Self-enforcing agreements and forward induction reasoning[*]

Emiliano Catonini[†]

June 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 subgame perfect equilibrium outcome may

[*] 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, Carlo Cusumano, Nicodemo De Vito, Alfredo Di Tillio, Andrew Ellis, Francesco Fabbri, Amanda Friedenberg, Giacomo Lanzani, Gilat Levy, Matthew Levy, Francesco Nava, Andres Perea, Burkhard Schipper, Madhav Shrihari Aney, Balazs Szentes, 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, Maastricht University, and London School of Economics.

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

not. The incompleteness of the agreement can be crucial to implement an outcome.

Keywords: Self-enforcing agreements, Incomplete agreements, Forward induction, Extensive-Form Rationalizability.

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 will play as an *agreement*. In many cases, players only reach 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 induces 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 will believe in and comply with. Moreover, in an implementation perspective, I investigate 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 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 disbelief in some features of the equilibrium. Can we expect the deviator to

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

best reply to threats that are meant precisely to deter the deviation? Often, the deviation 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, anticipated distrust or objective impossibility of credible (re-)coordination: these are some of the reasons why players may be unable or unwilling to reach a complete agreement. In economic applications, absence of a (intuitive) SPE solution is often blamed on a misspecification of the model, rather than on the objective impossibility to reach a precise agreement among players. A classical example is the two-stage Hotelling model with linear transportation cost, which has no SPE in pure strategies.[2] A quadratic transportation cost has been introduced by D'Aspremont et al. [21] to obtain a unique SPE solution, where firms, contrary to Hotelling's conjecture, locate at the extremes of the spectrum $(0, 1)$. In a separate paper [17], I obtain the transportation-efficient $(1/4, 3/4)$ as the unique symmetric locations pair that can be induced by a self-enforcing agreement.

To illustrate these insights in a simple but 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

[2] A SPE in mixed strategies has been found numerically by Osborne and Pitchik [34]. In this equilibrium, counterintuitively, firms locate at a distance that puts them at risk of a "price war", whereby a slightly higher distance would prevent this possibility.

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)),[3] 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 agree to play (N, W) and that Ann should play U in case Bob deviates.[4] Is the agreement credible? If Bob is rational[5] 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 com-

[3] 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.

[4] 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 modify the game in such a way that the Nash outcome is sustained by a credible threat which differs from the unique equilibrium action of the subgame.

[5] 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.

pletely 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.

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, a plan of actions (also known as *reduced strategy*) already features a basic form of incompleteness: it does 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,[6] and no chance moves. However, the methodology can be applied/adapted to all dynamic games with perfect recall.

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 the 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.[7] 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

[6] 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.

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

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.[8] This instance of forward induction reasoning is 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 prescribed by a SES. For a single outcome of a two-players game, SES's boil down to Nash equilibria in extensive-form rationalizable[9] plans that satisfy a strictness condition.[10] To complete the search for implementable outcomes in games with more than two players and stages, *tight agreements* augment SES's by restricting the behavior of deviators. An outcome set is implementable if and only if it is prescribed by a tight agreement. Since a tight agreement (like a SES) induces exactly the outcome set

[8] 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 modified game in the Supplemental Appendix.

[9] The original notion of extensive-form-rationalizability is due to Pearce [35] and was later refined by Battigalli [5] and Battigalli and Siniscalchi [11].

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

it prescribes, we have a "revelation principle" for agreements design: players need not be vague about the set of outcomes they want to achieve.

Tight agreements and SES's have the double value of *solution concepts* and "soft mechanisms" for implementation,[11] because they prescribe directly the outcome set they induce. They provide to the analyst (or a mediator) all possible 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-*Δ*-Rationalizability* (Battigalli, [7]; Battigalli and Siniscalchi, [12]). Strong-Δ-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 opposite hypothesis of this paper, another elimination procedure with belief restrictions, *Selective Rationalizability*, is constructed and analyzed epistemically in the companion paper [18]. 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 *epistemic priority* attributed to the beliefs in rationality, the structure given by agreements to the belief restrictions, and the requirement of self-enforceability greatly increase the predictive power with respect to *Extensive-Form Best Response Sets* (Battigalli and Friedenberg [8]), which capture the predictions of Strong-Δ-Rationalizability across all first-order belief restrictions.[12] In Section 5, I

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

[12] For instance, competition among firms on price, quantity, or quality often leads to a precise outcome under common belief in rationality (see cobweb stability or Cournot

expand on this comparison and revise the results of Section 4 under Strong-Δ-Rationalizability.

When the agreement prescribes a single outcome, a possible way to interpret deviations is that the deviator believed in the agreed-upon path (i.e., that the co-players would have complied with it), but does not believe in the threats. In Section 6, I provide an example where imposing this particular 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 (thus less agnostic) strategic reasoning hypotheses, which further increase the refinement power.

Strategic stability à la Kohlberg and Mertens [30] 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. Stability is hard to interpret and verify, and does not offer an implementation strategy: what should players exactly agree on/believe in?[13] Later refinements focus exclusively on sequential equilibrium and, to simplify the analysis, sacrifice depth of reasoning (e.g., forward induction equilibria of Govindan and Wilson [25] capture only strong belief in rationality[14]) or scope (e.g., the intuitive criterion of Cho and Kreps [20] and divine equilibrium of Banks and Sobel [3] are tailored on signaling games). Moreover, the equilibrium 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 threats after the deviation) is at odds precisely with this kind of forward induction reasoning.

duopoly), but this predictive power is lost in subgames where orders of belief in rationality are dropped. In [17] I show that in Hotelling almost every symmetric locations pair is induced by some extensive-form best response set.

[13] An interesting critique of this kind to strategic stability has been put forward by Van Damme [40].

[14] See [25], pagg. 11 and 21. An explicit example of this fact is provided by Perea ([36], pag. 509).

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 further examples and technical remarks that 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. Each individual $j \in [0, 48]$ either does not buy, or buys one unit from firm i that maximizes $u_{ji} = 72 - p_i - t \cdot d_{ji}$, where p_i is the price fixed by firm i, $t = 1/2$ is the transportation cost, and d_{ji} is the distance from firm i: $d_{j1} = j$ and $d_{j2} = 48 - j$. Then, each firm $i = 1, 2$ faces demand

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

There are two technologies: $k = A$, with marginal cost $mc = 48$ and no fixed cost; and $k = B$, with no marginal cost and fixed cost F such that firm i is indifferent between the two technologies for $p_{-i} = 48$. Suppose that firms choose technology and price simultaneously; then, for $p_{-i} \in [24, 72]$, firm i's best response correspondence is

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

For $p_{-i} < 24$, firm i has no incentive to produce. For $p_{-i} > 72$, firm i's best reply is $p_i = 48$ with $k = B$. Probabilistic conjectures do not expand the set of rational prices, which thus are $[36, 60]$. For each $(p_i, p_{-i}) \in [36, 60]^2$, firm i's demand is $24 - p_i + p_{-i}$, hence the best replies to $\mu \in \Delta([36, 60])$ are

$\widehat{p}_i(\mathbb{E}_\mu(p_{-i}))$. Then, the rationalizable prices of each firm (*in the static game*) are $[36, 42]$ (with $k = B$) and $[54, 60]$ (with $k = A$). The pure equilibrium price pairs are $(40, 56)$ and $(56, 40)$, and the only mixed equilibrium assigns probability 1/2 to 36 and 60 for both firms. Profits are increasing in the other firm's (expected) price, so let $\overline{\overline{\pi}} > \overline{\pi} > \underline{\pi}$ denote the profit of firm 2 in the three equilibria ($\overline{\pi}$ is the expected profit in the mixed equilibrium), and let $\underline{\underline{\pi}}$ denote the optimal profit of firm 2 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?[15] I am going to tackle this question for different values of the entry cost.

Case 1) $\overline{\pi} < E < \overline{\overline{\pi}}$ (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 believes that firm 1 is rational, firm 2 will enter only if it expects $p_1 \geq \widetilde{p}$ with $\widetilde{p} \in (48, 56)$ that depends on E, and then fix $p_2 \in \left[12 + \frac{1}{2}\widetilde{p}, 42\right]$ with $k = B$. Understanding this, firm 1 has the incentive to fix $p_1 \in \left[42 + \frac{1}{4}\widetilde{p}, 57\right]$ with $k = A$, thus $p_1 > \widetilde{p}$. So, firm 2 has always the incentive to enter.

Case 2) $\underline{\pi} < E < \overline{\pi}$ (agreement incompleteness). According to SPE, entry is deterred by equilibrium $(40, 56)$. But believing in $p_2 = 56$ is incompatible with forward induction reasoning. If firm 2 is rational and believes that firm 1 is rational, firm 2 will enter only if it expects $p_1 \geq \widetilde{p}$ with $\widetilde{p} \in (40, 48)$ that depends on E, and then fix either $p_2 \in \left[36 + \frac{1}{2}\widetilde{p}, 60\right]$ with $k = A$, or $p_2 \in [36, 42]$ with $k = B$, thus not $p_2 = 56$. However, every $p_1 \in [36, 42] \cup [54, 60]$ is a best

[15] For the purpose of the example, the incumbent has no commitment power or switching costs. Instead, Dixit [21] 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 interactions. [...] 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."

response to a belief over these entry plans of firm 2. Hence, it is credible that firm 1 will react to entry with $p_1 = 40$. But $p_1 = 40$ is motivated by uncertainty over values of p_2 that do not best respond to it. So, it must be formulated as a unilateral threat and not as part of a complete agreement. Furthermore, firm 1 does not actually need to specify p_1: it is enough to announce the use of technology $k = B$. 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 $k = A$. But then, realizing this, firm 1 would best reply with $p_1 \in \left[30 + \frac{1}{4}\widetilde{p}, 40.5\right]$ and $k = B$. 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 $k = B$ by the incumbent is credible and deters entry. Such a coarse announcement can have real-life advantages; for instance it may be illegal to state future prices.[16] I discuss another advantage of this announcement over the precise SPE threat in Section 4.1, where I analyze it formally.

Case 3) $\underline{\underline{\pi}} < E < \underline{\pi}$ **(SPE is too restrictive).** Now, firm 2 enters in every SPE. But, similarly to Case 2, there is $\widetilde{p} \in (36, 40]$ such that every $p_2 \in [36, 42] \cup \left[36 + \frac{1}{2}\widetilde{p}, 60\right]$ is compatible with forward induction reasoning, and then every $p_1 \in [36, 42] \cup [54, 60]$ as well. So, firm 2 can credibly threaten to fix $p_1 \in [36, \widetilde{p})$ and deter entry.[17] The arguments for the credibility of this threat are identical to the arguments for the SPE threat in Case 2. For instance, $p_1 = 36$ is justified by a uniform distribution over the three equilibrium (expected) prices, which are now all compatible with strategic reasoning.

[16] Harrington [28] 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.

[17] One could argue that alternated best responses from p_1 would lead to the $(40, 56)$ equilibrium in the long run. If firms are impatient, this is immaterial for the analysis. If firms are patient, 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.

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. Or, the "outside option" could also be 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 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 only. This kind of forward induction reasoning is modeled explicitly in Section 6.

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 $X_J := \times_{j\in J}X_j$, $X := X_I$, $X_{-i} := X_{I\backslash\{i\}}$. 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 $H:=\overline{H}\backslash Z$ 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

[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.

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 $z \in Z$. (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

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

For any $J \subseteq I$ and $\widehat{S}_J \subset S_J$, the set of histories compatible with $\widehat{S}_J$ is

$$H(\widehat{S}_J) := \left\{h \in H : \widehat{S}_J \cap S_J(h) \neq \emptyset\right\}.$$

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 of probability: 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]

[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*

For any player i and any set of co-players $J \subseteq I \setminus \{i\}$, I say that a CPS μ_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 the 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.

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

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

The set of sequential best replies to μ_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 say that a plan s_i is *justifiable* if $s_i \in \rho_i(\mu_i)$ for some $\mu_i \in \Delta^H(S_{-i})$.

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]

independence (Battigalli [5]) could be assumed throughout the paper and the results would not change. See the companion paper for details.

[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).

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^∞ are called rationalizable.

3.3 Agreements, belief in the agreement, and Selective Rationalizability

All the notions introduced in this section are illustrated with 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 sets of rationalizable plans:*

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

However, as far as extensive-form-rationalizability is concerned, all the classical definitions (Pearce [35], Battigalli [5], Battigalli and Siniscalchi [11]) are equivalent in the framework of this paper. See the companion paper for details.

[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 induced anyway. As Pearce [35] puts it, "this indeterminacy is an accurate reflection of the difficult situation faced by players in a game."

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 in case she 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] 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 instead specify only partially, or not at all, what a player plans to do from some history onwards. This is obtained as follows. First, e_i^0, and each subsequent e_i^n, needs not be a singleton. Second, some 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 regarding 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 just require players to 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".

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 $i \in I$, $e_i = (e_i^0) = (S_i^\infty(z))$.*

A path agreement on z features all rationalizable plans of player i compatible with z to remain silent regarding i's reactions to co-players' deviations.[25]

[24] In light of this, agreements (and in particular, tight agreements) could be given a more compact representation with just one set of *strategies* (as opposed to plans of actions) for each player. However, the current representation is way more handy for the definition of "belief in the agreement" (Definition 6).

[25] The restriction to rationalizable plans has no bite: players would only expect rationalizable reactions to deviation anyway by the beliefs in rationality.

Instead, like any other reduced agreement, a path agreement remains silent regarding i's continuation plans after own deviations by not introducing alternative sets. 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 — see footnote 30.

I say that 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 $j \neq i$, μ_i strongly believes $e_j^0, ..., e_j^{k_j}$.*

Let Δ_i^e be the set of all μ_i's 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 [18], 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 $e = (e_i)_{i \in I}$.

[26] This is reminiscent of the agreement being a *basis* for player i's CPS: see Siniscalchi [38].

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

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^∞ are called selectively-rationalizable.

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 μ_i always has to belong to Δ_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, $S_{j,e}^n \cap S_j(h) \neq \emptyset$, but $S_{j,e}^n \cap e_j^m \cap S_j(h) = \emptyset$ 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 some order of belief in rationality — that is, at each $h \notin H(S_{j,e}^1)$. 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 $S_e^\infty \neq \emptyset$.*

A credible agreement induces each player i to strongly believe in the agreed-upon plans that are compatible with strategic reasoning ($S_{-i,e}^\infty \cap e_{-i}^0$). But this belief may be contradicted by the actual play, because credibility does not imply that players will comply with the agreement, it only implies that they *may* do so *everywhere in the game.* In particular, the plans that are compatible with strategic reasoning (S_e^∞) may induce a larger set of outcomes with respect to those that are also compatible with the agreement ($S_e^\infty \cap e^0$). When instead they induce the same outcomes, I say that the agreement is *self-enforcing.*

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

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

So, a self-enforcing agreement may induce a strict subset of the outcomes it prescribes. When instead the agreement is not more permissive, in terms of outcomes, than the behavior it induces, I say that the agreement is *truthful.*[28]

Definition 10 *A self-enforcing agreement is **truthful** if $\zeta(S_e^\infty) = \zeta(e^0)$.*

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

In the introductory game, all plans are justifiable, hence they are all rationalizable: $S_e^0 = S$. Consider first the path agreement on (S, E): $e_A^0 = \{S\}$, $e_B^0 = \{E.L, E.R\}$. We have $\Delta_A^e = \{\mu_A : \mu_A(\{E.L, E.R\} \,|h^0) = 1\}$ and $\Delta_B^e = \{\mu_B : \mu_B(S|h^0) = 1\}$. So, $S_e^1 = \{S, N.D\} \times \{E.L, E.R\}$: Ann plays either S, or $N.D$ if she gives sufficiently high probability to $E.R$; Bob plays E and either

[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.

L or R depending on his new belief after being surprised by Ann's deviation. Then, we have $S_e^2 = \{S, N.D\} \times \{E.L\}$, and finally $S_e^3 = \{S\} \times \{E.L\} = S_e^\infty$: e is truthful.

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

Agreement	Reduced	"Unilateral"	Path on (N, W)
e_A	$(\{N.U\})$	(S_A)	$(S_A \backslash \{S\})$
e_B	$(\{W\})$	$(\{W\}, \{W, E.L\})$	$(\{W\})$
$S_{A,e}^1 \times S_{B,e}^1$	$(S_A \backslash \{S\}) \times \{W\}$	$\{N.U\} \times S_B$	$(S_A \backslash \{S\}) \times S_B$
$S_{A,e}^\infty \times S_{B,e}^\infty$	$(S_A \backslash \{S\}) \times \{W\}$	$\{N.U\} \times \{W\}$	$(S_A \backslash \{S\}) \times S_B$
Conclusion	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 two agreements are self-enforcing and induce (N, W). The "unilateral" agreement (which is not reduced) has 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 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 because of 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, agreement incompleteness triggers steps of reasoning that refine players' beliefs up to a point where the restrictions play no role anymore (but they crucially initiated the reasoning process). This can be seen from one of the agreements of Section 2. Consider the announcement of technology $k = B$ by the incumbent (firm 1). Formally, this is a reduced agreement where e_1^0 is the set of all technology-price pairs with $k = B$, 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	$[36,48]\times\{k=B\}\cup[48,60]\times\{k=A\}$
S_2^1	$\{\textit{no-entry}\}\cup[36,48]\times\{k=B\}\cup[56.5,60]\times\{k=A\}$
$S_1^2=S_1^\infty$	$[36,42]\times\{k=B\}\cup[54,60]\times\{k=A\}$
$S_2^2=S_2^\infty$	$\{\textit{no-entry}\}\cup[36,42]\times\{k=B\}\cup[56.5,60]\times\{k=A\}$
S_e^1	$S_{1,e}^1=S_1^\infty,\ \ S_{2,e}^1=\{\textit{no-entry}\}\cup[56.5,57]\times\{k=A\}$
S_e^2	$S_{2,e}^2=S_{2,e}^1,\ \ S_{1,e}^2=[40.25,40.5]\times\{k=B\}$
$S_e^3=S_e^\infty$	$S_{1,e}^2=S_{1,e}^3,\ \ S_{2,e}^3=\{\textit{no-entry}\}$

The announcement of $k=B$, per se, is not sufficient to deter entry, but it entails sufficiently low prices by the incumbent for the entrant to employ $k=A$ and exclude the highest rationalizable prices. Anticipating this, the incumbent has the strict incentive to use $k=B$ and exclude the highest prices compatible with $k=B$ as well. In turn, this provides to firm 2 the strict incentive not to enter. So, the agreement is credible and it deters entry. Moreover, 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^∞) 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 on 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 the 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 the first 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 μ_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 μ_i that strongly believes $(S_j^*)_{j\neq i}$ and $(S_j^\infty)_{j\neq i}$ such that $s_i \in \rho_i(\mu_i)$.[29]

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 every player i, every 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 the co-players carry on plans in S_{-i}^*. Analogously, say that a Nash

[29] The focus will always be on rationalizable plans that can be justified under strong belief in the rationalizable plans of the co-players. Basically, it is as if the game is reduced to $(S_i^\infty)_{i\in I}$. Then, one could in principle take this reduced strategic form and reformulate the analysis in terms of lexicographic beliefs instead of CPS's. This approach would be generically equivalent to the present one.

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 every $i \in I$. Then, when S^* induces a unique outcome, 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, Rationalizability is performed. This is a standard elimination procedure that does not depend on agreements. 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 μ_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 μ_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 the 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 the co-players form beliefs in the same way. That is, when they refine their beliefs with forward induction reasoning based on 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", and indeed boil down to *sets closed under rational behavior* (Basu and Weibull [4]) in static games. By Forward Induction, the closure of the SES cannot be refined with forward induction considerations. Therefore, the agreement on the SES implements precisely the SES outcomes ($\zeta(S^*)$).

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).*

A SES is constructed in the first example of 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 Realization-strictness and Self-Justifiability. By Self-Justifiability, $S^* \subseteq \overline{S}^*$, so strong belief in each $\overline{S}_j^*$ can have additional bite with respect to strong belief in S_j^* only at a history not compatible with S_j^* ($h \notin H(S_j^*)$). Realization-strictness implies that a deviation by player j from the SES paths cannot be rationalized under belief in the SES ($h \notin H(\overline{S}_j^*)$). Then, if there are no other co-players, or if there is no time for subsequent deviations by other players, strong belief in S_j^* implies strong belief in $\overline{S}_j^*$, and Forward Induction holds by definition of closure. 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 satisfies 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 Corollary 3 for the "if" part. Statement 3 follows from statement 2 and Proposition 4. ■

Moreover, in two-players games, Realization-strictness implies Self-Justifiability when there is only one path to follow.

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, the 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 equilibrium itself.*

Proof. "Only if" comes from 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 candidate SES would induce further deviations by other players down the line. Possibly, this can be avoided by restricting the continuation plans of the deviators, 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 from 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 reach under belief in the agreement, $H(\rho_i(\Delta_i^e) \cap S_i^\infty)$, must also be reached by a set of agreed-upon plans e_i^n that can be justified under belief in the agreement (this is T3). Then, all the agreed-upon 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]

Proof. "If" comes from Proposition 7. "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. A tight agreement is constructed in this way in the second example of 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.

[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 of the rationalizable plans of a player are equivalent ways not to restrict beliefs, but the first is convenient for tight agreements, the second for SES's.

[31] The proof of 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 with infinite horizon the set of reached histories may squeeze at each of infinite steps of reasoning. A characterization of implementable outcomes with truthful agreements in games with infinite horizon, if possible, is subject for future research.

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 though an important difference: while a direct mechanism requires players to specify *only* their type, a truthful agreement, beside the outcome(s), typically needs to specify off-path behavior. This is 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 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 plans of the deviator (as allowed by tight agreements) can be necessary for implementation. This complements the example of Section 2, where the incumbent can credibly specify a precise 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, Ann the row, and Bob the column (payoffs are in alphabetical order).

$$\begin{array}{rl} \text{DAVE} & \text{---}\ Out \longrightarrow 0,0,0,0 \\ Instigate \downarrow & \end{array}$$

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 an *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 intervention is 3 for Dave and 1 for Cleo; moreover, if Ann or Bob engages in the arms race and the other does not, Cleo has to spend 1 additional unit to defend the unarmed player, who in turn has to share its 6 units of resources with Cleo. If Cleo does not intervene and Ann or Bob engage in the arms race, a war starts. If they are both armed, the war comes to costly impasse; if one is unarmed, it gets conquered and loses all its resources to the other.

The game has only one SPE, where Dave instigates, Cleo does not intervene, and Ann and Bob engage in the arms race.[33] However, Cleo could intervene in the hope that Ann and Bob do not coordinate, and the unarmed player falls under her influence.[34] I show that there is a SES where Cleo threatens Dave to intervene, Ann and Bob remain silent, and Dave does not instigate.

[32] This game is freely inspired by the leading example in Greenberg [26].

[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 (without correlation) Cleo would rather not intervene.

[34] In view of Cleo's intervention, coordinating is not an obvious task for Ann and Bob: coordinating on Peaceful dominates coordinating on the Arms Race, but the Arms Race is a way less risky action. Moreover, to justify Cleo's threat to Dave, it is in the interest of Ann and Bob not to establish any form of coordination, if, as assumed, it would not remain secret to Cleo's intelligence.

All plans are justifiable, hence rationalizable. Let $S^* = \{AR, P\} \times \{AR, P\} \times \{Int\} \times \{Out\}$. 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 μ_D that strongly believes S_C^*. For every $i = A, B, C$, Realization-strictness trivially follows from $\zeta(S_i \times e_{-i}^0) = \{(Out)\}$. There remains to show Self-Justifiability. For Cleo, Int is justified by any μ_C with $\mu_C(s_A \neq s_B|(Inst)) \geq 1/2$, and let $\mu_C(S_A \times S_B \times \{Out\}\,|h^0) = 1$ for μ_C to strongly believe $(S_j^*)_{j \neq C}$. For Ann, AR (resp., P) is justified by any μ_A with $\mu_A((AR, Int, Inst)|(Inst)) \geq 1/3$ (resp., with $\mu_A((P, Int, Inst)|(Inst)) \geq 2/3$), and let $\mu_A(S_B \times \{Int\} \times \{Out\}\,|h^0) = 1$ for μ_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. 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 have to pay cost $g < w - r$ to *Go* on the market (they have a preference for staying). If they both stay, the game ends. If one stays and the other does not, say Bob, the game continues 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 no 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. As deadlines approach, all players must make their choices without knowing the choices of others. If Ann hires a new game theorist at w, she will keep only Cleo if she stays, or Bob if he stays and Cleo leaves, in both cases at salary r. If Ann does not hire and Bob and Cleo do not leave, they will bargain a salary t; if one leaves and the other stays, the latter bargains a salary u with $t < u < v$. Ordinal payoffs compatible with this story are in the figure (cardinal payoffs will not matter for the analysis). In the last stage, Bob chooses the row.

Bob	— *Go* $\longrightarrow$	A\C	*Stay*	*Go*	
Stay $\downarrow$	(Cleo	*Open*	$5,4,3$	$\cdot -$	$\longrightarrow \Gamma$
$6,3,3$	stays)	*Shut*	$4,2,3$	$2,2,6$	

	Hire	*Stay*	*Go*	*Not*	*Stay*	*Go*
Γ :	*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 a meeting to clarify her intention to shut down a game theorist position if one of them, say Bob, does not accept the offer. But this will induce Cleo to bargain a higher salary by going on the market. In turn, this may induce Ann to increase her bargaining power by keeping Bob's position open and looking for potential new hires who can fill it. However, Ann has no real intention to hire, and Cleo has no real intention to leave. Understanding this with forward induction reasoning, Cleo will stay, and Ann will keep the position open and not hire. This leaves the position to Bob at a higher wage than initially offered. How to solve the impasse? Bob and Cleo, who want to avoid the cost of going on the market for an uncertain gain, convene with Ann that if they will all be on the market, they will go their separate ways: Bob and Cleo will leave and Ann will hire.

We are going to construct such an agreement and show it is tight through the roadmap of Section 4.2. We look for an agreement that implements outcome (*Stay*) in the game of the figure; by symmetry, it can be extended to the whole game. All plans are justifiable, hence rationalizable. Thus, 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, respectively, $G.G \in e^0_C$ and $S \notin e^0_C$, so we have $\{G.G\} \subseteq e^0_C \subseteq \{G.S, G.G\}$. The second case boils down to the first, because Ann's Self-Justifiability requires $O.H \notin e_A^0$ as well. Thus, we focus on agreements with $e_A^0 = \{S\}$, $e_B^0 = \{S\}$, and either $e_C^0 = \{G.G\}$, or $e_C^0 = \{G.S, G.G\}$. Does any of the two constitute a SES? No. In both cases, the closure of e^0 for Ann is $\{S, O.N\}$: under belief that Cleo goes on the market, $O.H$ is never optimal. But then, Forward Induction is violated for Cleo, because the only sequential best reply under strong belief in $\{S, O.N\}$ is $G.S$. Therefore, we look for a tight agreement with $e^0 = \{(S, S, G.G)\}$. Restrict Bob's behavior after his deviation by imposing $e_B^1 = \{S, G.G\}$. Also let $e_A^1 = \{S, O.H\}$, 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.

Note that the tight agreement is a "complete agreement", in that it specifies one action for each player and history, and it corresponds to a SPE.

5 Epistemic priority to the agreement

The literature on strategic reasoning with first-order belief restrictions is mostly based on the use of Strong-Δ-Rationalizability (Battigalli [7], Battigalli and Siniscalchi [12]). Strong-Δ-Rationalizability is here denoted by $((S^q_{i,\Delta^e})_{i\in I})_{q=0}^{\infty}$ and defined like Selective Rationalizability but without requiring that plans are rationalizable; i.e., as in Definition 7 with S_i in place of S_i^∞. The differences between this paper and the aforementioned literature are due to (i) the adoption of Selective Rationalizability in place of Strong-Δ-Rationalizability, (ii) the structure on the first-order belief restrictions imposed by the notion of agreement, and (iii) the focus on self-enforceability rather than just credibility.

Differences and similarities between Selective Rationalizability and Strong-Δ-Rationalizability are analyzed in depth in the companion paper. 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-Δ-Rationalizability captures the hypothesis that, upon observing such move, the 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-Δ-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 justifiable price; then, entry would be considered a sign of the entrant's irrationality, and the incumbent could have any belief about the entrant's price. This is a typical loss of predictive power that the inversion of epistemic priority entails. In all other examples, all plans are rationalizable; then, Selective Rationalizability and Strong-Δ-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 (S_e^∞) with strongly-Δ-rationalizable plans ($S_{\Delta^e}^\infty$);*
2. *rationalizable plans (S^∞) with justifiable plans (S^1) in Corollary 2, Proposition 6, and Theorem 2, and with all plans (S) elsewhere.*

To verify Remark 2, one can follow the proofs for Section 4 with the appropriate 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 the companion paper [18], Selective Rationalizability does not refine Strong-Δ-Rationalizability for given 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.*

For instance, by Remark 2.2, under priority to the agreement any realization-strict Nash equilibrium in justifiable plans of a two-players game is a self-enforcing agreement, also when incompatible with strong belief in rationality.[35]

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

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

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 belief must be in the EFBRS. Generically, maximality has no bite, thus SES's refine EFBRS's.[36] This can be seen already in generic static games (where the epistemic priority has no bite): EFBRS's boil down to best response sets, SES's boil down to sets closed under rational behavior. The reasons are the following. First, EFBRS's can be induced by 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 are induced precisely by all the beliefs compatible with the SES itself, thus, as all agreements, they align any two players' belief restrictions about a third player's moves. Second, an EFBRS may induce a larger set of outcomes with respect to what players expect under the restrictions that yields the EFBRS. Realization-strictness, and more generally self-enforceability, rule this out.[37]

[35] As shown by the introductory example of the companion paper, Strong-Δ-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 only extensive-form-rationalizable one: see Battigalli [6], Heifetz and Perea [29], Chen and Micali [19], and Perea [37] for proofs of this result.

[36] Generically, every justifiable plan can be justified by a CPS that has not other sequential best reply. In terms of outcomes SES's refine EFBRS in all games: S_e^∞ is an EFBRS.

[37] Relatedly, Battigalli and Siniscalchi [12] show that when Strong-Δ-Rationalizability is

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 she does not believe that he plays as in the SPE. In the baseline analysis of Section 4, Bob was free to believe 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 trusts him, in the following sense: she believes that he would not violate the agreement before herself. 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 in the second stage. 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 that he would have complied on path, and drops the belief that Ann believes that he will comply

non-empty under belief in a particular outcome, the outcome is induced by a self-confirming equilibrium (Fudenberg and Levine [24]). Also under priority to the agreement, 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, so, in particular, under one common profile of plans of all players.

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 order 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 finer epistemic priority orderings with a generalization of Selective Rationalizability, which I specialize 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. Fix an agreement $e = (e_i)_{i\in I}$ with $e^0 \subseteq S(z)$ and $\times_{i\in I} e^{k_i}_i \subseteq S^\infty_z$.

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 $(S^\infty_{i,z})_{i\in I}$ are refined using the belief in the whole agreement. So, the agreement must be compatible with strategic reasoning around the path.[38]

The analysis of Section 4 can be replicated under priority to the path for single outcomes z. 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;*

[38] 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.

2. *selectively-rationalizable (S_e) with z-selectively-rationalizable plans (S_{e^z});*

3. *rationalizable plans (S^∞) with z-rationalizable plans (S^∞_z).*

To verify Remark 3, one can follow the proofs for Section 4 with the appropriate substitutions, as highlighted in the Appendix. Although a self-enforcing agreement under priority to the path needs not be self-enforcing under priority to rationality, the following holds.

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

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 here. In the example of this section, the agreement on the SPE plans is self-enforcing under priority to rationality but not to the path, because the corresponding path agreement is not credible. Such path resembles[39] a "*path that can be upset by a convincing deviation*", a notion proposed by Osborne [33] for repeated coordination games. Osborne proves that such paths are not stable, in the sense of Kohlberg and Mertens [30]. 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-Δ-Rationalizability with belief restrictions on the equilibrium outcome distribution, and Cho and Kreps [20] show that an equilibrium that does not satisfy the intuitive criterion is not strategically stable. Sobel et al. [39] provide similar arguments both for the intuitive criterion and for divine equilibria (Banks and Sobel [3]).

In this equilibrium refinement literature, the focus is kept on sequential equilibria. Already under the baseline hypotheses of Section 4, the distinction

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

between subgame perfect and non-subgame perfect equilibria appears meaningless for self-enforceability (see the example of Section 2).[40] Further, subgame perfection seems even 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. I elaborate on this through the first example in the Supplemental Appendix.

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). Let $S_{-i,j}:=S_{I\setminus\{i,j\}}$.

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. Realization-strictness: Fix $i\in I$ and μ_i that strongly believes $S_{-i}^*=S_{-i,e}^\infty\cap e_{-i}^0$. 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 $h\in H(S_{-i}^*)$. By standard arguments, 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 $h\in H(S_{-i}^*)\cap H(s_i)$. So, $\zeta(\rho_i(\mu_i)\times S_{-i}^*)\subseteq\zeta(\rho_i(\mu_i')\times S_{-i}^*)$. Moreover, by $\rho_i(\mu_i')\subseteq S_{i,e}^\infty$ and self-enforceability of e, $\zeta(\rho_i(\mu_i')\times S_{-i}^*)\subseteq\zeta(S_e^\infty)=\zeta(S^*)$.

Self-Justifiability: Fix $s_i\in S_i^*\subseteq S_{i,e}^\infty$. By finiteness of the game,[41] every $s_i\in S_{i,e}^\infty$ is a sequential best reply to some $\mu_i\in\Delta_i^e$ that strongly believes

[40] Interestingly, Man [32] finds that also the invariance argument, used to motivate the notions of forward induction of Kohlberg and Mertens [30] and Govindan and Wilson [25], does not imply sequential equilibrium.

[41] 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" defined in Battigalli and Tebaldi, 2017)

$((S^q_{j,e})_{j\neq i})^\infty_{q=0}$, thus that strongly believes $(S^\infty_{j,e})_{j\neq i}$, $(S^\infty_j)_{j\neq i}$, and $(e^0_j)_{j\neq i}$. Then, μ_i strongly believes also $(S^\infty_{j,e} \cap e^0_j)_{j\neq i} = (S^*_j)_{j\neq i}$. ■

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 μ_i that strongly believes S^*_{-i}, the set of continuation best replies to $\mu_i(\cdot|h^0)$ coincides with $S_i(z)$. Thus, we have $\rho_i(\mu_i) \subseteq S_i(z)$. With $S^*_{-i} \subseteq S_{-i}(z)$, we obtain $\zeta(\rho_i(\mu_i) \times S^*_{-i}) = \{z\}$.

Only if: For each $(s^*_i)_{i\in I} \in S^*$, $i \in I$, and μ_i that strongly believes S^*_{-i} with $\mu_i(s^*_{-i}|h^0) = 1$, by Realization-strictness $\rho_i(\mu_i) \subseteq S_i(z)$. By standard arguments, for every continuation best reply 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 $\mu_i(S_{-i}(h)|h^0) > 0$. If we had $s_i \notin S_i(z)$, since $\mu_i(S_{-i}(h)|h^0) > 0$ for all $h \prec z$, there would be $s'_i \in \rho_i(\mu_i)\backslash S_i(z)$. But this contradicts $\rho_i(\mu_i) \subseteq S_i(z)$, so we must have $s_i \in S_i(z)$. Therefore, $\arg\max_{s_i} u_i(\zeta(s_i, s^*_{-i})) \subseteq S_i(z)$. For each $s_i \in S_i(z)$, $u_i(\zeta(s_i, s^*_{-i})) = u_i(z)$. So, $\arg\max_{s_i} u_i(\zeta(s_i, s^*_{-i})) = S_i(z)$. ■

Proof of Proposition 4. By definition, $\overline{S}^* = S^1_e$. Thus, by Forward Induction, $\overline{S}^* \subseteq S^2_e$, and obviously $S^1_e \supseteq S^2_e$. Hence, $S^1_e = S^2_e$. So, (i) $S^1_e = S^\infty_e$.

For each $i \in I$ and $s_i \in S^1_{i,e}$, there is μ_i that strongly believes $(S^*_j)_{j\neq i}$ and thus S^*_{-i} such that $s_i \in \rho_i(\mu_i)$. For each $h \in H(S^*) \cap H(s_i)$, we must have $s_i(h) = s'_i(h)$ for some $s'_i \in S^*_i \cap S_i(h)$, otherwise we would have $\zeta(s_i, s_{-i}) \notin \zeta(S^*)$ for any $s_{-i} \in S^*_{-i} \cap S_{-i}(h)$, violating Realization-strictness. Then, $\zeta(S^1_e) \subseteq \zeta(S^*)$. By Self-Justifiability, $S^* \subseteq S^1_e$. So, (ii) $\zeta(S^1_e) = \zeta(S^*)$.

By Self-Justifiability $S^* \subseteq S^1_e$. With (i), we get (iii) $S^* = S^\infty_e \cap S^* \neq \emptyset$.

By (i), (ii), and (iii), $\zeta(S^\infty_e) = \zeta(S^1_e) = \zeta(S^*) = \zeta(S^\infty_e \cap S^*) \neq \emptyset$: e is self-enforcing and truthful. ■

Proof of Proposition 5. As shown in the proof of Proposition 4, by Realization-Strictness $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$. By Self-Justifiability, $S^*_j \subseteq \overline{S}^*_j$.

In games with two stages, for each $j \in I$, $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$ implies $H(\overline{S}^*_j) \subseteq H(S^*_j)$, because every move allowed by $\overline{S}^*_j$ at h^0 must be allowed also by S^*_j.

Then, strong belief in $S_j^* \subseteq \overline{S}_j^*$ implies strong belief in $\overline{S}_j^*$. Thus, for each $i \in I$, $\overline{S}_i^*$ satisfies Forward Induction by definition.

In two-players games, for each $i \in I$ and $s_i \in \overline{S}_i^*$, fix μ_i that strongly believes S_j^* and S_j^∞ such that $s_i \in \rho_i(\mu_i)$. By $S_j^* \subseteq \overline{S}_j^* \subseteq S_j^\infty$, I can construct μ_i' that strongly believes S_j^*, $\overline{S}_j^*$, and S_j^∞ such that $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S_j^*)$ and all $h \notin H(\overline{S}_j^*)$. For each $h \in H(\overline{S}_i^*)$, either $h \notin H(\overline{S}_j^*)$, or $h \in H(\overline{S}^*) \subseteq H(S^*) \subseteq H(S_j^*)$. Thus, $\mu_i(\cdot|h) = \mu_i'(\cdot|h)$ for all $h \in H(\overline{S}_i^*)$. Since $\rho_i(\mu_i) \subseteq \overline{S}_i^*$, $\rho_i(\mu_i) = \rho_i(\mu_i')$. Hence, $s_i \in \rho_i(\mu_i')$. So, $\overline{S}_i^*$ satisfies Forward Induction. ■

Proof of Proposition 6. Let $\zeta(S^*) = \{z\}$. By Proposition 3, S^* is a set of realization-strict Nash equilibria. Fix $i \in I$ and $s_i \in S_i^*$. Since $s_i \in S_i^\infty$,[42] by finiteness of the game there exists μ_i that strongly believes S_j^∞ such that $s_i \in \rho_i(\mu_i)$.[43] Fix $s_j \in S_j^*$ and construct μ_i' that strongly believes S_j^* and S_j^∞ 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 $\mu_i'(\cdot|h) = \mu_i'(\cdot|h^0)$ coincides with $S_i(z)$. For every $h \in H(s_i)$ with $h \nprec z$, $h \notin H(S_j(z))$, hence $h \notin H(S_j^*)$, so s_i is a continuation best reply to $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$. So, $s_i \in \rho_i(\mu_i')$. ■

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

Proof of 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 μ_i strongly believes e_j^n, $1 = \mu_i(e_j^n \times S_{-i,j}|h) \leq \mu_i(S_j^\infty \times S_{-i,j}|h)$. Thus, μ_i strongly believes $(S_j^\infty)_{j\neq i}$. Therefore, for each $i \in I$, $\rho_i(\Delta_i^e) \cap S_i^\infty = S_{i,e}^1$.

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 $\emptyset \neq e_j^n \cap S_j(h) \subseteq \rho_j(\Delta_j^e) \cap S_j^\infty = S_{j,e}^1$. Then, since μ_i strongly believes e_j^n, $1 = \mu_i(e_j^n \times S_{-i,j}|h) \leq \mu_i(S_{j,e}^1 \times S_{-i,j}|h)$.

[42] Under priority to the agreement, here S_i^∞ must be substituted by S_i^1 in place of S_i. In the rest of the proof, substitute S_j^∞ with S_j as usual.

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

Thus, μ_i strongly believes $(S^1_{j,e})_{j\neq i}$, besides $(S^\infty_j)_{j\neq i}$. Hence, $S^2_{i,e} \supseteq \rho_i(\Delta^e_i) \cap S^\infty_i = S^1_{i,e}$. With $S^2_{i,e} \subseteq S^1_{i,e}$, we get $S^1_{i,e} = S^2_{i,e}$ for each $i \in I$. So, (i) $S^1_e = S^\infty_e$.

For each $i \in I$ and $s_i \in S^1_{i,e}$, there is μ_i that strongly believes $(e^0_j)_{j\neq i}$ and thus e^0_{-i} such that $s_i \in \rho_i(\mu_i)$. By T1 (Realization-strictness), for each $h \in H(e^0) \cap H(s_i)$, we have $s_i(h) = s'_i(h)$ for some $s'_i \in e^0_i \cap S_i(h)$. Then, $\zeta(S^1_e) \subseteq \zeta(e^0)$.

For each $i \in I$, by T3, $e^0_i = e^0_i \cap S_i(h) \subseteq \rho_i(\Delta^e_i) \cap S^\infty_i = S^1_{i,e}$. With $\zeta(S^1_e) \subseteq \zeta(e^0)$, we get (ii) $\zeta(S^1_e) = \zeta(e^0)$; with (i), we get (iii) $e^0 = S^\infty_e \cap e^0 \neq \emptyset$.

By (i), (ii), and (iii), $\zeta(S^\infty_e) = \zeta(S^1_e) = \zeta(e^0) = \zeta(S^\infty_e \cap e^0) \neq \emptyset$: 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 smallest $m \geq 0$ such that $S^\infty_e = S^m_e$ (it exists by finiteness of the game)[44].

The proof is constructive. For each $i \in I$, let $e^{k_i+1}_i := S^\infty_i$. For each $q = 0, ..., M + k_i + 1$, let

$$\overline{e}^q_i = \bigcup_{(n,m)\in\{0,k_i+1\}\times\{0,M\}:n+m=q} (e^n_i \cap S^{M-m}_{i,e}).$$

In the table, I show graphically the construction of each $\overline{e}^q_i$. Each box represents the intersection of its coordinates, and the union of the boxes marked with "x" represents $\overline{e}^q_i$ for some $q \leq \min\{k_i + 1, M\}$:

$\cap$	$S^M_{i,e}$	...	$S^{M-q}_{i,e}$	...	...	$S^0_{i,e}$
e^0_i			x			
...		x				
e^q_i	x					
...						
$e^{k_i+1}_i$						

So, $\overline{e}^q_i$ is the union of all boxes along the line that connects box $e^q_i \cap S^M_{i,e}$ with

[44] Here finiteness cannot be substituted by milder assumptions; for instance, in an infinitely repeated game, convergence to S^∞ and S^∞_e may require an infinite number of steps.

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 the line by 1 to the right, until $\overline{e}_i^{k_i+M+1} = e_i^{k_i+1} \cap S_{i,e}^0 = S_i^\infty$. The boxes above the line are subsets of the boxes along the line.

Without loss of generality, suppose that $\overline{e}_i^n \subsetneq \overline{e}_i^{n+1}$ for each $n = 0, ..., k_i + M$.[45] Then, $\overline{e} = (\overline{e}_i)_{i \in I}$ is an agreement, and it prescribes P because

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

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

For each $j \in I$, $S_{j,e}^M$ is the set of all $s_j \in S_j^\infty$ such that $s_j \in \rho_j(\mu_j)$ for some μ_j that strongly believes $((S_{i,e}^q)_{i \neq j})_{q=0}^M$ and $((e_i^q)_{q=0}^{k_i})_{i \neq j}$. I am going to show that, for each $i \neq j$, 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 $(e_i^q)_{q=0}^{k_i}$. Then, $S_{j,e}^M = \rho_j(\Delta_j^{\overline{e}}) \cap S_j^\infty$, which will be useful later.

First, I show that every μ_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 $e_i^{k_i+1} = S_i^\infty = S_{i,e}^0$, μ_j 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 and m with $e_i^n \cap S_{i,e}^m \subseteq \overline{e}_i^q$. Since μ_j strongly believes e_i^n and $S_{i,e}^m$, we have $1 = \mu_j((e_i^n \cap S_{i,e}^m) \times S_{-j,i}|h) \leq \mu_j(\overline{e}_i^q \times S_{-j,i}|h)$. Hence, μ_j strongly believes $\overline{e}_i^q$.

Second, I show that every μ_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 $S_{i,e}^0 = S_i^\infty \supseteq e_i^n$). By credibility of e, there exists μ_j' that strongly believes $(S_{i,e}^q)_{q=0}^M$ and $(e_i^q)_{q=0}^{k_i}$, and thus $\mu_j'(e_i^n \times S_{-j,i}|h) = \mu_j'(S_{i,e}^m \times S_{-j,i}|h) = 1$. Hence, $e_i^n \cap S_{i,e}^m \cap S_i(h) \neq \emptyset$. By construction, $e_i^n \cap S_{i,e}^{M-(M-m)} \subseteq \overline{e}_i^{M-m+n}$. So, $\overline{e}_i^{M-m+n} \cap S_i(h) \neq \emptyset$. For every $n' \leq n$, $e_i^{n'} \subseteq e_i^n$. For every $n' > n$, if $m = M$, $\overline{e}_i^{M-m+n} \cap (e_i^{n'} \backslash e_i^n) = \emptyset$; else, $\overline{e}_i^{M-m+n} \cap (e_i^{n'} \backslash e_i^n) \subseteq S_{i,e}^{m+1}$, but then, by definition of m, $S_{i,e}^{m+1} \cap S_i(h) = \emptyset$. (Graphically: all the boxes of $\overline{e}_i^{M-m+n}$ to the left of $e_i^n \cap S_{i,e}^m$ do not allow h.) So, $\overline{e}_i^{M-m+n} \cap S_i(h) \subseteq e_i^n$. Since μ_j strongly believes $\overline{e}_i^{M-m+n}$, we have $1 = \mu_j(\overline{e}_i^{M-m+n} \times S_{-j,i}|h) \leq$

[45] 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.

$\mu_j(e_i^n \times S_{-j,i}|h)$. So, μ_j strongly believes 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 $e_i^{k_i+1} = S_i^\infty \supseteq S_{i,e}^m$). By credibility of e, there exists μ'_j that strongly believes $(S_{i,e}^q)_{q=0}^M$ and $(e_i^q)_{q=0}^{k_i}$, and thus $\mu'_j(e_i^n \times S_{-j,i}|h) = \mu'_j(S_{i,e}^m \times S_{-j,i}|h) = 1$. Hence, $e_i^n \cap S_{i,e}^m \cap S_i(h) \neq \emptyset$. By construction, $e_i^n \cap S_{i,e}^{M-(M-m)} \subseteq \overline{e}_i^{M-m+n}$. So, $\overline{e}_i^{M-m+n} \cap S_i(h) \neq \emptyset$. For every $m' \geq m$, $S_{i,e}^{m'} \subseteq S_{i,e}^m$. For every $m' < m$, if $n = 0$, $\overline{e}_i^{M-m+n} \cap (S_{i,e}^{m'} \backslash S_{i,e}^m) = \emptyset$; else $\overline{e}_i^{M-m+n} \cap (S_{i,e}^{m'} \backslash S_{i,e}^m) \subseteq e_i^{n-1}$, but then, by definition of n, $e_i^{n-1} \cap S_i(h) = \emptyset$. (Graphically: all the boxes of $\overline{e}_i^{M-m+n}$ above $e_i^n \cap S_{i,e}^m$ do not allow h.) So, $\overline{e}_i^{M-m+n} \cap S_i(h) \subseteq S_{i,e}^m$. Since μ_j strongly believes $\overline{e}_i^{M-m+n}$, we have $1 = \mu_j(\overline{e}_i^{M-m+n} \times S_{-j,i}|h) \leq \mu_j(S_{i,e}^m \times S_{-j,i}|h)$. So, μ_j strongly believes $S_{i,e}^m$.

With $m = M$, the last paragraph shows that, for each $i \in I$ and $h \in H(S_{i,e}^M)$, there is n such that $\emptyset \neq \overline{e}_i^n \cap S_i(h) \subseteq S_{i,e}^M$. So, since $S_{i,e}^M = \rho_i(\Delta_i^{\overline{e}}) \cap S_i^\infty$, $\overline{e}$ satisfies T3. By $\overline{e}_i^{k_i+M+1} = S_i^\infty$ for every $i \in I$, $\overline{e}$ satisfies T2. It remains to show that $\overline{e}$ satisfies T1. Fix $i \in I$ and μ_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 $((S_j^q)_{j \neq i})_{q=0}^\infty$[46] (thus $\rho_i(\mu'_i) \subseteq S_i^\infty$) such that $\mu'_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(\overline{e}_{-i}^0)$ (it exists by $\overline{e}_{-i}^0 \subseteq S_{-i}^\infty$). Then,

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

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

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

■

[46] Under priority to the agreement, μ'_i needs not strongly believe $((S_j^q)_{j \neq i})_{q=0}^\infty$; under priority to the path, μ'_i needs to strongly believe $((S_{j,z}^q)_{j \neq i})_{q=0}^\infty$ in place of $((S_j^q)_{j \neq i})_{q=0}^\infty$.

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*, DOI 10.1007/s00199-018-1142-8.

[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., “A simple solution to the Hotelling problem”, working paper, 2019.

[18] Catonini, E. “Rationalizability and epistemic priority orderings,” *Games and Economic Behavior*, **114**, 2019, 101-117.

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

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

[21] D’Aspremont, C, J. J. Gabszewicz, J. F. Thisse “On Hotelling’s stability in competition,” *Econometrica*, **47**, 1979, 1145-1150.

[22] Hotelling, H. “Stability in competition,” *Economic Journal*, **39**, 1929, 41-57.

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

[24] Fudenberg, D., and D. Levine, “Self-confirming equilibrium”, *Econometrica,* **61**, 1993, 523-546.

[25] Govindan, S., and R. Wilson, “On forward induction,” *Econometrica*, **77**, 2009, 1-28.

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

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

[28] Harrington, J. “A Theory of Collusion with Partial Mutual Understanding”, **71(1)**, 2017, 140-158.

[29] 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.

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

[31] Kreps, D. M. and R. Wilson, “Sequential equilibria”, *Econometrica*, **50**, 1982, 863-94.

[32] Man, P. “Forward Induction Equilibrium”, *Games and Economic Behavior*, **75**, 2012, 265-276.

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

[34] Osborne, M. and C. Pitchik “Equilibrium in Hotelling’s model of spatial competition,” *Econometrica*, **55**, 1987, 911-922.

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

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

[37] 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.

[38] Siniscalchi, M., “Structural Rationality in Dynamic Games”, working paper, 2018.

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

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