HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Emiliano Catonini

SELF-ENFORCING AGREEMENTS AND FORWARD INDUCTION REASONING

Working Paper WP9/2018/01
Series WP9
Research of economics and finance

Moscow
2018

Editor of the Series WP9
“Research of economics and finance”
A. Belianin

Catonini, E.
Self-Enforcing Agreements and Forward Induction Reasoning* [Electronic resource] : Working paper WP9/2018/01 / E. Catonini ; National Research University Higher School of Economics. – Electronic text data (500 Kb). – Moscow : Higher School of Economics Publ. House, 2018. – (Series WP9 “Research of economics and finance”). – 84 p.

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 not. The incompleteness of the agreement can be crucial to implement an outcome. A particular way to rationalize deviations allows to establish connections with strategic stability (Kohlberg and Mertens, 1986).

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

J.E.L. Classification: C72, C73, D8

Emiliano Catonini, National Research University Higher School of Economics, International College of Economics and Finance.

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

1 Introduction

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

I take the view that players believe in the agreement only if compatible with the beliefs in rationality[2] and their interaction with the beliefs in the agreement of all orders. Ann will believe in the agreement only if Bob may comply with it in case he is rational, he believes in the agreement, he believes that Ann is rational and

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

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

believes in the agreement (which may add non-agreed upon restrictions on what Bob expects Ann to do), and so on. Moreover, I take the view that deviations, or more generally past actions, are not interpreted as mistakes but as intentional choices. Suppose that for Bob, in case he is rational and believes in the agreement, some move makes sense only if he plans to play a certain action thereafter. Ann, upon observing such move, will believe that Bob will play that action (and Bob may use the move to signal this). This instance of forward induction reasoning is based not just on the belief in Bob's rationality, but also on its interaction with the belief that Bob believes in the agreement. Example 3 in Section 2 is a case in point. Consider now a move that Bob, if he is rational and believes in the agreement, cannot find profitable whatever he plays thereafter. Example 1 in Section 2 illustrates a situation of this kind. Then Ann cannot keep believing that Bob is rational and, at the same time, that he believes in the agreement. Which belief will she maintain? Given the cheap talk nature of the agreement, I take the view that Ann will keep believing that Bob is rational (if this is per se compatible with Bob's behavior). However, in Section 5 I argue that the main insights of the paper go through under the opposite assumption. In addition, if compatible with Bob's behavior, Ann may maintain the belief that Bob believes that she would have not violated the agreement *before him*. In Section 6 I argue that the main insights go through also under this additional assumption.

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

In two-players games, I find that an outcome is implementable if and only if it is induced by a Nash equilibrium in extensive-form rationalizable strategies (Pearce [26]; Battigalli and Siniscalchi [8]) that satisfies "*realization-strictness*": all the normal-form best replies to co-players' equilibrium strategies induce the equilibrium outcome. Therefore, standard elimination procedure and fixed point

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

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

condition provide to the analyst (or to a mediator) the set of outcomes that can be achieved through pre-play coordination. Also in games with more than two players, an implementable outcome is not necessarily induced by a subgame perfect equilibrium (henceforth, SPE). This result may be surprising for two reasons. First, it is obtained under all the orders of belief in rationality which are compatible with the observed behavior, also after deviations from the agreed-upon path. Second, the literature has always assigned to subgame perfection a dominating role. At the end of Section 6 I will elaborate further on why I find this emphasis misplaced.[5]

In games with more than two players, not all realization-strict Nash equilibria in extensive-form rationalizable strategies induce an implementable outcome: the threats of two players towards a deviator may be mutually incompatible. Thus, further conditions on off-the-path behavior are required. To accomplish this task, I define a new, set-valued solution concept in reduced strategies: *Self-*

[5] The relationship between subgame perfection and strategic reasoning in absence of agreements has already been extensively studied for perfect information games (i.e. without simultaneous moves) with no relevant ties. Reny [27] shows that backward and forward induction strategies do not coincide. Nonetheless, Battigalli [4] proves that backward and extensive-form rationalizability yield the same unique outcome. This result is proved also by Heifetz and Perea [19] and by Chen and Micali [12]. The latter show that in all games with perfect recall, extensive-form rationalizability refines backward induction without equilibrium reasoning in terms of outcomes. In a previous work I find an overlapping between extensive-form rationalizability and SPE outcomes in games with observable actions.

Enforcing Set (henceforth, SES). Differently than in a SPE, in a SES the plans of deviators are not exogenously given, but are determined by forward induction. To implement a SES outcome, players can agree on the SES itself. Hence, they do not need to promise (and co-players trust) what they would do after an own violation of the agreement. That SES's are set-valued reflects the incompleteness of the agreement, which may be crucial for the implementation of an outcome: see Example 2 in Section 2.

Sometimes, the implementation of an outcome is possible only if players declare in advance what they would do after a own deviation. To fully characterize implementable outcomes, SES's are enriched through the notion of *tight agreement.* Like SES's, tight agreements only require to verify one-step conditions instead of many steps of reasoning, and implement exactly the outcomes they allow. In this sense, tight agreements are *truthful.* Hence, the characterization of implementable outcomes with tight agreements provides a revelation principle for agreements design: players need not be vague about the outcome they want to achieve.[6]

In many contexts, there are limitations to which agreements players can actually reach. On the one hand, players may be unable (or unwilling) to coordinate on a precise outcome.[7] On the other hand,

[6] Thus, agreement incompleteness and the related uncertainty are useful only off-path.

[7] For instance, Harrington [18] documents instances of "mutual partial understanding" among firms which leaves the exact path of price increase un-

in some contexts it may be natural to agree simply on an outcome to reach, without discussing what to do in case of a deviation. The methodology developed in the paper allows to evaluate agreements with any kind of incompleteness.

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, [5]; Battigalli and Siniscalchi, [9]). Strong-Δ-Rationalizability is based on the hypothesis that players *do not* maintain the belief in the rationality of the co-players when they display behavior which cannot be optimal under their first-order belief restrictions. Battigalli and Prestipino [7] show that Strong-Δ-Rationalizability captures indeed transparency of the first-order belief restrictions, i.e. the assumption that all orders of belief in the restrictions always hold in the game. Battigalli and Friedenberg [6] interpret the restrictions as the context in which the game takes place; for instance, a well-established convention.

To characterize the different hypotheses of this paper, another rationalizability procedure with first-order belief restrictions, *Selective Rationalizability*, is constructed and analyzed epistemically in [11]. Selective rationalizability captures *common strong belief in ra-*

determined 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.

tionality (Battigalli and Siniscalchi [8]), i.e. the assumption that any order of belief in rationality holds as long as not contradicted by the observed behavior. Thus, it combines unconstrained strategic reasoning (i.e. based only on beliefs in rationality) and constrained strategic reasoning (i.e. based also on first-order belief restrictions). In Section 5, I show how the assumptions and the notions adopted in this paper explain the differences in the results with respect to this literature.

Kohlberg and Mertens [20] were the first to introduce forward induction considerations into equilibrium reasoning, through the set-valued notion of *strategically stable* equilibria. Govindan and Wilson [15] refine sequential equilibrium with a notion of forward induction. However, these two prominent works and the related literature share the two same shortcomings. First, they never question subgame perfection as a must-have for a "strategically stable" solution. Second, the strategic reasoning that leads to playing such equilibria is unclear or limited.[8] The rationalizability approach adopted in this paper, which is backed by epistemic foundations, allows to eliminate both shortcomings. First, there is no constraint about how precisely and on which kind of equilibrium behavior players agree. Second, there is transparency about which particular agreements, beliefs, and epistemic assumptions induce different lines of reasoning, with

[8] A similar critique to strategic stability has been put forward by Van Damme [29].

a clear demarcation between unconstrained and constrained forward induction reasoning (missing in this literature).

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

Section 2 discusses the three simple examples mentioned above. More elaborate examples where the main ideas interact (Examples 4 and 5), along with an applied example, are presented in the Supplemental Appendix. Section 3 introduces the theoretical framework

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

and the analytic tools for the formal treatment of Section 4. Sections 5 and 6 discuss the relationship with the literatures on rationalizability and on equilibrium in dynamic games, and the robustness of the analysis to different kinds of forward induction reasoning. The Appendix collects all the proofs.

2 Examples

Example 1 Consider the following game.

$A\backslash B$	W	E		$A\backslash B$	L	R
N	$3,3$	$\cdot -$	$\longrightarrow$	U	$1,1$	$2,2$
S	$0,0$	$2,2$		D	$0,6$	$3,5$

The subgame has only one equilibrium, where all actions are played with equal probability. Hence, the unique SPE of the game induces outcome (S, E), which is Pareto-dominated by (N, W). Suppose, Ann and Bob agree to play (N, W) and that Ann should play U in case of deviation of Bob. Is the agreement credible? If Bob is rational, he may deviate only if he does not believe in N, or U, or both. Then, after the deviation, Ann cannot believe at the same time that Bob is rational and believes in the agreement. If she drops the belief that Bob believes in the agreement and maintains the belief that Bob is rational, she can believe that Bob does not believe in U and that he will play L. Hence, she can react with U. Anticipating this, Bob can expect N and U, and refrain from deviating. Further steps of reasoning do not modify the conclusion: the agreement is credible and, once believed, players will comply with it.

Example 4 provides a similar game where the unique SPE outcome is Pareto-dominated by a just Nash one. While here the Nash threat U is played with positive probability also in the SPE, the credible threat that sustains the Nash outcome in Example 4 differs from the unique equilibrium action of the subgame. Moreover, while here the SPE outcome can be achieved without an explicit threat,[10] in Example 4 the unique SPE outcome cannot be secured without explicit threats, just like the Nash.

Example 2. In this 4-players game,[11] in the subgame, Cleo chooses the matrix, Ann the row, and Bob the column (payoffs are

[10] If players agree on (S, E) and Ann deviates to N, Bob can still believe that she is rational and believes in the agreement. In this case, Ann would deviate only under beliefs about Bob's reaction which make her play D. But then, Bob would always react with L.

[11] This game is freely inspired by the leading example in Greenberg [16], with a fundamental difference: in that example, the mediating country remains silent, and there simply *exist* beliefs about its behavior that make the warring countries behave as desired; here the warring countries remain silent and the mediating country speaks, and this suffices to pin down beliefs that *all* induce the desired behavior by the first mover (here a fourth country).

in alphabetical order).

$$\mathit{Dave} \quad \text{—}\, \mathit{Out} \longrightarrow \quad 5,5,1,3$$

$$\downarrow \mathit{Inst}$$

Int	*Arm*	*Not*	*Not*	*Arm*	*Not*
Arm	$4,4,0,2$	$4,3,2,1$	*Arm*	$0,0,1,9$	$6,1,1,6$
Not	$3,4,2,1$	$5,5,0,0$	*Not*	$1,6,1,6$	$5,5,1,3$

Dave, a weapons producer, can *instigate* a conflict between Ann and Bob. If he does, Cleo can *intervene* to avoid an escalation and retaliate against Dave, with a cost of 1 for herself and 3 for him. By doing so, if only Ann or only Bob participates to the *arms race*, Cleo can extract 2 utils from the other for protection. Under Cleo's peacekeeping, the arms race transfers 1 from Ann/Bob to Dave, and Ann and Bob prefer to adopt the same strategy; in case of escalation, the arms race transfers 3, and Ann and Bob have the incentive to be belligerent when the other is peaceful, and vice versa. The unique equilibrium of the subgame assigns equal probability to all actions and induces Dave to instigate. However, instead of looking for a diplomatic solution that involves all parties, Cleo can simply threaten Dave to intervene, while Ann and Bob remain silent. This is credible: under belief in the intervention, both actions are rational for Ann and Bob, who may fail to coordinate. Once Dave believes in Cleo's intervention, he has the incentive not to instigate.

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

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

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

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

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

This example is analyzed formally in the Supplemental Appendix.

3 Agreements, beliefs and strategic reasoning

3.1 Preliminaries

Primitives of the game.[12] Let I be the finite set of *players*. For any profile $(X_i)_{i\in I}$ and any $\emptyset \neq J \subseteq I$, I write $X_J := \times_{j\in J} X_j$, $X := X_I$, $X_{-i} := X_{I\setminus\{i\}}$. Let $(\overline{A}_i)_{i\in I}$ be the finite sets of *actions* potentially available to each player. Let $\overline{H} \subseteq \cup_{t=1,...,T}\overline{A}^t \cup \{\emptyset\}$ be the set of histories, where $h^0 := \emptyset \in \overline{H}$ is the start of the game and T is the finite horizon. For any $h = (a^1, ..., a^t) \in \overline{H}$ and $l < t$, it holds $h' =$

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

$(a^1, ..., a^l) \in \overline{H}$, and I write $h' \prec h$.[13] Let $Z := \{z \in \overline{H} : \forall h \in \overline{H}, z \nprec h\}$ be the set of terminal histories (henceforth, *outcomes* or *paths*)[14], and $H := \overline{H} \backslash Z$ the set of non-terminal histories (henceforth, just *histories*). Let $A_i : H \rightrightarrows \overline{A}_i$ be the correspondence that assigns to each history h, always observed by player i, the set of actions $A_i(h) \neq \emptyset$[15] available at h to player i; as standard, for all $h \in H$, $(h, a) \in \overline{H}$ 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 strategy of player i is a function $s_i : h \in H \mapsto s_i(h) \in A_i(h)$. Let S_i denote the set of all strategies of i. A strategy *profile* $s \in S$ naturally induces a unique outcome $z \in Z$. Let $\zeta : S \to Z$ be the function that associates each strategy profile with the induced outcome. For any $h \in \overline{H}$, the set of strategies of i compatible with h is:

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

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

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

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

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

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

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

Histories and outcomes of $\Gamma(h)$ will be identified by those that follow h in the whole game, and not redefined as shorter sequences of action profiles.

Realization-strictness. A Nash equilibrium $s = (s_i)_{i\in I} \in S$ is *realization-strict* (r-strict) if, for all $i \in I$ and $s_i' \notin S_i(\zeta(s))$, $u_i(\zeta(s)) > u_i(\zeta(s_i', s_{-i}))$.

3.2 Agreements

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

- Players do not coordinate explicitly as the game unfolds: All the opportunities for coordination are discussed beforehand.
- No subset of players can reach a private agreement, secret to co-players.
- Players do not agree on the use of randomization devices. Players would lack the incentive to (set the agreed-upon odds and) stick to the output of a (artificial) randomization device over the own actions.[16] Players also lack the ability to commit, otherwise it would not make sense to talk of non-binding agreements. Agreeing on the use of joint randomization devices, instead, would expand the set of outcomes players can achieve,[17] and could be analyzed with the methodology developed in this paper.

Players can leave two kinds of strategic uncertainty, i.e. *agreement incompleteness*. First, and more importantly, players can be

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

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

vague about which action they intend to play at some history. Second, players can claim to be planning a certain action at only one of two unordered histories, without revealing at which one. This second kind of vagueness (which can also arise naturally from rationality: see Section 3.4) can be profitably exploited in agreements: see Example 5. A player can also declare what she plans to do in case she deviates from her initial plans. And so on. Also the trust in a player who has already violated the agreement can be strategically exploited:[18] see again Example 5. Thus, agreements are formally modeled as follows.

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

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

Starting from the root of the game, an agreement can assign to a player a non-empty set of continuation plans only at histories that immediately follow a deviation by the player from the plans already assigned.[19] However, (i) the agreement may be empty at all

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

[19] This is reminiscent of the notion of *basis* of a CPS by Siniscalchi [28]: new theories are introduced only at histories that are not deemed as "*plausible*" as the previous ones under the theories already introduced.

such histories. Moreover, (ii) it may de facto not restrict a player's behavior also at histories that follow a deviation by anyone else. Agreements are particularly simple when (iii) players declare which actions they may play at each history, independently of what they plan to do at other histories.

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

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

ii *a path agreement on $z \in Z$ if it is reduced and for every $i \in I$, $e_i^0 = S_i(z)$;*[20]

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

A reduced agreement corresponds to a profile of strategy sets.[21] A path agreement corresponds to just agreeing on an outcome to achieve. Within the formalism of agreements, agreements on actions are expressed through *vetos* V_i^h cast by players on outcomes. In the examples of the paper, where most agreements are reduced and on actions, agreements on actions are equivalently expressed declaring actions instead of continuation plans at each individual history. Non-reduced agreements can be found in Examples 4 and 5. An

[20] The term path agreement was first used by Greenberg et al. [17]: see also footnote 34.

[21] Recall that all strategies are reduced.

agreement which is not on actions is discussed in Example 5. Path agreements can be found in Examples 3 and 4.

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

3.3 Belief in the agreement

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

Definition 3 *Fix $i \in I$. An array of probability measures $(\mu_i(\cdot|h))_{h \in H}$ over co-players strategies S_{-i} is a* Conditional Probability System *if for all $h \in H$, $\mu_i(S_{-i}(h)|h) = 1$, and for all $h' \succ h$ and $\overline{S}_{-i} \subseteq S_{-i}(h')$,*

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

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

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

For any player i and any set of co-players $J \subseteq I \backslash \{i\}$, I say that a CPS μ_i *strongly believes* $\overline{S}_J \subseteq S_J$ if for every $h \in H(\overline{S}_J)$,

$\mu_i(\overline{S}_J \times S_{I\setminus(J\cup\{i\})}|h) = 1$.[22] In formulae and proofs, I will write "that s.b." for "that strongly believes".

Note that a player can have correlated beliefs about the strategies of different co-players. This is not in contradiction with the absence of joint randomization devices in the agreement: players can believe in spurious correlations among co-players' strategies (see, for instance, Aumann [1] and Brandenburger and Friedenberg [10]).[23] However, *strategic independence* (Battigalli [3])[24] could be assumed throughout the paper and the results would not change.

I say that a player believes in the agreement if, at each history and for each co-player, she assigns probability 1 to strategies of the co-player which comply with the agreement from her most recent violation of the agreement onwards.

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

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

[22] In the original meaning of "strong belief", due to Battigalli and Siniscalchi [8], $\overline{S}_J \times S_{I\setminus(J\cup\{i\})}$ and not $\overline{S}_J$ is "strongly believed". The slight difference in the use of the term is only for notational convenience.

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

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

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

Note that every $\mu_i \in \Delta_i^e$ strongly believes $(e_j^0)_{j\neq i}$ (and not just $\times_{j\neq i} e_j^0$).

3.4 Rationality and Rationalizability

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

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

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

The set of sequential best replies to μ_i (resp., to some $\mu_i \in \Delta_i^e$) is denoted by $\rho(\mu_i)$ (resp., by $\rho(\Delta_i^e)$). The set of normal-form best replies to a probability measure ν_i on S_{-i} is denoted by $r_i(\nu_i)$. I say that a strategy s_i is *rational* if it is a sequential best reply to some $\mu_i \in \Delta^H(S_{-i})$. An important remark: Even when no rational strategy prescribes action a_i at two unordered histories h and h', there might be other two rational strategies, both compatible with h and h', which prescribe a_i only at, respectively, h and h'.

Here I take the view that players refine their first-order beliefs through strategic reasoning based on beliefs in rationality and beliefs in the belief in the agreement. In particular, I assume that every player, as long as not contradicted by observation, believes that each co-player is rational and believes in the agreement; that each co-player believes that each other player is rational and believes in the agreement; and so on. At histories where common belief in, jointly, rationality and belief in the agreement is contradicted by observation,[25] I assume that players maintain all orders of belief in rationality that are per se compatible with the observed behavior, and drop the incompatible orders of belief in the agreement. I will call *independent rationalization* the hypothesis that players maintain a order of belief in rationality or in the agreement about a co-player when her *individual* behavior allows, as opposed to the hypothesis that players maintain such order of belief about all co-players only until none of them contradicts it.[26] The adoption of independent ra-

[25] In [11] I show how anticipating which beliefs are kept at these histories refines also first-order beliefs at histories where all orders of belief in rationality and agreement hold.

[26] This is not in contradiction with the absence of strategic independence: players can believe in spurious correlations among co-players' strategies, although

tionalization shows better the robustness of the main insights. After a deviation that displays the disbelief of the deviator in the agreement, without independent rationalization co-players' threats would not be required any degree of coordination. In Example 5, independent rationalization makes it much more challenging for players to find an effective agreement.

As shown in [11], the behavioral consequences of this kind of strategic reasoning are captured by Selective Rationalizability. Selective Rationalizability refines the following version of Extensive Form Rationalizability[27] (henceforth just **Rationalizability**).

Definition 6 *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 $s_i \in \rho(\mu_i)$ 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.

Next, **Selective Rationalizability**. Fix an agreement $e = (e_i)_{i\in I}$.

they are ready to believe that different co-players have different orders of belief in rationality or in the agreement. For instance, the beliefs of a more and a less sophisticated players can be affected by weather in the same way.

[27] This notion of Extensive-Form-Rationalizability is the adaptation of Strong Rationalizability (Battigalli and Siniscalchi, [8]) to independent rationalization. Independent rationalization is also a feature of Independent Rationality Orderings (Battigalli [3]), where strategic independence is adopted. The original notion of Extensive-Form-Rationalizability, due to Pearce [26], adopts instead structural consistency (Kreps and Wilson [21]).

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 $s_i \in S_i$, let $s_i \in S_{i,e}^n$ if and only if $s_i \in \rho(\mu_i)$ for some $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^{n-1}$ such that:*

S3: μ_i strongly believes $((S_j^q)_{j\neq i})_{q=0}^{\infty}$.

Finally, let $S_{i,e}^\infty := \cap_{n\geq 0} S_{i,e}^n$. The profiles S_e^∞ are called selectively-rationalizable.

S3 guarantees that a player always believes in co-players' strategies which are compatible with the highest possible order of belief in rationality. Among those, a player believes in co-players' strategies which are compatible with the agreement and with the highest possible order of belief in the agreement. Note that first-order belief in the agreement, as required by Δ_i^e, is mandatory at all histories. Then, the empty set is obtained when at some step some co-player can reach a history only with strategies that do not comply with the agreement from the history on. In this way, the compatibility of the belief in the agreement with the strategic reasoning hypotheses is tested.

Consider now the following class of rationalizable continuation plans, which are "realization equivalent" under the assumption that the opponents play rationalizable plans too. For any $h \in H(S^\infty)$ and $\overline{s}_i^h \in S_i^h$, let $\left[\overline{s}_i^h\right]^\infty$ be the set of all $s_i^h \in S_i^\infty|h$ such that $\zeta(s_i^h, s_{-i}^h) = \zeta(\overline{s}_i^h, s_{-i}^h)$ for all $s_{-i}^h \in S_{-i}^\infty|h$. For any $\overline{S}_i^h \subseteq S_i^h$ (possibly empty), let $[\overline{S}_i^h]^\infty := \cup_{\overline{s}_i^h \in \overline{S}_i^h} \left[\overline{s}_i^h\right]^\infty$. I say that $e = (e_i)_{i \in I}$ is a **rationalizable agreement** if for all $i \in I$, $e_i^h = \left[e_i^h\right]^\infty$ for all $h \in H(S^\infty)$ and $e_i^h = \emptyset$ for all $h \notin H(S^\infty)$. By Definition 12 and Theorem 1, rationalizable agreements suffice to induce all the implementable outcome sets (and also the agreements that correspond to a Self-Enforcing Set are rationalizable, see Definition 13). S3 can be substituted by $s_i \in S_i^\infty$ for all rationalizable agreements: see Lemma 3 in the Appendix. However, for any agreement, Rationalizability and Selective Rationalizability can be merged into one elimination procedure, where the belief in the agreement kicks in once the rationalizable profiles are obtained (see footnote 49). Finally, strong belief in $((S_{j,e}^q)_{j \neq i})_{q=0}^{n-2}$ can be replaced by $s_i \in S_{i,e}^{n-1}$ in 2-players games or dropping independent rationalization: see [11] for details.

Only in the applied example in the Supplemental Appendix, the game features non-rationalizable strategies. To see Selective Rationalizability at work, check the formalization of Example 3 in the Supplemental Appendix.

I will refer to $\zeta(S_e^\infty)$ as the set of outcomes *induced* by e, and to histories in $H(S^\infty)$ as "rationalizable histories".

4 Self-enforceability and implementability

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

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

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

Definition 9 *A credible agreement is self-enforcing if*

$$\zeta(S_e^\infty) = \zeta(S_e^\infty \cap e^0).$$

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

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

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

In Examples 1 and 2, the reduced agreements with, respectively, $e_A^0 = \{N.U\}, e_B^0 = \{W\}$, and $e_C^0 = \{Int\}, e_i^0 = S_i, i = A, B, D$ are self-enforcing. All strategies are rationalizable. At the first step of Selective Rationalizability, in Example 1 Ann eliminates S and Bob selects W, while in Example 2 Dave selects O and the other players do not eliminate any strategy. In both cases, Selective Rationalizability is over at the first step. Example 3 provides two non-credible agreements, as formally shown in the Supplemental Appendix.

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

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

For these reasons, in the remainder of the paper, the focus will be on self-enforcing agreements. Which outcomes of the game can be achieved through self-enforcing agreements?

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

With "implementable outcomes" I will refer specifically to implementable singletons. The set of outcomes prescribed by a self-enforcing agreement may be larger than the outcome set it induces (i.e. $\zeta(e^0) \supset \zeta(S_e^\infty)$). So, a natural question arises: for each implementable outcome set, is there an implementing agreement that prescribes precisely that set of outcomes? The answer is not obvious because simply restricting the initial plans of some self-enforcing agreement to those that allow the implemented outcome set may not work: see Example 4. Therefore, consider the following classes of agreements.

Definition 11 *A self-enforcing agreement is truthful if*

$$\zeta(S_e^\infty) = \zeta(e^0).$$

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

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

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

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

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

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

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

Proposition 3 *A tight agreement is truthful.*

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

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

Then, by Remark 1, the following holds.

Corollary 1 *Every implementable outcome is induced by a r-strict Nash equilibrium in rationalizable strategies.*[28]

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

Theorem 1 and Proposition 3 answer to the original question.

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

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

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

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

Example 5 illustrates an interesting tight agreement, which prescribes an outcome that cannot be implemented without restrictions to the behavior of a deviator, nor by an agreement on actions. But usually, tight agreements are more complex than needed for the implementation of an outcome set. For a single outcome, the simplest and more natural agreement is the corresponding path agreement. Yet, very few path agreements are self-enforcing. In Example 4, not even the path agreement on the unique SPE outcome is self-enforcing. Therefore, one may wonder which outcome sets can be implemented with reduced agreements and agreements on actions.

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

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

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

♣ *Self-Justifiability:*

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

♡ *Forward Induction:*

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

$\diamondsuit$ *Self-Enforceability*: *For each* μ_i *that s.b.* S^*_{-i},

$$\zeta(\rho(\mu_i) \times S^*_{-i}) \subseteq \zeta(S^*).$$

Rationalizability says that the SES prescribes rationalizable plans without further restricting behavior at the non-rationalizable histories. Consider now players who strongly believe that *each* co-player will play as the SES prescribes and, alternatively, as rationalizability prescribes. Self-Justifiability says that they may play any strategy in the SES. Forward Induction says that all the strategies such players may play, thus including the SES strategies, are compatible with strong belief that co-players form beliefs in the same way. At each history h that follows a deviation by player j from S^*_j, the logics of Forward Induction differ from the logics of subgame perfection in the following way. Forward Induction determines the expected continuation plans of j with forward induction reasoning, based on her belief in the SES if possible ($\overline{S}_j$) or just the beliefs in rationality otherwise (S^∞_j). Subgame perfection prescribes exogenously the continuation plans of deviators, and imposes that they *always* best reply to the planned reactions of co-players. The best response condition of Forward Induction, imposed after one step of reasoning instead of just at the start, suffices to guarantee credibility after all steps of reasoning, which players do not actually need to perform when they agree on a SES.

On top of this, Self-Enforceability[29] guarantees that players will not leave the paths induced by the SES if they strongly believe that *all* co-players will play as the SES prescribes.[30] Self-Justifiability further guarantees truthfulness of the agreement that corresponds to the SES.

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

Conversely, one could think that, for any self-enforcing agreement, $S_e^\infty \cap e^0$ is a SES. While $S_e^\infty \cap e^0$ satisfies Self-Enforceability and Self-Justifiability, and restrictions to behavior at non-rationalizable histories can always be eliminated as to satisfy Rationalizability, $S_e^\infty \cap e^0$ may not satisfy Forward Induction. The sequential best replies of player i under strong belief in $(S_{j,e}^\infty \cap e_j^0)_{j \neq i}$ may not be, at some history, what co-players expect after all steps of reasoning under the agreement. Such refinement of beliefs may be crucial to sustain the threats. For this reason, not every implementable outcome set is induced by some SES, not even if implemented by a truthful, reduced agreement: see Example 5. However, a SES always exists.

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

[30] This is reminiscent of the notion of "strategy subsets closed under rational behavior" by Basu and Weibull [2], but in the context of dynamic games and with focus on the realized paths instead of the strategies.

Remark 2 *S^∞ is a SES.*

The search for candidate SES's conveniently coincides with the search of the initial plans of a tight agreement. Then, Forward Induction must be checked. If no candidate SES for the implementation of an outcome set satisfies Forward Induction, then one can try to transform a candidate SES into a tight agreement, by declaring the behavior of deviators as to satisfy T1 and T2 off-path. This whole procedure is performed in Example 5. Also the reduced agreement of Example 2 is an interesting SES where set-valuedness, i.e. agreement incompleteness, plays a crucial role (while forward induction is immaterial because each player moves only once.)

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

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

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

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

Casting unilateral vetos on outcomes is equivalent to exclude actions instead of strategies. The candidate SES is then the set of rationalizable strategies that do not prescribe the excluded actions. The implementing reduced agreement on actions is the set of *all* strategies that allow the SES outcomes.[31] Note that S^{∞} always satisfies Rationalizable Vetos.

Let us focus now on implementable outcomes. By Rationalizability and Self-Enforceability, every SES that induces a unique outcome is a set of r-strict Nash equilibria in rationalizable strategies. Does the opposite hold? The answer is no: the threats of two different players towards a potential deviator may be incompatible with each other. But this cannot happen in a two-players game.

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

Moreover, for each $s \in S^*$*, the reduced agreement* e *with* $e^0 = \{s\}$ *implements* z*.*[32]

Together with Corollary 1, the following holds.

[31] With $V_i^0 = W_i$, the agreement may be not credible: for some $h \in H(S_i^{\infty}) \cap H(S_i \backslash \cup_{z \in W_i} S_i(z))$, it may hold $S_i^{\infty}(h) \backslash \cup_{z \in W_i} S_i(z) = \emptyset$, so strong belief in both S_i^{∞} and $S_i \backslash \cup_{z \in W_i} S_i(z)$ is impossible.

[32] The agreement on a SES that induces the outcome instead of on a precise Nash may be however more natural: see the applied example in the Online Appendix.

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

Together with Proposition 5, the following holds.

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

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

5 Comparison with rationalizability literature

The literature on strategic reasoning with first-order belief restrictions is mostly based on the use of Strong-Δ-Rationalizability ([5], [9]). The definition of Strong-Δ-Rationalizability with independent rationalization coincides with Definition 7 without S3 and with $S^0 = S$. The differences between the results of this paper and the results in this literature are due to (i) the adoption of Selective Rationalizability in place of Strong-Δ-Rationalizability, (ii) the structure on the first-order belief restrictions imposed by the notion of agreement, and (iii) the focus on self-enforceability rather than just credibility.

Differences and similarities between Selective Rationalizability and Strong-Δ-Rationalizability are deeply analyzed in [11]. Here I only recall the main conceptual difference behind the two solution concepts. Fix 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, co-players *drop* the belief that the player is rational. This hypothesis is called in [11] "*(epistemic) priority to the agreement*" (as opposed to *rationality*). So, the question is: how would the adoption of Strong-Δ-Rationalizability instead of Selective Rationalizability affect the results?

In every example except the applied example in the Supplemental

Appendix, all strategies are rationalizable, thus Selective Rationalizability and Strong-Δ-Rationalizability coincide. Hence, the insights from the examples are robust to a shift of epistemic priority from rationality to the agreement.

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

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

1. *selectively-rationalizable strategies (S^∞_e) with strongly-Δ-rat. strategies ($S^\infty_{\Delta^e}$) everywhere;*
2. *rationalizable strategies (S^∞) with all strategies (S) in the definitions of $[\cdot]^\infty$,*[33] *tight agreement, and SES, and with rational strategies (S^1) in the statements of Corollary 1, Proposition 5, and Theorem 3.*

To verify Remark 3, the required modifications to the proofs of the results are highlighted in the Appendix. A credible agreement under priority to rationality needs not be credible under priority to the agreement: as shown in [11], Selective Rationalizability is not a refinement of Strong-Δ-Rationalizability for the same first-order

[33] This is just to adapt to the formalism of Section 4: the equivalence classes become singletons (in the sense of one reduced strategy).

belief restrictions. Across all agreements, instead, under priority to the agreement more outcome sets can be implemented.

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

However, since agreements originate from mere pre-play cheap talk, epistemic priority to rationality appears in my view as a more considerate hypothesis. Else, for instance, any Nash equilibrium in rational strategies of a two-players game would correspond to a self-enforcing agreement, also when incompatible with just strong belief in rationality.

Battigalli and Friedenberg [6] capture the implications of Strong-Δ-Ratio- nalizability without independent rationalization *across all* first-order belief restrictions with the notion of Extensive Form Best Response Set. An EFBRS is a Cartesian set of strategy profiles $\overline{S} = \times_{i \in I} \overline{S}_i$ satisfying the following:

EFBRS: for every $i \in I$ and $s_i \in \overline{S}_i$, $s_i \in \rho(\mu_i)$ for some μ_i that strongly believes $\overline{S}_{-i}$ with $\rho(\mu_i) \subseteq \overline{S}_i$.

The EFBRS Condition is the analogue of Self-Justifiability in absence of priority to rationality and independent rationalization, but with an additional "maximality" requirement: all the sequential best replies to some justifying beliefs must be in the EFBRS. These beliefs are not expressed by the EFBRS itself, whereas a SES directly provides the first-order belief restrictions that yield the SES outcomes. The restrictions that yield the EFBRS may impose belief in specific randomizations, or, more fundamentally, differ across two players regarding the moves of a third player.[34] An agreement, instead, aligns any two player's beliefs about a third player's moves. For this reason, even with randomizations in agreements and without independent rationalization, EFBRS's would still be insufficient for implementability of the induced outcomes under priority to the agreement, calling for Self-Enforceability in place of maximality.

Battigalli and Siniscalchi [9] find out that, for first-order belief restrictions which correspond to the belief in an outcome, Strong-Δ-Rationalizability yields a non-empty set only if there exists a self-confirming equilibrium (Fudenberg and Levine [14]) inducing that outcome. Regardless of the epistemic priority choice, implementable outcomes are instead all Nash by Corollary 1 and Remark 3. Why is it the case? The reason lies in the difference between credibility and self-enforceability. Under a self-enforcing agreement, players have

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

the incentive to stay on path for *all* their refined beliefs. This allows to find strategies of co-players against which there is no incentive to deviate. Credibility, instead, may be granted just by some particular (correlated) belief about the reactions of co-players to the deviation.

Conversely, in signaling games, Battigalli and Siniscalchi [9] show that when an equilibrium outcome satisfies the Iterated Intuitive Criterion (Cho and Kreps [13]), Strong-Δ-Rationalizability yields a non-empty set under belief in that outcome. Yet, even in the simplest examples of this paper, off-the-path restrictions are usually needed for self-enforceability. What does strategic reasoning under path restrictions represent when the agreement is richer than the path agreement? The next section sheds light on this point.

6 Comparison with equilibrium literature

Kohlberg and Mertens [20] motivate their analysis in a similar way to this paper: "*A noncooperative game is played without any possibility of communication between the players. However, we may think of the actual play as being preceded by a more or less explicit process of preplay communication (the course of which has to be common knowledge to all players), which gives rise to a particular choice of strategies.*" ([20], page 1004) Then, they introduce forward induction as implicit communication *during* the game, based on actual moves:

"*Essentially what is involved here is an argument of "forward induction": a subgame should not be treated as a separate game, because it was preceded by a very specific form of preplay communication — the play leading to the subgame.*" ([20], page 1013) Finally, they claim that the "forward induction" property of their notion of strategic stability, "*captures the "forward-induction" logic of our basic example.*" ([20], page 1029) The two examples of forward induction in the paper refer to a player who gives up an outside option. The consequent reasoning is not based on pre-play communication: unconstrained forward induction reasoning suffices for players to coordinate on the strategically stable solutions of two examples.

Govindan and Wilson [15] use the Beer-Quiche game (Cho and Kreps, [13]) to show a different kind of forward induction reasoning. In Beer-Quiche, one of the two pure equilibria can be ruled out with a story of interactive beliefs in its outcome distribution. That is, constrained forward induction reasoning. However, both kinds of reasoning are hard to detect in their formal definition of forward induction, while depth of reasoning and scope of the analysis remain limited. As acknowledged by the authors themselves, their notion of forward induction only captures rationality and strong belief in rationality in two-players games ([15], page 11),[35] and fails in games

[35] I suggest that the two steps limitation (rationality and strong belief in rationality) on uncontrained reasoning extends to the constrained reasoning captured by forward induction. Moreover, I suggest that, once forward induction is immerged in sequential equilibrium, a further step of reasoning is captured *at the*

with more than two players ([15], page 21). Moreover, it applies only to sequential equilibrium.

Osborne [24] identifies a class of non strategically stable SPE in two-players, finitely repeated, coordination games: those with an *equilibrium path that can be upset by a convincing deviation*. Differently than for the general definition of strategic stability, it is easy to identify a precise line of strategic reasoning that rules out these equilibria: forward induction about the path agreement. Indeed, equilibrium paths that can be upset by a convincing deviation can be characterized as non-credible path agreements. This is proved in the Supplemental Appendix. Thus, also strategic stability captures (at least to some extent) constrained forward induction reasoning about the beliefs in an outcome (distribution).[36]

However, very few path agreements implement the outcome they prescribe. Off-path restrictions are usually needed for implementation. Analogously, also strategic stability entails restrictions on off-path continuation strategies. So, what does strategic reasoning under the path agreement represent when off-the-path threats are actually in place? It represents a particular way to rationalize deviations, transparent to players. This rationalization of deviations relies

beginning of the game. Indeed, the equilibrium selection in Beer-Quiche also requires a further step of reasoning at the start.

[36] Indeed, also Kohlberg and Mertens [20], in the applications section, refine equilibria in Beer-Quiche with strategic stability, without discussing the connection with forward induction.

on the belief that the deviator believes that no deviation by a co-player would have occured had she stayed on path. If the deviation does not contradict this belief, the co-players, instead of dropping the belief that the deviator believes in the *whole* agreement, drop the belief that the deviator believes in the post-deviation threats, and save the belief that the deviator believed in the agreement on-path. (So, they believe that the deviator will try to achieve a higher payoff than under the agreed-upon path.) In other words, the beliefs in the compliance with the agreement on-path have higher epistemic priority than the beliefs in the compliance with the agreement off-path. Assigning the highest epistemic priority to the beliefs in rationality, I call this finer epistemic priority order "*(epistemic) priority to the path*". Its behavioral consequences are captured by an extension of Selective Rationalizability, epistemically characterized in [11]. With this, I will show the robustness of the insights of the paper to this kind of strategic reasoning, and provide a general and transparent approach to the forward induction stories in the background of the equilibrium literature.

For simplicity, I restrict the analysis to agreements which prescribe a unique outcome z. Let $((S^q_{j,z})_{j\in I})^{\infty}_{q=0}$ denote Selective Rationalizability under the path agreement on z, and call $(S^{\infty}_{j,z})_{j\in I}$ z-*rationalizable*. Fix an agreement $e=(e_i)_{i\in I}$ with $\zeta(e^0)=\{z\}$.

Definition 14 *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* $s_i\in S_i$, *let* $s_i\in S^n_{i,e^z}$ *if and* *only $s_i \in \rho(\mu_i)$ for some $\mu_i \in \Delta_i^e$ that strongly believes $((S^q_{j,e^z})_{j\neq i})_{q=0}^{n-1}$ such that:*[37]

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

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.

E3 captures the interpretation of deviations depicted above. On top of this, players refine their beliefs according to the whole agreement. Then, for the agreement to be credible, the off-the-path threats have to be compatible with the rationalization of deviations based on the beliefs in the path.

So, the credibility of the path agreement only constitutes a preliminary test for the implementability of z under the hypotheses of this section. If the outcome passes the test, there exist off-the-path beliefs, compatible with the rationalization of deviations depicted above, which induce players to stay on path. However, no agreement may be able to restrict players' beliefs to those, like for the beliefs that sustain an EFBRS. An example of this is provided in [11], and it motivates the consideration of different belief restrictions in an epistemic priority order, instead of just turning to path restrictions and using credibility in place of self-enforceability.

[37] Although typically $\Delta_i^e \not\subseteq \Delta_i^z$, requiring $\mu_i \in \Delta_i^e$ is equivalent to requiring $\mu_i \in \Delta_i^e \cap \Delta_i^z$, thus $S^1_{e^z} \subseteq S^\infty_z$; see the manuscript "On non-monotonic strategic reasoning" (Catonini, 2017).

Analogously to Selective Rationalizability, E3 can be substituted by $s_i \in S^{\infty}_{i,z}$ for all the agreements $e = (e_i)_{i \in I}$ such that, redefining $[\cdot]^{\infty}$ with S^{∞}_z in place of S^{∞}, $e^h_i = \left[e^h_i\right]^{\infty}$ for all $i \in I$ and $h \in H(S^{\infty}_z)$, and $e^h_i = \emptyset$ otherwise.[38] And again, this class of agreements suffices to induce all the implementable outcome sets under priority to the path. Indeed, the analysis of Section 4 can be replicated under this finer epistemic priority order.

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

1. *selectively-rationalizable strategies (S_e) with z-selectively-rat. strategies (S_{e^z});*
2. *rationalizable strategies (S^{∞}) with z-rationalizable strategies (S^{∞}_z).*

To verify Remark 4, the required modifications to the proofs of the results are highlighted in the Appendix. Although z-Selective Rationalizability does not refine Selective Rationalizability for a fixed agreement, the following holds.

Proposition 7 *If an outcome is implementable under priority to the path, then it is implementable under priority to rationality.*

[38] I do not provide formal proof of this fact. However, E3 is maintained in the proofs.

In all the examples, the self-enforcing agreements remain self-enforcing under priority to the path. Hence, the insights are robust to the finer epistemic priority order adopted in this section. Strategic stability does not eliminate every non subgame perfect equilibrium either;[39] yet, in the attempt to do so, equilibria that are compatible with forward induction are disregarded.[40]

The final question is: does subgame perfection perform a meaningful further refinement under these strategic reasoning hypotheses? My answer is no. Subgame perfection is at deep contradiction with the interpretation of deviations behind this kind of forward induction reasoning. Fix a r-strict SPE. After any deviation from the SPE path, co-players will believe that the deviator believed in the path but does not believe in the threat. Then, they will not expect the deviator to best reply to the threat. But then, that the threat is a best reply to a plan of the deviator which is a best reply to the threat itself is of no additional value. This breaks down the logics of

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

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

subgame perfection. Example 4 illustrates this intuition. Thus, the insistence on subgame perfection in the forward induction literature is, in my view, particularly misplaced.[41]

7 Appendix - Proofs

The results of Section 4 are proved explicitly. To prove the same results under priority to the agreement (Remark 3), substitute

$$((S^q_{j,e})_{j\in I})^\infty_{q=0} \text{ with } ((S^q_{j,\Delta^e})_{j\in I})^\infty_{q=0},$$
$$((S^q_j)_{j\in I})^\infty_{q=0} \text{ with } (S_j)_{j\in I},$$

and see the footnotes; under priority to the path (Remark 4), substitute

$$((S^q_{j,e})_{j\in I})^\infty_{q=0} \text{ with } ((S^q_{j,e^z})_{j\in I})^\infty_{q=0},$$
$$((S^q_j)_{j\in I})^\infty_{q=0} \text{ with } (((S^q_j)_{j\in I})^Q_{q=0}, ((S^q_{j,z})_{j\in I})^\infty_{q=0}),$$

where Q is the smallest q such that $S^q = S^{q+1}$.

Throughout, let $H^\infty := H(S^\infty)$ and

$$H_\infty := \{h \notin H^\infty : p(h) \in H^\infty\}.$$

[41] Interestingly, Man [23] finds out that also the "invariance" argument, used to motivate the notions of forward induction of Kohlberg and Mertens [20] and Govindan and Wilson [15], does not imply sequential equilibrium.

For any $\mu_i \in \Delta_i^H(S_{-i})$, let

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

Proof of Proposition 1. "Only if": trivial. "If": e is credible by $\zeta(S_e^\infty) \neq \emptyset$, and $\zeta(S_e^\infty) \supseteq \zeta(S_e^\infty \cap e^0)$ is obvious; for the opposite inclusion I show that for every $s = (s_i)_{i\in I} \in S_e^\infty$, there is $s^* \in S_e^\infty \cap e^0$ such that $\zeta(s^*) = \zeta(s)$. Fix $i \in I$ and $\mu_i \in \Delta_i^e$ that s.b. $((S_{j,e}^q)_{j\neq i})_{q=0}^\infty$ and $((S_j^q)_{j\neq i})_{q=0}^\infty$ with $s_i \in \rho(\mu_i)$. By $\zeta(S_e^\infty) \subseteq \zeta(e^0)$, for each $h \in H(s_i) \cap H(S_e^\infty)$, $s_i(h) = \overline{s}_i(h)$ for some $\overline{s}_i \in e_i^0(h)$. Since the agreement is on actions, there is $\overline{s}_i \in e_i^0$ such that $\overline{s}_i(h) = s_i(h)$ for all $h \in H(s_i) \cap H(S_e^\infty)$. Fix $h \in H' := \{h' \in H(s_i)\backslash H(S_e^\infty) : p(h') \in H(S_e^\infty)\}$. Since $p(h) \in H(s_i) \cap H(S_e^\infty)$, $h \in H(\overline{s}_i) \subseteq H(e_i^0)$. Thus, since $h \in H(s_i) \subseteq H(S_{i,e}^\infty)$ and e is credible, $e_i^0 \cap S_{i,e}^\infty(h) \neq \emptyset$. Fix $s_{i,h} \in e_i^0 \cap S_{i,e}^\infty(h)$ and $\mu_{i,h} \in \Delta_i^e$ that s.b. $((S_{j,e}^q)_{j\neq i})_{q=0}^\infty$ and $((S_j^q)_{j\neq i})_{q=0}^\infty$ with $s_{i,h} \in \rho(\mu_{i,h})$. Since μ_i strongly believes $S_{-i,e}^\infty$, $\mu_i(S_{-i}(h)|p(h)) = 0$. Thus, there exists $\mu_i^* \in \Delta_i^e$ that s.b. $((S_{j,e}^q)_{j\neq i})_{q=0}^\infty$ and $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $\mu_i^*(\cdot|\widetilde{h}) = \mu_i(\cdot|\widetilde{h})$ for all $\widetilde{h} \in H(S_e^\infty)$, and $\mu_i^*(\cdot|\widetilde{h}) = \mu_{i,h}(\cdot|\widetilde{h})$ for all $h \in H'$ and $\widetilde{h} \succeq h$. So, there is $s_i^* \in \rho(\mu_i^*) \subseteq S_{i,e}^\infty$ such that $s_i^*(\widetilde{h}) = s_i(\widetilde{h}) = \overline{s}_i(\widetilde{h})$ for all $\widetilde{h} \in H(s_i) \cap H(S_e^\infty)$, and $s_i^*|h = s_{i,h}|h$ for all $h \in H'$. Since the agreement is on actions, $s_i^* \in e_i^0$, and by $H(s^*) \subseteq H(S_e^\infty)$, $\zeta(s^*) = \zeta(s)$. ■

Proof of Proposition 2. Since e is credible, $S_e^\infty \cap e^0 \neq \emptyset$. Since

$\zeta(S_e^\infty)$ is a singleton and $\zeta(S_e^\infty) \supseteq \zeta(S_e^\infty \cap e^0)$, $\zeta(S_e^\infty) = \zeta(S_e^\infty \cap e^0)$. ■

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

Proof. First, I show that $\zeta(S_e^\infty) \subseteq \zeta(e^0)$. Fix $s = (s_i)_{i \in I} \in S_e^\infty$ and $h \in H(s) \cap H(e^0)$. Since $e^0 = \times_{i \in I} e_i^0$ is Cartesian, so is $A_e^h := \{a \in A : (h, a) \in H(e^0) \cup \zeta(e^0)\}$. For each $i \in I$, since $s_i \in \rho(\Delta_i^e) \cap S_i(h)$ and $e_{-i}^0(h) \neq \emptyset$, by T3 $s_i(h) \in A_{i,e}^h$. Thus $(h, s(h)) \in H(e^0) \cup \zeta(e^0)$. By induction, $\zeta(s) \in \zeta(e^0)$.

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

Lemma 2 *Fix $i \in I$, $\overline{h} \in H^\infty$, $s_i^{\overline{h}} \in S_i^\infty|\overline{h}$, and $h \in H(s_i^{\overline{h}}) \cap H_\infty$. Then, $[s_i^{\overline{h}}]^\infty|h = S_i^\infty|h$.*[42]

Proof. Fix $s_i, s_i' \in S_i^\infty(h)$ with $s_i|\overline{h} = s_i^{\overline{h}}$. Fix μ_i, μ_i' that s.b. $((S_j^q)_{j \neq i})_{q=0}^\infty$ with $s_i \in \rho(\mu_i)$ and $s_i' \in \rho(\mu_i')$. Since μ_i strongly believes S_{-i}^∞, $\mu_i(S_{-i}(h)|p(h)) = 0$. Then, there is μ_i^* that s.b. $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that $\mu_i^*(\cdot|\widetilde{h}) = \mu_i(\cdot|\widetilde{h})$ for all $\widetilde{h} \not\succeq h$, and $\mu_i^*(\cdot|\widetilde{h}) = \mu_i'(\cdot|\widetilde{h})$ for all $\widetilde{h} \succeq h$. Thus, there is $s_i^* \in \rho(\mu_i^*) \subseteq S_i^\infty$ such that $s_i^*|h = s_i'|h$ and $s_i^*(\widetilde{h}) = s_i(\widetilde{h})$ for all $\widetilde{h} \not\succeq h$ with $\widetilde{h} \in H(s_i)$. So, $s_i^*|\overline{h} \in [s_i^{\overline{h}}]^\infty$. Hence, $s_i'|h \in [s_i^{\overline{h}}]^\infty|h$. ■

[42] This lemma and the next are not needed under priority to the agreement.

Lemma 3 *Fix a rationalizable agreement* $e = (e_i)_{i \in I}$*. For each* $i \in I$ *and* $\mu_i \in \Delta_i^e$ *that s.b.* $(S_j^\infty)_{j \neq i}$*,* $[\rho(\mu_i)]^\infty \subseteq S_{i,e}^1$*.*[43]

Proof. Fix $s_i \in [\rho(\mu_i)]^\infty \subseteq S_i^\infty$ and $\overline{s}_i \in \rho(\mu_i)$ with $\overline{s}_i(h) = s_i(h)$ for all $h \in H^\infty \cap H(s_i)$. For each $h \in H^\infty \cap H(\overline{s}_i) = H^\infty \cap H(s_i)$, by $\mu_i(S_{-i}^\infty|h) = 1$, $s_i \in S_i^\infty$, and $s_i(h) = \overline{s}_i(h)$ for all $h \in H^\infty$, also s_i is a continuation best reply to $\mu_i(\cdot|h)$. Fix μ_i' that s.b. $((S_j^q)_{j \neq i})_{q=0}^\infty$ with $s_i \in \rho(\mu_i')$. Fix $h \in H(s_i) \cap H_\infty$ and $s_{-i} = (s_j)_{j \neq i} \in S_{-i}(h)$. Fix $j \neq i$. If $s_j \notin S_j^\infty$ or $\cup_{\overline{h} \prec h} e_j^{\overline{h}}(h) = \emptyset$, let $s_j' = s_j$. Else, fix $\overline{h} \prec h$ with $e_j^{\overline{h}}(h) \neq \emptyset$. Since e is rationalizable, $e_j^{\overline{h}} = [e_j^{\overline{h}}]^\infty$, and by Lemma 2, $[e_j^{\overline{h}}]^\infty|h = S_j^\infty|h$. Thus, there is $s_j' \in S_j^\infty$ such that $s_j'|\overline{h} \in e_j^{\overline{h}}$ and $s_j'|h = s_j|h$. Let $\eta^h(s_{-i}) := (s_j')_{j \neq i}$. Since μ_i strongly believes S_{-i}^∞, $\mu_i(S_{-i}(h)|p(h)) = 0$. Then, there exists $\mu_i^* \in \Delta_i^e$ that s.b. $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that $\mu_i^*(\cdot|\widetilde{h}) = \mu_i(\cdot|\widetilde{h})$ for all $\widetilde{h} \in H^\infty$, and $\mu_i^*(s_{-i}|\widetilde{h}) = \mu_i'((\eta^h)^{-1}(s_{-i})|\widetilde{h})$ for all $h \in H(s_i) \cap H_\infty$, $\widetilde{h} \succeq h$, and $s_{-i} \in S_{-i}(\widetilde{h})$. Thus, $s_i \in \rho(\mu_i^*) \subseteq S_{i,e}^1$. ■

Proof of Proposition 3. For each $i \in I$, let $\overline{S}_i := \rho(\Delta_i^e) \cap S_i^\infty$. I show that $e^0 \subseteq S_e^\infty$; then, by T3, the result follows from Lemma 1.

By T2, $e_i^0 \subseteq \overline{S}_i$ for all $i \in I$. Now I show that $\overline{S}_i \subseteq S_{i,e}^1$.[44] Fix $s_i \in \overline{S}_i$, $\mu_i \in \Delta_i^e$, and μ_i' that s.b. $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that $s_i \in \rho(\mu_i) \cap \rho(\mu_i')$.

[43] This also implies that for rationalizable agreements, S3 can be substituted by $s_i \in S_i^\infty$ at the first step. An easy induction argument extends this fact to all steps.

[44] Under priority to the agreement, $\overline{S}_i = S_{i,e}^1$ by definition, but the construction is still needed for the second step.

Fix $h \in H^{\mu_i}$ and $s_{-i} = (s_j)_{j \neq i}$ with $\mu_i(s_{-i}|h) > 0$. Fix $j \neq i$. By T1, there is $\overline{h} \preceq h$ such that $\emptyset \neq e_j^{\overline{h}}(h) \subseteq S_j^\infty|\overline{h}$. By $\mu_i \in \Delta_i^e$, $s_j|\overline{h} \in e_j^{\overline{h}}$. If $h \in H(\overline{S}_j)$, by T2, $e_j^{\overline{h}} \subseteq \overline{S}_j|\overline{h}$. Thus, there is $s'_j \in S_j^\infty$ such that $s'_j|\overline{h} = s_j|\overline{h} \in e_j^{\overline{h}}$ and, if $h \in H(\overline{S}_j)$, $s'_j \in \overline{S}_j$. Let $\eta^h(s_{-i}) := (s'_j)_{j \neq i}$. Fix $h \in H(s_i) \cap H_\infty$[45] and $s_{-i} = (s_j)_{j \neq i} \in S_{-i}(h)$. Fix $j \neq i$. If (1) $s_j \in S_j^\infty$ and $\cup_{\overline{h} \prec h} e_j^{\overline{h}}(h) \neq \emptyset$, fix $\overline{h} \prec h$ such that $e_j^{\overline{h}}(h) \neq \emptyset$. By T1, $e_j^{\overline{h}} = [e_j^{\overline{h}}]^\infty$, and by Lemma 2, $[e_j^{\overline{h}}]^\infty|h = S_j^\infty|h$. If $h \in H(\overline{S}_j)$, by T2, $e_j^{\overline{h}} \subseteq \overline{S}_j|\overline{h}$. Thus, there is $s'_j \in S_j^\infty$ such that $s'_j|\overline{h} \in e_j^{\overline{h}}$, $s'_j|h = s_j|h$, and, if $h \in H(\overline{S}_j)$, $s'_j \in \overline{S}_j$. If (2) $s_j \in S_j^\infty$, $\cup_{\overline{h} \prec h} e_j^{\overline{h}}(h) = \emptyset$, and $s_j|h \in \overline{S}_j|h$, pick $s'_j \in \overline{S}_j$ such that $s'_j|h = s_j|h$. Else (3), let $s'_j := s_j$. Let $\eta^h(s_{-i}) := (s'_j)_{j \neq i}$. Since $h \in H(S_i^\infty) \backslash H^\infty$, $p(h) \in H^\infty$, and, by $\mu_i \in \Delta_i^e$ and T1, $\mu_i(\{s_{-i} : s_{-i}|p(h) \in S_{-i}^\infty|p(h)\}\,|p(h)) = 1$, then $\mu_i(S_{-i}(h)|p(h)) = 0$. Thus, there exists $\mu_i^* \in \Delta_i^e$ that s.b. $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that (i) $\mu_i^*(s_{-i}|\widetilde{h}) = \mu_i((\eta^h)^{-1}(s_{-i})|\widetilde{h})$ for all $\widetilde{h} \in H^{\mu_i}$ and s_{-i} with $\mu_i((\eta^h)^{-1}(s_{-i})|\widetilde{h}) > 0$, and (ii)

$$\mu_i^*(s_{-i}|\widetilde{h}) = \mu'_i((\eta^h)^{-1}(s_{-i})|\widetilde{h})$$

for all $h \in H(s_i) \cap H_\infty$, $\widetilde{h} \succeq h$, and $s_{-i} \in S_{-i}(\widetilde{h})$. Clearly, $s_i \in \rho(\mu_i^*) \subseteq S_{i,e}^1$. Obviously, $\overline{S}_i \supseteq S_{i,e}^1$. So, $\overline{S} = S_e^1$.

Fix $j \neq i$. For each $s_j \in S_{j,e}^1$, $s_j \in \rho(\mu_j)$ for some $\mu_j \in \Delta_j^e$ that s.b. $(S_k^\infty)_{k \neq j}$. By T1, e is rationalizable. So, by Lemma 3, $[S_{j,e}^1]^\infty \subseteq S_{j,e}^1$. Moreover, by $S_{j,e}^1 \subseteq S_j^\infty$, $S_{j,e}^1 \subseteq [S_{j,e}^1]^\infty$. So, $S_{j,e}^1 = [S_{j,e}^1]^\infty$.

[45] H_∞ is empty under priority to the agreement.

Thus, by $\overline{S} = S^1_e$, $\overline{S}_j = [\overline{S}_j]^\infty$. For each $h \in H_\infty \cap H(\overline{S}_j)$, since $\overline{S}_j \subseteq S^\infty_j$, by Lemma 2 $[\overline{S}_j]^\infty|h = S^\infty_j|h$. So, $\overline{S}_j|h = S^\infty_j|h$. Then, for each $s_j \in S^\infty_j \supseteq \overline{S}_j$, $s_j|h \in \overline{S}_j|h$; so, if (1) is not verified, (2) is. Then, μ^*_i strongly believes also $(\overline{S}_j)_{j\neq i} = (S^1_{j,e})_{j\neq i}$. So, $s_i \in S^2_{i,e}$. Thus, $e^0 \subseteq \overline{S} = S^1_e = S^2_e = S^\infty_e$. ■

Proof of Theorem 2. Define $\overline{S}$ like in Definition 13. I show that $e^0 = S^* \subseteq S^\infty_e$;[46] then, since Self-Enforceability implies T3, the result follows from Lemma 1. By Self-Justifiability, $S^* \subseteq \overline{S}$. By $\overline{S} \subseteq S^\infty$, $\overline{S} \subseteq [\overline{S}]^\infty$. Since e is rationalizable (by Rationalizability), by Lemma 3, $[\overline{S}]^\infty \subseteq S^1_e$. Obviously, $\overline{S} \supseteq S^1_e$. So, $S^* \subseteq \overline{S} = [\overline{S}]^\infty = S^1_e$. It remains to show that $S^1_e = S^\infty_e$.

Fix $i \in I$ and $s_i \in \overline{S}_i \subseteq S^\infty_i$. Fix μ'_i that s.b. $((S^q_j)_{j\neq i})^\infty_{q=0}$ and μ_i that s.b. $(S^*_j, \overline{S}_j, S^\infty_j)_{j\neq i}$ such that $s_i \in \rho(\mu'_i) \cap \rho(\mu_i)$ (μ_i exists by Forward Induction). Fix $h \in H(s_i) \cap H_\infty$ and $s_{-i} = (s_j)_{j\neq i} \in S_{-i}(h)$. Fix $j \neq i$. If $s_j \notin S^\infty_j$ or $h \notin H(\overline{S}_j)$, let $s'_j := s_j$. Else, $s_j|h \in S^\infty_j|h = \overline{S}_j|h$ (by $\overline{S} = [\overline{S}]^\infty$ and Lemma 2), and if $h \in H(S^*_j)$, $s_j|h \in S^\infty_j|h = S^*_j|h$ (by Rationalizability and Lemma 2). Then, there is $s'_j \in \overline{S}_j$ such that $s'_j|h = s_j|h$ and, if $h \in H(S^*_j)$, by $S^* \subseteq \overline{S}$, $s'_j \in S^*_j$. Let $\eta^h(s_{-i}) := (s'_j)_{j\neq i}$. Since μ_i strongly believes S^∞_{-i}, $\mu_i(S_{-i}(h)|p(h)) = 0$. Thus, there exists μ^*_i that s.b. $(S^*_j, \overline{S}_j)_{j\neq i} = (S^*_j, S^1_{j,e})_{j\neq i}$ and $((S^q_j)_{j\neq i})^\infty_{q=0}$ such that $\mu^*_i(\cdot|\widetilde{h}) = \mu_i(\cdot|\widetilde{h})$ for all $\widetilde{h} \in H^\infty$, and $\mu^*_i(s_{-i}|\widetilde{h}) = \mu'_i((\eta^h)^{-1}(s_{-i})|\widetilde{h})$ for all $h \in H(s_i) \cap H_\infty$, $\widetilde{h} \succeq h$,

[46] Under priority to the agreement, $S^* \subseteq \overline{S}$ by Self-Justifiability, $\overline{S} = S^1_{\Delta^e}$ by definition, and then $S^1_{\Delta^e} = S^2_{\Delta^e} = S^\infty_{\Delta^e}$ by Forward Induction.

and $s_{-i} \in S_{-i}(\widetilde{h})$. Clearly, $s_i \in \rho(\mu_i^*) \subseteq S_{i,e}^2$. Thus, $S_e^1 = S_e^2 = S_e^\infty$.
■

Proof of Proposition 4.[47] First, I show that S^* is a SES, i.e. that Rationalizable Vetos implies Rationalizability. Fix $i \in I$, $s_i \in S_i^*$, and $s_i' \in [s_i]^\infty$. For each $z \in W_i$, by $z \in \zeta(S^\infty)$, $s_i(h) = s_i'(h)$ for all $h \prec z$. Thus $s_i \notin S_i(z)$ implies $s_i' \notin S_i(z)$. So, $s_i' \in S_i^*$.

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

Proof of Proposition 5.[48] Fix $i \in I$. For each $s_i \in S_i^\infty$ and $s_{-i} \in r_{-i}(s_i)$, $\zeta(s_i, s_{-i}) \in \zeta(S^\infty)$. Then, S_i^* is the set of all $s_i \in S_i^\infty(z)$ such that $s_i \notin S_i(\widehat{z})$ for all $\widehat{z} \in \zeta(S^\infty) \setminus \{z\}$ with $u_{-i}(\widehat{z}) \geq u_{-i}(z)$. So, Rationalizable Vetos holds. Define $\overline{S}$ like in Definition 13. Fix $s_i^* \in S_i^* \cup \overline{S}_i$ and μ_i that s.b. S_{-i}^∞ with $s_i^* \in \rho(\mu_i)$. By r-strict Nash, $\overline{S} \subseteq S(z)$. Thus, there exists μ_i^* that s.b. S_{-i}^* and S_{-i}^∞ such that (i)

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

[48] Under priority to the agreement, substitute $(S_i^\infty(z))_{i \in I}$ with $(S_i^1(z))_{i \in I}$ (while still substituting $(S_i^\infty)_{i \in I}$ with $(S_i)_{i \in I}$).

$\mu_i^*(\cdot|h) = \mu_i(\cdot|h)$ for all $h \notin H(S_{-i}(z))$, and (ii) $\mu_i^*(\overline{S}_{-i}|h) = 1$ for all $h \in H(\overline{S}_{-i}) \backslash H(S_{-i}^*)$. By r-strict Nash and (i), $s_i^* \in \rho(\mu_i^*) \subseteq \overline{S}_i$. So, $S^* \subseteq \overline{S}$, i.e. Self-Justifiability holds. Thus, μ_i^* strongly believes also $\overline{S}_{-i}$. So, Forward Induction holds. R-strict Nash implies Self-Enforceability.

Fix $s \in S^*$. Let e be the reduced agreement with $e^0 = \{s\}$. Fix $i \in I$, $s_i' \in S_i^\infty(z)$, and μ_i' that s.b. $(S_{-i}^q)_{q=0}^\infty$ with $s_i' \in \rho(\mu_i')$. Fix any μ_i that s.b. $S_{-i}^\infty(z)$ and $(S_{-i}^q)_{q=0}^\infty$ such that $\mu_i(s_{-i}|h^0) = 1$ and $\mu_i(\cdot|h) = \mu_i'(\cdot|h)$ for all $h \notin H(S_{-i}(z))$. By r-strict Nash, $s_i' \in \rho(\mu_i) \subseteq S_{i,e}^1$. Fix any μ_i'' with $\mu_i''(s_{-i}|h^0) = 1$ that s.b. $(S_{-i}^q)_{q=0}^\infty$. By r-strict Nash, $\rho(\mu_i'') \subseteq S_i^\infty(z)$. So, $S_e^1 = S^\infty(z) \ni s$. Then, μ_i strongly believes $S_{-i,e}^1$. Thus, $S^\infty(z) = S_e^1 = S_e^2 = S_e^\infty$. By Proposition 2, e is self-enforcing. Thus, e implements z. ■

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

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

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

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

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

$$\begin{aligned} H' &: = \left\{h \in H^L : \cup_{\overline{h} \prec h} e_i^{n-1,\overline{h}}(p(h)) \neq \emptyset = \cup_{\overline{h} \preceq h} e_i^{n-1,\overline{h}}(h)\right\}; \\ m(h) &: = \max\left\{q \geq L : h \in H(\overline{S}_i^q)\right\}, \quad \forall h \in H'. \end{aligned}$$

For each $h \notin H'$, let $e_i^{n,h} := e_i^{n-1,h}$. Now fix $h \in H'$. If there is $\overline{h} \preceq h$ with $e_i^{\overline{h}}(h) \neq \emptyset$, let $e_i^{n,h} := ((\overline{S}_i^{m(h)}|\overline{h}) \cap e_i^{\overline{h}})|h$, which is non-empty because $\overline{S}^M \neq \emptyset$ and 1. imply the existence of $j \neq i$ and $\mu_j \in \Delta_j^e$ that s.b. $\overline{S}_i^{m(h)}$. Else, let $e_i^{n,h} := \overline{S}_i^{m(h)}|h$. Since histories in H' are unordered, e^n is an agreement. By finiteness, $e^K = e^{K+1}$ for some $K \in \mathbb{N}$. Define $\overline{e}$ as, for each $i \in I$ and $h \in H$, $\overline{e}_i^h = [e_i^{K,h}]^L$ if $h \in H^L$ and $\overline{e}_i^h = \emptyset$ else. By construction, $\overline{e}$ satisfies T1^L.

Fix $i \in I$ and let $\overline{S}_{-i} \subseteq \overline{S}_{-i}^L$ and $k = L$, or $\overline{S}_{-i} \subseteq \overline{S}_{-i}^M \cap e_{-i}^0$ and $k = M$. I show that (▲) for each μ_i that s.b. $\overline{S}_{-i}$, $\zeta(\rho(\mu_i) \times \overline{S}_{-i}) \subseteq \zeta(\overline{S}^k)$. Suppose not. Fix μ_i' that s.b. $((\overline{S}_j^q)_{j \neq i})_{q=0}^k$, with $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(\overline{S}_{-i})$. Then, $\zeta(\rho(\mu_i') \times \overline{S}_{-i}) \nsubseteq \zeta(\overline{S}^k)$ too. But by 2. for $k = L$ and by 1. for $k = M$, $\rho(\mu_i') \subseteq \overline{S}_i^k$, a contradiction.

Fix $\overline{\mu}_i$ that s.b. $\overline{e}_{-i}^0 \subseteq \overline{S}_{-i}^L$. By construction of $\overline{e}$, there exists μ_i

that s.b. $e^{K,0}_{-i} = \overline{S}^{M}_{-i} \cap e^{0}_{-i} \subseteq S^{L}_{-i}$ with $\mu_i(S_{-i}(z)|h) = \overline{\mu}_i(S_{-i}(z)|h)$ for all $h \in H^L$ and $z \in \zeta(\overline{S}^{L})$ with $z \succeq h$. By (▲), $\zeta(\rho(\overline{\mu}_i) \times \overline{e}^{0}_{-i}), \zeta(\rho(\mu_i) \times e^{K,0}_{-i}) \subseteq \zeta(\overline{S}^{L})$. Thus, $\zeta(\rho(\overline{\mu}_i) \times \overline{e}_{-i}) = \zeta(\rho(\mu_i) \times e^{K,0}_{-i})$. By (▲), $\zeta(\rho(\mu_i) \times e^{K,0}_{-i}) \subseteq \zeta(\overline{S}^{M})$. Since $\zeta(\overline{S}^{M}) = \zeta(\overline{S}^{M} \cap e^0) = \zeta(e^{K,0}) = \zeta(\overline{e}^0)$, $\overline{e}^0$ satisfies T3^L.

Finally, I will show that $\rho(\Delta^{\overline{e}}_i) \cap \overline{S}^{L}_i = [\overline{S}^{M}_i]^L$. Then, for each $\overline{h} \in H(\rho(\Delta^{\overline{e}}_i) \cap \overline{S}^{L}_i)$ with $\overline{e}^{\overline{h}}_i \neq \emptyset$, since by T1^L $\overline{h} \in H^L$, $\overline{h} \in H(\overline{S}^{M}_i)$. By construction of $\overline{e}$, $\overline{e}^{\overline{h}}_i \subseteq [\overline{S}^{M}_i|\overline{h}]^L = [\overline{S}^{M}_i]^L|\overline{h}$. So T2^L holds.

Fix $s'_i \in [\overline{S}^{M}_i]^L \subseteq \overline{S}^{L}_i$. If $L \neq 0$, by 2. there is μ'_i that s.b. $((\overline{S}^{q}_j)_{j\neq i})^{L}_{q=0}$ with $s'_i \in \rho(\mu'_i)$. Fix $s_i \in \overline{S}^{M}_i$ with $s_i(h) = s'_i(h)$ for all $h \in H^L \cap H(s_i)$. By 1., there is $\mu_i \in \Delta^{e}_i$ that s.b. $((\overline{S}^{q}_j)_{j\neq i})^{M}_{q=0}$ with $s_i \in \rho(\mu_i)$. Fix $h \in H^L$, $s_{-i} = (s_j)_{j\neq i}$ with $\mu_i(s_{-i}|h) > 0$, and $j \neq i$. By T1^L, there is h'' with $\overline{e}^{h''}_j(h) \neq \emptyset$. If there is $h' \preceq h$ with $e^{h'}_i(h) \neq \emptyset$, by $\mu_i \in \Delta^{e}_i$, $s_j|h' \in e^{h'}_j$, and by construction of $\overline{e}$, $h'' \succeq h'$. Since μ_i strongly believes $((\overline{S}^{q}_j)_{j\neq i})^{M}_{q=0}$, $s_j \in \overline{S}^{m}_j$ for all m with $\overline{S}^{m}_j(h) \neq \emptyset$. So, $s_j|h'' \in \overline{e}^{h''}_j$. Fix $h \in H(s'_i) \cap H_L$ and $s_{-i} = (s_j)_{j\neq i} \in S_{-i}(h)$. Fix $j \neq i$. If $s_j \notin \overline{S}^{L}_j$ or $\cup_{h' \prec h} \overline{e}^{h'}_j(h) = \emptyset$, let $s'_j := s_j$. Else, fix $h' \prec h$ with $\overline{e}^{h'}_j(h) \neq \emptyset$. By T1^L, $\overline{e}^{h'}_j = [\overline{e}^{h'}_j]^L$, and by Lemma 2 with L in place of ∞, $[\overline{e}^{h'}_j]^L|h = \overline{S}^{L}_j|h$. Thus, there is $s'_j \in \overline{S}^{L}_j$ such that $s'_j|h' \in \overline{e}^{h'}_i$ and $s'_j|h = s_j|h$. Let $\eta^h(s_{-i}) := (s'_j)_{j\neq i}$. Since μ_i strongly believes $\overline{S}^{L}_{-i}$, $\mu_i(S_{-i}(h)|p(h)) = 0$. Then, there exists $\mu^*_i \in \Delta^{\overline{e}}_i$ that s.b. $((\overline{S}^{q}_j)_{j\neq i})^{L}_{q=0}$ such that $\mu^*_i(\cdot|\widetilde{h}) = \mu_i(\cdot|\widetilde{h})$ for all $\widetilde{h} \in H^L$, and $\mu^*_i(s_{-i}|\widetilde{h}) = \mu'_i((\eta^h)^{-1}(s_{-i})|\widetilde{h})$ for all $h \in H'$, $\widetilde{h} \succeq h$, and $s_{-i} \in S_{-i}(\widetilde{h})$. Thus, $s'_i \in \rho(\mu^*_i) \subseteq \rho(\Delta^{\overline{e}}_i)$.

Fix $s_i \in \rho(\Delta_i^{\overline{e}}) \cap \overline{S}_i^L$ and $\mu_i \in \Delta_i^{\overline{e}}$ with $s_i \in \rho(\mu_i)$. Let

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

For each $h \in H^{L,\mu_i}$ and $s_{-i} = (s_j)_{j\neq i}$ with $\mu_i(s_{-i}|h) > 0$, by construction of $\overline{e}$, there is $\eta^h(s_{-i}) = (s'_j)_{j\neq i}$ such that, for all $j \neq i$: (i) $s'_j(h') = s_j(h')$ for all $h' \in H^L \cap H(s_j)$ with $h' \succeq h$; (ii) $s'_j \in \overline{S}_j^m$ for all $m \geq L$ with $\overline{S}_j^m(h) \neq \emptyset$; (iii) if there is $\overline{h} \preceq h$ with $e_j^{\overline{h}}(h) \neq \emptyset$, $s'_j|\overline{h} \in e_j^{\overline{h}}$. Fix any $\mu_i^* \in \Delta_i^e$ that s.b. $((\overline{S}_j^q)_{j\neq i})_{q=0}^M$ such that $\mu_i^*(s_{-i}|h) = \mu_i((\eta^h)^{-1}(s_{-i})|h)$ for all $h \in H^{L,\mu_i}$ and s_{-i} with $\mu_i((\eta^h)^{-1}(s_{-i})|h) > 0$. By (i), $\mu_i(S_{-i}(z)|h) = \mu_i^*(S_{-i}(z)|h)$ for all $h \in H^L$ and $z \in \zeta(\overline{S}^L)$ with $z \succeq h$. By ($\blacktriangle$), $\zeta(\{s_i\} \times \overline{S}_{-i}^L) \subseteq \zeta(\overline{S}^L)$. Thus, there is $s_i^* \in \rho(\mu_i^*) \subseteq \overline{S}_i^M$ (by 1.) such that $s_i^*(h) = s_i(h)$ for all $h \in H^L \cap H(s_i)$. So, with $s_i \in \overline{S}_i^L$, $s_i \in [\overline{S}_i^M]^L$. $\blacksquare$

Proof of Theorem 1. "If": it coincides with Proposition 3. "Only if": fix an implementable outcome set $P \subseteq Z$ and an agreement e with $\zeta(S_e^\infty) = \zeta(S_e^\infty \cap e^0) = P$. Apply Lemma 4 with[49] $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^L, (S_e^q)_{q=1}^K)$, where L and K are the smallest l and k such that $S^l = S^{l+1}$ and $S_e^k = S_e^{k+1}$. $\blacksquare$

Proof of Proposition 6 [7]. Fix an implementable outcome set $P \subseteq Z$ under priority to rationality [to the path], and an implementing agreement e. Since e is self-enforcing under priority to rationality

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

[to the path], I can apply Lemma 4 with $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^D, (S_e^q)_{q=1}^K)$ and $L = 0$ [with $(\overline{S}^q)_{q=0}^M = ((S^q)_{q=0}^L, (S_z^q)_{q=1}^D, (S_{e^z}^q)_{q=1}^K)$], where D and K are the smallest d and k such that $S^d = S^{d+1}$ and $S_e^k = S_e^{k+1}$ [where L, D and K are the smallest l, d, and k such that $S^l = S^{l+1}$, $S_z^d = S_z^{d+1}$, and $S_{e^z}^k = S_{e^z}^{k+1}$]. The obtained agreement $\overline{e}$ is tight under priority to the agreement [to rationality]. Thus, by Proposition 3 and Remark 3 [by Proposition 3], $\overline{e}$ implements P under priority to the agreement [to rationality]. ■

References

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

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

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

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

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

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

[7] 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.

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

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

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

[11] Catonini, E., "Rationalizability and epistemic priority orderings", working paper, 2017.

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

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

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

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

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

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

[18] Harrington, J. “A Theory of Collusion with Partial Mutual Understanding”, *Research in Economics,* forthcoming.

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

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

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

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

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

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

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

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

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

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

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

8 Supplemental Appendix

8.1 An applied example

Consider a linear city model of monopolistic competition between two firms, $i = 1, 2$.[50] Each firm i sets price p_i and, up to some prices, faces demand function

$$D_i(p_i, p_{-i}) = \begin{cases} 0 & \text{if } p_i > p_{-i} + 28 \\ 28 - p_i + p_{-i} & \text{if } p_i \in [p_{-i} - 28, p_{-i} + 28] \\ 56 & \text{if } p_i < p_{-i} - 28. \end{cases}$$

There are two production technologies: $k = 1, 2$. Technology $k = 1$ entails no fixed cost and a constant marginal cost $c^1 = 56$. Technology $k = 2$ entails a fixed cost $F = 2128$ and no marginal cost: $c^2 = 0$. Conditional on employing $k = 1, 2$, the best response function of firm i reads:

$$p_i^k(p_{-i}) = 14 + \frac{1}{2}c^k + \frac{1}{2}p_{-i}.$$

Conditional on employing $k = 1$, the unique rationalizable (hence, equilibrium) price vector is $(84, 84)$. Yet, if firm i can freely choose the technology, it is indifferent between the two technologies for $\overline{p}_{-i} = 76$,[51] and the best reply to $p_{-i} = 84$ is $p_i = 56$ with the

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

[51] Because $(p_i^1(76) - c^1) \cdot (28 - p_i^1(76) + 76) = 24^2 = 52^2 - 2128 = p_i^2(76) \cdot (28 - p_i^2(76) + 76) - F$.

use of $k = 2$. No pure equilibrium with free choice of technology exists.

Suppose now that firms play for two periods. Suppose that firms cannot upgrade from $k = 1$ to $k = 2$ between the two periods, while they can costlessly revert from $k = 2$ to $k = 1$.[52] Can firms agree on $(84, 84)$ in both periods?

Suppose that in the first period firm i employs $k = 1$ and firm $-i$ deviates to $k = 2$. Then, the best response correspondences in the second period read:

$$\begin{aligned} \widehat{p}_i(p_{-i}) &= 42 + \frac{1}{2}p_{-i}; \\ \widehat{p}_{-i}(p_i) &= \begin{cases} 42 + \frac{1}{2}p_i & \text{if } p_i < 76 \\ \left\{14 + \frac{1}{2}p_i, 42 + \frac{1}{2}p_i\right\} & \text{if } p_i = 76 \\ 14 + \frac{1}{2}p_i & \text{if } p_i > 76 \end{cases} . \end{aligned}$$

In the subgame, the set of rationalizable price vectors is $[68, 82] \times ([52, 55] \cup [76, 80])$. Each $p_i \in [68, 82]$ is a best reply to a conjecture over 52 and 80. Each $p_{-i} \in [52, 55]$ is a best reply to some $p_i \in$

[52] This can represent asymmetric switching time or cost (e.g., installation time, firing costs, if $k = 1, 2$ are interpreted as labor and capital intensive technologies, or domestic production vs FDI).

However, this assumption is merely needed for firms to be able to agree on an equilibrium price vector in the second period on path; it would be unneeded if firms were allowed to agree on mixed actions, or the game was infinitely repeated. Such extensions would be rather straightforward, but would complicate the analysis without providing any different insight.

$[76, 82]$ and each $p_{-i} \in [76, 80]$ is a best reply to some $p_i \in [68, 76]$. Each $p_i > 82$ can be best reply only to $p_{-i} > 80$, which can be best reply only to $p_i > 132$, until the highest price at which consumers buy is hit. Analogous arguments prove that all other p_1, p_2 are not rationalizable. There is a unique equilibrium where firm i sets $p_i = 76$ and firm 2 sets $p_2 = 52$ with probability $3/7$ and $p_2 = 80$ with probability $4/7$.

Note preliminarily that the path

$$z := (((1, 84), (1, 84)), ((1, 84), (1, 84)))$$

is not induced by any SPE of the game. The unilateral deviation in the first period to $k = 2$ and $p_{-i} = 56$ followed by the equilibrium of the subgame is profitable for firm $-i$;[53] the same applies to deviations to $p_{-i} = 52, ..., 60$

Suppose instead that firm i reacts to these deviations with price $p_i = 68$. Then, the deviations are not profitable. Can firm i credibly threaten to fix $p_i = 68$ after these deviation? The answer is yes. To be rigorous, assume from now on that firms can pick only integer prices. The price vectors that are prescribed by the rationalizable strategy profiles of the whole game at a pre-terminal history compatible with them must constitute a best response set. Then, after a rationalizable deviation of firm $-i$ to $k = 2$, some $p_{-i} \in [52, 55]$

[53] Because $2 \cdot (84 - c^1) \cdot 28 < p^2_{-i}(84) \cdot (28 - p^2_i(84) + 84) + p^2_{-i}(76) \cdot (28 - p^2_i(76) + 76) - 2F$.

and some $p_{-i} \in [76, 80]$ must both be possible. But then, firm i can react with $p_i = 76$. So, whenever expecting $p_i = 76$ makes the deviation profitable, firm $-i$ can fix $p_{-i} = 52$, and the best reply of firm i to $p_{-i} = 52$ is precisely $p_i = 68$. After the other rationalizable deviations, firm i can fix $p_i = 76$.

The set of rationalizable strategy profiles that induce z and where players react to rationalizable deviations with 68 if the deviation is profitable against 76 and with 76 otherwise is indeed a SES. This is straightforward to see once the existence of rationalizable strategies with these characteristics is established. Rationalizability is a simple algorithm that can be performed by a computer; nonetheless, a formal construction of the SES through the steps of Rationalizability is provided below. By Theorem 2, the agreement on the SES implements z.

Is z implementable also under priority to the path? Yes: by displaying the intention to gain a higher profit than under the path, firm $-i$ is not able to re-coordinate on a more profitable subpath with firm i, who may always react with a lower price than firm $-i$ hoped for. In particular, if the least optimistic belief of $-i$ that justifies the deviation is $\widetilde{p}_i > 76$, the best reply to the best reply to $\widetilde{p}_i$ is smaller than $\widetilde{p}_i$ itself ($p_i^1(p_{-i}^2(p_i^1)) < \widetilde{p}_i$); if $68 < \widetilde{p}_i \leq 76$, $-i$ may fix $p_{-i} = 52$, and i can react with $p_i = 68$. The construction below of the SES is valid also under priority to the path. By Remark 4, the corresponding agreement implements z under priority to the path.

Now I construct formally the SES. For brevity, I will omit the technology choice in the description of strategies. For each $i = 1, 2$, let $H_i^* := \{((1, 84), (2, p_{-i}))\}_{p_{-i}=52,\ldots,60}$, and for each $n \geq 0$, let:

$$\begin{aligned} S_i^{*,n} &: = \{s_i \in S_i^n(z) : \forall h \in H_i^*, s_i(h) = 68\}; \\ \widehat{S}_{i,h}^{n,x} &: = \{s_i \in S_i^n(h) : s_i(h) = x\}, \ h \in H_{-i}^*, x = 52, 80; \\ \overline{S}_{i,h}^n &: = \{s_i \in S_i^n(z) : s_i(h) = 76 \wedge \forall h' \in H_i^* \backslash \{h\}, s_i(h') = 68\}, \end{aligned}$$

for all $h \in H_i^*$. Fix $n \geq 0$ and suppose that all these sets are non-empty and that for each $i = 1, 2$ and $\overline{h} \in H_i^*$, there are $s_i^* \in S_i^{*,n}$ and $\overline{s}_{i,\overline{h}} \in \overline{S}_{i,\overline{h}}^n$ such that $s_i^*(h) = p_i^1(\min_{s_{-i} \in S_{-i}^n(h)} s_{-i}(h)) = \overline{s}_{i,\overline{h}}(h)$ for all $h \in H(S_i(z) \times S_{-i}^n) \backslash H_i^*$ with $h \not\prec z$. For each $\overline{h} \in H_i^*$ and $\widetilde{s}_i = s_i^*, \overline{s}_{i,\overline{h}}$, fix μ_{-i} that s.b. $(S_i^q)_{q=0}^n$ with $\mu_{-i}(\widetilde{s}_i | h^0) = 1$. For each $h = ((1, 84), (k, p_{-i})) \in H(S_i(z) \times S_{-i}^n) \backslash H_i^*$, $\rho(\mu_{-i}) \cap S_{-i}(h) = \emptyset$, otherwise $\widetilde{s}_i(h) > 84$ if $k = 1$ and $\widetilde{s}_i(h) > 76$ if $k = 2$, but then $\widetilde{s}_i(h) > p_i^1(p_{-i}^k(\widetilde{s}_i(h)))$ and $p_{-i}^k(\widetilde{s}_i(h)) = s_{-i}(h)$ for some $s_{-i} \in \rho(\mu_{-i}) \subseteq S_{-i}^n$, contradicting $\widetilde{s}_i(h) = \widehat{p}_i(\min_{s_{-i} \in S_{-i}^n(h)} s_{-i}(h))$. Thus, if $\widetilde{s}_i = \overline{s}_{i,\overline{h}}$, $\rho(\mu_{-i}) \cap \widehat{S}_{-i,\overline{h}}^{n,x} \neq \emptyset$ for $x = 52, 80$. If $\widetilde{s}_i = s_i^*$, $\rho(\mu_{-i}) \subseteq S_{-i}(z)$. Since $s_i^* \in S_i(z)$, for each $\overline{h} \in H_{-i}^*$ there is μ_{-i} that s.b. $(S_i^q)_{q=0}^n$ with $\mu_{-i}(s_i^* | h^0) = 1$ such that $\mu_{-i}(\arg\min_{s_i \in S_i^n(h)} s_i(h) | h) = 1$ for all $h \in H(S_i^n \times S_{-i}(z)) \backslash H_{-i}^*$ with $h \not\prec z$, $\mu_{-i}(\widehat{S}_{i,h}^{n,52} | h) = 1$ for all $h \in H_{-i}^* \backslash \{\overline{h}\}$, and either $\mu_{-i}(\widehat{S}_{i,h}^{n,52} | \overline{h}) = 1$, or $\mu_{-i}(\widehat{S}_{i,h}^{n,52} | \overline{h}) = 3/7$ and $\mu_{-i}(\widehat{S}_{i,h}^{n,80} | \overline{h}) = 4/7$. In the first case, $\rho(\mu_{-i}) \cap S_{-i}^{*,n} \neq \emptyset$, in the second case $\rho(\mu_{-i}) \cap \overline{S}_{-i,\overline{h}}^n \neq \emptyset$.

Then, by the observation above about pre-terminal histories,

$$\times_{i=1,2}\left\{s_i \in S_i^{*,\infty} : \forall h \in H(S_i^{\infty}(z) \times S_{-i}^{\infty} \backslash S_{-i}(z)) \backslash H_i^*, s_i(h) = 76\right\}$$

is non-empty too, and it is clearly a SES.[54]

All the employed μ_i strongly believe $S_{-i}(z)$. Thus, the procedure can be prolonged to obtain a SES $S^* \subseteq S_z^{\infty}$.

8.2 Games

Formalization of Example 3.

$$2\times \begin{array}{|c|c|c|}\hline A\backslash B & W & F \\ \hline W & 2,2 & 1,3 \\ \hline F & 3,1 & 0,0 \\ \hline \end{array}$$

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

[54] A formal proof would follow the line of the proof of Proposition 5.

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

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

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

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

Example 4. Consider the following game.

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

$\longrightarrow$

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

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

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

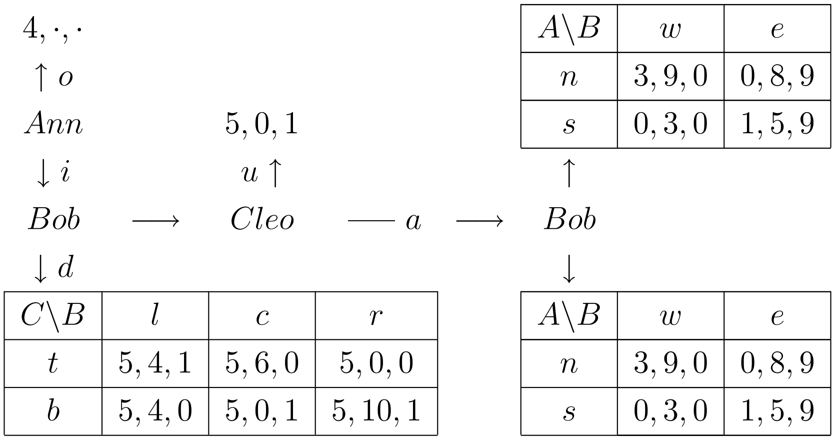
Players want instead to implement (N, W). Hence they reach the reduced agreement with $e_A^0 = \{N.U, N.M\}$ and $e_B^0 = \{W\}$. The agreement is self-enforcing: $S_e^1 = \{N.U, N.M, N.D\} \times \{W\}$, thus $S_e^\infty = S_e^1 = S((N, W))$. Also, the agreement is self-enforcing under priority to the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

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

Finally, consider the following, non-reduced, agreement: $e_A^0 = \{S\}, e_A^{(N,E)} = \{D\}, e_B^0 = S_B$. It implements (S, E): $S_e^1 = S_A \times$

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

Example 5. Consider the following game.



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

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

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

All strategies are rationalizable.

Let us search for the SES's of the game. Beside SES's where Bob may play d or not, which induce all possible payoff vectors,[55]

[55] If Bob may play d or not, then (i) Ann's Self-Enforceability requires her to play i and Bob's Self-Justifiability requires Cleo to play $t.a$, (ii) Cleo's Self-

Self-Enforceability and Self-Justifiability identify as candidate SES's all sets $S_A^* \times S_B^* \times S_C^*$ where:

$$S_B^* = \{d.r\}\,, \quad S_C^* \subseteq \{b.u, b.a\}\,, \quad o \notin S_A^*;$$

$$S_B^* = \{d.l, d.c, d.r\}\,, \quad \begin{array}{ll} b.a \notin S_C^* \supseteq \{t.a, b.u\}\,, & S_A^* = \{i.s.s\} \\ t.a \notin S_C^* \supseteq \{b.a, t.u\}\,, & o \notin S_A^* \end{array} ;$$

$$S_B^* \cap \{d.l, d.c, d.r\} = \emptyset,^{56} \quad S_C^* = \{t.a\}\,, \quad S_A^* = \{o\}\,.$$

Let us verify Forward Induction. All candidate SES's in the first and in the second group satisfy Forward Induction: there is no history where a player is active which is rationally reached by the player under strong belief in the SES but not under the SES itself. Let us consider now the third group. Under strong belief in the SES, Bob may play $d.c$ but not $d.l$. Then, Forward Induction requires $t.a$ to be rational against the belief that Bob at $(i.d)$ will play c, but this is not the case.

Therefore, outcome (o) is not induced by any SES. Are there restrictions to Ann's behavior after her deviation to i that transform

Justifiability requires Bob to play $d.l$, (iii) Bob's Self-Justifiability requires Cleo to play b and u, (iv) Bob's Self-Enforceability requires him to play $d.r$, Ann's Self-Enforceability requires her to play o, and Cleo's Self-Justifiability requires Bob to play w and e in one of the two subgames, (v) Ann's Self-Enforceability requires her to play n and s in that subgame and Bob's Self-Enforceability requires him to play $d.c$ (and e in the other subgame).

[56] And Cleo's Self-Justifiability further requires Bob to possibly play e in a subgame he reaches; however, this is immaterial for the discussion.

some candidate SES in the third group into a tight agreement? The answer is yes:

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

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

Note that the agreement is not on actions: Ann promises to play n in one of the two subgames, but she does not say in which one. Is there an agreement on actions that implements o? No. For Ann to select o, Bob and Cleo must exclude from the agreement, or eliminate through strategic reasoning, d and u. If u is excluded or eliminated, Bob expects a payoff of at least 5 by not playing d. Thus, Bob will eliminate $d.l$. If Bob still considers $d.c$ or $d.r$ when $d.l$ is eliminated, Cleo will best reply with b. But then Bob will select $d.r$, and o is not implemented. So, the agreement must make sure that Bob eliminates $d.c$ and $d.r$ no later than $d.l$.[57] For the elimination of $d.r$, it is necessary that Cleo excludes b from the agreement. Then Bob is confident that by playing $d.c$ he can get 6. So, for Bob to eliminate $d.c$, he must be confident of getting a higher payoff without playing d. So, he must be confident that in at least one of the two subgames, Ann will not play s. If this subgame was pinned down by the agreement or strategic reasoning, then Bob would play w in the subgame he moves to. Then, Cleo would select u, and o would not be implemented. Hence, Ann, through the agreement or strategic reasoning, does have to exclude planning s in both subgames, but at the same time she must not reveal in which subgame she is not planning s. In this game, she can do this only through the agreement: if she rationally plays i, she hopes for d or u, and if d and u are not played, she could plan s in both subgames. Thus, **agreements that are not on actions can be needed to implement an outcome**.

The tight agreement above is clearly equivalent to the following reduced agreement: $\overline{e}^0_A = \{o, i.n.n, i.n.s, i.s.n\}$, $\overline{e}^0_B = e^0_B$, $\overline{e}^0_C = e^0_C$. Thus, $\overline{e}$ is self-enforcing (but not truthful) and it implements (o). So, one may wonder whether reduced agreements suffice to implement all implementable outcomes. The answer is no. Imagine that at the

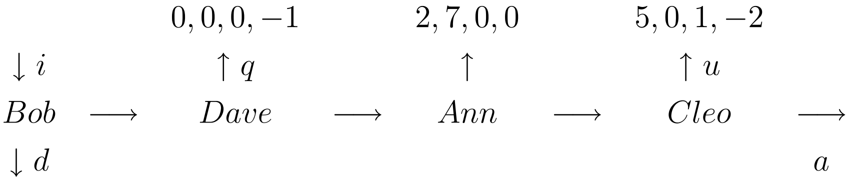
[57] And excluding $d.c$ or $d.r$ from the agreement is not viable if it survives longer than $d.l$, because it would bring to the empty set.

initial history, Ann plays simultaneously with Bob in the following way:[58]

$A\backslash B$	o	i
o	$4, 4, \cdot$	Γ'
i	Γ	$3, 3, \cdot$

where Γ is the game above and Γ' is the game above with roles and payoffs of Ann and Bob inverted. Then, for Ann and Bob to coordinate on (o, o), at least one of the two has to declare o (and then exclude $s.s$ after a deviation to i as above). Thus, **non-reduced agreements can be needed to implement an outcome.**

Back to the original game, add now the following component:



Let Dave get payoff 0 elsewhere. All strategies are still rationalizable. Consider the following reduced agreement on actions.

$$e_A^0 = \{o\}\,,\ e_B^0 = S_B \backslash \{d.c, d.r\}\,,\ e_C^0 = \{t.a, b.a\}\,,\ e_D^0 = S_D.$$

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

The agreement is truthful. At the first step, Ann eliminates the strategy that prescribes s in both subgames, Cleo eliminates $b.u$ and $b.a$, and Dave eliminates q. At the second step, Bob eliminates all strategies that prescribe d. At the third step, Ann selects o.

Despite the existence of a truthful reduced agreement on actions that implements (o), (o) is not induced by any SES. Ann's Self-Enforceability requires Bob not to play d and Cleo not to play u; thus, Bob's Self-Justifiability requires Cleo to play t and Dave's Self-Justifiability requires him not to play q; but then, believing all this, Bob can rationally play $d.c$ but not $d.l$, thus Cleo's Forward Induction is violated. This shows that **the reverse of Theorem 2 does not hold**.

Also, (o) is not induced by any tight agreement on actions. Ann's T3 requires Bob not to play d and Cleo not to play u; thus, Dave's T2 requires him not to play q; then, Bob's T2 requires Cleo to play t; thus, Cleo's T2 requires Bob to play l at $(i.d)$, which, since Bob expect at least 5 by not playing d, is compatible with Bob's T2 only if Bob does not rationally play d when he believes in the agreement. This can be accomplished in tight agreement only if Ann excludes playing $s.s$, for the same argument as in the previous game. Therefore, if one restricts the analysis to agreements on actions, a full characterization of implementable outcomes in the fashion of Theorem 1 cannot be made; i.e., **not all outcomes that are implemented by an agreement on actions are prescribed by a**

tight agreement on actions.

8.3 Equilibrium paths that can be upset by a convincing deviation

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

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

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

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

Example 3 provides two paths that can be upset by a convincing deviation,[59] although the agreements on the SPE that induce them are self-enforcing.

Proposition 8 *Let $\overline{z} = (\overline{a}^1, ..., \overline{a}^T)$ be a path that can be upset by a convincing deviation. The path agreement on $\overline{z}$ is not credible.*

Proof. Let $\widehat{h} := (\overline{a}^1, .., (\widehat{a}_i, \overline{a}_j^\tau))$ and $z := (\overline{a}^1, .., (\widehat{a}_i, \overline{a}_j^\tau), b^i, ..., b^i)$. Suppose that $S_e^1(\overline{z}) \neq \emptyset$, otherwise $S_e^2 = \emptyset$. Then, for each $k = i, j$, there exists $\overline{\mu}_k$ that s.b. $(S_{-k}^q)_{q=0}^\infty$ and $S_{-k}(\overline{z})$ such that $\rho(\overline{\mu}_k) \cap S_k(\overline{z}) \neq \emptyset$.

Fix $n \in \mathbb{N}$ and suppose that $S_i^{n-1}(z) \neq \emptyset$. Fix $s_j \in S_j$ with $\overline{\mu}_i(s_j|h^0) \neq 0$ and μ_j' that s.b. $(S_i^q)_{q=0}^\infty$ with $s_j \in \rho(\mu_j')$. Since $\overline{\mu}_j$ strongly believes $S_i(\overline{z})$, for each $h \notin H(S_i(\overline{z}))$ with $p(h) \prec \overline{z}$, $\overline{\mu}_j(S_i(h)|p(h)) = 0$. Thus, there exists μ_j t.s.b $(S_i^q)_{q=0}^{n-1}$ such that (i) $\mu_j(S_i(z)|\widehat{h}) = 1$, (ii) $\mu_j(\cdot|h) = \mu_j'(\cdot|h)$ for all $h \notin H(S_i(\overline{z}))$ with $p(h) \prec \overline{z}$ and $h \neq \widehat{h}$, and (iii) $\mu_j(\cdot|h) = \overline{\mu}_j(\cdot|h)$ for all $h \in H(S_i(\overline{z}))$. Then, there exists $\eta(s_j) \in \rho(\mu_j) \subseteq S_j^n$ such that by (iii) $s_j \in S_j(\overline{z}) \subseteq S_j(\widehat{h})$, by (i) and (J) $\eta(s_j) \in S_j(z)$, and by (ii) $\eta(s_j)|h = s_j|h$ for all $h \notin H(S_i(\overline{z}))$ with $p(h) \prec \overline{z}$ and $h \neq \widehat{h}$. Since $s_j \in S_j(\overline{z})$, $\eta(s_j)(h) = (s_j)(h)$ for all $h \in H(s_j)$ with $h \not\succeq \widehat{h}$. Construct μ_i that s.b. $(S_j^q)_{q=0}^n$ and $S_j(\overline{z})$ such that $\mu_i(s_j|h^0) = \overline{\mu}_i(\eta^{-1}(s_j)|h^0)$ for all s_j with $\overline{\mu}_i(\eta^{-1}(s_j)|h^0) \neq 0$. For each $\widetilde{z} \not\succeq \widehat{h}$, $\mu_i(S_j(\widetilde{z})|h^0) =$

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

$\overline{\mu}_i(S_j(\widetilde{z})|h^0)$, while $\mu_i(S_j(z)|h^0) = 1$. Thus, by $\rho(\overline{\mu}_i) \cap S_i(\overline{z}) \neq \emptyset$ and (I), $\emptyset \neq \rho(\mu_i) \cap S_i(z) \subseteq S_i^{n+1}(z)$. So, by induction, there exists μ_i that s.b. $(S_j^q)_{q=0}^{\infty}$ and $S_j(\overline{z})$ such that $\emptyset \neq \rho(\mu_i) \cap S_i(z) \subseteq S_{i,e}^1(z)$.

For each $\mu_i \in \Delta_i^e$, $\mu_i(S_j(\overline{z})|h^0) = 1$. So, for each $s_i \in S_i(\widehat{h}) \backslash S_i(z)$, by (I) $s_i \notin S_{i,e}^1$. Thus, for every $\mu_j \in \Delta_j^e$ that s.b. $S_{i,e}^1$, $\mu_j(S_i(z)|\widehat{h}) = 1$. So, by (J), $S_{j,e}^2(\widehat{h}) = S_{j,e}^2(z)$. Since $S_j(\overline{z}) \subseteq S_j(\widehat{h})$, for every $\mu_i \in \Delta_i^e$ that s.b. $S_{j,e}^2$, $\mu_i(S_j(z)|h^0) = 1$, so by (I) $\rho(\mu_i)(\overline{z}) = \emptyset$. Hence $S_{i,\Delta^e}^3(\overline{z}) = \emptyset$. So, $S_{j,\Delta^e}^4 = \emptyset$. ■

Катонини, Э.

Самоподдерживающиеся соглашения и принцип прямой индукции [Электронный ресурс] : препринт WP9/2018/01 / Э. Катонини ; Нац. исслед. ун-т «Высшая школа экономики». – Электрон. текст. дан. (500 Кб). – М. : Изд. дом Высшей школы экономики, 2018. – (Серия WP9 «Исследования по экономике и финансам»). – 84 с.

В динамических играх игроки могут сталкиваться с отклонениями от заранее заключенных, иногда неполных, но в любом случае не связывающих соглашений. Попытки объяснить подобные отклонения могут привести к тому, что игроки меняют свои убеждения о поведении партнера, отклонившегося от соглашения. Такие применения прямой индукции основаны не только на предположениях о рациональности, но и на предположениях о соответствии соглашению как таковому. В своей работе я изучаю влияние подобного рода рационализации на самоподдерживающиеся соглашения, при которых некоторые исходы игры могут быть осуществимы, а некоторые – нет. Результаты моего исследования существенно отличаются от результатов, полученных для традиционных усилений равновесий (equilibrium refinements). В частности, самоподдерживающиеся соглашения могут привести к исходам, не соответствующим равновесиям, совершенным по подыграм, а равновесия, совершенные по подыграм, могут не быть самоподдерживающимися. Подобная неполнота соглашений может играть решающую роль при достижении конкретных исходов. Конкретный способ рационализации нарушений соглашения позволяет также установить их связь с понятием стратегической стабильности (Кольберг и Мертенс, 1986).

Ключевые слова: соглашения, самоподдерживание, прямая индукция, рационализация игры в развернутой форме, стратегическая стабильность

Препринт WP9/2018/01
Серия WP9
Исследования по экономике и финансам

Катонини Эмилиано

Самоподдерживающиеся соглашения и принцип прямой индукции

(*на английском языке*)

Изд. № 2065