A deviation from an equilibrium path can safely be interpreted as

[1] Nevertheless, Aumann [2] provides an argument against this view.

disbelief in some features of the equilibrium. Often, it clearly displays confidence that none of the adverse re-coordination scenarios will realize. Then, credible threats are not the ones that rely on illusory re-coordination, but those that best respond to the potentially profitable continuation plans of the deviator. Indeed, compliance with non-binding agreements often relies on the threat/concern that a deviation will provoke the end of coordinated play, rather than less advantageous re-coordination. Moreover, agreements are often incomplete: differently than a SPE, they do not pin down exactly what to do in every contingency. Partially conflicting interests, legal constraints, social taboos, unilateral communication channels (e.g., the announcements of a central bank), anticipated distrust or objective impossibility of credible re-coordination after deviations: these are only some of the reasons why players may be unable or unwilling to reach a complete agreement. In economic applications, absence of equilibrium is often blamed on a misspecification of the model, rather than on the objective impossibility to reach a precise agreement among players, and the models have been modified to allow for a SPE solution. I will provide a solution method that, while retaining high predictive power, is flexible enough to avoid this means-ends inversion.

To illustrate these insights in a meaningful economic environment, in Section 2 I analyze an entry game in monopolistic competition. Depending on the value of the entry cost, SPE turns out to be too permissive, too restrictive, or simply inadequate to evaluate the credibility of the incumbent's threats.

That SPE can be too permissive is not a new observation. Classical examples, such as the battle of the sexes with an outside option (Ben Porath and Dekel [14]), have already shown this point. This paper captures these refinement arguments in a simple and general way.

That SPE can be too restrictive may instead sound surprising, therefore I sketch here the intuition behind this observation. Consider the following game.

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

$\longrightarrow$

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

In the first stage, Ann and Bob can potentially coordinate on two outcomes, $(N, W)$ and $(S, E)$. If they fail to coordinate, the game either ends (after $(S, W)$),[2] or moves to a second stage (after $(N, E)$), where in the unique equilibrium all actions are played with equal probability. So, the unique SPE of the game induces outcome $(S, E)$. But $(S, E)$ is Pareto-dominated by $(N, W)$; hence, Ann and Bob would rather reach an agreement that induces $(N, W)$. So, they agree to play $(N, W)$ and that Ann should play $U$ in case of deviation of Bob.[3] Is the agreement credible? If Bob is rational[4] and believes in the agreement, he has no incentive to deviate. Then, after a 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 $L$. Hence, she can react with $U$. Anticipating this, Bob can believe in $U$ 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. Therefore, the agreement is *self-enforcing*.

The further inadequacy of SPE comes from the intrinsic assumption of agreement completeness. In the entry game of Section 2, for intermediate values of the entry cost, the most realistic threat by the incumbent does not completely specify its plan, therefore its credibility cannot be evaluated through SPE. Moreover, the *complete* agreement on the SPE that deters entry is not credible, because the continuation plan of the entrant is not part of any rational entry plan. Yet, the SPE threat is credible, and its credibility relies on the (physiological) uncertainty regarding the behavior of the entrant. Sometimes, an outcome can be achieved only by not fully specifying the reactions to deviations either: see Section 4.3.

[2] This is just to keep the game small: it could continue in a symmetric way with respect to after $(N, E)$, and the analysis would not change.

[3] To keep the game small, the Nash threat $U$ that sustains $(N, W)$ is also played with positive probability in the SPE. This is by no means necessary for its credibility: in the Supplemental Appendix, I present a similar game where the Nash outcome is sustained by a credible threat that differs from the unique equilibrium action of the subgame.

[4] The notion of rationality employed in this paper simply requires expected utility maximization, without imposing by itself any restriction on beliefs. See Section 3 for details.

In section 3, I model agreements with *sets* of *plans of actions*, as opposed to one profile of strategies, from which players are expected to choose. Per se, plans of actions (also known as "reduced strategies") already feature a basic form of incompleteness: they do not prescribe moves after a deviation from the plan itself. However, an agreement can also specify alternative plans that players are expected to follow after deviations from the own primary plans (and so on, in a lexicographic fashion). For notational simplicity, I restrict the attention to the class of finite games with complete information, observable actions,[5] and no chance moves. However, the methodology can be applied to all dynamic games with perfect recall and countably many information sets, hence possibly infinite horizon.

In Section 4, I study credibility and self-enforceability of agreements starting from primitive assumptions about players' strategic reasoning.

An agreement is *credible* when players may comply with it in case they are rational, they believe in the agreement, they believe as long as possible that co-players are rational and believe in the agreement, and so on. When a player's move is not rational under belief in the agreement (such as Bob's deviation to $E$ in the example above), I assume that co-players keep the belief that the player is rational (if per se compatible with the observed move) and drop the belief that the player believes in the agreement.[6] Under this reasoning scheme, deviations, or more generally past actions, are not interpreted as mistakes but as intentional choices. To see this clearly, suppose that in the game above Ann and Bob agree on $(S, E)$, without specifying what to do in case of Ann's deviation. If Ann believes in $E$, she has the incentive to deviate to $N$ only if she expects $R$ with sufficiently high probability. Then, Bob expects her to play $D$ after the deviation.[7] This instance of forward induction reasoning is

[5] 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.

[6] This appears as the most sensible choice given the cheap-talk nature of the agreement.

[7] This induces Bob to play $L$ and thus Ann not to deviate from $S$. Therefore, the SPE outcome $(S, E)$ obtains without explicit threats. This is by no means a general property of a SPE, not even when unique: see the modification of the game in the Supplemental Appendix.

based not just on the belief in Ann's rationality, but also on the belief that Ann believes in the agreement.

Under a credible agreement, the outcomes players *should* reach (according to the agreement) and *might* reach (according to strategic reasoning) overlap but need not be nested. I will refer to the former as the outcome set the agreement *prescribes*, and to the latter as the outcome set the agreement *induces*. A credible agreement is *self-enforcing* when it induces a subset of the outcomes it prescribes.

A set of outcomes is *implementable* when it is induced by a self-enforcing agreement. I provide necessary and sufficient conditions for the implementability of an outcome set. An outcome set is implementable if it is prescribed by a *Self-Enforcing Set* of plans (henceforth, SES). SES's are self-enforcing agreements that do not require players to promise, and co-players trust, what they would do after a own deviation. Thus, they can be seen as a set-valued counterpart of SPE where the behavior of deviators is not exogenously given but determined by forward induction. In games with two players or two stages, every implementable outcome set is induced by a SES. For a single outcome of a two-players game, SES's boil down to Nash equilibria in extensive-form rationalizable[8] plans that satisfy a strictness condition.[9] To complete the search for implementable outcomes in games with more than two players and stages, *tight agreements* augment SES's by restricting the expected behavior of deviators. An outcome set is implementable if and only it is prescribed by a tight agreement. Since a tight agreement induces exactly the outcome set it prescribes, we have a "revelation principle" for agreements design: players need not be vague about the outcomes they want to achieve.

Tight agreements and SES's have the double value of *solution concepts* and "soft mechanisms" for implementation,[10] because they prescribe directly the outcome set they induce. They provide to the analyst (or a mediator) all possi-

[8] The original notion of extensive-form-rationalizability is due to Pearce [32], and has been later refined by Battigalli [5] and Battigalli and Siniscalchi [11].

[9] Every feature of this simple characterization is not assumed, but derived from first principles.

[10] "Soft" in the sense that they not modify the rules of the game, they only act via beliefs.

ble predictions (and an implementation strategy) under the non-binding agreements motivation, abstracting away from the foundations of self-enforceability. In particular, after a standard elimination procedure (extensive-form rationalizability), they only require to verify one-step conditions instead of doing all steps of reasoning under all candidate self-enforcing agreements.

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-$\Delta$-Rationalizability* (Battigalli, [7]; Battigalli and Siniscalchi, [12]). Strong-$\Delta$-Rationalizability does *not* require players to maintain belief in the rationality of the co-players when their behavior cannot be optimal under their first-order belief restrictions. To define self-enforceability under the different hypotheses of this paper, another elimination procedure with belief restrictions, *Selective Rationalizability*, is constructed and analyzed epistemically in the companion paper ([17]). Selective rationalizability captures *common strong belief in rationality* (Battigalli and Siniscalchi [11]), i.e., the hypothesis that each order of belief in rationality holds as long as not contradicted by the observed behavior. Thus, it combines "unrestricted" (i.e., based only on beliefs in rationality) and "restricted" (i.e., based also on first-order belief restrictions) forward induction reasoning. The structure given by agreements to the belief restrictions, the *epistemic priority* attributed to the beliefs in rationality, and the requirement of self-enforceability sharpen the predictions of this paper with respect to *Extensive-Form Best Response Sets* (Battigalli and Friedenberg [8]), which capture the predictions of Strong-$\Delta$-Rationalizability across all first-order belief restrictions. For instance, competition among firms on quality, quantity, or price often leads to a small set or a singleton solution only through strategic reasoning about rationality (see cobweb stability or Cournot duopoly). This predictive power would be lost with Strong-$\Delta$-Rationalizability in subgames that cannot be rationally reached under belief in the agreement. In Section 5, I expand on this comparison and I revise the results of Section 4 under Strong-$\Delta$-Rationalizability.

When the agreement prescribes a specific outcome, another reasonable (but

less agnostic) way to interpret deviations is that the deviator believed in the agreed-upon path (i.e., that co-players would have complied with it), but does not believe in the threats. In Section 6, I provide an example where this stricter rationalization of deviations matters, and I show that a simple revision of the methodology accommodates it. All the general insights of the paper are robust to these stricter strategic reasoning hypotheses, although they refine the set of implementable outcomes.

Strategic stability à la Kohlberg and Mertens [27] and related refinements are often justified with stories of forward induction reasoning that involve the equilibrium path as a focal point. However, understanding and applying stability and related refinements presents various difficulties.[11] Stability is hard to interpret and verify, and does not offer an implementation strategy: what should players exactly agree on/believe in? Later refinements focus exclusively on sequential equilibria and, to simplify the analysis, sacrifice depth of reasoning (e.g., forward induction equilibria of Govindan and Wilson [22] capture only strong belief in rationality[12]) or scope (e.g., the intuitive criterion of Cho and Kreps [19] and divine equilibrium of Banks and Sobel [3] are tailored on signaling games). Overall, the motivation for equilibrium is unspecified and the language does not allow to talk of incomplete agreements. Then, the analysis of Section 6 can also be seen as a general and transparent approach to the forward induction stories in the background of this literature. It turns out that the spirit of subgame perfection (i.e., the idea that a deviator will best reply to the equilibrium strategies after the deviation) is at contradiction precisely with this kind of forward induction reasoning.

The Appendix collects the proofs of the results of Section 4, which can be replicated under the alternative strategic reasoning hypotheses of Sections 5 and 6. Other results from Sections 5 and 6 are proved in the Supplemental Appendix, which also contains other examples and technical remarks that

[11] An interesting critique of this kind to strategic stability has been put forward by Van Damme [37].

[12] See [22], pag. 11 and 21. An explicit example of this fact is provided by Perea ([33], pag. 509).

can be useful to whoever wishes to develop (as opposed to just apply) the methodology.

# 2 An example

Consider the following linear city model of monopolistic competition. Two firms, $i = 1, 2$, sell the same good at the extremes of a continuum of potential buyers of measure 48. The payoff of buyer $j \in [0, 48]$ when she buys from firm $i$ is $u - p_i - t \cdot d_{ij}$, where $u = 72$ is the utility from the good, $p_i$ is the price fixed by firm $i$, $t = 1/2$ is the transportation cost, and $d_{ij}$ is the distance from firm $i$: $d_{1j} = j$ and $d_{2j} = 48 - j$. Then, firm $i$ faces demand

$$D_i(p_i, p_{-i}) = \max\{0, \min\{48, 24 - p_i + p_{-i}, 2 \cdot (72 - p_i)\}\}$$

(see Green et al. [29] for more details). There are two technologies: $k = 1$, with marginal cost $mc = 48$ and no fixed cost; and $k = 2$, with no marginal cost and fixed cost $F$ such that firm $i$ is indifferent between the two technologies for $p_{-i} = 24$.[13] Suppose that firms choose technology and price simultaneously (or equivalently, do not observe each other's technology choice before fixing prices). Then, for $p_{-i} \in [24, 72]$, firm $i$'s the best response correspondence is

$$\widehat{p}_i(p_{-i}) = \begin{cases} 36 + \frac{1}{2}p_{-i} \quad (\text{with } k = 1) \quad \text{if } p_{-i} < 48 \\ \{36, 60\} \quad (\text{with } k = 2, 1) \quad \text{if } p_{-i} = 48 \\ 12 + \frac{1}{2}p_{-i} \quad (\text{with } k = 2) \quad \text{if } p_{-i} > 48 \end{cases}.$$

For $p_{-i} < 24$, firm $i$ has no incentive to produce. For $p_{-i} > 72$, firm $i$'s demand depends only on $p_i$ and the optimal value is 48 with $k = 2$. With $k = 1$, every price above 60 is dominated by 60 and every price below 48 generates losses; with $k = 2$, every price above 48 is dominated by 48. Finally, for any conjecture

[13] In the long run, this generates a cost function with economies of scale, linear up to $q = 24$ and flat thereafter. I will focus on the short run, where the technology cannot be modified after observing prices, thus the demand. However, the analysis would be almost identical using the long run cost function in place of the technology choice.

$\mu \in \Delta(p_{-i})$, no $p_i < \mathbb{E}_\mu(p_{-i})$ can do better than $\widehat{p}_i(\mathbb{E}_\mu(p_{-i}))$, because demand cannot increse above the upper bound and is linear in $p_{-i}$ below it. Hence, the rational prices are $[36, 60]$. For each $(p_i, p_{-i}) \in [36, 60]^2$, firm $i$'s demand is $24 - p_i + p_{-i}$, thus the best replies to $\mu \in \Delta(p_{-i})$ are $\widehat{p}_i(\mathbb{E}_\mu(p_{-i}))$. Hence, the rationalizable prices of each firm (*in the static game*) are $[36, 42]$ (with $k = 2$) and $[54, 60]$ (with $k = 1$). The pure equilibrium price pairs are $(40, 56)$ and $(56, 40)$, and the unique mixed equilibrium assigns probability $1/2$ to 40 and 56 for both firms. Profits are increasing in the other firm's price, so let $\pi^1 > \pi^2 > \pi^3$ denote the profit of firm 2 in the three equilibria ($\pi^2$ is the expected profit in the mixed equilibrium), and let $\pi^4$ denote firm 2's profit when $p_1 = 36$.

Suppose now that firm 1 is already in the market, while firm 2 still has to pay an entry cost $E$. If firm 2 does not enter, its profit is 0. Can firm 1 deter the entry of firm 2 by announcing how it plans to react?[14] I am going to tackle this question for different values of the entry cost $E$.

**Case 1)** $\pi^2 < E < \pi^1$ **(SPE is too permissive).** According to SPE, entry is deterred by two equilibria of the subgame that follows it. But if firm 2 is rational and expects firm 1 to react rationally, firm 2 will enter only if the expected $p_1$ is not lower than some $\widetilde{p} > 48$ (that depends on $E$). Then, entry displays firm 2's intention to employ $k = 2$ with some $p_2 \in \left[12 + \frac{1}{2}\widetilde{p}, 42\right]$. Given this, firm 1 has the incentive to choose $k = 1$ and $p_1 \in \left[42 + \frac{1}{4}\widetilde{p}, 60\right]$, thus $p_1 > \widetilde{p}$. So, firm 2 has always the incentive to enter. With further steps of

[14] Dixit [20] studies entry deterrance through an irreversible investment in productive capacity. Interestingly, Dixit motivates his analysis with the following observations: "The theory of large-scale entry into an industry is made complicated by its game-theoretic aspects. Even in the simplest case of one established firm facing one prospective entrant, there are subtle strategic itneractions. [...] In reality, there may be no agreement about the rules of the post-entry duopoly, and there may be periods of disequilibrium before any order is established."

However, an incumbent may want to avoid wasteful investment and only threaten a reaction without costly commitment actions. Then, the credibility of her claims must be assessed. Note that our incumbent would use $k = 2$ already as a monopolist; it would then be interesting to consider an intermediate situation between Dixit [20] and this example where dismissal of productive capacity is possible but costly.

reasoning, $(p_1, p_2)$ converges to the equilibrium $(56, 40)$, which does not deter entry.

**Case 2)** $\pi^3 < E < \pi^2$ **(agreement incompleteness).** According to SPE, entry is deterred by equilibrium $(40, 56)$. But $p_2 = 56$ is incompatible with forward induction reasoning. If firm 2 is rational and believes that firm 1 is rational, firm 2 will enter only if its expectation about $p_1$ is not lower than some $\widetilde{p} \in (40, 48]$ (that depends on $E$). Then, entry displays firm 2's intention to fix either $p_2 \in \left[36 + \frac{1}{2}\widetilde{p}, 60\right]$ with $k = 1$, or $p_2 \in [36, 42]$ with $k = 2$, but not $p_2 = 56$. Note however that every $p_1 \in [36, 42] \cup [54, 60]$ is a best response to a belief over such entry plans of firm 2. Hence, it is credible that firm 1 will react to entry with $p_1 = 40$. But credibility of $p_1 = 40$ requires uncertainty about $p_2$, thus it must be formulated as a unilateral threat and not as part of a complete agreement with firm 2. Even more interestingly, firm 1 does not actually need to specify $p_1$: it is enough to announce the use of technology $k = 2$.[15] Then, firm 2 will expect firm 1 to fix $p_1 \in [36, 42]$. If $\widetilde{p} > 42$, this is sufficient to deter entry. If $\widetilde{p} \in (40, 42]$, firm 2 may believe that entry will be profitable and fix $p_2 \in \left(36 + \frac{1}{2}\widetilde{p}, 57\right]$ with $k = 1$. But then, realizing this, firm 1 would best reply with $p_1 \in \left(12 + \frac{1}{4}\widetilde{p}, 40.5\right]$ and $k = 2$. This realization is based not just on the belief that firm 2 is rational, believes that firm 1 is rational, and so on, but also on the belief that firm 2 believes in firm 1's announcement, which is not at odds with rational entry. If needed, further steps of reasoning eventually bring the highest possible $p_1$ below $\widetilde{p}$. Hence, the announcement of $k = 2$ by the incumbent is credible and deters entry. Such a parsimonious announcement can have real-life advantages; for instance it may be illegal to state future prices.[16] Moreover, under this announcement, all reactions to entry that are

[15] Although firm 1 has the incentive to threaten $k = 2$ to deter entry, also in a scenario where technologies must be irreversibly chosen *before* setting prices, firm 1 has no incentive to display $k = 2$ once entry is established and technologies are chosen, because it would just create the expectation of a low $p_1$ and induce firm 2 to lower $p_2$ as well. This justifies the unobservability of each other's technology before setting prices also in this scenario.

[16] Harrington [25] 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.

compatible with strategic reasoning turn out to deter entry. This is not true when firm 1 announces exactly $p_1 = 40$: entry cannot be rationalized under belief in this announcement, and every $p_1 \in [36, 42] \cup [54, 60]$ remains an equally justifiable reaction. In Section 4.1 I will expand on this point after analyzing formally the $k = 2$ announcement.

**Case 3)** $\pi^4 < E < \pi^3$ **(SPE is too restrictive).** Now, firm 2 enters in every SPE. But, as in Case 2, there is $\widetilde{p} \in (36, 40]$ such that both $p_2 \in \left[36 + \frac{1}{2}\widetilde{p}, 60\right]$ and $p_2 \in [36, 42]$ are compatible with forward induction reasoning, and then all $p_1 \in [36, 42] \cup [54, 60]$ as well. So, firm 2 can credibly threaten to fix $p_1 < \widetilde{p}$ and deter entry.[17] The arguments for the credibility of this threat are identical to the arguments for the credibility of the SPE threat in Case 2, and break the logics of subgame perfection: credibility is not granted by coordination after entry but by beliefs over potentially profitable continuation plans of the entrant that are compatible with forward induction reasoning.

**Case 4) When** $E$ **is an agreed-upon payoff.** The game would be strategically equivalent if entry was costless and $E$ was the value of an exogenously given outside option. What if $E$ is firm 2's payoff from an agreement with firm 1 that comes into place if firm 2 does not enter? (For instance, a collusive agreement on another market.) Then, entry could be interpreted as disbelief in firm 1's threat, *or* as disbelief in the firm 1's promises in case of no entry. The analysis of Cases 1-2-3 remains valid if firms commonly believe that entry would be interpreted as disbelief in the threat and not as disbelief in the agreed-upon path (that grants payoff $E$ to firm 2). This kind of forward induction reasoning is modeled explicitly in Section 6.

[17] One could argue that alternated best responses from $p_1$ would lead to the $(68, 100)$ equilibrium in the long run. But, if firms are impatient, this is immaterial for the analysis. If firms are patient, these dynamics do not seem compelling in this multiple equilibria scenario with conflicting interests: firm 2 could try to upset this trajectory by switching to $k = 2$ at any time. The choice of $p_1 < \widetilde{p}$ is justified precisely by this uncertainty.

# 3 Agreements, beliefs and strategic reasoning

## 3.1 Framework

**Primitives of the game.** Let $I$ be the finite set of *players*. For any profile of sets $(X_i)_{i\in I}$ and any $J \subseteq I$, I write $X_J := \times_{j\in J} X_j$, $X := X_I$, $X_{-i} := X_{I\setminus\{i\}}$. Let $(\overline{A}_i)_{i\in I}$ be the finite sets of *actions* potentially available to each player. Let $\overline{H} \subseteq \cup_{t=1,...,T} \overline{A}^t \cup \{h^0\}$ be the set of histories, where $h^0 \in \overline{H}$ is the empty initial history and $T$ is the finite horizon. The set $\overline{H}$ must have the following properties. First property: For any $h = (a^1, ..., a^t) \in \overline{H}$ and $l < t$, it holds $h' = (a^1, ..., a^l) \in \overline{H}$, and I write $h' \prec h$.[18] Let $Z := \{z \in \overline{H} : \nexists h \in \overline{H}, z \prec h\}$ be the set of terminal histories (henceforth, *outcomes* or *paths*)[19], and $H := \overline{H}\setminus Z$ be the set of non-terminal histories (henceforth, just *histories*). Second property: For every $h \in H$, there exists a non-empty set $A_i(h) \subseteq \overline{A}_i$ for each $i \in I$[20] such that $(h, a) \in \overline{H}$ if and only if $a \in A(h)$. Let $u_i : Z \to \mathbb{R}$ be the *payoff function* of player $i$. 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 *plan of actions* (henceforth, just "plan") of player $i$ is a function $s_i$ that assigns an action $s_i(h) \in A_i(h)$ to each history $h$ that can be reached if $i$ plays $s_i$. Let $S_i$ denote the set of all plans of player $i$. A profile of plans $s \in S$ naturally *induces* a unique outcome $z \in Z$. Note that, when referring to profiles of plans rather than to agreements, the word "induce" will still be used with this traditional meaning. Let $\zeta : S \to Z$ be the function that associates each profile of plans with the induced outcome. For any $h \in \overline{H}$, the set of plans of $i$ compatible with $h$ is:

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

Fix subsets of plans $(\widehat{S}_j)_{j\in I}$. For each $i \in I$, let $\widehat{S}_i(h) := S_i(h) \cap \widehat{S}_i$. For

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

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

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

any $J \subseteq I$, let $H(\widehat{S}_J) := \left\{h \in H : \widehat{S}_J(h) \neq \emptyset\right\}$ denote the set of histories compatible with $\widehat{S}_J$.

## 3.2 Beliefs, Rationality, and Rationalizability

A player's beliefs over co-players' plans are modeled as a Conditional Probability System (henceforth, CPS).

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

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

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

A CPS is an array of beliefs, one for each history, that satisfies the chain rule: whenever possible, the belief at a history is an update of the belief at the previous history based on the observed co-players' moves.[21]

For any player $i$ and any set of co-players $J \subseteq I\setminus\{i\}$, I say that a CPS $\mu_i$ *strongly believes* (Battigalli and Siniscalchi [11]) $\widehat{S}_J \subseteq S_J$ if for every $h \in H(\widehat{S}_J)$, $\mu_i(\widehat{S}_J \times S_{I\setminus(J\cup\{i\})}|h) = 1$. I say that a CPS strongly believes a profile or a sequence of sets when it strongly believes each set of the profile or sequence.

I consider players who respond rationally to their beliefs. A rational player, at every history, chooses an action that maximizes her expected payoff given her belief about how co-players will play and the expectation to choose rationally again in the continuation of the game. By standard arguments, this is equivalent to playing a *sequential best reply* to the CPS.

[21] Note that a player can have correlated beliefs over the plans of different co-players, although players will not make use of joint randomization devices. The two things are not at odds, because players can believe in spurious correlations among co-players' plans (see, for instance, Aumann [1] and Brandenburger and Friedenberg [16]). However, *strategic independence* (Battigalli [5]) could be assumed throughout the paper and the results would not change. See the companion paper for details.

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

$$\sum_{s_{-i} \in S_{-i}(h)} u_i(\zeta(s_i, s_{-i}))\mu_i(s_{-i}|h) \geq \sum_{s_{-i} \in S_{-i}(h)} u_i(\zeta(\widetilde{s}_i, s_{-i}))\mu_i(s_{-i}|h). \quad (1)$$

I say that a plan $s_i$ is *justifiable* if $s_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta^H(S_{-i})$. The set of sequential best replies to $\mu_i$ (resp., to some $\mu_i \in \overline{\Delta}_i \subset \Delta^H(S_{-i})$) is denoted by $\rho_i(\mu_i)$ (resp., by $\rho_i(\overline{\Delta}_i)$).

I consider players who always ascribe to each co-player the highest order of strategic sophistication that is compatible with her past behavior. This means that players *strongly believe* that each co-player is rational; strongly believe that each co-player is rational and strongly believes that everyone else is rational; and so on. This form of *common strong belief in rationality* (Battigalli and Siniscalchi 2002) is captured by the following version of extensive-form-rationalizability, which I will call **Rationalizability** for brevity.[22]

**Definition 3** *Let $S^0 := S$. Fix $n > 0$ and suppose to have defined $((S_j^q)_{j \in I})_{q=0}^{n-1}$. For each $i \in I$ and $s_i \in S_i$, let $s_i \in S_i^n$ if and only if $s_i \in \rho_i(\mu_i)$ 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 $S^\infty$ are called* rationalizable.

[22] For conceptual coherence with the notion of Selective Rationalizability (see Section 3.3), this definition of extensive-form-rationalizability combines *strong rationalizability* as in Battigalli [7] (i.e., with memory of all previous steps) with *independent rationalization* as in Battigalli and Siniscalchi [10] (i.e., with strong belief in each $S_j^q$ instead of just $S_{-i}^q$). However, as far as extensive-form-rationalizability is concerned, all the classical definitions (Pearce [32], Battigalli [5], Battigalli and Siniscalchi [11]) are equivalent for the scope of this paper. See the companion paper for details.

## 3.3 Agreements, belief in the agreement, and Selective Rationalizability

All the notions introduced in this section are illustrated with concrete examples in the next section.

Players talk about 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 (joint) randomization devices.[23]

Under these assumptions, agreements can be modeled as follows:

**Definition 4** *An **Agreement** is a profile* $e = (e_i)_{i \in I}$ *where each* $e_i = (e_i^0, e_i^1, ..., e_i^{k_i})$ *is a chain of subsets of rationalizable plans:*

$$e_i^0 \subset e_i^1 \subset ... \subset e_i^{k_i} \subseteq S_i^\infty.$$

First, an agreement specifies for each player $i \in I$ a set of plans $e_i^0$ that player $i$ promises to follow. Second, the agreement can also specify alternative sets of plans $e_i^n$ $(n = 1, ..., k_i)$ that player $i$ promises to follow once he fails to follow any of the plans in $e_i^{n-1}$. So, the plans in $e_i^n \backslash e_i^{n-1}$ will be relevant for co-players' beliefs only after a deviation by player $i$ from the plans in $e_i^{n-1}$.[24]

[23] The use of randomization devices can be easily introduced in the methodology. Note however that a player would lack the strict incentive to use an individual randomization device over the own actions. Therefore, in absence of joint randomization devices, only sets of outcomes instead of outcome distributions can be enforced anyway. As Pearce [32] puts it, "this indeterminacy is an accurate reflection of the difficult situation faced by players in a game."

[24] In light of this, agreements could be given a more synthetic but less handy representation with just one set of *strategies* (as opposed to plans of actions) for each player. In particular, tight agreements (see Definition 14) could also be expressed in this more familiar language for a solution concept. However, the chosen formulation of agreements is more transparent regarding players' beliefs, see Definition 6.

The focus on rationalizable plans is without loss of generality: agreements that feature non-rationalizable plans can be analyzed in the same way, but do not offer any additional opportunity in terms of outcomes they can induce.

With respect to a strategy profile, which can be seen as a *complete* agreement, an agreement can involve two forms of incompleteness. First, $e_i^0$ may not be a singleton. Hence, player $i$ can have more than one plan which is compatible with the agreement. The same applies to each further $e_i^n$. Second, when a history $h$ may not be allowed by any plan in $e_i^{k_i}$, thus the agreement does not say anything about what player $i$ should do from $h$ onwards.

I will often focus on *reduced agreements*, where each player $i$ is silent about how she would play after a own deviation from the plans in $e_i^0$. Reduced agreements do not require players to trust the promises of a co-player who has already violated the agreement. *Path agreements* are reduced agreements that represent players who simply agree on an outcome to achieve. So, players do not specify how they will react to someone else's deviation either. Path agreements are to be expected, for instance, when discussing deviations is "taboo". Formally:

**Definition 5** *An agreement $e = (e_i)_{i \in I}$ is:*

**i** *reduced if for every $i \in I$, $e_i = (e_i^0)$;*

**ii** *a path agreement on $z \in Z$ if for every $i \in I$, $e_i = (e_i^0) = (S_i^\infty(z))$.*

Note that a path agreement on $z$ must feature all rationalizable plans of player $i$ compatible with $z$ to remain silent regarding $i$'s reactions to co-players' deviations.[25] Instead, like any other reduced agreement, a path agreement remains silent regarding $i$'s continuation plans after the own deviations by not introducing alternative plans. Introducing all rationalizable plans (as $e_i^1 = S_i^\infty$) would be equivalent: these two ways of not specifying a player's behavior from some history onwards will be convenient in different contexts.

[25] The restriction that $i$ will react in a rationalizable way will not have any actual bite, because players are able to conclude this already from the beliefs in rationality.

I say that a player $i$ believes in the agreement if she believes as long as possible that each co-player $j$ is carrying on a plan in $e_j^0$; and when this is no more possible, she believes as long as possible that $j$ is carrying on a plan in $e_j^1$; and so on.[26]

**Definition 6** *Fix an agreement $e = (e_i)_{i \in I}$. I say that player $i$ believes in the agreement when, for each $j \neq i$, $\mu_i$ strongly believes $e_j^0, ..., e_j^{k_j}$.*

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

I take the view that players refine their beliefs about co-players' plans through strategic reasoning based on beliefs in rationality and beliefs 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 rationality and 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. In the companion paper (Catonini [17]), I provide the details of this reasoning scheme, and I show that its behavioral implications are captured by an elimination procedure called **Selective Rationalizability**.[27] Fix an agreement $e = (e_i)_{i \in I}$.

**Definition 7** *Let $S_e^0 := S^\infty$. Fix $n > 0$ and suppose to have defined $((S_{j,e}^q)_{j \in I})_{q=0}^{n-1}$. For each $i \in I$ and $s_i \in S_i^\infty$, let $s_i \in S_{i,e}^n$ if and only if $s_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j \neq i})_{q=0}^{n-1}$.*
*Finally, let $S_{i,e}^\infty := \cap_{n \geq 0} S_{i,e}^n$. The profiles $S_e^\infty$ are called selectively-rationalizable.*

[26] This is reminiscent of the agreement being a *basis* for the player's CPS: see Siniscalchi [35].

[27] See the Supplemental Appendix for the equivalence between this definition of Selective Rationalizability and the more complicated one that is given an epistemic characterization in the companion paper.

Selective Rationalizability refines Rationalizability with the belief in the agreement and strategic reasoning about it. In particular, the first step refines Rationalizability with the belief in the agreement; the second step refines a player's plans further with strong belief that each co-player refines her rationalizable plans with the belief in the agreement as well; and so on.

Player $i$ is required to believe in the agreement everywhere in the game and at all steps of reasoning, because $\mu_i$ always has to belong to $\Delta_i^e$. Hence, Selective Rationalizability yields the empty set whenever a co-player $j$, at some step $n$, allows a history $h$ only with plans that violate the agreement; that is, $h \in H(S_{j,e}^n)$, but $S_{j,e}^n(h) \cap e_j^m = \emptyset$ for some $m$ with $e_j^m \cap S_j(h) \neq \emptyset$. Then, no $\mu_i \in \Delta_i^e$ strongly believes $S_{j,e}^n$, thus $S_{i,e}^{n+1} = \emptyset$: the belief in the agreement is incompatible with strategic reasoning and it is rejected as a whole. The belief that $j$ believes in the agreement, instead, is imposed by strong belief in $S_{j,e}^1$, thus it is abandoned as soon as not compatible with all orders of belief in rationality of $j$. The same applies to higher order beliefs in the agreement.

Recall that I will refer to $\zeta(e^0)$ as the outcome set that the agreement *prescribes*, and to $\zeta(S_e^\infty)$ as the outcome set the agreement *induces*. For a set of plans $S^* \subset S$, I will still say that $\zeta(S^*)$ are the outcomes the set induces, as customary.

# 4 Self-enforceability and implementability

## 4.1 Credibility and Self-Enforceability

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

induces each player $i$ to strongly believe in the (non-empty) set of agreed-upon plans that are compatible with strategic reasoning, namely $S^{\infty}_{-i,e} \cap e^0_{-i}$. 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* $\zeta(S^{\infty}_e) = \zeta(S^{\infty}_e \cap e^0)$.

Self-enforceability implies that, for *all* their refined beliefs, players will comply with the agreement *on the induced paths*, so that no violation of the agreement will actually occur. That is, $\zeta(S^{\infty}_e) \subseteq \zeta(e^0)$.

So, a self-enforcing agreement may induce a strict subset of the outcomes it prescribes. When instead the agreement is not vague about the outcome(s) it is going to induce, I say that the agreement is *truthful*.[28]

**Definition 10** *A self-enforcing agreement is* ***truthul*** *if* $\zeta(S^{\infty}_e) = \zeta(e^0)$.

To illustrate the whole methodology, I am going to analyze several agreements for the game in the Introduction and one agreement from the game of Section 2.

In the introductory game, all plans are justifiable, hence they are all rationalizable: $S^0_e = S$. Consider first the path agreement on $(S, E)$: $e^0_A = \{S\}$, $e^0_B = \{E.L, E.R\}$. We have $\Delta^e_A = \{\mu_A : \mu_A(\{E.L, E.R\}\,|h^0) = 1\}$ and $\Delta^e_B = \{\mu_B : \mu_B(S|h^0) = 1\}$. So, $S^1_e = \{S, N.D\} \times \{E.L, E.R\}$: Ann plays either $S$, or $N.D$ if she gives sufficiently high probability to $E.R$; Bob plays $E$ and either $L$ or $R$ depending on his new belief after being surprised by Ann's deviation. Then, we have $S^2_e = \{S, N.D\} \times \{E.L\}$, and finally $S^3_e = \{S\} \times \{E.L\} = S^{\infty}_e$: $e$ is truthful.

The following agreements require only one step of reasoning, except for the "unilateral" agreement that requires two.

[28] The choice of the term "truthful" is clearly inspired by the implementation literature, although an important caveat applies: see the end of Section 4.2.

| Agreement | Reduced | "Complete" | "Unilateral" | Path on $(N, W)$ |
|---|---|---|---|---|
| $e_A$ | $e_A^0 = \{N.U\}$ | $e_A^0 = \{N.U\}$ | $e_A^0 = S_A$ | $e_A^0 = S_A \backslash \{S\}$ |
| $e_B$ | $e_B^0 = \{W\}$ | $(\{W\}, \{W, E.L\})$ | $(\{W\}, \{W, E.L\})$ | $e_B^0 = \{W\}$ |
| $S_{A,e}^1 \times S_{B,e}^1$ | $(S_A \backslash \{S\}) \times \{W\}$ | $\{N.U\} \times \{W\}$ | $\{N.U\} \times S_B$ | $(S_A \backslash \{S\}) \times S_B$ |
| $S_{A,e}^\infty \times S_{B,e}^\infty$ | $(S_A \backslash \{S\}) \times \{W\}$ | $\{N.U\} \times \{W\}$ | $\{N.U\} \times \{W\}$ | $(S_A \backslash \{S\}) \times S_B$ |
| Conclusion | Truthful | Truthful | Self-enforcing | Credible |

The path agreement on $(N, W)$ is not self-enforcing, while the path agreement on the SPE path $(S, E)$ is, but this is far from true in general, even when the SPE is unique: see the Supplemental Appendix. The other three agreements are self-enforcing and induce $(N, W)$. The first agreement has the advantage of being reduced: Ann is not required to trust a statement of Bob about what he will do after deviating. The second and third agreements have the seeming advantage that only $N.U$ and not $N.D$ is compatible with strategic reasoning for Ann ($S_{A,e}^\infty = \{N.U\}$). But in both cases, after $E$, all beliefs about Bob's next move are equally compatible with strategic reasoning, and Ann believes in $L$ (and thus plays $U$) only by Bob's post-deviation promise. This is why requiring $S_e^\infty \subseteq e^0$ does not seem to be a compelling strengthening of self-enforceability.

Sometimes, agreements incompleteness induces steps of reasoning that refine players' beliefs after deviations, so that all remaining beliefs induce reactions that discourage the deviation. I illustrate this fact with a concrete example from the game of Section 2. (The game is not finite, but the methodology can be applied as is to all games with a countable number of *non-terminal* histories.) Consider the announcement by the incumbent (firm 1) of technology $k = 2$. Formally, this is a reduced agreement where $e_1^0$ is the set of all technology-price pairs with $k = 2$, and $e_2^0 = S_2$. Compatibly with Case 2, suppose that entry is profitable only if, in expectation, $p_1 \geq 41$. Omitting

"entry" in the description of firm 2's plans, we have:

| | |
|---|---|
| $S_1^1$ | $[36,48] \times \{k=2\} \cup [48,60] \times \{k=1\}$ |
| $S_2^1$ | $\{no\text{-}entry\} \cup [36,48] \times \{k=2\} \cup [56.5,60] \times \{k=1\}$ |
| $S_1^2 = S_1^\infty$ | $[36,42] \times \{k=2\} \cup [54,60] \times \{k=1\}$ |
| $S_2^2 = S_2^\infty$ | $\{no\text{-}entry\} \cup [36,42] \times \{k=2\} \cup [56.5,60] \times \{k=1\}$ |
| $S_e^1$ | $S_{1,e}^1 = S_1^\infty,\ \ S_{2,e}^1 = \{no\text{-}entry\} \cup [56.5,57] \times \{k=1\}$ |
| $S_e^2$ | $S_{2,e}^2 = S_{2,e}^1,\ \ S_{1,e}^2 = [40.25,40.5] \times \{k=2\}$ |
| $S_e^3 = S_e^\infty$ | $S_{1,e}^2 = S_{1,e}^3,\ \ S_{2,e}^3 = \{no\text{-}entry\}$ |

The announcement of $k=2$ is not per se sufficient to deter entry, but it entails sufficiently low prices by the incumbent for the entrant to employ $k=1$ and exclude the highest rationalizable prices. Anticipating this, the incumbent has the strict incentive to use $k=2$ and exclude the highest prices compatible with $k=2$ as well. In turn, this provides the strict incentive to firm 2 not to enter. So, the agreement is credible and it deters entry. Note that strategic reasoning (in particular, belief in $S_{2,e}^1$ after entry) always induces the incumbent to choose technology-price pairs that deter entry, absent any restriction on the entrant's continuation plan. Under the SPE threat $p_1 = 40$, instead, entry cannot be rationalized under belief in the threat ("entry" would not be in $S_{2,e}^1$), therefore any belief over the entrant's rationalizable plans ($S_2^\infty$) remains possible.

## 4.2 Implementability and agreements design

I say that an agreement *implements* a set of outcomes $P \subseteq Z$ when it is self-enforcing and it induces $P$.

**Definition 11** *A set of outcomes $P \subseteq Z$ is **implementable** if there exists a self-enforcing agreement such that $\zeta(S_e^\infty) = P$.*

A set of outcomes induced by a merely credible agreement does not correspond to what players agreed upon and believe in. For this reason, implementation requires the agreement to be self-enforcing. All in all, only self-enforcing agreements are able to induce a specific outcome.

**Proposition 1** *If $\zeta(S_e^\infty)$ is a singleton, then $e$ is self-enforcing.*

Which sets of outcomes are implementable? How to design agreements that implement them? This section aims to answers these questions.

By definitions of self-enforceability and implementability, every implementable outcome set is induced by $S_e^\infty \cap e^0$ for some self-enforcing agreement $e$. This provides some useful necessary conditions for implementability.

**Proposition 2** *For every self-enforcing agreement $e = (e_i)_{i\in I}$, the set $S^* = \times_{i\in I} S_i^* := S_e^\infty \cap e^0$ satisfies the following properties:*
*Realization-strictness: For each $i \in I$ and $\mu_i$ that strongly believes $S_{-i}^*$,*

$$\zeta(\rho_i(\mu_i) \times S_{-i}^*) \subseteq \zeta(S^*);$$

*Self-Justifiability: For each $i \in I$ and $s_i \in S_i^*$, there exists $\mu_i$ that strongly believes $(S_j^*)_{j\neq i}$ and $(S_j^\infty)$ such that $s_i \in \rho_i(\mu_i)$.*

**Corollary 1** *If a set of outcomes is implementable, then it is induced by a Cartesian set of rationalizable profiles that satisfies Realization-strictness and Self-Justifiability.*

Self-Justifiability says that, for each player $i$, each plan in $S_i^*$ is justifiable under strong belief that each co-player $j$ carries on a plan in $S_j^*$, and some other rationalizable plan otherwise. Realization-strictness says that players have the strict incentive to stay on the paths induced by $S^*$ whenever they strongly believe that co-players carry on plans in $S_{-i}^*$. Analogously, say that a Nash equilibrium $s^* = (s_i^*)_{i\in I}$ is *realization-strict* when it provides strict incentive to stay on path, that is, $\arg\max_{s_i\in S_i} u_i(\zeta(s_i, s_{-i}^*)) = S_i(\zeta(s^*))$ for all $i \in I$. Then, when $S^*$ induces a unique outcome $z = \zeta(S^*)$, Realization-strictness boils down to $S^*$ being a set of realization-strict Nash equilibria.

**Proposition 3** *A Cartesian set of rationalizable profiles that induce the same outcome satisfies Realization-strictness if and only if each element is a realization-strict Nash equilibrium.*

**Corollary 2** *If an outcome is implementable, then it is induced by a realization-strict Nash equilibrium in rationalizable plans.*

These necessary conditions simplify the search for implementable outcome sets. First, a standard elimination procedure like Rationalizability is performed. Then, for each candidate outcome set, one can look for a set of rationalizable profiles that induces it and satisfies Realization-strictness and Self-Justifiability. If the set satisfies a further forward induction condition, I call it a Self-Enforcing Set.

**Definition 12** *Fix a Cartesian set of rationalizable profiles $S^* = \times_{i\in I} S_i^* \subseteq S^\infty$ that satisfies Self-Justifiability. The closure of $S^*$ (under rationalizable behavior), denoted by $\overline{S}^* = \times_{i\in I}\overline{S}_i^*$, is, for each $i \in I$, the set of all $s_i \in S_i^\infty$ such that $s_i \in \rho_i(\mu_i)$ for some $\mu_i$ that strongly believes $(S_j^*)_{j\neq i}$ and $(S_j^\infty)_{j\neq i}$.*

**Definition 13** *A Cartesian set of rationalizable profiles $S^*$ is a **Self-Enforcing Set** if it satisfies Realization-strictness, Self-Justifiability, and:*
*Forward Induction: For each $i \in I$ and $s_i \in \overline{S}_i^*$, there exists $\mu_i$ that strongly believes $(S_j^*)_{j\neq i}$, $(\overline{S}_j^*)_{j\neq i}$, and $(S_j^\infty)_{j\neq i}$ such that $s_i \in \rho_i(\mu_i)$.*

The closure of a set includes all rationalizable plans that players who strongly believe in the set and in the rationalizable plans of co-players may play. Forward Induction says that such players need not change their behavior when they also strongly believe that co-players form beliefs in the same way. That is, when they predict the behavior of deviators from the set with forward induction reasoning (i.e., when they strongly believe in the closure of the set).

A SES and its closure are realization-equivalent: by Self-Justifiability, $S^* \subseteq \overline{S}^*$, and by Realization-strictness, $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$. So, in terms of induced outcomes, SES's are "closed under rationalizable behavior",[29] and indeed boil down to "sets closed under rational behavior" (Basu and Weibull [4]) in static

[29]Note that closedness in terms of plans ($S^* = \overline{S}^*$) would be very hard to satisfy: as already stressed, there are typically many rationalizable continuation plans of deviators, that justify many possible reactions.

games. By Forward Induction, the closure of the SES cannot be further refined with forward induction considerations. Therefore, the agreement on the SES implements precisely the SES outcome set ($\zeta(S^*)$).

**Proposition 4** *Fix a SES $S^* = \times_{i \in I} S_i^*$. The reduced agreement $e = ((S_i^*))_{i \in I}$ is truthful.*

**Corollary 3** *If an outcome set is induced by a SES, then it is implementable (with a truthful, reduced agreement).*

An example of SES is provided in the next section.

The current gap between necessary and sufficient conditions for implementation is given by a seemingly strong condition: Forward Induction. But the power of Forward Induction is mitigated by Self-Justifiability and Realization-strictness. Realization-strictness implies that deviations from the SES paths cannot be rationalized under belief in the SES. Therefore, Forward Induction does not restrict the beliefs about the first player who deviates more than Self-Justifiability. Then, if there are no other co-players, or if there is no time for the first deviation to trigger other deviations by other players, Forward Induction has no bite. I say that a game has two stages when $Z \subseteq \overline{A} \cup \overline{A}^2$.

**Proposition 5** *In games with 2 players or 2 stages, any Cartesian set of rationalizable profiles that satisfies Realization-strictness and Self-Justifiability also satisfies Forward Induction.*

Hence, in these games, SES's fully characterize implementable outcome sets and provide truthful, reduced agreements that implement them.

**Theorem 1** *In games with 2 players or 2 stages, the following hold:*

1. *a Cartesian set of rationalizable profiles is a Self-Enforcing Set if and only if it satifies Realization-strictness and Self-Justifiability;*
2. *an outcome set is implementable if and only if it is induced by a Self-Enforcing Set;*

3. *every implementable outcome set is implemented by a truthful, reduced agreement.*

**Proof.** Statement 1 follows from Proposition 5. Statement 2 follows from Corollary 1 and statement 1 for the "only if" part, and from Proposition 4 for the "if" part. Statement 3 follows from statement 2 and Proposition 4. ■

Moreover, in two-players games, Realization-strictness implies Self-Justifiability for a set of profiles that all induce the same outcome.

**Proposition 6** *In 2-players games, any Cartesian set of rationalizable profiles that induces a unique outcome and satisfies Realization-strictness also satisfies Self-Justifiability.*

Then, in two-players games, implementable outcomes are fully characterized by realization-strict Nash equilibrium in rationalizable plans.

**Theorem 2** *In 2-players games, an outcome is implementable if and only if it is induced by a realization-strict Nash equilibrium in rationalizable plans, and it is implemented by the truthful, reduced agreement on the Nash itself.*

**Proof.** "Only if" coincides with Corollary 2. For "if" and the final statement: let $s^* = (s_i^*)_{i \in I} \in S^\infty$ be a realization-strict Nash equilibrium. By Proposition 3, the singleton $\{s^*\}$ satisfies Realization-strictness. By Proposition 6, it also satisfies Self-Justifiability. By Proposition 5, it also satisfies Forward Induction, thus it is a SES. Then, by Proposition 4, $\zeta(s^*)$ is implemented by the reduced agreement $e = (\{s_i^*\})_{i \in I}$. ■

How to fill the gap between necessary and sufficient conditions in games with more than two players and stages? Forward Induction may be violated because a deviation from a SES induces further deviations down the line. Possibly, this can be avoided by restricting the behavior of the deviator, compatibly with forward induction reasoning. This is what *tight agreements* do.

**Definition 14** *An agreement $e = (e_i)_{i \in I}$ is tight when:*

**T1** $e^0$ *satisfies Realization-strictness;*

**T2** *for every* $i \in I$ *and* $h \in H(S_i^\infty)$*, there is* $n \leq k_i$ *such that*

$$e_i^n \cap S_i(h) \neq \emptyset;$$

**T3** *for every* $i \in I$ *and* $h \in H(\rho_i(\Delta_i^e) \cap S_i^\infty)$*, there is* $n \leq k_i$ *such that*

$$\emptyset \neq e_i^n \cap S_i(h) \subseteq \rho_i(\Delta_i^e).$$

**Remark 1** *If* $e = (e_i)_{i \in I}$ *is tight,* $e^0$ *satisfies Self-Justifiability.*

Like a SES, a tight agreement initially specifies plans that satisfy Realization-strictness (by T1) and Self-Justifiability (by Remark 1). Differently than a SES, a tight agreement also specifies alternative plans $e_i^1, ..., e_i^{k_i}$ that each player $i$ should follow, until all histories compatible with her rationalizable plans, $H(S_i^\infty)$, are reached by some $e_i^n$ (this is T2). All histories that player $i$ can rationally reach under belief in the agreement, $H(\rho_i(\Delta_i^e) \cap S_i^\infty)$, must first be reached by some $e_i^n$ only with plans that can be justified under belief in the agreement (this is T3). Then, all such plans can be believed by co-players who reason by forward induction based on rationality and the agreement.[30] This makes the agreement credible and, by T1, self-enforcing. Self-Justifiability of $e^0$ adds truthfulness.

**Proposition 7** *Tight agreements are truthful.*

**Theorem 3** *An outcome set is implementable if and only if it is prescribed by a tight agreement.*[31]

[30] By imposing belief in these alternative plans, the Forward Induction condition of SES's becomes unnecessary. A SES $S^*$ can indeed be transformed into the following tight agreement $e = (e_i)_{i \in I}$: for each $i \in I$, $e_i^0 = S_i^*$, $e_i^1 = \overline{S}_i^*$, $e_i^2 = S_i^\infty$. Introducing $e_i^2$ is immaterial for the agreement but verifies T2: introducing all or none rationalizable plans of a player are equivalent ways not to restrict behavior, but the first is convenient for tight agreements, the second for SES's.

[31] The "only if" statement, and thus also Corollary 4 rely on the game being finite, in particular on finite horizon. This is because agreements are finite sequences, but in games

**Proof.** "If" comes from Proposition 7. For "only if": see the Appendix. ■

Tight agreements close the gap between necessary and sufficient conditions for implementability in all games, and the roadmap for the joint search of implementable outcome sets and agreements that implement them. If a candidate set of outcomes is implementable, a tight agreement that implements it can be found by following the search for SES's first, and introducing alternative plans if Forward Induction cannot be satisfied. All the games in the paper have two stages; therefore, by Theorem 1, the search can stop at SES's, without the need to verify Forward Induction. A case where Forward Induction cannot be satisfied but the search for a tight agreement is completed succesfully is presented in the next section.

Since tight agreements are truthful and fully characterize implementable outcomes, we have the following "revelation principle" for agreements design.

**Corollary 4** *Every implementable outcome set is implemented by a truthful agreement.*

This means that if players want to implement an outcome $z$ (or a set $P$), there is no use of being vague about it in the agreement.

The use of the terms "truthful" and "implementation" is indeed inspired by an analogy with robust implementation (Bergemann and Morris [15]). A robust mechanism implements the outcome assigned by the social choice function to players' types for all their beliefs about co-players' types; a self-enforcing agreement implements (a subset of) the agreed-upon outcome(s) for all players' refined beliefs. When players use direct mechanisms, they truthfully reveal their types and the corresponding outcome obtains; when players use truthful agreements, they declare precisely the outcome(s) they want to achieve. Both direct mechanisms and truthful agreements suffice for implementation. Note

with finite horizon the set of induced paths may squeeze for infinite steps of reasoning. A characterization of implementable outcomes with truthful agreements in games with infinite horizon, if possible, is subject for future research.

though an important difference: while a direct mechanism requires player to specify *only* their type, a truthful agreement, beside the outcome(s), can also specify off-the-path(s) behavior. This can be seen as the price to pay for the agreement being a "soft mechanism", which does not change the rules of the game.

## 4.3 Further examples

The aim of this section is two-fold. First, it provides concrete examples of (the search for) a SES and of a tight agreement where, respectively, realization-strict Nash and SES's do not implement the desired outcome. Second, it shows that, after a deviation from the desired path, agreement incompleteness regarding the reaction of co-players (as allowed by SES) or restrictions to the continuation plan of the deviator (as allowed by tight agreements) can be necessary for implementation. This complements the example of Section 2, where the incumbent can specify precisely a credible reaction that deters entry, while specifying the behavior of the entrant is unneeded or even precludes the implementation of no-entry.

**Peacekeeping game** The example of Section 2 showed why the behavior of deviators may need to be left unspecified. Here I show by example that also leaving the behavior of co-players partially unspecified may be needed for implementation. This form of agreement incompleteness is enabled by $|e^0| > 1$, even when $\zeta(e^0)$ is a singleton, and it is allowed by SES's. Consider the following 4-players game,[32] where in the subgame, Cleo chooses the matrix,

[32] This game is freely inspired by the leading example in Greenberg [23], with a fundamental difference: in that example, the superpower remains silent, and there simply *exist* beliefs about its behavior that make the warring countries behave as desired; here the warring countries remain silent and the superpower speaks, and this suffices to pin down beliefs that *all* induce the desired behavior by the first mover (here a fourth country).

Ann the row, and Bob the column (payoffs are in alphabetical order).

DAVE — $Out \longrightarrow 0,0,0,0$

$Instigate \downarrow$

| $Int$ | $Arms\ Race$ | $Peaceful$ | $Not$ | $Arms\ Race$ | $Peaceful$ |
|---|---|---|---|---|---|
| $AR$ | $-1,-1,-1,-1$ | $-1,-3,\ 1,-2$ | $AR$ | $-3,-3,\ 0,\ 2$ | $5,-6,\ 0,\ 1$ |
| $P$ | $-3,-1,\ 1,-2$ | $0,\ 0,-1,-3$ | $P$ | $-6,\ 5,\ 0,\ 1$ | $0,\ 0,\ 0,\ 0$ |

Dave, a weapons producer, can *Instigate* a conflict between Ann and Bob. If he does, Ann and Bob can engage in the *Arms Race*, or remain *Peaceful*. Engaging in the arms race transfers 1 util to Dave. Cleo, a superpower, can *Intervene* to avoid an escalation of the conflict and impose sanctions against Dave. The cost of the sanctions for Dave is 3, and the cost of peacekeeping for Cleo is 1 if both Ann and Bob both engage in the arms race and 2 if only one does. In the first case, without Cleo's intervention, the war comes to a costly impass. In the second case, without Cleo's intervention, the peaceful player gets conquered and loses all its resources (6) to the other; with Cleo's intervention, the defended player has to share its resources with Cleo.

The game has only one SPE, where Dave instigates, Cleo does not intervene, and Ann and Bob engage in the arms race.[33] Even allowing for randomizations, there is no credible specification of Ann, Bob, and Cleo's actions that induces Dave not to instigate. However, Cleo can threaten Dave to intervene, and Ann and Bob remain silent about their plans. I show that $S^* = \{AR, P\} \times \{AR, P\} \times \{Int\} \times \{Out\}$, inducing outcome $(Out)$, is a SES. All plans are justifiable, hence rationalizable. Since the game has 2 stages, by Theorem 1 it suffices to show Realization-strictness and Self-Justifiability. For Dave, they both follow from the fact that $\rho_D(\mu_D) = \{Out\}$ for every $\mu_D$ that strongly believes $S^*_C$. For every $i = A, B, C$, Realization-strictness trivially follows from $\zeta(S_i \times e^0_{-i}) = \{(Out)\}$. Self-Justifiability for Cleo: $Int$ is justified by $\mu_C$ with $\mu_C((P, AR, Inst)|(Inst)) = 1$, and let

[33] Ann and Bob may have the incentive to be peaceful only if they assign probability at least 2/3 to the other being peaceful and Cleo intervening. But if both are peaceful with probability at least 2/3, Cleo would rather not intervene.

$\mu_C(S_A \times S_B \times \{Out\}\,|h^0) = 1$ for $\mu_C$ to strongly believe $(S_j^*)_{j\neq C}$; for Ann, $AR$ (resp., $P$) is justified by any $\mu_A$ with $\mu_A((AR, Int, Inst)|(Inst)) = 1$ (resp., with $\mu_A((P, Int, Inst)|(Inst)) = 1$), and let $\mu_A(S_B \times \{Int\} \times \{Out\}\,|h^0) = 1$ for $\mu_A$ to strongly believe $(S_j^*)_{j\neq A}$; likewise for Bob.

**Should I stay or should I go?** In the department of dean Ann there are two game theorists, Bob and Cleo, who are up for midterm review. Ann maximizes the benefit from game theorists to the department, which is marginally decreasing, minus the opportunity cost of their salaries, which is marginally increasing. At the end of the year, Ann offers to Bob and Cleo the renewal at salary $r$, lower than the market salary $w$, but sufficient to make them prefer to $Stay$ if they still have to pay cost $g < w - r$ to $Go$ on the market. If they both stay, the game ends. If one stays and the other does not, say Bob, the game continues in the following year as in the figure. (What happens if they both go will not matter for the analysis.) Cleo can $Stay$ or $Go$ on the market as well; Ann can $Shut$ down Bob's position, or keep it $Open$. If Cleo stays, she has zero bargaining power and her salary remains $r$. If Cleo is on the market and Ann has shut down Bob's position, Ann is in a weak bargaining position and Cleo obtains a raise to $v > r + g$ ($v < w$). If Ann keeps Bob's position open and Cleo stays, Bob bargains a salary $t > r + g$ ($t < v$). If both Bob and Cleo are on the market, bargaining is complicated and gets delayed to the market stage. Ann can $Hire$ or $Not$; Bob and Cleo can $Stay$ or $Go$ for good. If Ann hires a new game theorist at $w$, bargaining is terminated because she is not willing to pay to Bob or Cleo a salary that makes them prefer to stay, so she will keep at salary $r$ only Cleo if she stays, or Bob if he stays and Cleo leaves. If Ann does not hire and Bob and Cleo do not leave, they will reach an agreement at $t$; if one leaves and the other stays, the latter obtains a salary $u$ with $t < u < v$. As deadlines approach, all players must make their choices without knowing the choices of others. Ordinal payoffs compatible with this

story are in the figure (cardinal payoffs will not matter for the analysis).

| Bob | — *Go* ⟶ | A\C | *Stay* | *Go* |
|---|---|---|---|---|
| *Stay* ↓ | (Cleo | *Open* | 5, 4, 3 | ·– |
| 6, 3, 3 | stays) | *Shut* | 4, 2, 3 | 2, 2, 6 |

⟶

| *Hire* | *Stay* | *Go* | *Not* | *Stay* | *Go* |
|---|---|---|---|---|---|
| *Stay* | 1, 0, 1 | 1, 1, 2 | *Stay* | 3, 4, 4 | 3, 5, 2 |
| *Go* | 1, 2, 1 | 1, 2, 2 | *Go* | 3, 2, 5 | 0, 2, 2 |

When Ann offers the renewal to Bob and Cleo, she calls for a meeting to set expectations for the next year and induce them to accept her offer. Is there an agreement that achieves this goal? We will look for an agreement that implements outcome (*Stay*) in the game above; by symmetry, it can be extended to the whole game. All plans are justifiable, hence rationalizable. Following the roadmap of Section 4.2, we look for $e^0$ that induces (*Stay*) and satisfies Realization-strictness and Self-Justifiability. Bob's Realization-strictness is satisfied if $e^0_A = \{S\}$, or if $O.N \notin e^0_A$ and $S \notin e^0_C$. In the first case, Ann's and Cleo's Self-Justifiability require $G.G \in e^0_C$ and $S \notin e^0_C$, so we have $\{G.G\} \subseteq e^0_C \subseteq \{G.S, G.G\}$. In the second case, Ann's and Cleo's Self-Justifiability require $O.H \notin e^0_A$ and $G.S \notin e^0_C$, so we are back to the first case. Thus, let us focus on agreements with $e^0_A = \{S\}$, $e^0_B = \{S\}$, and either $e^0_C = \{G.G\}$, or $e^0_C = \{G.S, G.G\}$. Does any of the two constitute a SES? No. The closure of $e^0$ for Ann includes $O.N$ but not $O.H$: under belief that Cleo plays $G$ in the second stage, $H$ is not a best reply to any belief in the last stage that induces Ann to play $O$. But then, Forward Induction is violated for Cleo, because the only sequential best reply under strong belief in $S$ and $\{S, O.N\}$ is $G.S$. Therefore, we try to find a tight agreement with $e^0$ as above. Pick $e^0 = (S, S, G.G)$ and restrict Bob's behavior after his deviation by imposing $e^1_B = \{S, G.G\}$. Also let $e^1_A = \{S, O.H\}$, so that all histories are reached by all players and T2 is satisfied. (T1 is Realization-strictness of $e^0$.) Is T3 verified? Under belief in the agreement, players play exactly $e^0$, so it immediate to check that T3 is satisfied.

Informally speaking, at the meeting Ann claims that if Bob or Cleo will not stay, she will shut one position down. Everybody is aware that this will induce the remaining game theorist to obtain a higher salary by going on the market, and that Ann could then be tempted to keep the other position open

and threaten to hire. However, Ann, Bob and Cleo convene that in the market scenario, they will take their less risky options: Ann will hire and Bob and Cleo will leave.[34] By stating what he would do after the own deviation, Bob helps making it unprofitable. This can be in the own interest of a player who agrees on a desired path. When this interpretation is not plausible, the restriction to the behavior of a deviator can also be seen as an "agreement of beliefs" among the other players, remaining agnostic about how it is originated. Be as it may, the belief that Bob will finally leave prevents a "signaling war" between Ann and Cleo that would not allow them to coordinate on a threat. Cleo's claim to leave in the last stage is credible because Ann's deviation can be interpreted as disbelief that Cleo goes on the market instead of belief that her move will reassure Cleo that she has optimistic beliefs about succesful bargaining. Note also that the tight agreement above is a "complete agreement" in that it specifies one action for each player and history, and it corresponds to a SPE.

If the true game was the one in the figure, specifying that Bob goes at the initial history is immaterial, and outcome $(Stay)$ would be implemented also by the reduced agreement with $e_0 = (S, \{S, G.G\}, G.G)$. This is not a truthful agreement, hence not a SES. For this reason, SES's do not fully characterize the outcomes implemented by reduced agreements, and not all outcomes that can be implemented by a reduced agreement can be implemented by a *truthful* reduced agreement, calling for restrictions to the behavior of deviators and tight agreements.

# 5 Epistemic priority to the agreement

The literature on strategic reasoning with first-order belief restrictions is mostly based on the use of Strong-$\Delta$-Rationalizability (Battigalli [7], Battigalli and Siniscalchi [12]). Strong-$\Delta$-Rationalizability is here denoted by $((S^q_{i,\Delta^e})_{i\in I})^\infty_{q=0}$ and defined like Selective Rationalizability without requiring that plans are rationalizable; i.e., as in Definition 7 with $S_i$ in place of $S^\infty_i$. The differences be-

[34] These can also be seen as their less risky options.

tween the results of this paper and the results in this literature are due to (i) the adoption of Selective Rationalizability in place of Strong-$\Delta$-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-$\Delta$-Rationalizability are deeply analyzed in [17]. Here I only recall the main conceptual difference between the two solution concepts. Consider a move that a player would not rationally make under belief in the agreement. Contrary to Selective Rationalizability, Strong-$\Delta$-Rationalizability captures the hypothesis that, upon observing such move, co-players *drop* the belief that the player is rational. I call this hypothesis *(epistemic) priority to the agreement* (as opposed to *rationality*). So, the question is: how would the adoption of Strong-$\Delta$-Rationalizability instead of Selective Rationalizability affect the results of this paper?

In the example of Section 2, the incumbent could deter entry also in Case 1 by threatening a low rational price; then, entry would be considered a sign of the entrant's irrationality, and the incumbent could expect a high entrant's price which does not best reply to any belief that justifies entry. This is a typical loss of refinement power that the inversion of epistemic priority entails. In all other examples, all plans are rationalizable; then, Selective Rationalizability and Strong-$\Delta$-Rationalizability coincide and the insights are robust to the inversion of epistemic priority.

The formal analysis of Section 4 can be replicated under priority to the agreement as follows. Allow agreements to feature non-rationalizable plans.

**Remark 2** *Under priority to the agreement, the results of Section 4 hold through verbatim after substituting everywhere:*

1. *selectively-rationalizable plans ($S_e^\infty$) with strongly-$\Delta$-rationalizable plans ($S_{\Delta^e}^\infty$);*

2. *rationalizable plans ($S^\infty$) with justifiable plans ($S^1$) in Proposition 6 and*

*Theorem 2, and with all plans ($S$) elsewhere.*[35]

To verify Remark 2, one can follow the proofs for Section 4 with the opportune substitutions, as highlighted in the Appendix. A credible agreement under priority to rationality needs not be credible under priority to the agreement: as shown in [17], Selective Rationalizability does not refine Strong-$\Delta$-Rationalizability for the same first-order belief restrictions. Across all agreements, instead, more outcome sets can be implemented under priority to the agreement.

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

Note in particular that, by Remark 2.2, under priority to the agreement any Nash equilibrium in justifiable plans of a two-players game is a self-enforcing agreement, also when incompatible with just strong belief in rationality.[36]

Battigalli and Friedenberg [8] capture the implications of Strong-$\Delta$-Rationalizability across *all* first-order belief restrictions with the notion of Extensive Form Best Response Set. An EFBRS is a Cartesian set of plans profiles $S^* = \times_{i \in I} S_i^*$ satisfying the following:

**EFBRS:** for each $i \in I$ and $s_i \in S_i^*$, $s_i \in \rho_i(\mu_i)$ for some $\mu_i$ that strongly believes $S_{-i}^*$ with $\rho_i(\mu_i) \subseteq S_i^*$.

With $(S_j^*)_{j \neq i}$ in place of $S_{-i}^*$, the EFBRS condition corresponds to Self-Justifiability under priority to the agreement, plus a "maximality" requirement: all the sequential best replies to some justifying beliefs must be in the EFBRS. Generically, maximality has no bite,[37] thus SES's generically refine EFBRS's. The

[35] Where Self-Justifiability is among the hypotheses, that plans are justifiable is implied.

[36] As shown by the introductory example of the companion paper, Strong-$\Delta$-Rationalizability can yield the outcome of a non-subgame perfect equilibrium in sequentially rational plans even in a perfect information game (i.e., a game where players move one at a time), where the unique backward induction outcome is also the unique extensive-form-rationalizable one: see Battigalli [6], Heifetz and Perea [26], Chen and Micali [18], Perea [34].

[37] Generically, every plan can be justified by a CPS that has not other sequential best reply.

three reasons are the anticipated ones. Under priority to rationality, SES's are in rationalizable plans; however, SES's generically refine EFBRS's also under priority to the agreement. This can be seen already in static games, where EFBRS's boil down to best response sets, while SES's boil down to sets closed under rational behavior. First, EFBRS's can be based on first-order belief restrictions that impose belief in specific randomizations, or, more fundamentally, differ across two players regarding the moves of a third player. Instead, SES's/agreements align any two players' beliefs about a third player's moves. (Relatedly, the restrictions are not expressed by the EFBRS itself, while SES's provide directly the restrictions that induce the outcomes they prescribe.) Second, an EFBRS may induce a larger set of outcomes with respect to what players expect under the restrictions that yield it. Realization-strictness/self-enforceability rule this out.[38]

# 6 Epistemic priority to the path

Consider the twofold repetition of the following game. All plans are justifiable, hence also rationalizable.

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

Suppose that Ann and Bob agree on the SPE where Bob works in the first period and Ann works in the second period. Then, if Bob observes that Ann works in the first period and believes that she is rational, he must believe that

[38] Relatedly, Battigalli and Siniscalchi [12] show that when Strong-$\Delta$-Rationalizability is non-empty under belief in a particular outcome, the outcome is induced by a self-confirming equilibrium (Fudenberg and Levine [21]). Regardless of the epistemic priority choice, implementable outcomes are instead all Nash by Corollary 2 and Remark 2. This is because under a self-enforcing agreement, players have the incentive to stay on path for *all* their refined beliefs. This allows to find plans of co-players against which there is no incentive to deviate. Credibility, instead, may be granted just by some correlated belief about the reactions of co-players to the deviation.

she does not believe that he plays as agreed. In the baseline analysis of this paper, Bob was free to believe, for instance, that Ann did not believe that he would have worked in the first period. Then, Bob could think that Ann is going to work also in the second period, and best-respond by free-riding, as agreed.

Suppose now instead that Bob believes that Ann trusted him, in the following sense: she believes that he would have not violated the agreement before her. Then, Bob must interpret Ann's deviation as an attempt to gain a higher payoff than under the agreement, and the only way for her to do so is to free ride after the deviation. If Ann anticipates that Bob will interpret the deviation in this way, she expects him to work after the deviation, and therefore she has incentive to deviate. The agreement is not credible.

When this way of interpreting a deviation is transparent to players, the interactive beliefs about (compliance with) the agreed-upon path receive a higher epistemic priority in players' strategic reasoning than the beliefs in the rest of the agreement: when Bob cannot believe anymore that Ann believes in the whole agreement, he keeps the belief that Ann believed in (his compliance with) the path, and drops the belief that Ann believes that he will comply off-path. Giving for granted that rationality keeps the highest epistemic priority, I call this finer epistemic priority ordering "priority to the path". First, each orders of belief in rationality is maintained as long compatible with the observed behavior. Second, each order of belief in the path is maintained as long as compatible with all orders of belief in rationality. Third, each order of belief in the *whole* agreement is maintained as long as compatible with all the aforementioned beliefs. In the companion paper, I capture epistemic priority orderings among different theories of players' behavior with a generalization of Selective Rationalizability. Such procedure is specialized here for the problem at hand. Fix a path $z \in Z$ and let $((S^q_{j,z})_{j\in I})^\infty_{q=0}$ denote Selective Rationalizability under the path agreement on $z$.[39] Fix an agreement $e = (e_i)_{i\in I}$ with $e^0 \subseteq S(z)$ and $\times_{i\in I} e_i^{k_i} \subseteq S_z^\infty$.

[39] The method can be generalized to any agreement, by analyzing first strategic reasoning under strong belief in the paths induced by $e^0$ (i.e., strong belief in $\cup_{z\in\zeta(e^0)} S_j(z)$).

**Definition 15** *Let $S^0_{e^z} = S^\infty_z$. Fix $n > 0$ and suppose to have defined $((S^q_{j,e^z})_{j\in I})^{n-1}_{q=0}$. For each $i \in I$ and $s_i \in S^\infty_{i,z}$, let $s_i \in S^n_{i,e^z}$ if and only $s_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta^e_i$ that strongly believes $((S^q_{j,e^z})_{j\neq i})^{n-1}_{q=0}$.*

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

The first two levels of epistemic priority are captured by Selective Rationalizability under the path agreement on $z$. Thus, the credibility of the path agreement is a preliminary test for the self-enforceability of an agreement that prescribes $z$ under priority to the path. Then, the "$z$-rationalizable" plans $(S^\infty_{j,z})_{j\in I}$ are refined using the belief in the whole agreement. So, the agreement must be compatible with strategic reasoning around the path.[40]

The analysis of Section 4 can be replicated under priority to the path for single outcomes. Allow agreements (including SES's) to prescribe only $z$ and feature only $z$-rationalizable plans. Then, the following holds.

**Remark 3** *Under priority to the path, the results of Section 4 hold through verbatim after substituting everywhere:*

1. *outcome sets $P$ with single outcomes $z$;*
2. *selectively-rationalizable plans ($S_e$) with $z$-selectively-rationalizable plans ($S_{e^z}$);*
3. *rationalizable plans ($S^\infty$) with $z$-rationalizable plans ($S^\infty_z$).*

To verify Remark 3, one can follow the proofs for Section 4 with the opportune substitutions, as highlighted in the Appendix. Although the set of $z$-selectively-rationalizable plans does not refine the set of selectively rationalizable plans for a given agreement, the following holds.

**Proposition 9** *If an outcome is implementable under priority to the path, then it is implementable under priority to rationality.*

[40] In the companion paper, I provide an example of a SPE whose path constitutes a credible path agreement, but no explicit threats are credible under priority to the path.

For all the agreements analyzed in the previous sections that prescribe a precise outcome $z$, the conclusions do not change under priority to the path. Hence, the insights from the examples are robust to the finer epistemic priority order adopted in this section. In the example of this section, the agreement on the SPE plans is self-enforcing under priority to rationality (by Theorem 2) but not to the path because the corresponding path agreement is not credible. Such path resembles[41] a "*path that can be upset by a convincing deviation*", a notion proposed by Osborne [31] for repeated coordination games. Osborne proves that such paths are not stable, in the sense of Kohlberg and Mertens [27]. In the Supplemental Appendix, I prove that agreements on such paths are not credible. Analogously, in signaling games, Battigalli and Siniscalchi [12] show that a violation of the intuitive criterion implies emptiness of Strong-$\Delta$-Rationalizability with belief restrictions on the equilibrium outcome distribution, and Cho and Kreps [19] show that an equilibrium that does not satisfy the intuitive criterion is not strategically stable. Sobel et al. [36] provide similar arguments both for the intuitive criterion and for divine equilibria (Banks and Sobel [3]).

In all these refinements, and in Govindan and Wilson [22], the focus is kept on sequential equilibria. Already under the baseline hypotheses of Section 4, the distinction between subgame perfect and non-subgame perfect equilibria appears meaningless for self-enforceability (see the example of Section 2).[42] Further, subgame perfection even seems at odds with the stricter interpretation of deviations that these refinements aim to capture: if the deviator is trying to achieve a higher payoff than under the path, she will *certainly not* best reply to the threat (see the first example in the Supplemental Appendix).

[41] Osborne's definition is more restrictive. The epistemic approach of this paper allows to capture precisely the hypotheses that inspire Osborne's solution concept.

[42] Interestingly, Man [30] finds out that also the invariance argument, used to motivate the notions of forward induction of Kohlberg and Mertens [27] and Govindan and Wilson [22], does not imply sequential equilibrium.

# 7 Appendix - Proofs

To prove all the results of Section 4 under priority to the agreement (i.e., to prove Remark 2), substitute $(S_i^\infty)_{i\in I}$ with $(S_i)_{i\in I}$ (or $S_i^1$ where indicated in footnote) and $((S_{j,e}^q)_{j\in I})_{q=0}^\infty$ with $((S_{j,\Delta^e}^q)_{j\in I})_{q=0}^\infty$; under priority to the path (i.e., to prove Remark 3), substitute $P \subseteq Z$ with $z \in Z$, $(S_i^\infty)_{i\in I}$ with $(S_{i,z}^\infty)_{i\in I}$, and $((S_{j,e}^q)_{j\in I})_{q=0}^\infty$ with $((S_{j,e^z}^q)_{j\in I})_{q=0}^\infty$ (recalling that only agreements and SES's that prescribe a single $z$ are considered).

**Proof of Proposition 1.** Since $e$ is credible, $S_e^\infty \cap e^0 \neq \emptyset$. Since $\zeta(S_e^\infty)$ is a singleton and $\zeta(S_e^\infty) \supseteq \zeta(S_e^\infty \cap e^0)$, $\zeta(S_e^\infty) = \zeta(S_e^\infty \cap e^0)$. ■

**Proof of Proposition 2.** Let $e = (e_i)_{i\in I}$ be an agreement that implements $P$. Let $S^* = S_e^\infty \cap e^0$.

Fix $i \in I$ and $\mu_i$ that strongly believes $S_{-i}^*$. Fix $\mu_i' \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^\infty$ such that $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S^*)$. Thus, $\rho_i(\mu_i') \subseteq S_{i,e}^\infty$. So, by self-enforceability of $e$, $\zeta(\rho_i(\mu_i') \times S_{-i}^*) \subseteq \zeta(S^*)$. For every $s_i \in \rho_i(\mu_i)$, there is $s_i' \in \rho_i(\mu_i')$ such that $s_i(h) = s_i'(h)$ for all $h \in H(S^*)$. Then, $\zeta(\rho_i(\mu_i) \times S_{-i}^*) \subseteq \zeta(S^*)$ as well.

Fix $s_i \in S_i^* \subseteq S_{i,e}^\infty$. By finiteness of the game,[43] for every $s_i \in S_{i,e}^\infty$ there exists $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^\infty$, thus that strongly believes $(S_{j,e}^\infty)_{j\neq i}$, $(S_j^\infty)_{j\neq i}$, and $(e_j^0)_{j\neq i}$, such that $s_i \in \rho_i(\mu_i)$. Then, $\mu_i$ strongly believes also $(S_{j,e}^\infty \cap e_j)_{j\neq i}$. ■

**Proof of Proposition 3.** Let $z := \zeta(S^*)$. If: For each $i \in I$ and $s_{-i} \in S_{-i}^*$, by definition of realization-strict Nash we have $u_i(z) = u_i(\zeta(s_i', s_{-i})) > u_i(\zeta(s_i'', s_{-i}))$ for all $s_i' \in S_i(z)$ and $s_i'' \notin S_i(z)$. Then, for each $\mu_i$ that strongly believes $S_{-i}^*$, since $\mu_i(S_{-i}^*|h^0) = 1$, every continuation best reply to $\mu_i(\cdot|h^0)$ is in $S_i(z)$. Thus, we have $\rho_i(\mu_i) \subseteq S_i(z)$. Hence, $\zeta(\rho_i(\mu_i) \times S_{-i}^*) = \{z\}$. Only if: For each $(s_i^*)_{i\in I} \in S^*$, $i \in I$, and $\mu_i$ that strongly believes $S_{-i}^*$ with $\mu_i(s_{-i}^*|h^0) = 1$, by Realization-strictness $\rho_i(\mu_i) \subseteq S_i(z)$. For every continuation best reply

[43] The vast majority of infinite dynamic games used in applications (such as infinitely repeated games) satisfy this property too (see, for instance, the class of "simple dynamic games" defiined in Battigalli and Tebaldi, 2017)

$s_i$ to $\mu_i(\cdot|h^0)$, there exists $s_i' \in \rho_i(\mu_i)$ such that $s_i'(h) = s_i(h)$ for all $h \in H(s_i)$ with $\mu_i(S_{-i}(h)|h^0) > 0$. Hence, $\rho_i(\mu_i) \subseteq S_i(z)$ and $\mu_i(S_{-i}(z)|h^0) = 1$ require $s_i \in S_i(z)$. Thus, $\arg\max_{s_i} u_i(\zeta(s_i, s_{-i}^*)) = S_i(z) \ni s_i^*$: $(s_i^*)_{i\in I}$ is a realization-strict Nash. ■

**Proof of Proposition 4.** By Self-Justifiability, $S^* \subseteq S_e^1$. By definition, $S_e^1 = \overline{S}^*$. Thus, by Forward Induction, $S_e^1 \subseteq S_e^2$, and obviously $S_e^1 \supseteq S_e^2$. So, $S^* \subseteq S_e^1 = S_e^2 = S_e^\infty$.

For each $i \in I$ and $s_i \in S_{i,e}^1$, there is $\mu_i$ that strongly believes $(S_j^*)_{j\neq i}$ and thus $S_{-i}^*$ such that $s_i \in \rho_i(\mu_i)$. By Realization-strictness, for each $h \in H(S^*) \cap H(s_i)$, $s_i(h) = s_i'(h)$ for some $s_i' \in S_i^*(h)$. Then, $\zeta(S_e^1) \subseteq \zeta(S^*)$.

So, $\zeta(S_e^\infty) = \zeta(S_e^1) \subseteq \zeta(S^*) = \zeta(S_e^\infty \cap S^*)$, and obviously $\zeta(S_e^\infty \cap S^*) \subseteq \zeta(S_e^\infty)$. Thus, $\zeta(S_e^\infty) = \zeta(S_e^\infty \cap S^*) = \zeta(S^*)$: $e$ is self-enforcing and truthful. ■

**Proof of Proposition 5.** Suppose that $S^* = \times_{i\in I} S_i^* \subseteq S^\infty$ satisfies Realization-strictness and Self-Justifiability. For each $j \in I$ and $s_j \in \overline{S}_j^*$, there is $\mu_j$ that strongly believes $(S_i^*)_{i\neq j}$ and thus $S_{-j}^*$ such that $s_j \in \rho_j(\mu_j)$. Hence, by Realization-strictness, for each $h \in H(S^*) \cap H(s_j)$ there is $s_j' \in S_j^*(h)$ such that $s_j(h) = s_j'(h)$. Then $H(\overline{S}^*) \subseteq H(S^*)$. Moreover, by Self-Justifiability, $S_j^* \subseteq \overline{S}_j^*$ for all $j \in I$.

In games of depth 2, $H(\overline{S}^*) \subseteq H(S^*)$ implies $H(\overline{S}_j^*) \subseteq H(S_j^*)$ for all $j \in I$.[44] So, strong belief in $S_j^* \subseteq \overline{S}_j^*$ implies strong belief in $\overline{S}_j^*$. Then, for each $i \in I$, Definition 12 implies that $\overline{S}_i^*$ satisfies Forward Induction.

In two-players games, for each $\mu_i$ that strongly believes $S_j^* \subseteq \overline{S}_j^*$ and $S_j^\infty$ $(j \neq i)$, there is $\mu_i'$ that strongly believes $S_j^*$, $\overline{S}_j^*$, and $S_j^\infty$ such that $\mu_i(\cdot|h) = \mu_i'(\cdot|h)$ for all $h \in H(\overline{S}^*) \subseteq H(S^*)$ and all $h \in H(\overline{S}_i^*) \backslash H(\overline{S}_j^*)$, thus for all $h \in H(\overline{S}_i^*)$. Since $\rho_i(\mu_i) \subseteq \overline{S}_i^*$ by definition of $\overline{S}_i^*$, $\rho_i(\mu_i) = \rho_i(\mu_i')$. Then, Definition 12 implies that $\overline{S}_i^*$ satisfies Forward Induction. ■

**Proof of Proposition 6.** Let $S^*$ be a Cartesian set of rationalizable profiles with $\zeta(S^*) = \{z\}$. By Proposition 3, it is a set of realization-strict

[44] Because it implies that the moves allowed by $\overline{S}_j^*$ and $S_j^*$ at $h^0$ are the same.

Nash equilibria. Fix $i \in I$ and $s_i \in S_i^*$. Since $s_i \in S_i^\infty$,[45] by finiteness of the game there exists $\mu_i$ that strongly believes $S_j^\infty$ $(j \neq i)$ such that $s_i \in \rho_i(\mu_i)$.[46] Fix $s_j \in S_j^*$ and construct $\mu_i'$ that strongly believes $S_j^*$ and $S_j^\infty$ such that $\mu_i'(s_j|h^0) = 1$ and $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \notin H(S_j^*)$. Since $(s_i, s_j)$ is a realization-strict Nash that induces $z$, for every $h \prec z$ the set of continuation best replies to $\mu_i'(\cdot|h)$ coincides with $S_i(z)$. For every $h \in H(s_i)$ with $h \not\prec z$, $h \notin H(S_j^*)$, thus $s_i$ is a continuation best reply to $\mu_i'(\cdot|h)$ by $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$. So, $s_i \in \rho_i(\mu_i')$. ■

**Proof of Remark 1.** Fix $i \in I$. By T3, $e_i^0 = e_i^0 \cap S_i(h^0) \subseteq \rho_i(\Delta_i^e)$, and by T2, every $\mu_i \in \Delta_i^e$ strongly believes $(S_j^\infty)_{j \neq i}$ (beside $(e_j^0)_{j \neq i}$). ■

**Proof or Proposition 7.** Fix $i \in I$ and $\mu_i \in \Delta_i^e$. For each $j \neq i$ and $h \in H(S_j^\infty)$, by T2 there is $n \leq k_j$ such that $e_j^n \cap S_j(h) \neq \emptyset$. Then, since $\mu_i$ strongly believes $e_j^n$, $\mu_i(S_j^\infty \times S_{-i,j}|h) \geq \mu_i(e_j^n \times S_{-i,j}|h) = 1$. Thus, $\mu_i$ strongly believes $(S_j^\infty)_{j \neq i}$. Therefore, $\rho_i(\Delta_i^e) \cap S_i^\infty = S_{i,e}^1$ for all $i \in I$.

Now, fix again $i \in I$ and $\mu_i \in \Delta_i^e$. For each $j \neq i$ and $h \in H(S_{j,e}^1) = H(\rho_j(\Delta_j^e) \cap S_j^\infty)$, by T3 there is $n \leq k_j$ such that $e_j^n \cap S_j(h) \subseteq \rho_j(\Delta_j^e) \cap S_j^\infty = S_{j,e}^1$. Then, since $\mu_i$ strongly believes $e_j^n$, $\mu_i(S_{j,e}^1 \times S_{-i,j}|h) \geq \mu_i(e_j^n \times S_{-i,j}|h) = 1$. Thus, $\mu_i$ strongly believes $(S_{j,e}^1)_{j \neq i}$, beside $(S_j^\infty)_{j \neq i}$. Therefore, $\rho_i(\Delta_i^e) \cap S_i^\infty \subseteq S_{i,e}^2$, and obviously $S_{i,e}^2 \subseteq S_{i,e}^1$. So, $S_{i,e}^1 = S_{i,e}^2$ for all $i \in I$. Thus, $S_e^1 = S_e^\infty$.

For each $i \in I$ and $s_i \in S_{i,e}^1$, there is $\mu_i$ that strongly believes $(e_j^0)_{j \neq i}$ and thus $e_{-i}^0$ such that $s_i \in \rho_i(\mu_i)$. By Realization-strictness, for each $h \in H(e^0) \cap H(s_i)$, we have $s_i(h) = s_i'(h)$ for some $s_i' \in e_i^0 \cap S_i(h)$. Then, $\zeta(S_e^1) \subseteq \zeta(e^0)$.

For each $i \in I$, by T3 at $h^0$, $e_i^0 \subseteq \rho_i(\Delta_i^e) \cap S_i^\infty = S_{i,e}^1 = S_{i,e}^\infty$.

So, $\zeta(S_e^\infty) = \zeta(S_e^1) \subseteq \zeta(e^0) = \zeta(S_e^\infty \cap e^0)$, and obviously $\zeta(S_e^\infty \cap e^0) \subseteq \zeta(S_e^\infty)$. Thus, $\zeta(S_e^\infty) = \zeta(e^0) = \zeta(S_e^\infty \cap e^0)$: $e$ is self-enforcing and truthful. ■

**Proof of Theorem 3 (Only if).** Fix an implementable outcome set $P$ and a self-enforcing agreement $e = (e_i)_{i \in I}$ that implements it. Let $M$ be the

[45] Under priority to the agreement, here $S_i^\infty$ must be substituted by $S_i^1$ in place of $S_i^0 = S_i$. In the rest of the proof, substitute $S_j^\infty$ with $S_j$ as usual.

[46] Much milder conditions than finitess guarantee this fact. For instance, simple games as defined by Battigalli and Tebaldi [13].

smallest $m \geq 0$ such that $S_e^\infty = S_e^m$ (it exists by finiteness of the game)[47]. Note that, for each $i \in I$, $S_{i,e}^M$ is the set of all $s_i \in S_i^\infty$ such that $s_i \in \rho_i(\mu_i)$ for some $\mu_i$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^M$ and $((e_j^q)_{j\neq i})_{q=0}^{k_j}$.

The proof is constructive. Let $e_i^{k_i+1} := S_i^\infty$. For each $q = 0, ..., M + k_i + 1$, let

$$\overline{e}_i^q = \bigcup_{(n,m)\in\{0,k_i+1\}\times\{0,M\}:n+m\leq q} (e_i^n \cap S_{i,e}^{M-m}).$$

To see graphically the construction of each $\overline{e}_i^q$, suppose that $M \geq k_i$. In the table below, each box represents the intersection of its coordinates, and the union of the boxes marked with "x" represents $\overline{e}_i^q$ for some $q \leq k_i$:

| $\cap$ | $S_{i,e}^M$ | ... | $S_{i,e}^{M-q}$ | ... | ... | $S_{i,e}^0$ |
|---|---|---|---|---|---|---|
| $e_i^0$ | x | x | x | | | |
| ... | x | x | | | | |
| $e_i^q$ | x | | | | | |
| ... | | | | | | |
| $e_i^{k_i+1}$ | | | | | | |

So, $\overline{e}_i^q$ is the union of all boxes above the line that connects box $e_i^q \cap S_{i,e}^M$ with box $e_i^0 \cap S_{i,e}^{M-q}$. Starting from $\overline{e}_i^0 = e_i^0 \cap S_{i,e}^M$, every increase of $q$ by 1 shifts such line by 1 to the right, until $\overline{e}_i^{k_i+M+1} = e_i^{k_i+1} \cap S_{i,e}^0 = S_i^\infty$ (all other boxes are subsets). Note that for $q > k_i$, $\overline{e}_i^q$ includes the whole columns from $S_{i,e}^{M-q+k_i+1}$ to $S_{i,e}^M$; for $q > M$, $\overline{e}_i^q$ includes the whole rows from $e_i^0$ to $e_i^{q-M-1}$. If $M < k_i$, the table will have more rows than colums and rows will start to be filled earlier than columns.

Without loss of generality, suppose that $\overline{e}_i^n \subsetneq \overline{e}_i^{n+1}$ for each $n = 0, ..., k_i + M$;[48] then, $\overline{e} = (\overline{e}_i)_{i\in I}$ is an agreement. To see that $\overline{e}$ prescribes $P$, note that

$$P = \zeta(S_e^\infty) = \zeta(S_e^M \cap e^0) = \zeta(\overline{e}^0),$$

[47] Here finiteness cannot be substituted by milder assumptions; for instance, in an infinitely repeated game, convergence to $S^\infty$ and $S_e^\infty$ may require an infinite number of steps.

[48] If $\overline{e}_i^n = \overline{e}_i^{n+1}$ for some $n$, $\overline{e}_i^{n+1}$ can simply be eliminated from the chain.

where the first equality is by implementation of $P$, the second by self-enforceability of $e$, and the third by construction.

I am going to show that strong belief in $(\overline{e}_i^q)_{q=0}^{k_i+M+1}$ is equivalent to strong belief in $(S_{i,e}^q)_{q=0}^M$ and $(e_i^q)_{q=0}^{k_i}$. This yields $S_{j,e}^\infty = \rho_j(\Delta_j^{\overline{e}}) \cap S_j^\infty$ for each $j \in I$.

First, I show that every $\mu_j$ that strongly believes $(S_{i,e}^q)_{q=0}^M$ and $(e_i^q)_{q=0}^{k_i}$ strongly believes also $(\overline{e}_i^q)_{q=0}^{k_i+M+1}$. Since $e_i^{k_i+1} = S_{i,e}^0 = S_i^\infty$, formally $\mu_i$ strongly believes also $e_i^{k_i+1}$. Fix $q \in \{0, ..., k_i + M + 1\}$. For each $h \in H(\overline{e}_i^q)$, by construction $h \in H(e_i^n \cap S_{i,e}^m)$ for some $n \in \{0, ..., k_i + 1\}$ and some $m \in \{0, ..., M\}$ with $e_i^n \cap S_{i,e}^m \subseteq \overline{e}_i^q$. Since $\mu_j$ strongly believes $e_i^n$ and $S_{i,e}^m$, it strongly believes also $e_i^n \cap S_{i,e}^m$, thus $1 = \mu_j((e_i^n \cap S_{i,e}^m) \times S_{-j,i}|h) \leq \mu_j(\overline{e}_i^q \times S_{-j,i}|h)$. Hence, $\mu_j$ strongly believes $\overline{e}_i^q$.

Second, I show that every $\mu_j$ that strongly believes $(\overline{e}_i^q)_{q=0}^{k_i+M+1}$ strongly believes also $(e_i^q)_{q=0}^{k_i}$ and $(S_{i,e}^q)_{q=0}^M$.

Fix $n = 0, ..., k_i$ and $h \in H(e_i^n)$. Fix the highest $m \in \{0, ..., M\}$ such that $h \in H(S_{i,e}^m)$ (it exists because $S_{i,e}^0 = S_i^\infty \supseteq e_i^n$). I show that $\emptyset \neq \overline{e}_i^{M-m+n} \cap S_i(h) \subseteq e_i^n$. Note that $e_i^n \cap S_{i,e}^m \subseteq \overline{e}_i^{M-m+n}$. Since $h \in H(e_i^n) \cap H(S_{i,e}^m)$, by credibility of $e$ there exists $\mu_j'$ that strongly believes $(S_{i,e}^q)_{q=0}^M$ and $(e_i^q)_{q=0}^{k_i}$ such that $\mu_j'(e_i^n \times S_{-j,i}|h) = \mu_j'(S_{i,e}^m \times S_{-j,i}|h) = 1$. Hence, $\emptyset \neq e_i^n \cap S_{i,e}^m \cap S_i(h) \subseteq \overline{e}_i^{M-m+n}$. Moreover, for every $m'$ with $m < m' \leq M$, $S_i^{m'} \cap S_i(h) = \emptyset$ by construction of $m$, and for every $m' \leq m$, $S_i^{m'} \cap e_i^{n'} \subseteq \overline{e}_i^{M-m+n}$ only if $n' \leq n$, which implies $e_i^{n'} \subseteq e_i^n$. So, $\overline{e}_i^{M-m+n} \cap S_i(h) \subseteq e_i^n$. (Graphically: $e_i^n \cap S_{i,e}^m$ is along the diagonal that identifies $\overline{e}_i^{M-m+n}$, all the boxes to the left of column $S_{i,e}^m$ do not reach $h$, and the triangle to the right occupies only rows above row $e_i^n$.)

Fix $m = 0, ..., M$ and $h \in H(S_{i,e}^m)$. Fix the lowest $n \in \{0, ..., k_i + 1\}$ such that $h \in H(e_i^n)$ (it exists because $e_i^{k_i+1} = S_i^\infty \supseteq S_{i,e}^m$). I show that $\emptyset \neq \overline{e}_i^{M-m+n} \cap S_i(h) \subseteq S_{i,e}^m$. Note that $e_i^n \cap S_{i,e}^m \subseteq \overline{e}_i^{M-m+n}$. Since $h \in H(e_i^n) \cap H(S_{i,e}^m)$, by credibility of $e$ there exists $\mu_j'$ that strongly believes $(S_{i,e}^q)_{q=0}^M$ and $(e_i^q)_{q=0}^{k_i}$ such that $\mu_j'(e_i^n \times S_{-j,i}|h) = \mu_j'(S_{i,e}^m \times S_{-j,i}|h) = 1$. Hence, $\emptyset \neq e_i^n \cap S_{i,e}^m \cap S_i(h) \subseteq \overline{e}_i^{M-m+n}$. Moreover, for every $n'$ with $0 \leq n' < n$, $e_i^{n'} \cap S_i(h) = \emptyset$ by construction of $n$, and for every $n' \geq n$, $S_{i,e}^{m'} \cap e_i^{n'} \subseteq \overline{e}_i^{M-m+n}$ only if $m' \geq m$, which implies $S_{i,e}^{m'} \subseteq S_{i,e}^m$. So, $\overline{e}_i^{M-m+n} \cap S_i(h) \subseteq S_{i,e}^m$. (Graphically: $e_i^n \cap S_{i,e}^m$ is

along the diagonal that identifies $\overline{e}_i^{M-m+n}$, all the boxes above line $e_i^n$ do not reach $h$, and the triangle below occupies only columns left of column $S_{i,e}^m$.)

With $m = 0$, the last paragaph also shows that $\overline{e}$ satisfies T2. With $m = M$, since $S_{i,e}^M = \rho_i(\Delta_i^{\overline{e}}) \cap S_i^\infty$, it shows that $\overline{e}$ satisfies T3. It remains to show that $\overline{e}$ satisfies T1. Fix $i \in I$ and $\mu_i$ that strongly believes $\overline{e}_{-i}^0$. I am going to show that $\zeta(\rho_i(\mu_i) \times \overline{e}_{-i}^0) \subseteq \zeta(\overline{e}^0)$. Fix $\mu_i' \in \Delta_i^{\overline{e}}$ that strongly believes $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that for each $h \in H(\overline{e}_{-i}^0)$, $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ (it exists because agreements are in rationalizable plans). Thus,

$$\zeta(\rho_i(\mu_i) \times \overline{e}_{-i}^0) = \zeta(\rho_i(\mu_i') \times \overline{e}_{-i}^0) = \zeta((\rho_i(\mu_i') \cap S_i^\infty) \times \overline{e}_{-i}^0).$$

By $\mu_i' \in \Delta_i^{\overline{e}}$, $\rho_i(\mu_i') \cap S_i^\infty \subseteq S_{i,e}^M$. So, by $\overline{e}^0 = S_e^M \cap e^0$ and self-enforceability of $e$,

$$\zeta((\rho_i(\mu_i') \cap S_i^\infty) \times \overline{e}_{-i}^0) \subseteq \zeta(S_e^M) = \zeta(S_e^M \cap e^0) = \zeta(\overline{e}^0).$$

■

# References


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

[2] Aumann, R., "Nash-Equilibria are not Self-Enforcing", in Economic Decision Making: Games, Econometrics and Optimisation (J. Gabszewicz, J.-F. Richard, and L. Wolsey, Eds.), Amsterdam, Elsevier,1990, 201-206.

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

[4] Basu, K. and J. W. Weibull, "Strategy subsets closed under rational behavior", *Economic Letters*, **36**, 1991, 141-146.

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

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

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

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

[9] 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*, **13(1)**, 2013, 79-130.

[10] Battigalli, P. and M. Siniscalchi, "Interactive Beliefs, Epistemic Independence and Strong Rationalizability", *Research in Economics,* **53** (1999), 247-273.

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

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

[13] Battigalli, P. and P. Tebaldi, "Interactive Epistemology in Simple Dynamic Games with a Continuum of Strategies," *Economic Theory*, (2018), 1-27.

[14] Ben Porath, E. and E. Dekel, "Signaling future actions and the potential for sacrifice," *Journal of Economic Theory*, **57**, 1992, 36-51.

[15] Bergemann, D. and S. Morris, "Robust Implementation in Direct Mechanisms", *Review of Economic Studies*, **76**, 2009, 1175–1204

[16] Brandenburger, A., and A. Friedenberg, "Intrinsic correlation in games", *Journal of Economic Theory*, **141,** 2008, 28-67.

[17] Catonini, E., "Rationalizability and epistemic priority orderings", working paper, 2018.

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

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

[20] Dixit, A., "The Role of Investment in Entry-Deterrence", *The Economic Journal*, 1980, 90-95.

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

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

[23] Greenberg, J., "The right to remain silent", *Theory and Decisions*, **48(2)**, 2000, 193-204.

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

[25] Harrington, J. "A Theory of Collusion with Partial Mutual Understanding", *Research in Economics,* forthcoming.

[26] 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.

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

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

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

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

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

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

[33] Perea, A., "Forward Induction Reasoning and Correct Beliefs", *Journal of Economic Theory*, **169**, 2017, 489-516.

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

[35] Siniscalchi, M., "Structural Rationality in Dynamic Games", working paper, 2018.

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

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

# 8 Supplemental Appendix

## 8.1 On SPE and self-enforcing agreements

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 plans are justifiable, hence they are all 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 outcome $(S,E)$.

The SPE outcome $(S,E)$ is implementable, but differently from the game in the Introduction, only with an agreement that features also the off-the-path threat $R$ by Bob. For instance, the reduced agreement on the strict Nash $(S,E.R)$ is self-enforcing by Theorem 2. Instead, the path agreement on $z=(S,E)$ is not self-enforcing because Ann may rationally deviate and then play $U$ or $M$, hence Bob could best reply with any action, and not just with $R$. Formally, we have $S^{\infty}_{A,z}=S^{1}_{A,z}=\{S,N.U,N.M\}$ and $S^{\infty}_{B,z}=S^{1}_{B,z}=\{E.L,E.C,E.R\}$.

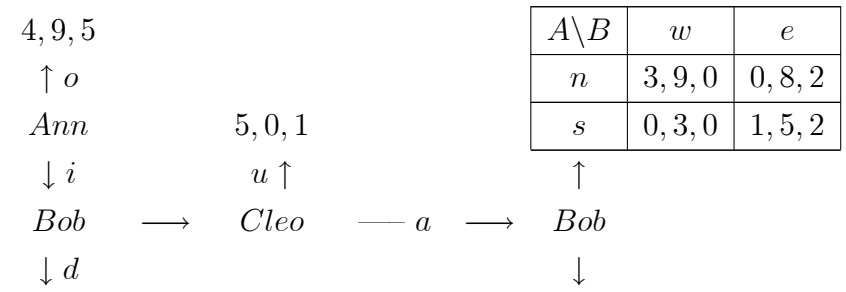
Note moreover that if Ann believes in the SPE path, it is not rational for her to deviate and then play $D$. Thus, if Bob interprets the deviation of Ann as an attempt to increase her payoff with respect to the equilibrium (as implicitly assumed by strategic stability and related refinements, see Section 6), the fact that $R$ is a best reply to $D$ which is best reply to $R$ itself is of no value: $R$ is a credible reaction of Bob only by virtue of other beliefs he may have.

Players can also implement the outcome $(N, W)$, and differently from the game in the Introduction, only with an agreement that *does not* feature a threat played with positive probability in an equilibrium of the subgame (here, just $D$). For instance, the reduced agreement with $e_A^0 = \{N.U, N.M\}$ and $e_B^0 = \{W\}$ is self-enforcing: we have $S_e^1 = \{N.U, N.M, N.D\} \times \{W\}$, thus $S_e^\infty = S_e^1 = S((N, W))$.

To conclude, note that there is no conceptual difference behind the reasons for self-enforceability of the SPE and of the Pareto-superior Nash outcome.

## 8.2 Another form of agreement incompleteness

Consider the following game.




| $A\backslash B$ | $w$ | $e$ |
|---|---|---|
| $n$ | $3, 9, 0$ | $0, 8, 2$ |
| $s$ | $0, 3, 0$ | $1, 5, 2$ |

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

| $A\backslash B$ | $w$ | $e$ |
|---|---|---|
| $n$ | $3, 9, 0$ | $0, 8, 2$ |
| $s$ | $0, 3, 0$ | $1, 5, 2$ |

All plans are justifiable, hence they are all rationalizable. Players want to implement outcome $(o)$. As suggested in Section 4, we first look for the sets $S^* = S_A^* \times S_B^* \times S_C^* \subseteq S^\infty = S$ that induce $(o)$ and satisfy Self-Enforceability and Self-Justifiability. Ann's Self-Enforceability requires Bob not to play $d$ and Cleo not to play $u$. Then, Bob's Self-Justifiability requires that Cleo may play $t$, and Cleo's Self-Justifiability requires that Bob may play $e$ in a subgame he allows. Hence, calling $S_B^w$ and $S_B^e$ the binary sets of plans of Bob where the last move is $w$ and $e$ respectively, the required sets $S^*$ coincide with those that

satisfy

$$S_A^* = \{o\}, \quad S_B^* \subseteq S_B^w \cup S_B^e,\ S_B^e \cap S_B^* \neq \emptyset, \quad S_C^* \subseteq \{t.a, b.a\},\ t.a \in S_C^*.$$

Does any of these sets satisfy Forward Induction? No. Under belief in $S_C^*$, it is irrational for Bob to play $d.l$, because both plans in $S_B^e$ guarantee a higher payoff. Yet, it is rational to play $d.c$, because $t.a \in S_C^*$. Therefore, Forward Induction requires Cleo to play $b$ and not $t$, a contradiction. Thus, there is no SES that implements $(o)$.

So, we look for a tight agreement $e$ where $e^0$ satisfies the conditions above and alternative plans of Ann and Bob, $e_A^1$ and $e_B^1$, are introduced to reach all histories (for T2) and restrict their behavior after deviations to $i$ and $d$. First, observe that we need $e_C^0 = \{t.a\}$. If $b.a \in e_C^0$, then, regardless of $e_A^1$, we have $d.r \in \rho_B(\Delta_B^e) \cap S_B((i,d))$, but $d.l \notin \rho_B(\Delta_B^e)$. So, for Bob, T3 imposes $d.l \notin e_B^1 \cap S_B((i,d)) \neq \emptyset$, but then $t.a \notin \rho_C(\Delta_C^e)$, a violation of T3. Still, without restrictions on $e_A^1$, we have $d.c \in \rho_B(\Delta_B^e) \cap S_B((i,d))$, so again $d.l \notin e_B^1 \cap S_B((i,d)) \neq \emptyset$ and $t.a \notin \rho_C(\Delta_C^e)$. Hence, we must obtain $d.c \notin \rho_B(\Delta_B^e)$. So, we must impose $i.s.s \notin e_A^1$. If Ann guarantees to play $n$ in a specific subgame, then we have $\rho_i(\Delta_B^e) \subseteq S_B^w$; hence, T3 imposes $e_B^0 \subseteq S_B^w$, a contradiction of the conditions on $e_B^0$. So, the only remaining option is $e_A^1 = \{i.n.n, i.n.s, i.s.n\}$. Then, on the one hand there is $\mu_B \in \Delta_B^e$ with $\mu_B(i.n.s|(i)) = \mu_B(i.s.n|(i)) = 1/2$ and $\rho_B(\mu_B) = S_B^e$; on the other hand, for every $\mu_B \in \Delta_B^e$, there is $s_B \in \rho_B(\mu_B) \cap (S_B^w \cup S_B^e)$ that gives to Bob an expected payoff of at least 6.5, so $d.c \notin \rho_B(\Delta_B^e) \cap S_B((i,d)) = \emptyset$.

$$e_A^0 = \{o\}, \quad e_B^0 = S_B^w \cup S_B^e, \quad e_C^0 = \{t.a\};$$
$$e_A^1 = \{i.n.n, i.n.s, i.s.n\}, \quad e_B^1 = \{d.l, d.c, d.r\}.$$

The vagueness of Ann about in which subgame she is going to play $n$ is a kind of agreement incompleteness that, like here, can be necessary to implement an outcome. It can be interpreted as Ann doing the following speech: "I guarantee that I will be prepared to play $n$ in at least one contingency, but I

cannot guarantee that I will be prepared to play $n$ in both."

This kind of strategic uncertainty also arises naturally from strategic reasoning. Example ?? in Battigalli [6] (provided by Gul and Reny) shows that already the set of justifiable plans of a player is not a Cartesian product of sets of actions at different information sets. This is the reason why extensive-form rationalizability is defined as an elimination procedure of plans instead of actions at different information sets, and agreements are defined in terms of plans as well.

## 8.3 Proofs for Sections 5 and 6

For any $h \in \overline{H} \setminus \{h^0\}$, let $p(h) \in H$ be the immediate predecessor of $h$.

**Proof of Proposition 8**

Fix an outcome set $P \subseteq Z$ that is implementable under priority to rationality. Then, by Theorem 3, $P$ is implemented by an agreement $e = (e_i)_{i \in I}$ which is tight under priority to rationality. The proof is constructive. Let $M$ be the smallest $m$ such that $S^m = S^\infty$ (it exists by finiteness of the game). For each $i \in I$ and $n = 0, ..., k_i$, let

$$\overline{e}_i^n := \{s_i \in S_i^\infty : \exists s_i' \in e_i^n, \forall h \in H(s_i') \cap H(S^\infty), s_i'(h) = s_i(h)\};$$

for each $n = k_i + 1, ..., k_i + M + 1$, let $\overline{e}_i^n = S_i^{k_i + M + 1 - n}$. Assume without loss of generality that $\overline{e}_i^n \subsetneq \overline{e}_i^{n+1}$ for each $n = 0, ..., k_i + M$,[49] so that $\overline{e} = (\overline{e}_i)_{i \in I}$ is an agreement. I am going to show that $\overline{e}$ is tight under priority to the agreement. Indicate with $\text{T1}^a$, $\text{T2}^a$ and $\text{T3}^a$ the conditions of tightness under priority to the agreement (i.e., with $S_i$ in place of $S_i^\infty$).

First, I show that $\overline{e}$ satisfies $\text{T1}^a$ (which is identical to T1). Fix $i \in I$ and $\mu_i$ that strongly believes $\overline{e}_{-i}^0 \subseteq S_{-i}^\infty$. For each $j \in I$ and $s_j \in e_j^0$, there is $s_j' \in \overline{e}_j^0$ such that $s_j'(h) = s_j(h)$ for all $h \in H(S^\infty) \cap H(s_j')$, and vice versa. Hence, (i) $\zeta(S_i \times \overline{e}_{-i}^0) \cap \zeta(S^\infty) = \zeta(S_i \times e_{-i}^0) \cap \zeta(S^\infty)$, and there exists $\mu_i'$ that strongly

[49] If $\overline{e}_i^n = \overline{e}_i^{n+1}$ for some $n$, $\overline{e}_i^{n+1}$ can simply be eliminated from the chain.

believes $e^0_{-i}$ such that (ii) $\mu_i(S_{-i}(z)|h) = \mu'_i(S_{-i}(z)|h)$ for all $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$. By T1, $\zeta(\rho_i(\mu'_i) \times e^0_{-i}) \subseteq \zeta(e^0) \subseteq \zeta(S^\infty)$. Then, by (i), $\zeta(\rho_i(\mu'_i) \times e^0_{-i}) = \zeta(\rho_i(\mu'_i) \times \overline{e}^0_{-i}) \subseteq \zeta(S^\infty)$. Note that $\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}) \subseteq \zeta(S^\infty)$ as well, because $\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}) = \zeta(\rho_i(\widetilde{\mu}_i) \times \overline{e}^0_{-i}) \subseteq \zeta(S^\infty)$ for any $\widetilde{\mu}_i$ that strongly believes $((S^q_j)_{j\neq i})^\infty_{q=0}$ with $\widetilde{\mu}_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(\overline{e}^0_{-i})$. Hence, by (ii), $\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}) = \zeta(\rho_i(\mu'_i) \times \overline{e}^0_{-i})$. So, we obtain

$$\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}) = \zeta(\rho_i(\mu'_i) \times \overline{e}^0_{-i}) = \zeta(\rho_i(\mu'_i) \times e^0_{-i}) \subseteq \zeta(e^0) = \zeta(\overline{e}^0).$$

Moreover, $\overline{e}$ satisfies T2$^a$ by $\overline{e}^{k_i+M+1}_i = S_i$. It remains to show that $\overline{e}$ satisfies T3$^a$.

Fix $\mu_i \in \Delta^{\overline{e}}_i$ and $\overline{h} \in H(\rho_i(\mu_i) \cap S_i)$.[50] Suppose first that $p(\overline{h}) \in H(S^\infty)$.

By construction of $\overline{e}$, I can construct $\mu'_i \in \Delta^e_i$ that strongly believes[51] $((S^q_j)_{j\neq i})^\infty_{q=0}$ such that for each $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$, $\mu'_i(S_{-i}(z)|h) = \mu_i(S_{-i}(z)|h)$. Since $\mu_i$ and $\mu'_i$ strongly believe $S^\infty_{-i}$ (by $\times_{j\neq i}\overline{e}^{k_j+1}_j = S^\infty_{-i}$ and by T2 respectively), $\zeta(\rho_i(\mu_i) \times S^\infty_{-i}), \zeta(\rho_i(\mu'_i) \times S^\infty_{-i}) \subseteq \zeta(S^\infty)$. Then,

$$\zeta(\rho_i(\mu_i) \times S^\infty_{-i}) = \zeta(\rho_i(\mu'_i) \times S^\infty_{-i}) = \zeta((\rho_i(\mu'_i) \cap S^\infty_i) \times S^\infty_{-i}).$$

Thus, since $p(\overline{h}) \in H(\rho_i(\mu_i)) \cap H(S^\infty)$, $p(\overline{h}) \in H(\rho_i(\mu'_i) \cap S^\infty_i) \cap H(S^\infty)$ as well; moreover, at $p(\overline{h})$, plans in $\rho_i(\mu_i)$ and in $\rho_i(\mu'_i) \cap S^\infty_i$ must prescribe the same moves. Hence, $\overline{h} \in H(\rho_i(\mu'_i) \cap S^\infty_i)$. Then, by T3, there is $n \leq k_i$ such that

$$\emptyset \neq e^n_i \cap S_i(\overline{h}) \subseteq \rho_i(\Delta^e_i).$$

So, by construction of $\overline{e}$, $\overline{e}^n_i \cap S_i(\overline{h}) \neq \emptyset$.

Fix $s_i \in \overline{e}^n_i \cap S_i(\overline{h}) \subseteq S^\infty_i$. By construction of $\overline{e}$, there is $s'_i \in e^n_i \cap S_i(\overline{h}) \subseteq \rho_i(\Delta^e_i)$ such that $s'_i(h) = s_i(h)$ for all $h \in H(s'_i) \cap H(S^\infty)$. Fix $\widehat{\mu}_i \in \Delta^e_i$ such that $s'_i \in \rho_i(\widehat{\mu}_i)$ and $\widetilde{\mu}_i$ that strongly believes $((S^q_j)_{j\neq i})^\infty_{q=0}$ such that $s_i \in \rho_i(\widetilde{\mu}_i)$.

Fix $\widehat{h} \notin H(S^\infty_{-i})$ with $p(\widehat{h}) \in H(S^\infty_{-i})$, $j \neq i$, and $s_j \in S_j(\widehat{h})$. If $s_j \notin S^\infty_j$, let

[50] Of course, the intersection with $S_i$ is superfluous here. It will be substituted with $S^\infty_i$ in the next proof.

[51] Since agreements are in rationalizable plans, adding strong belief in $((S^q_j)_{j\neq i})^\infty_{q=0}$ is always possible without modifying the CPS at the rationalizable histories.

$\eta_j^{\widehat{h}}(s_j) = s_j$. Else, let $m$ be the smallest $q \leq k_j + 1$ such that $\overline{e}_j^m \cap S_j^\infty(\widehat{h}) \neq \emptyset$ ($m$ exists because $e_j^{k_j+1} = S_j^\infty$) and fix $s'_j \in \overline{e}_j^m \cap S_j^\infty(\widehat{h})$. Define $\eta_j^{\widehat{h}}(s_j) \in S_j(\widehat{h})$ as $\eta_j^{\widehat{h}}(s_j)(h) = s'_j(h)$ for all $h \in H(s'_j)$ with $h \not\succeq \widehat{h}$, and $\eta_j^{\widehat{h}}(s_j)(h) = s_j(h)$ for all $h \in H(s_j)$ with $h \succeq \widehat{h}$. Since $s_j, s'_j \in S_j^\infty$, there are $\mu_j, \mu'_j$ that strongly believe $((S_k^q)_{k\neq j})_{q=0}^\infty$ such that $s_j \in \rho_j(\mu_j)$ and $s'_j \in \rho_j(\mu'_j)$. Since there is $k \neq j$ such that $\widehat{h} \notin H(S_k^\infty)$ and $p(\widehat{h}) \in H(S_k^\infty)$, there is $\overline{\mu}_j$ that strongly believes $((S_k^q)_{k\neq j})_{q=0}^\infty$ such that $\overline{\mu}_j(\cdot|h) = \mu'_j(\cdot|h)$ for all $h \not\succeq \widehat{h}$ and $\overline{\mu}_j(\cdot|h) = \mu_j(\cdot|h)$ for all $h \succeq \widehat{h}$. Then, $\eta_j^{\widehat{h}}(s_j) \in \rho_j(\overline{\mu}_j) \subseteq S_j^\infty$, and since $\eta_j^{\widehat{h}}(s_j)(h) = s'_j(h)$ for all $h \in H(S^\infty) \cap H(s'_j)$, by $s'_j \in \overline{e}_j^m$ and construction of $\overline{e}$, $\eta_j^{\widehat{h}}(s_j) \in \overline{e}_j^m$ as well.

Now, since $e_j^n \subseteq \overline{e}_j^n$ for all $j \neq i$ and $n \leq k_j$ and by T2 $H(e_j^{k_j}) = H(S_j^\infty)$, I can construct $\mu_i \in \Delta_i^{\overline{e}}$ such that $\mu_i(\cdot|h) = \widehat{\mu}_i(\cdot|h)$ for all $h \in H(S_{-i}^\infty)$ and $\mu_i(s_{-i}|h) = \widetilde{\mu}_i\left((\times_{j\neq i}\eta_j^{\widehat{h}})^{-1}(s_{-i}))|h\right)$ for all $\widehat{h} \notin H(S_{-i}^\infty)$ with $p(\widehat{h}) \in H(S_{-i}^\infty)$, $h \succeq \widehat{h}$, and $s_{-i} \in (\times_{j\neq i}\eta_j^{\widehat{h}})(S_{-i}(\widehat{h}))$. Then, $s_i \in \rho_i(\mu_i) \subseteq \rho_i(\Delta_i^{\overline{e}})$.

Suppose now that $\overline{h} \notin H(S^\infty)$. Then, there is $\overline{h}' \preceq \overline{h}$ such that $\overline{h}' \notin H(S^\infty)$ and $p(\overline{h}') \in H(S^\infty)$. As just shown, there is $n \leq k_i$ such that $\emptyset \neq \overline{e}_i^n \cap S_i(\overline{h}') \subseteq \rho_i(\Delta_i^{\overline{e}}) \cap S_i^\infty$. So, it suffices to show that $\overline{e}_i^n \cap S_i(\overline{h}) \neq \emptyset$. Since each $\mu_i \in \Delta_i^{\overline{e}}$ strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ and $\overline{h} \in H(\rho_i(\Delta_i^{\overline{e}}) \cap S_i)$, $\overline{h} \in H(S_i^\infty)$. Fix $s_i \in \overline{e}_i^n \cap S_i(\overline{h}') \subseteq S_i^\infty$, $s'_i \in S_i^\infty(\overline{h})$ and $\mu_i, \mu'_i$ that strongly believe $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $s_i \in \rho_i(\mu_i)$ and $s'_i \in \rho_i(\mu'_i)$. By $\overline{h}' \notin H(S^\infty)$, $p(\overline{h}') \in H(S^\infty)$, and $\overline{h}' \in H(\overline{e}_i^n) \subseteq H(S_i^\infty)$, there exists $\mu''_i$ that strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $\mu''_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \not\succeq \overline{h}'$, and $\mu''_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \succeq \overline{h}'$. So, there is $s''_i \in \rho_i(\mu''_i) \subseteq S_i^\infty$ such that $s''_i(h) = s_i(h)$ for all $h \not\succeq \overline{h}'$ with $h \in H(s_i)$ and $s''_i(h) = s'_i(h)$ for all $h \succeq \overline{h}'$ with $h \in H(s'_i)$. Thus, $s_i \in S_i(\overline{h}')$ implies $s''_i \in S_i(\overline{h}')$ and then $s'_i \in S_i(\overline{h})$ implies $s''_i \in S_i(\overline{h})$. Moreover, $s_i \in \overline{e}_i^n$ implies $s''_i \in \overline{e}_i^n$ by construction of $\overline{e}_i^n$. Hence, $\emptyset \neq \overline{e}_i^n \cap S_i(\overline{h})$. ■

**Proof of Proposition 9.** For each $P \subseteq Z$ which is implementable under priority to the path, a tight agreement $\overline{e}$ that implements $P$ under priority to rationality can be constructed exactly like in the proof of Proposition 8, substituting T1,T2,T3 with T1$^p$,T2$^p$,T3$^p$ (the requirements of tightness under priority to the path, that is, with $S_z^\infty$ in place of $S^\infty$), $S$ with $S^\infty$, $S^\infty$ with $S_z^\infty$, and $((S_i^q)_{i\in I})_{q=0}^\infty$ with $((S_{i,z}^q)_{i\in I})_{q=0}^\infty$. ■

**Proposition 10** *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.*

**Proof.** 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 $\overline{z} = (\overline{a}^1, .., \overline{a}^T)$ of Nash equilibria of the T-fold repetition of $G$ *can be upset by a convincing deviation* if there exist $\tau \in \{1, ..., T-1\}$ and $\widehat{a}_i \neq \overline{a}_i^\tau$ such that, letting $\overline{T} := T - \tau$,

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

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

Condition I says that player $i$ benefits from a unilateral deviation at $\tau$ only if followed by her preferred subpath.[52] Condition J says that player $j$ cannot benefit from a unilateral deviation from that subpath even if followed by her preferred subpath.[53]

Now I can prove the proposition. Let $e_i = (S_i(\overline{z}))$ and $e_j = (S_j(\overline{z}))$. Let $\widehat{h} := (\overline{a}^1, .., (\widehat{a}_i, \overline{a}_j^\tau))$ and $z := (\overline{a}^1, .., (\widehat{a}_i, \overline{a}_j^\tau), b^i, ..., b^i)$. Suppose that $S_e^1(\overline{z}) \neq \emptyset$, otherwise $S_e^2 = \emptyset$. Then, for each $k = i, j$, there exists $\overline{\mu}_k$ that strongly believes $S_{-k}^\infty$ and $S_{-k}(\overline{z})$ such that $\rho_k(\overline{\mu}_k) \cap S_k(\overline{z}) \neq \emptyset$.

Fix $n \in \mathbb{N}$ and suppose that $S_i^{n-1}(z) \neq \emptyset$. Fix $s_j \in S_j$ with $\overline{\mu}_i(s_j|h^0) \neq 0$. Since $\overline{\mu}_i$ strongly believes $S_j^\infty$ and $S_j(\overline{z})$, $s_j \in S_j^\infty(\overline{z})$. Fix $\mu_j$ that strongly.believes $(S_i^q)_{q=0}^\infty$ with $s_j \in \rho_j(\mu_j)$. Since $\overline{\mu}_j$ strongly believes $S_i(\overline{z})$, for each $h \notin H(S_i(\overline{z}))$ with $p(h) \prec \overline{z}$, $\overline{\mu}_j(S_i(h)|p(h)) = 0$. Thus, there exists $\mu_j'$ that strongly believes $(S_i^q)_{q=0}^{n-1}$ such that (i) $\mu_j'(\cdot|h^0) = \overline{\mu}_j(\cdot|h^0)$, (ii), $\mu_j'(S_i(z)|\widehat{h}) = 1$,

[52] In the example of Section 5, $i = Ann$, $j = Bob$, $(\overline{a}^1, \overline{a}^2) = ((FR, W), (W, FR))$, $b^i = (FR, W)$, $c^i = (W, W)$, $\tau = 1$, $\widehat{a}_i = W$, thus $\overline{T} - 1 = 0$. Formally, the first inequality in (I) is not satisfied (equality holds), but this is immaterial because $b^i$ and $c^i$ entail the same action for Bob, against which the best reply of Ann induces $b^i$.

[53] This implies that $i$'s preferred stage-outcome is Nash, reason why Osborne (1991) refers to coordination games.

and (iii) $\mu_j'(\cdot|h) = \mu_j(\cdot|h)$ for all $h \in H(S_j(\overline{z}))$ with $h \not\prec \overline{z}$ and $h \not\succeq \widehat{h}$. Then, there exists $s_j' \in \rho_j(\mu_j') \subseteq S_j^n$ such that: by $\rho_j(\overline{\mu}_j) \cap S_j(\overline{z}) \neq \emptyset$, $\overline{\mu}_j(S_i(z)|h^0) = 1$, and (i), $s_j' \in S_j(\overline{z}) \subseteq S_j(\widehat{h})$; by (ii) and (J), $s_j' \in S_j(z)$; by (iii) and $s_j, s_j' \in S_j(\overline{z})$, $s_j'(h) = s_j(h)$ for all $h \in H(S_j(\overline{z}))$ with $h \not\succeq \widehat{h}$. With these $s_j'$'s, I can construct $\mu_i$ that strongly believes $(S_j^q)_{q=0}^n$ such that $\mu_i(S_j(z)|h^0) = 1$, and $\mu_i(S_j(\widetilde{z})|h^0) = \overline{\mu}_i(S_j(\widetilde{z})|h^0)$ for all $\widetilde{z} \not\succ \widehat{h}$. Thus, by $\rho_i(\overline{\mu}_i) \cap S_i(\overline{z}) \neq \emptyset$, $\overline{\mu}_i(S_j(\overline{z})|h^0) = 1$, and (I), $\emptyset \neq \rho_i(\mu_i) \cap S_i(z) \subseteq S_i^{n+1}(z)$. So, by induction, there exists $\mu_i$ that strongly believes $(S_j^q)_{q=0}^\infty$ and $S_j(\overline{z})$ such that $\emptyset \neq \rho_i(\mu_i) \cap S_i(z) \subseteq S_{i,e}^1(z)$. On the other hand, for every $\mu_i$ that strongly believes $S_j(\overline{z})$, by (I) $\rho_i(\mu_i) \cap S_i(\widehat{h}) \subseteq S_i(z)$, so $S_{i,e}^1(\widehat{h}) \subseteq S_i(z)$. The two things combined imply that for every $\mu_j$ that strongly believes $S_{i,e}^1$ and $S_i(\overline{z})$, $\mu_j(S_i(z)|\widehat{h}) = 1$. So, by (J), $S_{j,e}^2(\widehat{h}) \subseteq S_j(z)$. Since $S_j(\overline{z}) \subseteq S_j(\widehat{h})$, for every $\mu_i$ that strongly believes $S_{j,e}^2$ and $S_j(\overline{z})$, $\mu_i(S_j(z)|h^0) = 1$, so by (I) $\rho_i(\mu_i)(\overline{z}) = \emptyset$. Hence $S_{i,e}^3(\overline{z}) = \emptyset$. So, $S_{j,e}^4 = \emptyset$. ■

## 8.4 On the definition of Selective Rationalizability.

Consider the following, alternative definition of Selective Rationalizability.

**Definition 16** *Let $((S_i^m)_{i\in I})_{m=0}^\infty$ denote the Rationalizability procedure. Consider the following procedure.*

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

(Step n>0) *For each $i \in I$ and $s_i \in S_i$, let $s_i \in \widehat{S}_{i,e}^n$ if and only if there is $\mu_i \in \Delta_i^e$ such that:*

*S1 $s_i \in \rho_i(\mu_i)$;*

*S2 $\mu_i$ strongly believes $\widehat{S}_{j,e}^q$ for all $j \neq i$ and $q < n$;*

*S3 $\mu_i$ strongly believes $\widehat{S}_j^q$ for all $j \neq i$ and $q \in \mathbb{N}$.*

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

This is the definition of Selective Rationalizability provided and charaterized epistemically in [17]. It differs from the definition used in this paper because of requirement S3 in place of the requirement that $s_i \in S_i^\infty$. Here I argue that the two definitions are equivalent for the analysis of agreements.

The two definitions are equivalent for the *same* agreement whenever the agreed-upon plans are chosen only according to what they prescribe at the rationalizable histories ($H(S^\infty)$).

**Proposition 11** *Fix an agreement $e = (e_i)_{i\in I}$ such that, for each $i \in I$, $n = 0, ..., k_i$, $s_i \in e_i^n$, and $s_i' \in S_i^\infty$, if $s_i'(h) = s_i(h)$ for all $h \in H(S^\infty) \cap H(s_i')$, then $s_i' \in e_i^n$. Then, $\widehat{S}_e^\infty = S_e^\infty$.*

**Proof.** By induction.

**Induction hypothesis:** for each $m \leq n$, $\widehat{S}_e^m = S_e^m$; moreover, unless $\widehat{S}_e^{n+1} = S_e^{n+1} = \emptyset$, for each $i \in I$ and $\overline{h} \notin H(S^\infty)$ with $p(\overline{h}) \in H(S^\infty)$, there exists a map $\eta_{i,n}^{\overline{h}} : S_i(\overline{h}) \to S_i(\overline{h})$ such that:

**a)** for each $\overline{s}_i \in S_i(\overline{h}) \backslash S_i^\infty(\overline{h})$, $\eta_{i,n}^{\overline{h}}(s_i) = \overline{s}_i$;

**b)** for each $\overline{s}_i \in S_i^\infty(\overline{h})$,

- **(i)** $\eta_{i,n}^{\overline{h}}(\overline{s}_i)(h) = \overline{s}_i(h)$ for all $h \in H(\overline{s}_i)$ with $h \succeq \overline{h}$,
- **(ii)** $\eta_{i,n}^{\overline{h}}(\overline{s}_i) \in S_{i,e}^m$ for all $m \leq n$ with $S_{i,e}^m(\overline{h}) \neq \emptyset$,
- **(iii)** if $e_i^q \cap S_i(\overline{h}) \neq \emptyset$ for some $q = 0, ..., k_i$, $\eta_{i,n}^{\overline{h}}(\overline{s}_i) \in e_i^q$.

**Basis step**: $S_e^0 = \widehat{S}_e^0 = S^\infty$, and the required maps exist by property of $e$ (in particular, (iii) can always be satisfied).

**Inductive step.** For $\widehat{S}_e^{n+1} = S_e^{n+1}$, since by the induction hypothesis $\widehat{S}_e^m = S_e^m$ for each $m \leq n$, it suffices to show that for every $i \in I$ and $s_i \in S_{i,e}^{n+1}$, there is $\widehat{\mu}_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^m)_{j\neq i})_{m=0}^n$ *and* $((S_j^m)_{j\neq i})_{m=0}^\infty$ such that $s_i \in \rho_i(\widehat{\mu}_i)$. So, fix $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^m)_{j\neq i})_{m=0}^n$ and $\mu_i'$ that strongly believes $((S_j^m)_{j\neq i})_{m=0}^\infty$ such that $s_i \in \rho_i(\mu_i) \cap \rho_i(\mu_i')$. By the induction hypothesis, I can construct $\widehat{\mu}_i$ such that $\widehat{\mu}_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S_{-i}^\infty)$ and $\widehat{\mu}_i(s_{-i}|h) = \mu_i'((\times_{j\neq i}\eta_{j,n}^{\overline{h}})^{-1}(s_{-i})|h)$ for all $\overline{h} \notin H(S_{-i}^\infty)$ with

$p(\overline{h}) \in H(S^\infty)$, $h \succeq \overline{h}$, and $s_{-i} \in \times_{j\neq i}\eta^{\overline{h}}_{j,n}(S_{-i}(\overline{h}))$. By (iii), $\widehat{\mu}_i \in \Delta^e_i$. By (ii), $\widehat{\mu}_i$ strongly believes $((S^m_{j,e})_{j\neq i})^n_{m=0}$ and, by (a), also $((S^m_j)_{j\neq i})^\infty_{m=0}$. By (i), $s_i \in \rho_i(\widehat{\mu}_i)$.

Now fix $\overline{h} \notin H(S^\infty)$ with $p(\overline{h}) \in H(S^\infty)$. If $S^{n+1}_{i,e}(\overline{h}) = \emptyset$, let $\eta^{\overline{h}}_{i,n+1} = \eta^{\overline{h}}_{i,n}$. Else, we need to update $\eta^{\overline{h}}_{i,n}(\overline{s}_i)$ for each $\overline{s}_i \in S^\infty_i(\overline{h})$. Fix $\overline{\mu}_i$ that strongly believes $((S^m_j)_{j\neq i})^\infty_{m=0}$ such that $\overline{s}_i \in \rho_i(\overline{\mu}_i)$. Unless $\widehat{S}^{n+1}_e = S^{n+1}_e = \emptyset$, there exists $\widehat{s}_i \in S^{n+1}_{i,e}(\overline{h}) = \widehat{S}^{n+1}_{i,e}(\overline{h})$ with $\widehat{s}_i \in e^q_i$ for all $q = 0, ..., k_i$ such that $e^q_i \cap S_i(\overline{h}) \neq \emptyset$, otherwise, for any $j \neq i$, there would not be any $\widehat{\mu}_j \in \Delta^e_j$ that strongly believes $S^{n+1}_{i,e} = \widehat{S}^{n+1}_{i,e}$. Fix $\widehat{\mu}_i \in \Delta^e_i$ that strongly believes $((S^m_{j,e})_{j\neq i})^n_{m=0}$ such that $\widehat{s}_i \in \rho_i(\widehat{\mu}_i)$. Since $\widehat{\mu}_i$ strongly believes $S^0_{-i,e} = S^\infty_{-i}$, by the induction hypothesis I can construct $\mu_i$ such that $\mu_i(\cdot|h) = \widehat{\mu}_i(\cdot|h)$ for all $h \not\succeq \overline{h}$ and $\mu_i(s_{-i}|h) = \overline{\mu}_i((\times_{j\neq i}\eta^{\overline{h}}_{j,n})^{-1}(s_{-i})|h)$ for all $h \succeq \overline{h}$ and $s_{-i} \in \times_{j\neq i}\eta^{\overline{h}}_{j,n}(S_{-i}(\overline{h}))$. By (iii), $\mu_i \in \Delta^e_i$. By (ii), $\mu_i$ strongly believes $((S^m_{j,e})_{j\neq i})^{n-1}_{m=0}$. By (i), there is $s_i \in \rho_i(\mu_i)$ such that $s_i(h) = \widehat{s}_i(h)$ for all $h \in H(s_i)$ with $h \not\succeq \overline{h}$ (thus $s_i \in S_i(\overline{h})$) and $s_i(h) = \overline{s}_i(h)$ for all $h \in H(\overline{s}_i)$ with $h \succeq \overline{h}$. So, $\eta^{\overline{h}}_{i,n+1}(\overline{s}_i) = s_i$ satisfies (i). If $s_i \in S^\infty_i$, then $s_i \in S^{n+1}_{i,e}$, satisfying (ii), and by the property of $e$, $s_i \in e^m_i$ for every $m$ such that $\widehat{s}_i \in e^m_i$, satisfying (iii). So, it only remains to show that $s_i \in S^\infty_i$. Since $\widehat{s}_i \in S^\infty_i$, there is also $\widehat{\mu}_i$ that strongly believes $((S^m_j)_{j\neq i})^\infty_{m=0}$ such that $\widehat{s}_i \in \rho_i(\widehat{\mu}_i)$. Thus, I can construct also $\mu_i$ that strongly believes $((S^m_j)_{j\neq i})^\infty_{m=0}$ such that $\mu_i(\cdot|h) = \widehat{\mu}_i(\cdot|h)$ for all $h \not\succeq \overline{h}$ and $\mu_i(\cdot|h) = \overline{\mu}_i(\cdot|h)$ for all $h \succeq \overline{h}$, so clearly $s_i \in \rho_i(\mu_i) \subseteq S^\infty_i$. ■

The intuition is the following: under an agreement in this class, all rationalizable plans can always be justified at the non-rationalizable histories under both definitions, while the two definitions do not differ in terms of beliefs they allow at the rationalizable histories. This class of agreements suffices for the implementation of all implementable outcome sets, for the following reason. Restricting behavior at the non-rationalizable histories cannot have a direct effect on the induced paths, which are always rationalizable. It can only have an indirect effect via a player's beliefs by combining an co-player's agreed behavior at rationalizable and non-rationalizable histories in a particular way. But given that the behavior of the co-player before and after our player leaves the rationalizable histories can always be "disentangled" (because the co-player

gets surprised by finding herself at the non-rationalizable histories and has to come up with new beliefs), this indirect effect can also be obtained directly by only restricting her behavior at the rationalizable histories. This can be proven formally with the same arguments of the proof of Proposition 8.

**Proposition 12** *Fix a self-enforcing agreement $e^* = (e_i^*)_{i\in I}$. Then, there exists an agreement $\overline{e} = (\overline{e}_i)_{i\in I}$ that satisfies the condition in Definition 11 such that $\zeta(S_{\overline{e}}^\infty) = \zeta(S_{e^*}^\infty)$.*

**Proof.** By Theorem 2, there exists a tight agreement $e = (e_i)_{i\in I}$ such that $\zeta(S_{e^*}^\infty) = \zeta(S_e^\infty)$. Define an agreement $\overline{e} = (\overline{e}_i)_{i\in I}$ by letting, for each $i \in I$ and $n = 0, ..., k_i$,

$$\overline{e}_i^n = \left\{ s_i \in S_i^\infty : \exists s_i' \in e_i^n, \forall h \in H(S^\infty) \cap H(s_i), s_i(h) = s_i'(h) \right\}.$$

I show that also $\overline{e}$ is tight, so that $\zeta(S_{\overline{e}}^\infty) = \zeta(\overline{e}^0) = \zeta(e^0) = \zeta(S_e^\infty) = \zeta(S_{e^*}^\infty)$. T2 is obvious.

To see T1, follow the proof for Proposition 8 that $\overline{e}$ satisfies T1$^a$ (which coincides with T1).

To see T3, fix $i \in I$, $\mu_i \in \Delta_i^{\overline{e}}$, and $\overline{h} \in H(\rho_i(\mu_i) \cap S_i^\infty)$, and follow the proof for Proposition 8 that $\overline{e}$ satisfies T3$^a$ (which coincides with T3 by $\rho_i(\Delta_i^{\overline{e}}) = \rho_i(\Delta_i^{\overline{e}}) \cap S_i^\infty$ in that proof). ■

The same is true if self-enforceability is defined using Definition 16 (and it can be proven in the same way). Therefore, we have the following.

**Corollary 5** *The implementable outcome sets under the two definitions of Selective Rationalizability coincide.*

This would not be true if agreements were allowed to feature non rationalizable plans. In this case, some $e_i^n$ could reach a history $h \notin H(S_i^\infty)$ with some plan $s_i \notin S_i^m$, although $h \in H(S_i^m)$, so that no $s_j \in S_j^\infty(h) \neq \emptyset$ is compatible with the belief in $e_i^n$. This can imply the elimination of a move by $j$ at a rationalizable history (possibly dominant within the rationalizable paths!) under Definition 7, whereas the agreement would not be credible at all under Definition 16.