Self-enforcing agreements and forward induction reasoning*

Emiliano Catonini†

June 1, 2020

Abstract

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

*A special thanks goes to Pierpaolo Battigalli, this paper would not exist without his mentoring. Thank you also to two anonymous referees, Alexei Belanin, Davide Bordoli, Adam Brandenburger, Shurojit Chatterji, Yi-Chun Chen, Carlo Cusumano, Nicodemo De Vito, Alfredo Di Tillio, Andrew Ellis, Francesco Fabbri, Amanda Friedenberg, Giacomo Lanzani, Gilat Levy, Matthew Levy, Julien Manili, Francesco Nava, Andres Perea, Burkhard Schipper, Madhav Shrihari Aney, Balazs Szentes, Satoru Takahashi, Elias Tsakas, and to all the attendants of the Stratemotions workshop (Università Bocconi, December 2017), of the ESEM-EEA conference 2017, of the Barcelona GSE Workshop (June 2018), and of the seminars at NUS, SMU, Oxford University, Maastricht University, and London School of Economics.

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

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

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

1 Introduction

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

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

Relevant economic decisions can seldom be interpreted as unintentional mistakes. A deviation from an equilibrium path can safely be interpreted as disbelief in some features of the equilibrium. Can we expect the deviator to

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

best reply to threats that are meant precisely to deter the deviation? Often, the deviation clearly displays confidence that none of the adverse re-coordination scenarios will realize. Then, credible threats are not the ones that rely on illusory re-coordination, but those that best respond to the potentially profitable continuation plans of the deviator. Indeed, compliance with non-binding agreements often relies on the threat/concern that a deviation will provoke the end of coordinated play, rather than less advantageous re-coordination. Moreover, agreements are often incomplete: differently than SPE, they do not pin down exactly what to do in every contingency. Partially conflicting interests, legal constraints, social taboos, unilateral communication channels, anticipated distrust or objective impossibility of credible (re-)coordination: these are some of the reasons why players may be unable or unwilling to reach a complete agreement.

In economic applications, absence of an intuitive SPE solution is often blamed on a misspecification of the model, rather than on the objective impossibility to reach a plausible agreement for every contingency. A classical example is the two-stage Hotelling model with linear transportation cost, which has no SPE in pure strategies. A quadratic transportation cost has been introduced by D'Aspremont et al. [19] to obtain a pure SPE where firms, contrary to Hotelling's conjecture, locate at the extremes of the spectrum (i.e., at 0 and 1). In the original model, Osborne and Pitchik [30] find a SPE in mixed strategies with numerical methods. In this solution, counterintuitively, firms locate at a distance that may induce a "price war", whereas a slightly larger distance would prevent this possibility. In a separate paper [16], I show that the transportation-efficient location pair $(1/4, 3/4)$ is the unique symmetric pair that is induced by a self-enforcing agreement between the two firms. This recovers the intuitive prediction that firms will locate at the smallest distance that prevents a price war. In Section 6, I replicate this result in a discretized version of the model.

To illustrate the main insights in a simpler but meaningful economic environment, in Section 2 I analyze an entry game in monopolistic competition. Depending on the value of the entry cost, SPE turns out to be too permis-

sive, too restrictive, or simply inadequate to evaluate the credibility of the incumbent's threats.

That SPE can be too permissive is not a new observation. Classical examples, such as the battle of the sexes with an outside option (Ben Porath and Dekel [13]), have already shown this point. This paper captures these refinement arguments in a simple and general way.

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

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

$\longrightarrow$

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

In the first stage, Ann and Bob can potentially coordinate on two outcomes, (N,W) and (S,E). If they fail to coordinate, the game either ends (after (S,W)),[2] or moves to a second stage (after (N,E)), where in the unique equilibrium all actions are played with equal probability. So, the unique SPE of the game induces outcome (S,E). But (S,E) is Pareto-dominated by (N,W), hence Ann and Bob would rather coordinate on (N,W). To do so, they agree that Bob should play W, and that Ann should play N in the first stage and U in case Bob deviates to E.[3] Is the agreement credible? If Bob is rational[4] and believes in the agreement, he has no incentive to deviate. Then, after a deviation to E, Ann cannot believe at the same time that Bob is rational and believes in the agreement. If she keeps the belief that Bob is rational, she has to drop the belief that Bob believes in the agreement. So, she can think that Bob gives higher probability to D than to U. Then, she expects Bob to play L and best replies with U. Anticipating this, Bob can believe in U and refrain from deviating. The conclusion is that the agreement is *credible*

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

[3] To keep the game small, the threat U that sustains (N,W) is played with positive probability also in the SPE. This is by no means necessary for its credibility, even when the SPE is unique: see the variation of this example in Supplemental Appendix I.

[4] The notion of rationality employed in this paper simply requires expected utility maximization, without imposing by itself any restriction on beliefs. See Section 3.2 for details.

and, once believed, players will comply with it. Therefore, the agreement is *self-enforcing*.

The further inadequacy of SPE comes from the intrinsic assumption of agreement completeness. In the entry game of Section 2, for intermediate values of the entry cost, the most realistic threat by the incumbent does not completely specify its plan, therefore its credibility cannot be evaluated through SPE. Moreover, the *complete* agreement on the SPE that deters entry is not credible, because the continuation plan of the entrant is not part of any rational entry plan. Yet, the SPE threat is credible, and its credibility relies on the (physiological) uncertainty regarding the behavior of the entrant. I will show in Section 5.2 that, sometimes, an outcome can be achieved only by not fully specifying the reactions to deviations either.

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

In Section 4, I study credibility and self-enforceability of agreements starting from primitive assumptions about players' strategic reasoning. An agreement is *credible* when players may comply with it in case they are rational, they believe in the agreement, they believe that the co-players are rational and believe in the agreement, and so on. When a player's move is not rational under belief in the agreement (such as Bob's deviation to E in the example above), I assume that the co-players keep the belief that the player is rational

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

and drop the belief that the player believes in the agreement.[6] Under this reasoning scheme, deviations, or more generally past actions, are not interpreted as mistakes but as intentional choices. To visualize this, suppose that in the game above Ann and Bob agree on (S, E), without specifying what to do in case Ann deviates. If Ann believes in E, she has the incentive to deviate to N only if she expects R with sufficiently high probability. Then, Bob will expect her to play D after the deviation.[7] This instance of forward induction reasoning is based not just on the belief that Ann is rational, but also on the belief that she believes in the agreement.

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

A set of outcomes is *implementable* when it is induced by a self-enforcing agreement. In Section 5 I provide necessary and sufficient conditions for implementability. A set of outcomes is implementable if it is prescribed by a *Self-Enforcing Set* of plans (henceforth, SES). SES's are self-enforcing agreements that do not require players to promise, and co-players trust, what they would do after a own deviation. Thus, they can be seen as a set-valued counterpart of SPE where the behavior of deviators is not exogenously given but determined by forward induction. In games with two players or two stages, every implementable outcome set is prescribed by a SES. In a two-player game, the SES's that induce a single outcome boil down to (sets of) Nash equilibria in extensive-form rationalizable plans with strict incentives.[8] To complete the search for implementable outcomes in games with more than two players and stages, *tight agreements* augment SES's by restricting the behavior of de-

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

[7] See Section 4.2 for the complete analysis of this and other agreements for this game.

[8] Every feature of this simple characterization is not assumed, but derived from first principles. The original notion of extensive-form-rationalizability is due to Pearce [31] and was further analyzed and clarified by Battigalli [6].

viators. An outcome set is implementable if and only if it is prescribed by a tight agreement. Since tight agreements induce exactly the outcomes they prescribe, we have a "revelation principle" for agreements design: players need not be vague about the set of outcomes they want to achieve.

Tight agreements and SES's have the double value of *solution concepts* and "soft mechanisms" for implementation,[9] because they specify the outcome set they induce. Thus, they provide to the analyst (or a mediator) all possible predictions (and an implementation strategy) under the non-binding agreements motivation, abstracting away from the foundations of self-enforceability. In particular, after a standard elimination procedure (extensive-form rationalizability), they only require to verify one-step conditions instead of doing all steps of reasoning under all candidate self-enforcing agreements.

This work is greatly indebted to the literature on rationalizability in dynamic games. In this literature, restrictions to first-order beliefs are usually accounted for through *Strong-Δ-Rationalizability* (Battigalli, [7]; Battigalli and Siniscalchi, [11]). Strong-Δ-Rationalizability does *not* require players to keep believing in the rationality of a co-player who displays behavior that is not optimal under her first-order belief restrictions. To capture the opposite hypothesis, in the companion paper I construct and analyze another elimination procedure with belief restrictions, *Selective Rationalizability*. Selective Rationalizability captures *common strong belief in rationality* (Battigalli and Siniscalchi [10]), i.e., the hypothesis that each order of belief in rationality holds as long as not contradicted by the observed behavior. Thus, it combines "unrestricted" (i.e., based only on beliefs in rationality) and "restricted" (i.e., based also on first-order belief restrictions) forward induction reasoning. In Section 3.3, I specialize Selective Rationalizability for the analysis of agreements, and I call it *Agreement-rationalizability*. The priority given to the beliefs in rationality, the structure given by agreements to the belief restrictions, and the requirement of self-enforceability greatly increase the predictive power with respect to the literature. In the Section 7.1, I expand on this comparison.

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

When the agreement prescribes a specific outcome, a possible way to interpret a deviation is that the deviator believed in the agreed-upon path (i.e., that the co-players would have complied with it), but does not believe in the threat. In Section 7.2, I provide an example where imposing this particular rationalization of deviations matters, and I argue that a simple revision of the methodology accommodates it. All the general insights of the paper are robust to these stricter strategic reasoning hypotheses, which further increase the refinement power.

Strategic stability à la Kohlberg and Mertens [27] and related refinements are often justified with stories of forward induction reasoning where deviators are believed to aim for a higher payoff than under the equilibrium path. However, understanding and applying stability and related refinements present various difficulties. Stability is hard to interpret and verify, and does not offer an implementation strategy: what should players exactly agree on/believe in?[10] Later refinements focus exclusively on sequential equilibrium and, to simplify the analysis, sacrifice depth of reasoning (e.g., forward induction equilibria of Govindan and Wilson [22] capture only strong belief in rationality[11]) or scope (e.g., the intuitive criterion of Cho and Kreps [18] and divine equilibrium of Banks and Sobel [3] are tailored on signaling games). Moreover, the equilibrium language does not allow to talk of incomplete agreements. Then, the methodology of this paper can also be seen as a general and transparent approach to the problems analyzed in this literature. It turns out that the spirit of subgame perfection, i.e., the idea that a deviator will best reply to the threat, is at odds precisely with the rationalization of deviations based on the belief in the path.

The Appendix collects the proofs omitted from Sections 5 and 6. The Supplemental Appendix formalizes the claims of Section 7 and contains further examples and technical remarks that can be useful to whoever wishes to develop (as opposed to just apply) the methodology.

[10] An interesting critique of this kind to stability was put forward by Van Damme [35].

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

2 An entry game

Consider the following linear city model of monopolistic competition. Two firms, $i = 1, 2$, sell the same good at the extremes of a continuum of potential buyers of measure 96. Each individual $j \in [0, 96]$ either does not buy, or buys one unit from the firm i that maximizes $u_{ji} = 120 - p_i - t \cdot d_{ji}$, where p_i is the price fixed by firm i, $t = 1/2$ is the transportation cost, and d_{ji} is the distance from firm i: $d_{j1} = j$ and $d_{j2} = 96 - j$. There are two production technologies: $k = A$, with marginal cost $mc = 48$ and no fixed cost; and $k = B$, with no marginal cost and fixed cost $F = 48^2$. Firms choose price and technology simultaneously.

Prices below 48 and above 96, with either technology, do not best reply to any conjecture about the competitor's price (see Supplemental Appendix II for the proof). For each $(p_i, p_{-i}) \in [48, 96]^2$, firm i's demand is

$$D_i(p_i, p_{-i}) = 48 - p_i + p_{-i},$$

and the best reply correspondence reads

$$\widehat{p}_i(p_{-i}) = \begin{cases} 48 + \frac{1}{2}p_{-i} & (\text{with } k = A) & \text{if } p_{-i} \in [48, 72) \\ \{60, 84\} & (\text{with } k = B, A) & \text{if } p_{-i} = 72 \\ 24 + \frac{1}{2}p_{-i} & (\text{with } k = B) & \text{if } p_{-i} \in (72, 96] \end{cases}.$$

Since demand is linear in p_{-i}, the best replies to a conjecture $\mu \in \Delta([48, 96])$ are $\widehat{p}_i(\mathbb{E}_\mu(p_{-i}))$. Then, only the price-technology pairs $[60, 72] \times \{B\}$ and $[72, 84] \times \{A\}$ can be best replies. In turn, only the pairs $[60, 66] \times \{B\}$ and $[78, 84] \times \{A\}$ best reply to some $\mu \in \Delta([60, 84])$. These are the rationalizable price-technology pairs in the static market game. The pure equilibrium prices are $(64, 80)$ and $(80, 64)$, and the only mixed equilibrium assigns probability $1/2$ to 60 and 84 for both firms. Let $\overline{\overline{\pi}} > \overline{\pi} > \underline{\pi}$ denote the (expected) profit of firm 2 in the $(80, 64)$, in the mixed, and in the $(64, 80)$ equilibrium. Let $\underline{\underline{\pi}}$ denote the optimal profit of firm 2 when $p_1 = 60$.

Suppose now that firm 1 is already in the market, while firm 2 still has to pay an entry cost E. If firm 2 does not enter, its profit is 0. Can firm 1 deter the entry of firm 2 by announcing how it plans to react?[12] I am going to tackle this question for different values of the entry cost. To avoid repeating uninteresting steps of reasoning, assume directly that firms cannot fix prices below 48 or above 96. The analysis is formalized in Section 4.2.

Case 1: $\overline{\pi} < E < \overline{\overline{\pi}}$ (SPE is too permissive). According to SPE, entry is deterred by two equilibria of the subgame. However, a rational firm 2 will enter only under the belief that, in expected terms, p_1 will be at least $\underline{p}$ for $\underline{p} \in (72, 80)$ that depends on E. Moreover, if firm 2 believes that firm 1 is rational, it will expect $p_1 \leq 84$. Then, in case of entry, firm 2 will fix $p_2 \in \left[24 + \frac{1}{2}\underline{p}, 66\right]$ (with $k = B$). Realizing this, firm 1 will react with $p_1 \in \left[60 + \frac{1}{4}\underline{p}, 57\right]$ (with $k = A$), thus $p_1 > \underline{p}$. So, firm 2 will enter. Firm 1 cannot credibly deter entry because any explicit threat would clash with strategic reasoning about rationality.

Case 2: $\underline{\pi} < E < \overline{\pi}$ (agreement incompleteness). According to SPE, entry is deterred by equilibrium $(64, 80)$. But believing in $p_2 = 80$ is incompatible with strategic reasoning. If firm 2 is rational and believes that firm 1 is rational, it will enter only when it expects $p_1 \in \left[\underline{p}, 84\right]$ with $\underline{p} \in (64, 72)$ that depends on E. Then, firm 2 will fix either $p_2 \in \left[48 + \frac{1}{2}\underline{p}, 84\right]$ with $k = A$, or $p_2 \in [60, 66]$ with $k = B$, thus not $p_2 = 80$. However, every $p_1 \in [60, 66]$ (with $k = B$) and every $p_1 \in [78, 84]$ (with $k = A$) are best replies to beliefs over these entry plans of firm 2. Hence, prices cannot be refined further with strategic reasoning about rationality. The SPE price $p_1 = 64$ is compatible with strategic reasoning, but it is justified by uncertainty over values of p_2

[12] For the purpose of the example, the incumbent has no commitment power or switching costs. Instead, Dixit [19] studies entry deterrance through an *irreversible* investment in productive capacity. Interestingly, Dixit motivates his analysis with the following observations: "The theory of large-scale entry into an industry is made complicated by its game-theoretic aspects. Even in the simplest case of one established firm facing one prospective entrant, there are subtle strategic interactions. [...] In reality, there may be no agreement about the rules of the post-entry duopoly, and there may be periods of disequilibrium before any order is established."

that do not best respond to it. So, to deter entry, it must be formulated as a unilateral threat and not as part of a complete agreement.

Furthermore, firm 1 does not actually need to specify p_1: it is enough to announce the use of technology $k = B$. If firm 2 believes that firm 1 (i) is rational, (ii) believes that firm 2 is rational, *and* (iii) adopts $k = B$, it expects firm 1 to fix $p_1 \in [60, 66]$. If $\underline{p} > 66$, this is sufficient to deter entry. If $\underline{p} \in (64, 66]$, firm 2 may enter and fix $p_2 \in \left[48 + \frac{1}{2}\underline{p}, 81\right]$ with $k = A$. But then, realizing this, firm 1 will react with $p_1 \in \left[48 + \frac{1}{4}\underline{p}, 64.5\right]$ and $k = B$. This realization is based not just on the beliefs in rationality, but also on the belief that firm 2 believes in firm 1's announcement, which is not at odds with rational entry. If needed, further steps of reasoning eventually bring the highest possible p_1 below $\underline{p}$. Hence, the announcement of $k = B$ by the incumbent is credible and deters entry. Such a coarse announcement can have real-life advantages; for instance it may be illegal to state future prices.[13]

Case 3: $\underline{\pi} < E < \underline{\pi}$ (SPE is too restrictive). Now firm 2 enters in every SPE. But, like in Case 2, there is $\underline{p} \in (60, 64]$ such that every $p_2 \in [60, 66] \cup \left[48 + \frac{1}{2}\underline{p}, 84\right]$ is compatible with strategic reasoning, and then every $p_1 \in [60, 66] \cup [78, 84]$ is compatible as well. So, firm 1 can credibly threaten to fix $p_1 \in \left[60, \underline{p}\right)$ and deter entry.[14] The arguments for the credibility of this threat are identical to the arguments for the SPE threat in Case 2. For instance, $p_1 = 60$ is justified by a uniform distribution over the three equilibrium (expected) prices, which are now all compatible with strategic reasoning.

[13] Harrington [26] documents instances of "mutual partial understanding" among firms which leaves the exact path of price increase undetermined to escape antitrust sanctions. Such mutual understanding can be modeled as an incomplete agreement, whose consequences can be studied with the methodology developed in this paper.

[14] One could argue that alternated best responses from p_1 would lead to the $(64, 80)$ equilibrium in the long run. If firms are impatient, this is irrelevant. If firms are patient, firm 2 could try to upset this trajectory by switching to $k = 2$ at any time. The choice of $p_1 = 60$ is justified precisely by this uncertainty.

3 Agreements, beliefs and strategic reasoning

3.1 Framework

Primitives of the game. Let I be the finite set of *players*. For any profile of sets $(X_i)_{i\in I}$ and any $J \subseteq I$, I write $X_J := \times_{j\in J} X_j$, $X := X_I$, $X_{-i} := X_{I\setminus\{i\}}$. Let $(\overline{A}_i)_{i\in I}$ be the finite sets of *actions* potentially available to each player. Let $\overline{H} \subseteq \cup_{t=1,...,T} \overline{A}^t \cup \{h^0\}$ be the set of histories, where $h^0 \in \overline{H}$ is the empty initial history and T is the finite horizon. The set $\overline{H}$ must have the following properties. First property: For any $h = (a^1, ..., a^t) \in \overline{H}$ and $l < t$, it holds $h' = (a^1, ..., a^l) \in \overline{H}$, and I write $h' \prec h$.[15] Let $Z := \{z \in \overline{H} : \nexists h \in \overline{H}, z \prec h\}$ be the set of terminal histories (henceforth, *outcomes* or *paths*)[16], and $H := \overline{H}\setminus Z$ be the set of non-terminal histories (henceforth just *histories*). Second property: For every $h \in H$, there exists a non-empty set $A_i(h) \subseteq \overline{A}_i$ for each $i \in I$[17] such that $(h, a) \in \overline{H}$ if and only if $a \in A(h)$. Let $u_i : Z \to \mathbb{R}$ be the *payoff function* of player i. The list $\Gamma = \left\langle I, \overline{H}, (u_i)_{i\in I} \right\rangle$ is a *finite game with complete information and observable actions*.

Derived objects. A *plan of actions* (henceforth, just "plan") of player i is a function s_i that assigns an action $s_i(h) \in A_i(h)$ to each history h that can be reached if i plays s_i. Let S_i denote the set of all plans of player i. A profile of plans $s \in S$ naturally *induces* a unique outcome $z \in Z$. (When referring to profiles of plans rather than to agreements, the word "induce" will still be used with this traditional meaning.) Let $\zeta : S \to Z$ be the function that associates each profile of plans with the induced outcome. For any $\overline{h} \in \overline{H}$, the set of plans of player i compatible with $\overline{h}$ is

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

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

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

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

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

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

3.2 Beliefs, Rationality, and Rationalizability

The beliefs of a player about the plans of the co-players are modeled as a Conditional Probability System (henceforth, CPS).

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

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

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

A CPS is an array of beliefs, one for each history, that satisfies the chain rule of probability: whenever possible, the belief at a history is an update of the belief at the previous history based on the observed co-players' moves.[18]

As put forward by Battigalli and Siniscalchi [10], a player *strongly believes* an event when she believes the event is true as long as not contradicted by observation. Here the events will correspond to sets of co-players' plans. Formally, for any player i and any set of co-players $J \subseteq I \setminus \{i\}$, I say that a CPS μ_i *strongly believes* $\widehat{S}_J \subseteq S_J$ if for every $h \in H(\widehat{S}_J)$, $\mu_i(\widehat{S}_J \times S_{I \setminus (J \cup \{i\})}|h) = 1$. I say that a CPS strongly believes a collection of sets when it strongly believes each set of the collection. I will often use the fact that strong belief in a collection of sets implies strong belief in their intersection.

[18] Note that a player can have correlated beliefs over the plans of different co-players, although players will not make use of joint randomization devices. The two things are not at odds, because players can believe in spurious correlations among co-players' plans (see, for instance, Aumann [1] and Brandenburger and Friedenberg [15]). However, *strategic independence* (Battigalli [5]) could be assumed throughout the paper and the results would not change. See the companion paper for details.

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

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

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

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

I consider players who always ascribe to each co-player the highest level of strategic sophistication that is compatible with her past behavior. This means that players *strongly believe* that each co-player is rational; strongly believe that each co-player is rational and strongly believes that everyone else is rational; and so on. This form of *common strong belief in rationality* (Battigalli and Siniscalchi [10]) is captured by the following version of extensive-form-rationalizability, which I will call **Rationalizability** for brevity.

Definition 3 *Let $S^0 := S$. Fix $n > 0$ and suppose to have defined $((S_j^q)_{j \in I})_{q=0}^{n-1}$. For each $i \in I$ and $s_i \in S_i$, let $s_i \in S_i^n$ if $s_i \in br_i(\mu_i)$ for some $\mu_i \in \Delta^H(S_{-i})$ that strongly believes $((S_j^q)_{j \neq i})_{q=0}^{n-1}$.*
Finally, let $S_i^\infty := \cap_{n \geq 0} S_i^n$. The profiles S^∞ are called rationalizable.

Definition 3 modifies *strong rationalizability* of Battigalli [7] by substituting strong belief in $(S_{-i}^q)_{q=0}^{n-1}$ with strong belief in $(S_j^q)_{q=0}^{n-1}$ for each $j \neq i$. Strong belief in $(S_j^q)_{j \neq i}$ is more restrictive than strong belief in S_{-i}^q, because it requires to believe that each co-player j is following a plan in S_j^q at every history compatible with just S_j^q, even if not compatible with S_k^q for some other co-player

k. This is irrelevant for Rationalizability,[19] but the analogous requirement will play an important role in Agreement-rationalizability, where it allows to better scrutinize the promises of each individual co-player (see the next section).

An example of Rationalizability is offered by the formal analysis of the entry game in Section 4.2.

3.3 Agreements, belief in the agreement, and Agreement-rationalizability

Players talk about their plans for the game before the start. I assume that:

- Players do not coordinate explicitly as the game unfolds: all the opportunities for coordination are discussed beforehand.
- No subset of players can reach a private agreement, secret to co-players.
- Players do not agree on the use of (joint) randomization devices.[20]

Under these assumptions, agreements can be modeled as follows:

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

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

First, an agreement specifies for each player i a set of plans e_i^0 that i promises to follow. Second, the agreement can also specify alternative sets of plans e_i^n $(n = 1, ..., k_i)$ that player i promises to follow in case she fails to

[19] All the classical definitions of extensive-form rationalizability (Pearce [31], Battigalli [6], Battigalli and Siniscalchi [10]) are outcome-equivalent in this context.

[20] The use of randomization devices can be easily introduced in the methodology. Note however that a player would lack the strict incentive to use an individual randomization device over the own actions. Therefore, in absence of joint randomization devices, only sets of outcomes instead of outcome distributions could be enforced anyway. As Pearce [31] puts it, "this indeterminacy is an accurate reflection of the difficult situation faced by players in a game."

follow any of the plans in e_i^{n-1}. So, the plans in $e_i^n \backslash e_i^{n-1}$ will be relevant for co-players' beliefs only after a deviation by player i from the plans in e_i^{n-1}.[21]

With respect to a strategy profile, which can be seen as a *complete* agreement, an agreement can instead specify only partially, or not at all, what a player should do from some history onwards. This is obtained as follows. First, e_i^0 and each $e_i^n \backslash e_i^{n-1}$ need not be singletons. Second, some history h may not be allowed by any plan in $e_i^{k_i}$; in this case, the agreement is silent regarding what player i should do from h onwards. Nonetheless, just like a strategy profile, an agreement can also pin down exactly one move for each player at each history: see the second example of Section 5.2, where the agreement coincides with a SPE.

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

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

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

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

A reduced agreement remains silent regarding a deviator's continuation plans by not introducing alternative sets. Introducing all rationalizable plans as $e_i^1 = S_i^\infty$ would be equivalent: these two ways of (essentially) not specifying a player's behavior from some history onwards will be convenient in different contexts — see footnote 28. A path agreement on z remains (essentially) silent regarding the behavior of all players after a deviation by featuring all

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

the rationalizable plans compatible with z. The formal analysis of the game in the Introduction, carried out in Section 4.2, offers examples of reduced, path, and non-reduced agreements.

I say that player i believes in the agreement if she believes as long as possible that each co-player j is carrying out a plan in e_j^0; and when this is no more possible, she believes as long as possible that j is carrying out a plan in e_j^1; and so on.[22]

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

If player j is observed to deviate from $e_j^{k_j}$, that is, a history $h \notin H(e_j^{k_j})$ is reached, from then on player i's beliefs about j's behavior are unrestricted. Let Δ_i^e denote the set of all the CPS's μ_i where player i believes in the agreement.

I take the view that players refine their beliefs about co-players' behavior through strategic reasoning based on rationality and the agreement. In particular, I assume that players, as long as not contradicted by observation, believe that each co-player is rational and believes in the agreement; that each co-player believes that all other players are rational and believe in the agreement; and so on. At histories where common belief in rationality and agreement is contradicted by observation, I assume that players maintain all orders of belief in rationality that are per se compatible with the observed behavior, and drop the incompatible orders of belief in the agreement. In the companion paper [17], I provide the details of this reasoning scheme, and I show that its behavioral implications are captured by an elimination procedure called Selective Rationalizability. The definition of Selective Rationalizability in [17] accommodates any kind of first-order belief restrictions, and it is equivalent to the following simpler definition for the analysis of agreements. (The equivalence is shown in Supplemental Appendix IV.) Fix an agreement $e = (e_i)_{i \in I}$.

Definition 7 *Let $S_e^0 := S^\infty$. Fix $n > 0$ and suppose to have defined $((S_{j,e}^q)_{j \in I})_{q=0}^{n-1}$.*

[22] This is reminiscent of the agreement being a *basis* for the CPS: see Siniscalchi [34].

For each $i \in I$ and $s_i \in S_i^\infty$, let $s_i \in S_{i,e}^n$ if $s_i \in br_i(\mu_i)$ for some $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^{n-1}$.
Finally, let $S_{i,e}^\infty := \cap_{n\geq 0} S_{i,e}^n$. The profiles S_e^∞ are called agreement-rationalizable.

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

Agreement-rationalizability requires each player i, at every step of reasoning n, to strongly believe both in $e_j^0, ..., e_j^{k_j}$ (by $\mu_i \in \Delta_i^e$) and in $S_{j,e}^{n-1}, ..., S_{j,e}^0$ for each co-player j. If $S_{j,e}^{n-1} \cap e_j^0 = \emptyset$ for some j, it means that j has come to the conclusion that she has no incentive to comply with the agreement. Then, $S_{i,e}^n$ is empty, because $\mu_i(\cdot|h^0)$ cannot give probability 1 to both $S_{j,e}^{n-1}$ and e_j^0 as required. But even if $S_{j,e}^{n-1} \cap e_j^0 \neq \emptyset$, there could be a history h that is compatible both with $S_{j,e}^{n-1}$ and with some e_j^m $(0 \leq m \leq k_j)$, but not with $S_{j,e}^{n-1} \cap e_j^m$. Then, $\mu_i(\cdot|h)$ cannot give probability 1 to both $S_{j,e}^{n-1}$ and e_j^m, and $S_{i,e}^n$ is empty. In light of this, non-emptiness of S_e^∞ means two things: complying with the agreement and believing in the agreement (from any history onwards) are both compatible with strategic reasoning.

Remark 1 *If $S_e^\infty \neq \emptyset$, then $S_e^\infty \cap e^0 \neq \emptyset$, and for each $i \in I$ and $s_i \in S_{i,e}^\infty$, there exists $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^\infty$ such that $s_i \in br_i(\mu_i)$.*

Proof. By finiteness of the game,[23] there exists M such that $S_e^M = S_e^{M+1} = S_e^\infty$. Then, for each $i \in I$ and $s_i \in S_{i,e}^\infty = S_{i,e}^{M+1}$, there exists $\mu_i \in \Delta_i^e$ that strongly believes $((S_{j,e}^q)_{j\neq i})_{q=0}^M$ such that $s_i \in br_i(\mu_i)$. For each $j \neq i$, we have $\mu_i((S_{j,e}^M \cap e_j^0) \times S_{I\setminus\{i,j\}}|h^0) = 1$, thus $S_{j,e}^\infty \cap e_j^0 \neq \emptyset$. Since both e^0 and S_e^∞ are Cartesian sets, we obtain $S_e^\infty \cap e^0 \neq \emptyset$. Moreover, μ_i strongly believes also $S_{j,e}^{M+1}, S_{j,e}^{M+2}, ...$ because all these sets are identical to $S_{j,e}^M$. ■

[23] The results of Battigalli and Tebaldi [12] imply that Remark 1 is true also in the large class of *infinite* dynamic games they study when the belief restrictions that arise from the agreement are compact (see [12], page 758).

Analogously, every $s_i \in S_i^\infty$ is a sequential best reply to some μ_i that strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$.

Examples of Agreement-rationalizability are offered by the analysis of the introductory game and of the entry game in Section 4.2. In the rest of the paper, recall that I will refer to $\zeta(e^0)$ as the outcome set that the agreement *prescribes*, and to $\zeta(S_e^\infty)$ as the outcome set the agreement *induces*. For a set of plans $S^* \subset S$, I will still say that S^* induces $\zeta(S^*)$, as customary.

4 Self-enforceability

4.1 Credible, self-enforcing, and truthful agreements

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

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

A credible agreement induces each player i to believe in the agreed-upon plans that are compatible with strategic reasoning, $S_{-i,e}^\infty \cap e_{-i}^0$ (cf. Remark 1). But this belief may be contradicted by the actual play, because credibility does not imply that players will comply with the agreement, it only implies that they *may* do so *everywhere in the game*. So, the set of outcomes induced by the agreement ($\zeta(S_e^\infty)$) may be larger than the set of outcomes players expect given the belief in the agreement ($\zeta\left(S_e^\infty \cap e^0\right)$). When instead the agreement induces only the outcomes players expect, I say that the agreement is *self-enforcing*.

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

In light of this, every agreement that induces one and just one outcome is self-enforcing.

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

Proof. Since $\zeta(S_e^\infty)$ is a singleton, $S_e^\infty \neq \emptyset$. Then, by Remark 1, $\zeta(S_e^\infty \cap e^0) \neq \emptyset$. So, since $\zeta(S_e^\infty)$ is a singleton, $\zeta(S_e^\infty \cap e^0) = \zeta(S_e^\infty)$. ■

Self-enforceability implies that, for *all* their refined beliefs, players will comply with the agreement *on the induced paths*, so that no violation of the agreement will actually occur. That is, $\zeta(S_e^\infty) \subseteq \zeta(e^0)$. This inclusion can be strict: a self-enforcing agreement may not explicitly exclude all the outcomes that will be ruled out by strategic reasoning. When instead the agreement specifies directly the outcomes it induces, I say that the agreement is *truthful*.[24]

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

The tools developed in Section 3 and the definitions above allow to assess the implications of a given agreement among players. In the next section, I use the game from the Introduction and the entry game of Section 2 to give a concrete illustration of this methodology.

4.2 Examples

The game in the Introduction It is easy to check that all plans are justifiable, hence they are all rationalizable: $S = S^1 = S^\infty$. The table summarizes the analysis of four possible agreements ($\{N.\cdot\}$ and $\{E.\cdot\}$ denote $\{N.U, N.D\}$ and $\{E.L, E.R\}$).

Agreement	Reduced	Path on (S, E)	"Unilateral"	Path on (N, W)
e_A	$\{N.U\}$	$\{S\}$	S_A	$\{N.\cdot\}$
e_B	$\{W\}$	$\{E.\cdot\}$	$\{W\}, \{W, E.L\}$	$\{W\}$
$S_{A,e}^1 \times S_{B,e}^1$	$\{N.\cdot\} \times \{W\}$	$\{S, N.D\} \times \{E.\cdot\}$	$\{N.U\} \times S_B$	$\{N.\cdot\} \times S_B$
$S_{A,e}^2 \times S_{B,e}^2$	$\{N.\cdot\} \times \{W\}$	$\{S, N.D\} \times \{E.L\}$	$\{N.U\} \times \{W\}$	$\{N.\cdot\} \times S_B$
$S_{A,e}^\infty \times S_{B,e}^\infty$	$\{N.\cdot\} \times \{W\}$	$\{S\} \times \{E.L\}$	$\{N.U\} \times \{W\}$	$\{N.\cdot\} \times S_B$
Conclusion	Truthful	Truthful	Self-enforcing	Credible

[24] The choice of the term "truthful" is clearly inspired by the implementation literature, although an important caveat applies: see the end of Section 5.1.

The reduced agreement is the one proposed in the Introduction. The belief in the agreement is given by $\Delta_A^e = \{\mu_A : \mu_A(W|h^0) = 1\}$ and $\Delta_B^e = \{\mu_B : \mu_B(N.U|h^0) = 1\}$. Since all plans are rationalizable ($S_e^0 = S$), we have $S_{A,e}^1 = br_A(\Delta_A^e) = \{N.U, N.D\}$ and $S_{B,e}^1 = br_B(\Delta_B^e) = \{W\}$. Strong belief in $S_{B,e}^1$ and in $S_{A,e}^1$ does not restrict Ann and Bob's conjectures further with respect to the belief in the agreement. Hence, $S_e^2 = S_e^1$. By induction, $S_e^\infty = S_e^1$. Since $\zeta(S_e^\infty) = \{(N, W)\}$, by Proposition 1 the agreement is self-enforcing, and since $\zeta(e^0) = \{(N, W)\}$ as well, the agreement is truthful.

The path agreement on (S, E) requires more steps of reasoning. We have $\Delta_A^e = \{\mu_A : \mu_A(\{E.L, E.R\}|h^0) = 1\}$ and $\Delta_B^e = \{\mu_B : \mu_B(S|h^0) = 1\}$. So, at the first step of reasoning, Ann plays either S, or $N.D$ if she gives sufficiently high probability to $E.R$; Bob plays E and either L or R depending on his new belief after being surprised by Ann's deviation. Strong belief in $S_{A,e}^1$ imposes belief in $N.D$ at history (N, E), so we get $S_{B,e}^2 = \{E.L\}$. Strong belief in $S_{B,e}^2$ imposes belief in $E.L$, so we obtain $S_e^3 = \{S\} \times \{E.L\} = S_e^\infty$: the agreement is self-enforcing and truthful.

In the "unilateral" agreement, Ann remains silent ($e_A^0 = S_A$), while Bob promises to play W ($e_B^0 = \{W\}$), and to play L in case he deviates to E ($e_B^1 = \{W, E.L\}$). Ann's belief in the agreement is given by

$$\Delta_A^e = \left\{\mu_A : \mu_A(W|h^0) = 1 = \mu_A(E.L|(N, E))\right\}.$$

Then, Ann plays $N.U$. Consequently, Bob plays W. The agreement induces (N, W), so by Proposition 1 it is self-enforcing. With respect to the reduced agreement, it has the seeming advantage that $N.D$ is not agreement-rationalizable for Ann. However, at $(N.E)$, both actions of Bob are equally compatible with strategic reasoning, and Ann believes in L and thus plays U only because of Bob's post-deviation promise. This is why requiring $S_e^\infty \subseteq e^0$ does not seem to be a compelling strengthening of self-enforceability.

The path agreement on (N, W) is only credible: beside the agreed outcome (N, W), also the outcomes where Bob plays E are compatible with strategic reasoning. Note that, while enforcing outcome (N, W) requires explicit threats,

the path agreement suffices to obtain the SPE outcome (S, E). This is far from true in general, even when the SPE is unique: the variation of the game in Supplemental Appendix I shows this point.

The entry game of Section 2 Let $\underline{p}$ be the smallest price of the incumbent that makes entry profitable. To fix ideas, I am going to consider three specific values of $\underline{p}$ that fall into the three cases analyzed in Section 2. The table illustrates the first four steps of Rationalizability: the prices in bold are associated with technology $k = B$, the prices in italics with $k = A$, action "entry" is omitted from the description of the entrant's plans, and the no-entry plan is denoted by N.

	Case 1 ($\underline{p} = 76$)	Case 2 ($\underline{p} = 65$)	Case 3 ($\underline{p} = 62$)
S_1^1	**[60,72],** *[72,84]*	**[60,72],** *[72,84]*	**[60,72],** *[72,84]*
S_2^1	**[62,72],** N	**[60,72],** *[80.5,84]*, N	**[60,72],** *[79,84]*, N
S_1^2	**60,***[79,84]*	**[60,66],** *[78,84]*	**[60,66],** *[78,84]*
S_2^2	**[62,66],** N	**[60,66],** *[80.5,84]*, N	**[60,66],** *[79,84]*, N
S_1^3	*[79,81]*	**[60,66],** *[78,84]*	**[60,66],** *[78,84]*
S_2^3	**[62,66],** N	**[60,66],** *[80.5,84]*, N	**[60,66],** *[79,84]*, N
S_1^4	*[79,81]*	"	"
S_2^4	**[63.5,64.5]**	"	"

For each firm, every step of reasoning n is entirely determined by the lowest and the highest prices of the competitor, $\underline{p}_{-i}^{n-1}$ and $\overline{p}_{-i}^{n-1}$, determined at step $n-1$, where "step-0" feasible prices range from 48 to 96 by assumption. Firm 1 best replies to any expected price between $\underline{p}_2^{n-1}$ and $\overline{p}_2^{n-1}$, whereas firm 2, in case of entry, best replies to any expected price between $\max\{\underline{p}, \underline{p}_1^{n-1}\}$ and $\overline{p}_1^{n-1}$ (see the first part of Section 2 for best replies).

In Case 1, we have $\max\{\underline{p}, \underline{p}_1^3\} = \underline{p}_1^3$; then, entry is always profitable and no-entry is eliminated from S_2^4. Note by induction that after infinite steps of reasoning prices converge to the $(80, 64)$ equilibrium. Any different announcement by the incumbent would conflict with some order of belief in rationality,

and thus is not credible (formally, it is not even allowed as an agreement).

In Cases 2 and 3, Rationalizability converges in two steps, because the lowest and highest prices of both firms are 60 and 84 for both the first two steps. Both entry and no-entry are rationalizable, therefore there is scope for the incumbent to deter entry with an announcement. The incumbent can announce any rationalizable price $p_1 < \underline{p}$. Formally, this translates into the reduced agreement with $e_1^0 = \{(p_1, B)\}$ and $e_2^0 = S_2^\infty$, and it induces $S_{2,e}^\infty = S_{2,e}^1 = \{N\}$ and $S_{1,e}^\infty = S_{1,e}^0 = S_1^\infty$. Alternatively, in Case 2, the incumbent can simply announce technology $k = B$. Formally, this translates into the reduced agreement with $e_1^0 = \{[60, 66] \times \{B\}\}$ and $e_2^0 = S_2^\infty$.[25] Agreement-rationalizability goes as follows ($S_{1,e}^1$ and $S_{2,e}^2$ are identical to $S_{1,e}^0$ and $S_{2,e}^1$).

$S_{2,e}^0 = S_2^\infty$	$S_{1,e}^0 = S_1^\infty$	$S_{2,e}^1$	$S_{1,e}^2$	$S_{2,e}^3$
[60,66],*[80.5,84]*,N	**[60,66],***[78,84]*	*[80.5,81]*,N	**[64.25,64.5]**	N

The prices in $S_{2,e}^1$ best reply to an expected price of the incumbent between $\underline{p} = 65$ (otherwise entry would not be rational) and 66 (because the incumbent is expected to fix a rationalizable price with $k = B$). The prices in $S_{1,e}^2$ best reply to beliefs over the prices in $S_{2,e}^1$ and leave no incentive to enter, so $S_{2,e}^3 = \{N\} = S_{2,e}^\infty$ and $S_{1,e}^2 = S_{1,e}^\infty$. Since the agreement induces no-entry as unique outcome, by Proposition 1 it is self-enforcing.

The incompleteness of the agreement triggers steps of reasoning that refine players' plans up to the point where every belief over these plans (here $S_{1,e}^2$) induces the desired behavior (no-entry). Under the SPE threat $p_1 = 64$, instead, entry cannot be rationalized under belief in the threat ("entry" is not in $S_{2,e}^1$), thus no rationalizable price of the incumbent can be eliminated.

[25] Note that the agreement, featuring only rationalizable plans, already incorporates strategic reasoning about rationality, which is convenient from an algorithmic viewpoint. In Section 2, I followed instead the equivalent but more natural reasoning scheme where the announcement does not talk of rationalizable prices, and belief in the announcement and beliefs in rationality interact from the first step of reasoning.

5 Implementability

5.1 Implementability and agreements design

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

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

A set of outcomes induced by a merely credible agreement does not correspond to what players agreed on and believe in. For this reason, implementation requires the agreement to be self-enforcing.

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

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

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

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

Self-Justifiability: For each $i \in I$ and $s_i \in S_i^*$, there exists μ_i that strongly believes $(S_j^*)_{j \neq i}$ and $(S_j^\infty)_{j \neq i}$ such that $s_i \in br_i(\mu_i)$.[26]

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

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

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

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

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

Corollary 1 simplifies the search for implementable outcome sets. First, Rationalizability is performed. This is a standard elimination procedure that does not depend on agreements. Then, one must look for sets of rationalizable plans that satisfy Realization-strictness and Self-Justifiability. However, there is no guarantee that the induced outcome set is implementable, because Realization-Strictness and Self-Justifiability are necessary but, in general, not sufficient conditions for implementability. The next step is finding additional conditions on the set of plans or conditions on the game that, together with Realization-Strictness and Self-Justifiability, ensure implementability. Definition 13 of a Self-Enforcing Set will provide sufficient conditions for all games. To this end, I must first define the "closure" of a self-justifiable set.

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

The closure of S^*, for each player i, consists of all the rationalizable plans that can be justified under strong belief that each co-player j follows a plan in S_j^*, and some other rationalizable plan otherwise. By Self-Justifiability of S^*, $\overline{S}^*$ includes S^* itself.

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

Forward Induction says that, for each player i, every plan in $\overline{S}_i^*$ is justifiable under strong belief that each co-player j follows a plan in S_j^*, *and some other plan in $\overline{S}_j^*$ otherwise* (or a rationalizable plan, as usual). Essentially, the closure of a SES is the set of plans that players may follow under belief in the SES, and Forward Induction requires these plans to remain justifiable after the additional consideration that the co-players believe in the SES as well.

Consider now the agreement on the SES; that is, the reduced agreement $e = (e_i)_{i\in I}$ with $e_i^0 = S_i^*$ for each $i \in I$. By definition of closure, $S_e^1 = \overline{S}^*$. By Forward Induction, $\overline{S}^*$ is not refined by forward induction reasoning based on the agreement. Hence, we obtain $S_e^\infty = \overline{S}^*$. A SES and its closure are outcome-equivalent: by Self-Justifiability, $S^* \subseteq \overline{S}^*$, and by Realization-strictness, $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$.[27] Therefore, the agreement on the SES implements precisely the SES outcomes $\zeta(S^*)$.

Proposition 4 *The reduced agreement on a SES is truthful.*

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

A simple SES is constructed in the first example of the next subsection. In Section 6 I uses SES's to solve the Hotelling problem.

[27] So, in terms of outcomes, SES's are "closed under rationalizable behavior", and indeed boil down to *sets closed under rational behavior* (Basu and Weibull [4]) in static games.

The current gap between necessary and sufficient conditions for implementability is given by a seemingly strong condition: Forward Induction. But the power of Forward Induction is mitigated by Realization-strictness and Self-Justifiability. In a nutshell, Self-Justifiability already captures forward induction reasoning based on the beliefs in rationality, and by Realization-strictness any deviation from the paths induced by S^* cannot be rationalized under the view that the deviator believes in S^*. Then, forward induction reasoning based on S^* kicks in only after *later* deviations by *other* co-players. With this, I am going to argue that, in a game with two players or two stages, Forward Induction is *implied* by Realization-strictness and Self-Justifiability. I say that a game has two stages when $Z \subseteq \overline{A} \cup \overline{A}^2$.

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

The proof of Proposition 5 is based on the following ideas. Recall that each $\overline{S}_i^*$ consists of all the rationalizable plans of player i that can be optimal under strong belief in $(S_j^*)_{j \neq i}$ and $(S_j^\infty)_{j \neq i}$. Forward Induction requires these plans to remain optimal when strong belief in $(\overline{S}_j^*)_{j \neq i}$ is also imposed. Now, Self-Justifiability yields $S^* \subseteq \overline{S}^*$, therefore strong belief in $\overline{S}_j^*$ can have additional bite with respect to strong belief in S_j^* only at histories that are compatible with $\overline{S}_j^*$ but not with S_j^*. However, Realization-Strictness yields $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$, therefore such histories are incompatible with $\overline{S}_i^*$ in two-player games, thus are irrelevant for Forward Induction, and do not exist in two-stage games, because the first-stage optimal moves of j must be compatible with S_j^*.

The important consequence of Proposition 5 is that, in games with two players or two stages, SES's fully characterize implementable outcome sets and provide truthful reduced agreements that implement them.

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

1. *a Cartesian set of rationalizable profiles is a Self-Enforcing Set if and only if it satisfies Realization-strictness and Self-Justifiability;*

2. *an outcome set is implementable if and only if it is induced by a Self-Enforcing Set;*

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

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

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

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

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

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

Proof. "Only if" comes from Corollary 2. For "if" and the final statement: let $s^* = (s_i^*)_{i \in I} \in S^\infty$ be a realization-strict Nash equilibrium. By Proposition 3, the singleton $\{s^*\}$ satisfies Realization-strictness. By Proposition 6, it also satisfies Self-Justifiability. By Proposition 5, it also satisfies Forward Induction, thus it is a SES. Then, by Proposition 4, $\zeta(s^*)$ is implemented by the reduced agreement $e = (e_i)_{i \in I}$ with $e_i^0 = \{s_i^*\}$ for each $i \in I$. ■

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

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

T1 e^0 *satisfies Realization-strictness;*

T2 *for every* $i \in I$ *and* $h \in H(S_i^\infty)$,

$$e_i^{k_i} \cap S_i(h) \neq \emptyset;$$

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

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

Remark 2 *If* $e = (e_i)_{i \in I}$ *is tight, then* e^0 *satisfies Self-Justifiability.*

Like a SES, a tight agreement initially specifies plans that satisfy Realization-strictness (by T1) and Self-Justifiability (by Remark 2). Differently from a SES, a tight agreement also specifies alternative plans $e_i^1, ..., e_i^{k_i}$ that each player i should follow, until all histories compatible with her rationalizable plans are reached (this is T2). All histories that player i can reach under belief in the agreement, $H(br_i(\Delta_i^e) \cap S_i^\infty)$, are also reached by a set of agreed-upon plans e_i^n that can be justified under belief in the agreement (this is T3). So, T2 and T3 guarantee that, when player j strongly believes that each co-player i follows plans in $e_i^0, ..., e_i^{k_i}$, her beliefs are also compatible with forward induction reasoning based on rationality and on the agreement.[28] This makes up for the fact that e^0 does not satisfy Forward Induction. Therefore, the agreement is credible. By Realization-strictness and Self-Justifiability of e^0, the agreement is self-enforcing and truthful.

Proposition 7 *Tight agreements are truthful.*

[28] Relatedly, given a SES S^*, one can anticipate these forward induction considerations in the agreement and transform it into a tight agreement as follows: for each $i \in I$, $e_i^0 = S_i^*$, $e_i^1 = \overline{S}_i^*$, $e_i^2 = S_i^\infty$. Introducing e_i^2 is immaterial for the agreement but verifies T2: introducing all or none of the rationalizable plans of a player are equivalent ways not to restrict beliefs, but the first is convenient for tight agreements, the second for SES's.

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

Proof. "If" comes from Proposition 7. "Only if": see the Appendix. ■

Tight agreements close the gap between necessary and sufficient conditions for implementability in all games,[29] and the roadmap for the joint search of implementable outcome sets and agreements that implement them. If a candidate set of outcomes is implementable, a tight agreement that implements it can be found by following the search for SES's first, and introducing alternative plans if Forward Induction cannot be satisfied. A tight agreement is constructed in this way in the second example of the next subsection.

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

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

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

The use of the terms "truthful" and "implementation" is indeed inspired by an analogy with robust implementation (Bergemann and Morris [14]). A robust mechanism implements the outcome assigned by the social choice function to players' types[30] for all their hierarchies of beliefs about co-players' types; a self-enforcing agreement implements (a subset of) the agreed-upon outcome(s) for all players' refined beliefs. When players use direct mechanisms, they truthfully reveal their types and the corresponding outcome obtains; when players use truthful agreements, they declare precisely the outcome(s) they want to

[29] A word of caution: extending the "only if" direction of Theorem 3 to games with infinite horizon probably requires to introduce agreements of infinite length.

[30] In this paragraph, I use the term types to mean *payoff-relevant types*. In robust implementation, precisely because it does not rely on a common prior, a given payoff-relevant type must be allowed to have different beliefs about the co-players' types, which become part of her full *epistemic type*: see, e.g., Penta [32] in a completely belief-free setting, and Ollar and Penta [30] in a setting with partial belief restrictions.

achieve. Both direct mechanisms and truthful agreements suffice for implementation. Note though an important difference: while a direct mechanism requires players to specify *only* their type, a truthful agreement, beside the outcome(s), typically needs to specify off-path behavior. This is the price to pay for the agreement being a "soft mechanism", which does not change the rules of the game.

5.2 Further examples

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

Peacekeeping game[31] Dave, a weapons producer, can *Instigate* a conflict between Ann and Bob. If he does, Ann and Bob can engage in an *Arms Race*, or remain *Peaceful*. Engaging in the arms race transfers 1 util to Dave. At the same time, Cleo, a superpower, can *Intervene* to prevent an escalation of the conflict and impose sanctions against Dave. The cost of the intervention is 3 for Dave and 2 for Cleo; however, if Ann or Bob engages in the arms race and the other does not, the unarmed player falls under Cleo's influence and has to share its 6 units of resources with Cleo. If Cleo does not intervene and Ann or Bob engage in the arms race, the conflict escalates. Cleo suffers a disutility of 1 from the war. If both Ann and Bob are armed, the war comes to a costly impasse; otherwise, the unarmed player gets conquered and loses

[31] This game is freely inspired by the leading example in Greenberg [23].

all its resources to the other. The game is represented in the figure, where the payoffs are in alphabetical order (Cleo chooses the matrix).

DAVE $\quad -Out \longrightarrow 0,0,0,0$

$Instigate \downarrow$

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

The game has only one SPE, where Dave instigates, Cleo does not intervene, and Ann and Bob engage in the arms race.[32] However, Cleo would rather stop Dave from instigating the conflict by threatening to intervene. Intervening is in Cleo's interest only if Ann and Bob do not coordinate.[33] This form of agreement incompleteness is enabled by $|e^0| > 1$ (even when $\zeta(e^0)$ is a singleton) and allowed by SES's. Hence, I show that there is a SES where Cleo threatens to intervene, Ann and Bob remain silent, and Dave does not instigate. All plans are justifiable, hence rationalizable. Let $S^* = \{AR, P\} \times \{AR, P\} \times \{Int\} \times \{Out\}$. To show that S^* is a SES, since the game has 2 stages, by Theorem 1 it is enough to show Realization-strictness and Self-Justifiability. For Dave, they both follow from the fact that $br_D(\mu_D) = \{Out\}$ for every μ_D that strongly believes S^*_C. For each $i = A, B, C$, Realization-strictness trivially follows from $\zeta(S_i \times e^0_{-i}) = \{(Out)\}$. There remains to show Self-Justifiability. For Cleo, Int is justified by any μ_C that strongly believes S^*_D such that (for instance) $\mu_C(s_A \neq s_B|(Inst)) \geq 1/2$. For Ann, AR (resp., P) is justified by any μ_A that strongly believes S^*_C and S^*_D such that $\mu_A(s_B = AR|(Inst)) \geq 1/3$ (resp., $\mu_A(s_B = P|(Inst)) \geq 2/3$); likewise for Bob.

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

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

Should I stay or should I go? In the department of dean Ann there are two game theorists, Bob and Cleo, who are up for midterm review. Ann maximizes the benefit from game theorists to the department, which is marginally decreasing, minus the opportunity cost of their salaries, which is marginally increasing. Ann wants to offer to Bob and Cleo the renewal at salary r, lower than the market salary w, but sufficient to make them prefer to *Stay* if they have to pay cost $g < w - r$ to *Go* on the market (they have a preference for staying). If they both accept, the game ends. If one accepts and the other does not, say Bob, the game continues in the next year as in the figure. Cleo can *Stay* or *Go* on the market as well; Ann can *Shut* down Bob's position, or keep it *Open*. If Ann shuts down the position and Cleo goes on the market, Ann is in a weak bargaining position and Cleo obtains a raise to $v > r + g$ ($v < w$). If Ann keeps Bob's position open and Cleo stays, Bob gets the position back and bargains a salary $t > r + g$ ($t < v$). With both game theorists on the market and the position open, Ann starts a job search and bargaining gets delayed to the market stage. Ann can *Hire* or *Not*; Bob and Cleo can *Stay* or *Go* for good. As deadlines approach, players must make their choices without knowing the choices of others. If Ann hires a new game theorist at salary w, she will keep only Cleo if she stays, or Bob if he stays and Cleo leaves, in both cases at salary r. If Ann does not hire and Bob and Cleo do not leave, they will bargain a salary t; if one leaves and the other stays, the latter bargains a salary u with $t < u < v$. Ordinal payoffs compatible with this story are in the figure (cardinal payoffs will not matter for the analysis). In the last stage, Bob chooses the row.

$8,3,3 \longleftarrow$ *Stay* — Bob — *Go* $\longrightarrow$ (Cleo stays)

A\C	*Stay*	*Go*
Open	$7,4,3$	$\cdot-$
Shut	$6,2,3$	$3,2,6$

$\longrightarrow \Gamma$

Γ :

Hire	*Stay*	*Go*	*Not*	*Stay*	*Go*
Stay	$2,0,1$	$2,1,2$	*Stay*	$5,4,4$	$4,5,2$
Go	$2,2,1$	$1,2,2$	*Go*	$4,2,5$	$0,2,2$

Before offering the renewal, Ann clarifies her intention to shut down a game theorist position if one of them, say Bob, will not accept the offer. But this will induce Cleo to bargain a higher salary by going on the market. In turn, this may induce Ann to increase her bargaining power by keeping Bob's position open to look for potential new hires. However, Ann has no real intention to hire. Understanding this with forward induction reasoning, Cleo will not leave. Then, Ann will actually keep the position open but not hire. This leaves the position to Bob at salary $t > r + g$. How to solve the impasse? Bob and Cleo, who are happy to secure their renewals at salary r, convene with Ann that if they will all be on the market, they will go their separate ways: Bob and Cleo will leave and Ann will hire.

We are going to construct this agreement and show it is tight through the roadmap of Section 5.1. We look for an agreement that implements outcome $(Stay)$ in the game of the figure; by symmetry, it can be extended to the whole game. All plans are justifiable, hence rationalizable. Thus, we look for e^0 that induces $(Stay)$ and satisfies Realization-strictness and Self-Justifiability. Bob's Realization-strictness is satisfied if $e^0_A = \{S\}$, or if $O.N \notin e^0_A$ and $S \notin e^0_C$. In the first case, Ann's and Cleo's Self-Justifiability require, respectively, $G.G \in e^0_C$ and $S \notin e^0_C$, so we have $\{G.G\} \subseteq e^0_C \subseteq \{G.S, G.G\}$. The second case boils down to the first, because Ann's Self-Justifiability requires $O.H \notin e^0_A$ as well. Thus, we focus on agreements with $e^0_A = \{S\}$, $e^0_B = \{S\}$, and either $e^0_C = \{G.G\}$, or $e^0_C = \{G.S, G.G\}$. Does any of the two constitute a SES? No. In both cases, the closure of e^0 for Ann is $\{S, O.N\}$: under belief that Cleo goes on the market, $O.H$ is never optimal. But then, Forward Induction is violated for Cleo, because the only sequential best reply under strong belief in $\{S, O.N\}$ is $G.S$. Therefore, we look for a tight agreement. Let $e^0 = \{(S, S, G.G)\}$. Restrict Bob's behavior after his deviation by imposing $e^1_B = \{S, G.G\}$. Also, let $e^1_A = \{S, O.H\}$, so that all histories are reached by all players and T2 is satisfied. T1 is Realization-strictness of e^0. Is T3 satisfied? Under belief in the agreement, players play exactly e^0, so it immediate to check that T3 holds.

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

6 Application - discretized Hotelling

In a separate paper [16], I show that in the original Hotelling model with two firms, two stages (location-pricing), and linear transportation cost, the transportation-efficient location pair $(1/4, 3/4)$ is the only symmetric location pair that is induced by a SES.

The intuition is simple. When firms locate at $(a_1, a_2) \in [0,1]^2$ with $a_1 \leq 1/4$ and $a_2 \geq 3/4$, there is only one rationalizable price pair. In this pricing solution, given the location of the competitor, the closer a firm is to the center, the higher its profit. This generates the incentive to move inwards, up to $(1/4, 3/4)$.

Suppose now firms are at $(1/4, 3/4)$. If a firm, say firm 1, deviates towards the middle, a multiplicity of rationalizable prices arises. This is because, as firms get closer to each other, it becomes cheaper to *undercut* the competitor's price by more than the transportation cost between the two locations, so to conquer the whole market. In particular, firm 1 has the incentive to undercut any price of firm 2 that makes the deviation profitable. In turn, firm 2 has the incentive to respond to an undercutting attempt with a low price, which firm 1 has no incentive to undercut. Hence, the SES inducing locations $(1/4, 3/4)$ is sustained by a very intuitive, incomplete, non-equilibrium threat: "if you deviate towards the middle, I will make sure you won't undercut me!"

The same uncertainty about prices prevents the existence of a SES where firms locate at (a_1, a_2) with $a_1 \in (1/4, 1/2)$ and $a_2 = 1 - a_1$. Differently from SPE, SES's do not reduce this uncertainty to one probability distribution. Any set of prices firms could credibly agree on includes undercutting attempts. Then, in case of pessimistic belief over this set, a firm has the incentive to "give in" and move outwards, to a location where undercutting is not rationalizable anymore for the competitor.

Here I replicate existence and uniqueness of the SES solution in a discretized version of the model. To simplify exposition, I will also impose that firm 1 locates in the first half and firm 2 in the second half of the spectrum, and I will consider only SES's that are symmetric also in prices, not just locations.

Model Each firm $i = 1, 2$ chooses a location a_i from the set of integers A_i, where $A_1 = \{0, ..., 49\}$ and $A_2 = \{51, ..., 100\}$. After observing the chosen locations, each firm chooses an integer price p_i below an arbitrarily large prohibitive threshold. Then each consumer $j \in [0, 100]$ buys one unit from the firm i that minimizes $p_i + |j - a_i|$, breaking ties at random. Firms maximize revenues. All the arguments will be provided from firm 1's viewpoint; for firm 2 symmetric arguments apply. I will write $S_i(a_i)$ for the set of plans s_i with $s_i(h^0) = a_i$, and I will write $s_i(a_{-i})$ for $s_i((s_i(h^0), a_{-i}))$.

Optimal prices Fix a location pair (a_1, a_2) and a price p_2 of firm 2. Let

$$D_1(p_1, p_2) = \frac{a_1 + a_2}{2} + \frac{1}{2}(p_2 - p_1).$$

When $p_2 - (a_2 - a_1) < p_1 < p_2 + (a_2 - a_1)$, $D_1(p_1, p_2)$ represents firm 1's demand; thus, among these values of p_1, the closer p_1 to

$$p_1^F(p_2) := \arg\max_{\widetilde{p}_1 \in \mathbb{R}} \widetilde{p}_1 D_1(\widetilde{p}_1, p_2) = \frac{a_1 + a_2}{2} + \frac{1}{2}p_2,$$

the higher firm 1's revenues. Let:

$$\begin{aligned} p_1^-(p_2) &: = p_2 - (a_2 - a_1) - 1, \\ p_1^+(p_2) &: = p_2 + (a_2 - a_1) - 1. \end{aligned}$$

Undercutting p_2 with $p_1^-(p_2)$ brings demand 100. The candidate best replies to p_2 are $p_1^-(p_2)$, and either the closest integers to $p_1^F(p_2)$, if $p_1^-(p_2) < p_1^F(p_2) < p_1^+(p_2)$, or $p_1^+(p_2)$, if $p_1^F(p_2) \geq p_1^+(p_2)$.[34] It is useful to record that the integer part of $p_1^F(p_2)$, denoted by $\lfloor p_1^F(p_2) \rfloor$, best replies to p_2 whenever $p_1^F(p_2) \leq p_1^+(p_2)$ and[35]

$$p_2 \leq \overline{p}_2 := 400 - a_1 - a_2 - 40\sqrt{100 - a_2}.$$

[34] When a_2 is close to 100, also $p_1 = p_2 - (a_2 - a_1)$ can best reply to p_2, however this situation will never materialize in the analysis. Instead, $p_1 = p_2 + (a_2 - a_1)$ is never optimal, because it brings demand $a_1/2$, while $p_1^+(p_2) = p_1 - 1$ brings $D_1(p_1^+(p_2), p_2) > a_1$.

[35] Since $p_1^F(p_2)$ is the average of two integers, either it is integer, or its integer part is one of the closest integer. Price $\overline{p}_2$ is defined by equation (4) in the Appendix. As apparent

For firm 2, the expressions for $p_2^-(p_1)$ and $p_2^+(p_1)$ are unchanged, and

$$\begin{aligned} p_2^F(p_1) &= 100 - \frac{a_1 + a_2}{2} + \frac{1}{2}p_1; \\ \overline{p}_1 &= 200 + a_1 + a_2 - 40\sqrt{a_1}. \end{aligned}$$

There is a unique pair $(p_1^*, p_2^*) \in \mathbb{R}^2$ such that $p_1^* = p_1^F(p_2^*)$ and $p_2^* = p_2^F(p_1^*)$:

$$(p_1^*, p_2^*) = \left(\frac{200 + a_1 + a_2}{3}, \frac{400 - a_1 - a_2}{3} \right).$$

When clear from the context, I will not say explicitly at which locations I am computing $\overline{p}_i$, p_i^*, $p_i^F(p_{-i})$, $p_i^-(p_{-i})$, and $p_i^+(p_{-i})$.

Existence Let $z := ((a_1, a_2), (p_1^*, p_2^*)) = ((25, 75), (100, 100))$. Let

$$S_i^* = \left\{ s_i \in S_i^\infty \cap S_i(z) \,\middle|\, \forall a'_{-i} \neq a_{-i}, s_i(a'_{-i}) < \overline{p}_i \right\}, \quad i = 1, 2.$$

I show that $S^* = S_1^* \times S_2^*$ is non-empty and satisfies Realization-Strictness. Then, by Proposition 6, S^* satisfies also Self-Justifiability, and thus by Theorem 1 it is a SES.

At locations $(25, 75)$, given $p_2^* = 100$, $p_1^* = 100$ brings revenues 5000, while $p_1^-(p_2^*) = 49$ brings revenues 4900. Hence, p_1^* best replies to p_2^*.

For every $a'_1 \neq 25$, at $(a'_1, 75)$ we have $\overline{p}_2 = 125 - a'_1$ and

$$p_1^F(\overline{p}_2) = 100 < 200 - 2a'_1 - 1 = p_1^+(\overline{p}_2).$$

Hence, $p_1^F(\overline{p}_2)$ best replies to $\overline{p}_2$ and brings demand 50 and revenues 5000. So, for every $s_2 \in S_2^*$, firm 1's revenues against $s_2(a'_1) < \overline{p}_2$ are lower than 5000. Then, for every μ_1 that strongly believes S_2^*, the set of continuation best replies

from that equation, in the continuous model $p_2 \leq \overline{p}_2$ is also a necessary condition for $p_1^F(p_2)$ to be a best reply; here it is not because undercutting requires to lower the price by an entire unit with respect to $p_2 - (a_2 - a_1)$. As a consequence, at $(a'_1, 75)$ with $a'_1 > 25$, for some $k > 1$, firm 1 has no incentive to undercut any $p_2 \in (\overline{p}_2, \overline{p}_2 + k)$, although it would make the deviation from the SES path profitable.

to $\mu_1(\cdot|h^0)$ (and $\mu_1(\cdot|(25,75))$) coincides with $S_1(z)$. Thus, $br_1(\mu_1) \subset S_1(z)$, establishing Realization-Strictness.

Now I show that $S_1^* \times S_2^*$ is non-empty. Suppose by induction that

$$S_{i,n}^* := \left\{ s_i \in S_i^n \cap S_i(z) \,\middle|\, \forall a'_{-i} \neq a_{-i}, s_i(a'_{-i}) < \overline{p}_i \right\} \neq \emptyset, \quad i = 1, 2.$$

For every μ_2 that strongly believes $S_{1,n}^*$, as shown above, the set of continuation best replies to $\mu_2(\cdot|h^0)$ and $\mu_2(\cdot|(25,75))$ coincides with $S_2(z)$, thus $br_2(\mu_2) \subset S_2(z)$. So, $S_2^{n+1} \cap S_2(z) \neq \emptyset$. Then, for each $a'_1 \neq 25$ we can define

$$p_2^{a'_1} := \min \left\{ p_2 \,\middle|\, \exists s_2 \in S_2^{n+1} \cap S_2(z), s_2(a'_1) = p_2 \right\},$$

and fix $\mu_2^{a'_1}$ that strongly believes $S_1^n, ..., S_1^0$ such that $s'_2(a'_1) = p_2^{a'_1}$ for some $s'_2 \in br_2(\mu_2^{a'_1})$. By $S_{1,n}^* \subset S_1(25)$, I can construct μ_2 that strongly believes $S_{1,n}^*, S_1^n, ..., S_1^0$ such that $\mu_2(\cdot|(a'_1,75)) = \mu_2^{a'_1}(\cdot|(a'_1,75))$ for each $a'_1 \neq 25$ (the chain rule is satisfied). Fix the unique $s_2 \in S_2(z)$ such that $s_2(a'_1) = p_2^{a'_1}$ for each $a'_1 \neq 25$; s_2 is a continuation best reply to $\mu_2(\cdot|h)$ for all $h \in H(s_2)$, thus $s_2 \in br_2(\mu_2) \subset S_2^{n+1}$. Suppose by contradiction that $S_{2,n+1}^* = \emptyset$, so $s_2 \notin S_{2,n+1}^*$. Thus, $s_2(a'_1) \geq \overline{p}_2$ for every a'_1 in a non-empty subset $\widetilde{A}_1$ of $A_1 \setminus \{25\}$. Fix μ_1 that strongly believes $S_2^n, ..., S_2^0$ with $\mu_1(s_2|h^0) = 1$. Recall that firm 1's revenues after z and at any $(a'_1, 75)$ against $\overline{p}_2$ are identical. Then, there exist $a'_1 \in \widetilde{A}_1$ and $s_1 \in br_1(\mu_1) \subset S_1^n$ such that $s_1 \in S_1(a'_1)$ and $p_1 := s_1(75)$ is the *smallest* best reply to $s_2(a'_1) = p_2^{a'_1} \geq \overline{p}_2$. With this I will show in the next paragraph that there is a best reply $p'_2 < p_2^{a'_1}$ to p_1. Then, for any μ'_2 that strongly believes $S_{1,n}^*, S_1^n, ..., S_1^0$ with $\mu'_2(s_1|(a'_1,75)) = 1$, there is $s'_2 \in br_2(\mu'_2) \subset S_2(z) \cap S_2^{n+1}$ such that $s'_2(a'_1) = p'_2$, contradicting the definition of $p_2^{a'_1}$.

If $p_1 = p_1^-(p_2^{a'_1})$, any best reply to p_1 is below $p_2^{a'_1}$. So, suppose $p_1 \neq p_1^-(p_2^{a'_1})$. Consider first $a'_1 < 25$. We have $p_2^{a'_1} \geq \overline{p}_2 = 125 - a'_1 > p_2^*$. So, $p_2^{a'_1} = p_2^* + k$ for some $k \in \mathbb{R}^+$. By $p_1 \neq p_1^-(p_2^{a'_1})$, p_1 is bounded above by $p_1^F(p_2^{a'_1}) = p_1^* + k/2$. So, the smallest best reply to p_1 is bounded above by either $p_2^-(p_1) < p_2^{a'_1}$, or

$$p_2^F(p_1^F(p_2^{a'_1})) = p_2^* + k/4 < p_2^{a'_1}.$$

Consider now $a_1' > 25$. Recall that firm 1's best reply to $\overline{p}_2$ is 100. Then, since p_1 best replies to $p_2^{a_1'} \geq \overline{p}_2$ without undercutting it, p_1 must be at least 100. For firm 2, against 100, recall that the optimal revenues are 5000 at $(25, 75)$, so when firm 1 is closer to the center they are lower than 5000 without undercutting, while they are at least 5000 with $p_2^-(100) \geq 50$. Thus, $p_2^-(100)$ best replies to 100, so a fortiori $p_2^-(p_1) < p_2^{a_1'}$ best replies to $p_1 \geq 100$.

Uniqueness Fix $(a_1, a_2) = (a_1, 100 - a_1)$ with $a_1 \neq 25$. Suppose by contradiction that there is a symmetric SES $S^* = S_1^* \times S_2^*$ inducing locations (a_1, a_2). By Self-Justifiability and Realization-strictness, S^* must prescribe at (a_1, a_2) a symmetric set of prices $P^* \times P^*$ closed under rational behavior (CURB), that is, P^* is exactly the set of prices that best reply to some belief over P^*. It is easy to see that S^∞ must prescribe a best response set (BRS) of prices $P_1 \times P_2$ at every location pair (a_1', a_2') that is compatible with S^∞, that is, P_i is contained in the set of prices that best reply to some belief over P_{-i}. In the Appendix, I show the existence of $a_1' \neq a_1$ such that for every symmetric CURB set $P \times P$ at (a_1, a_2) and every BRS $P_1 \times P_2$ at (a_1', a_2), firm 1's revenues against $\min P_2$ at (a_1', a_2) are not lower than against $\min P$ at (a_1, a_2). With this, I will show in the next paragraph that (a_1', a_2) is indeed compatible with S^∞. Then, for any $s_2 \in S_2^* \subset S_2^\infty$ with $s_2(a_1) = \min P^*$, firm i's revenues against $s_2(a_1')$ at (a_1', a_2) are not lower than against $s_2(a_1)$ at (a_1, a_2). Hence, for every μ_1 with $\mu_1(s_2|h^0) = 1$, we have $br_1(\mu_1) \not\subset S_1(a_1)$. Since μ_1 strongly believes S_2^*, this contradicts Realization-strictness.

Fix a BRS $P_1 \times P_2$ at (a_1', a_2). Fix $s_2 \in S_2^*$ such that $s_2(a_1) = \min P^*$. By Self-Justifiability, there is μ_2 that strongly believes S_1^* such that $s_2 \in br_2(\mu_2)$. By $s_2 \in S_2^\infty$, there is $\widetilde{\mu}_2$ that strongly believes $(S_1^q)_{q=0}^\infty$ such that $s_2 \in br_2(\widetilde{\mu}_2)$. So, I can construct $\widetilde{\mu}_2'$ that strongly believes $S_1^* \subset S_1^\infty \cap S_1(a_1)$ and $(S_1^q)_{q=0}^\infty$ as $\widetilde{\mu}_2'(\cdot|h^0) = \mu_2(\cdot|h^0)$ and $\widetilde{\mu}_2'(\cdot|(a_1'', a_2)) = \widetilde{\mu}_2(\cdot|(a_1'', a_2))$ for each $a_1'' \neq a_1$; clearly $s_2 \in br_2(\widetilde{\mu}_2')$. Let S_2' be the set of all $s_2' \in S_2(a_2)$ such that $s_2'(a_1') \in P_2$ and $s_2'(a_1'') = s_2(a_1'')$ for each $a_1'' \neq a_1'$. For each $p_1 \in P_1$, let $S_1'(p_1)$ be the set of all $s_1' \in S_1(a_1')$ with $s_1'(a_2) = p_1$.

Suppose by induction that $S_2' \subset S_2^n$ and $S_1'(p_1) \cap S_1^n \neq \emptyset$ for each $p_1 \in P_1$.

Fix $s_2' \in S_2'$ and a probability distribution ν over P_1 such that $s_2'(a_1')$ best replies to ν. Construct μ_2' that strongly believes $S_1^n, ..., S_1^0$ such that $\mu_2'(\cdot|h) = \widetilde{\mu}_2'(\cdot|h)$ for all $h \neq (a_1', a_2)$, $\mu_2'(S_1'(p_1)|(a_1', a_2)) = \nu(p_1)$ for all $p_1 \in P_1$. Thus, for each $h \in H(s_2) = H(s_2')$ with $h \neq (a_1', a_2)$, we have $\mu_2'(S_1(a_1')|h) = 0$. Note also that s_2' and s_2 yield the same outcome against each $s_1 \notin S_1(a_1')$. Then, since s_2 is a continuation best reply to $\widetilde{\mu}_2'(\cdot|h) = \mu_2'(\cdot|h)$, so is s_2'. Hence, $s_2' \in br_2(\mu_2') \subset S_2^{n+1}$.

Fix $p_1 \in P_1$ and a probability distribution ν over P_2 such that p_1 best replies to ν. Construct μ_1' that strongly believes $S_2^n, ..., S_2^0$ such that, for each $p_2 \in P_2$, $\mu_1'(s_2'|h^0) = \nu(p_2)$, where s_2' is the unique $\widetilde{s}_2 \in S_2'$ with $\widetilde{s}_2(a_1') = p_2$. Fix μ_1 with $\mu_1(s_2|h^0) = 1$. Since μ_1 strongly believes S_2^*, by Realization-strictness we have $br_1(\mu_1) \subset S_1(a_1)$. By construction of S_2', for each $a_1'' \neq a_1'$, $\mu_1'(\cdot|(a_1'', a_2))$ gives probability 1 to $s_2(a_1'')$ like $\mu_1(\cdot|(a_1'', a_2))$. Hence, $br_1(\mu_1') \subset S_1(a_1) \cup S_1(a_1')$. But then, since firm 1 has non-lower revenues against $\min P_2$ at (a_1', a_2) than against $s_2(a_1) = \min P^*$ at (a_1, a_2), we have $br_1(\mu_1') \cap S_1(a_1') \neq \emptyset$. So, there is $s_1' \in br_1(\mu_1') \subset S_1^{n+1}$ such that $s_1'(a_2) = p_1$.

7 Discussion - epistemic priority orderings

7.1 Epistemic priority to the agreement

The literature on strategic reasoning with first-order belief restrictions is mostly based on Strong-Δ-Rationalizability (Battigalli [7], Battigalli and Siniscalchi [11]). The predictions of this paper are typically stronger than in this literature for three reasons: (i) the adoption of Agreement-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.

Battigalli and Friedenberg [8] capture the implications of Strong-Δ-Rationalizability across *all* first-order belief restrictions with the notion of Extensive-Form Best Response Set. Essentially, an EFBRS is a set of plans where each plan can be justified under strong belief in co-players' plans. A comparison between

EFBRS and SES clarifies the three differences between the two approaches.

First, while a SES features only rationalizable plans, an EFBRS needs not. This is because Strong-Δ-Rationalizability allows players to stop believing in the rationality of a co-player who makes a move that cannot be optimal under her belief restrictions. Here I call this hypothesis *epistemic priority to the agreement* (as opposed to *rationality*).

Second, the belief restrictions that yield an EFBRS can impose specific randomizations, or differ across two players regarding the moves of a third player. SES's capture instead the material implications of the belief in the SES itself. Just like the belief in any agreement, this only restricts the support of beliefs about co-players' behavior, in the same way for every player.

Third, an EFBRS may induce a larger set of outcomes with respect to what players expect under the belief restrictions that yield the EFBRS itself. Realization-strictness, and more generally self-enforceability, rule this out.[36]

All this translates into a significant difference in predictive power. For instance, competition among firms on price, quantity, or quality often leads to a unique outcome under common belief in rationality (see cobweb stability or Cournot duopoly), but this predictive power is lost in subgames where orders of belief in rationality are dropped by Strong-Δ-Rationalizability. I show in [16] that in the Hotelling model almost every location pair is induced by some EFBRS. In Supplemental Appendix V, I show that the analysis of Sections 4 and 5 can be replicated verbatim under priority to agreement with mere rationality in place of rationalizability and Strong-Δ-Rationalizability in place of Agreement-rationalizability, and that this expands the collection of implementable outcome sets.

[36] Relatedly, by Corollary 2, an implementable outcome is induced by a pure Nash equilibrium, whereas the outcome prescribed by a merely credible agreement can be just a self-confirming equilibrium outcome (Fudenberg and Levine [21]). This is because under a self-enforcing agreement players have the incentive to stay on path for *all* their refined beliefs, so in particular under belief in one profile of plans.

7.2 Epistemic priority to the path

Consider the twofold repetition of the following game, which is solved formally in Supplemental Appendix V.

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

Suppose that Ann and Bob agree on the SPE where Bob works in the first period and Ann works in the second period (no matter what happened in the first). However, suppose that not just Bob, but also Ann actually works in the first period. Then, if Bob believes that Ann is rational, he must conclude that she has not believed that he plays as in the SPE. At this point, in the baseline analysis, Bob is free to believe that Ann expected him to free ride in the first period. Then, Bob can believe that Ann expects him to free-ride in the second period and thus will work. So, he will free-ride as agreed. Suppose now instead that Bob believes that Ann did expect him to work in the first period. Then, Bob must believe that Ann expects him to work again after her deviation and thus will free-ride, for otherwise her deviation would not be profitable. So, he will work. If Ann anticipates that Bob will interpret the deviation in this way, she has incentive to deviate. The agreement is not credible.

In the example, when Bob cannot believe anymore that Ann believes in the whole agreement, he keeps the belief that Ann believed in the agreed-upon path, and drops the belief that Ann believes in the threat. Still assigning the highest epistemic priority to rationality, when this further epistemic priority choice is transparent to players, I say there is *epistemic priority to the path*. In Supplemental Appendix V, I operationalize this finer epistemic priority ordering with a variation of Agreement-rationalizability. With this, I show that the analysis of Sections 4 and 5 can be replicated verbatim under priority to the path, and that this refines the set of implementable outcomes.

8 Conclusion

I develop a novel methodology to assess the implications of pre-play communication in a dynamic game. I introduce a notion of agreement that, differently from equilibrium, is able to capture agreement incompleteness. In particular, while a subgame perfect equilibrium specifies one (mixed) action for each player in every contingency, an agreement can remain vague, or completely silent, regarding players' behavior in some contingencies. Leaving the behavior of deviators unspecified is a natural form of incompleteness. Then, a self-enforcing agreement needs not rely on coordinated play after deviations. While this rescues many intuitive threats (and consequently, paths), one may worry that it makes the notion of self-enforcing agreement too permissive. This is not the case, for two reasons. First, players' beliefs are refined with strategic reasoning. This restricts the beliefs regarding the continuation play of a deviator and challenges the belief in the agreement. Second, whenever players cannot coordinate on a precise path, or formulate precise threats, self-enforceability requires deviations to be suboptimal for all the refined beliefs, instead of neutralizing the uncertainty with one probability distribution. In the Hotelling model, the existence of an intuitive solution is guaranteed by a natural non-equilibrium threat, and its uniqueness by strategic reasoning and the ineliminable residual uncertainty.

In applications, an economist may need to carry out two types of analysis. Sometimes, there are obvious constraints to the extent of pre-play communication, or there are specific agreements of interest. For instance, an institution may be in the position to make a public announcement, while other players may not have this opportunity; or, some relevant agreement among parties has been reached and one would like to predict its consequences. Other times, an economist would like to start off by considering all possible agreements among players. The methodology developed in this paper satisfies both needs. The self-enforceability of a specific agreement can be determined with *Agreement-rationalizability*, a refinement of extensive-form rationalizability that captures strategic reasoning based not just on the beliefs in rationality but also on the

beliefs in the agreement. The search for all the outcomes induced by self-enforcing agreements, and for agreements that induce them, revolves around a set-valued solution concept called *Self-Enforcing Set*. A SES is a set of extensive-form rationalizable plans of actions that satisfies three simple conditions and coincides with a self-enforcing agreement that does not specify the behavior of deviators. In games with two players or two stages, all the outcomes induced by self-enforcing agreements are also induced by SES's. In the Hotelling model, it would be virtually impossible to establish uniqueness if one had to evaluate all possible agreements, instead of just the SES's. In games with more than two players and stages, some outcomes can be enforced only by augmenting SES's with restrictions to the behavior of deviators. *Tight agreements* offer a canonical way to do this, and fully characterize the outcomes induced by self-enforcing agreements in all games. Since for a tight agreement the announced outcomes and the induced outcomes coincide, a "revelation principle" for agreements design follows.

9 Appendix

9.1 Proofs for Section 5

Proof of Proposition 2. Realization-strictness: Fix $i \in I$ and μ_i that strongly believes $S^*_{-i} = S^\infty_{-i,e} \cap e^0_{-i}$. By Remark 1, there exists $\widetilde{\mu}_i$ that, for each $j \neq i$, strongly believes $e^0_j, ..., e^{k_j}_j$ and $(S^q_{j,e})^\infty_{q=0}$. By $e^{k_j}_j \subseteq S^\infty_j$, I can let $\widetilde{\mu}_i$ strongly believe also $(S^q_j)^\infty_{q=0}$. Since μ_i strongly believes S^*_{-i}, I can construct $\mu'_i \in \Delta^e_i$ that strongly believes $((S^q_{j,e})_{j\neq i})^\infty_{q=0}$ and $((S^q_j)_{j\neq i})^\infty_{q=0}$ as $\mu'_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S^*_{-i})$, $\mu'_i(\cdot|h) = \widetilde{\mu}_i(\cdot|h)$ for all $h \notin H(S^*_{-i})$. Thus, $br_i(\mu'_i) \subseteq S^\infty_{i,e}$. Every $s_i \in br_i(\mu_i)$ is a continuation best reply to $\mu'_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S^*_{-i}) \cap H(s_i)$. Then, for every $s_i \in br_i(\mu_i)$, there is $s'_i \in br_i(\mu'_i)$ such that $s'_i(h) = s_i(h)$ for all $h \in H(S^*_{-i}) \cap H(s_i)$.[37] So,

$$\zeta(br_i(\mu_i) \times S^*_{-i}) \subseteq \zeta(br_i(\mu'_i) \times S^*_{-i}) \subseteq \zeta(S^\infty_{i,e} \times S^\infty_{-i,e}) = \zeta(S^*),$$

where the last equality is by self-enforceability of e.

Self-Justifiability: Fix $i \in I$ and $s_i \in S^*_i \subseteq S^\infty_{i,e}$. By Remark 1, every $s_i \in S^\infty_{i,e}$ is a sequential best reply to some $\mu_i \in \Delta^e_i$ that strongly believes $((S^q_{j,e})_{j\neq i})^\infty_{q=0}$, thus that strongly believes $(S^\infty_{j,e})_{j\neq i}$, $(S^0_{j,e})_{j\neq i} = (S^\infty_j)_{j\neq i}$, and $(e^0_j)_{j\neq i}$. Then, μ_i strongly believes also $(S^\infty_{j,e} \cap e^0_j)_{j\neq i} = (S^*_j)_{j\neq i}$. ■

Proof of Proposition 3. Let $z := \zeta(S^*)$.

If: Fix $i \in I$ and μ_i that strongly believes S^*_{-i}. Fix $s_i \notin S_i(z)$ and $s^*_i \in S^*_i$. For every $s^*_{-i} \in S^*_{-i}$, since (s^*_i, s^*_{-i}) is a realization-strict Nash, $u_i(\zeta(s_i, s^*_{-i})) < u_i(z) = u_i(\zeta(s^*_i, s^*_{-i}))$. So, s_i is a worse reply than s^*_i to $\mu_i(\cdot|h^0)$. Hence, $br_i(\mu_i) \subseteq S_i(z)$. With $S^*_{-i} \subseteq S_{-i}(z)$, we obtain $\zeta(br_i(\mu_i) \times S^*_{-i}) = \{z\}$.

Only if: Fix $(s^*_j)_{j\in I} \in S^*$ and $i \in I$. Fix $s_i \in \arg\max_{s'_i} u_i(\zeta(s'_i, s^*_{-i}))$ and μ_i that strongly believes S^*_{-i} with $\mu_i(s^*_{-i}|h^0) = 1$. For each $h \prec \zeta(s_i, s^*_{-i})$, $\mu_i(s^*_{-i}|h) = 1$, so s_i is a continuation best reply to $\mu_i(\cdot|h)$. Then,[38] there exists

[37] If for every $h \in H(s'_i)$ there is a continuation best reply $\widetilde{s}_i$ to $\mu'_i(\cdot|h)$ such that $s'_i(h) = \widetilde{s}_i(h)$, then s'_i is a sequential best reply to μ'_i. This is because no matter which optimal actions are planned at future histories, the expected payoff of an action at the current history is always the same. This allows to construct the desired s'_i.

[38] See footnote 37 for the argument.

$s_i' \in br_i(\mu_i)$ such that $s_i'(h) = s_i(h)$ for every $h \prec \zeta(s_i, s_{-i}^*)$. Hence, $\zeta(s_i, s_{-i}^*) = \zeta(s_i', s_{-i}^*)$. By Realization-strictness, $\zeta(s_i', s_{-i}^*) = z$. Thus, $s_i \in S_i(z)$. For every $\widetilde{s}_i \in S_i(z)$, $\zeta(\widetilde{s}_i, s_{-i}^*) = z$. So, $\arg\max_{s_i'} u_i(\zeta(s_i', s_{-i}^*)) = S_i(z)$. ■

Proof of Proposition 4. By definition, $\overline{S}^* = S_e^1$. Then, by Forward Induction, for each $i \in I$ and $s_i \in \overline{S}_i^* = S_{i,e}^1$, there is $\mu_i \in \Delta_i^e$ that strongly believes $(\overline{S}_j^*)_{j \neq i} = (S_{j,e}^1)_{j \neq i}$ and $(S_j^\infty)_{j \neq i} = (S_{j,e}^0)_{j \neq i}$ such that $s_i \in br_i(\mu_i) \subseteq S_{i,e}^2$. Thus, $S_e^1 = S_e^2$. By induction, $S_e^1 = S_e^\infty$. Hence, (i) $\overline{S}^* = S_e^\infty$.

Fix $s = (s_i)_{i \in I} \in \overline{S}^*$. For all $i \in I$, $s_i \in br_i(\mu_i)$ for some μ_i that strongly believes $(S_j^*)_{j \neq i}$ and thus S_{-i}^*. Then, for all $h \in H(S^*) \cap H(s)$, we must have $s_i(h) = s_i'(h)$ for some $s_i' \in S_i^* \cap S_i(h)$, otherwise we would have $\zeta(s_i, s_{-i}') \notin \zeta(S^*)$ for any $s_{-i}' \in S_{-i}^* \cap S_{-i}(h)$, violating Realization-strictness. So, $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$. By Self-Justifiability, $S^* \subseteq \overline{S}^*$. With (i), we get (ii) $\zeta(S_e^\infty) = \zeta(S^*)$.

By $S^* \subseteq \overline{S}^*$ and (i), we get (iii) $S^* = S_e^\infty \cap S^*$.

By (ii) and (iii), $\zeta(S_e^\infty) = \zeta(S^*) = \zeta(S_e^\infty \cap S^*)$: e is credible, self-enforcing and truthful. ■

Proof of Proposition 5. Note preliminarly the following facts. As shown in the proof of Proposition 4, by Realization-strictness $\zeta(\overline{S}^*) \subseteq \zeta(S^*)$, thus $H(\overline{S}^*) \subseteq H(S^*)$. Moreover, by Self-Justifiability, $S_j^* \subseteq \overline{S}_j^*$ for each $j \in I$.

Fix $i \in I$ and $s_i \in \overline{S}_i^*$. By definition of $\overline{S}_i^*$, there is μ_i that strongly believes $(S_j^*)_{j \neq i}$ and $(S_j^\infty)_{j \neq i}$ such that $s_i \in br_i(\mu_i)$.

Consider first a game with two stages. Fix $j \neq i$. By $H(\overline{S}^*) \subseteq H(S^*)$, every move allowed by $\overline{S}_j^*$ at h^0 must be allowed also by S_j^*, that is, every history of length one compatible with $\overline{S}_j^*$ is compatible also with S_j^*. Since the game has two stages, all the longer histories are terminal, therefore $H(\overline{S}_j^*) \subseteq H(S_j^*)$. Then, strong belief in $S_j^* \subseteq \overline{S}_j^*$ implies strong belief in $\overline{S}_j^*$. Thus, μ_i strongly believes also $\overline{S}_j^*$. Hence, μ_i verifies Forward Induction.

Consider now a game with two players. Let j be i's co-player. By $S_j^* \subseteq \overline{S}_j^* \subseteq S_j^\infty$, I can construct μ_i' that strongly believes S_j^*, $\overline{S}_j^*$, and S_j^∞ such that $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S_j^*)$ and all $h \notin H(\overline{S}_j^*)$. For each $h \in H(\overline{S}_i^*)$, since j is the only co-player, either $h \notin H(\overline{S}_j^*)$, or $h \in H(\overline{S}^*) \subseteq H(S^*) \subseteq H(S_j^*)$. So, $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(\overline{S}_i^*) \supseteq H(s_i)$. Thus, $s_i \in br_i(\mu_i')$. ■

Proof of Proposition 6. Let $\zeta(S^*) = \{z\}$. Fix $i \in I$ and $s_i \in S_i^*$. By $s_i \in S_i^\infty$, there exists μ_i that strongly believes S_j^∞ such that $s_i \in br_i(\mu_i)$. Fix $s_j \in S_j^*$ and construct μ_i' that strongly believes S_j^* and S_j^∞ such that $\mu_i'(s_j|h^0) = 1$ and $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \notin H(S_j^*)$. By Realization-strictness, Proposition 3 implies that (s_i, s_j) is a Nash equilibrium. For every $h \prec z$, since $\mu_i'(s_j|h^0) = 1$ and $s_j \in S_j(z)$, $\mu_i'(s_j|h) = 1$ as well. Hence, s_i is a continuation best reply to $\mu_i'(\cdot|h)$. For every $h \in H(s_i) \subseteq H(S_i(z))$ with $h \not\prec z$, since j is the only co-player, $h \notin H(S_j(z)) \supseteq H(S_j^*)$, thus $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$. Hence, s_i is a continuation best reply to $\mu_i'(\cdot|h)$. So, $s_i \in br_i(\mu_i')$. ■

Proof of Remark 2. Fix $i \in I$ and $s_i \in e_i^0$. By T3 and $e_i^0 \subseteq S_i^\infty$, $e_i^0 = e_i^0 \cap S_i(h^0) \subseteq br_i(\Delta_i^e) \cap S_i^\infty$. So, $s_i \in S_i^\infty \cap br_i(\mu_i)$ for some $\mu_i \in \Delta_i^e$, which strongly believes $(e_j^0)_{j\neq i}$. There remains to show that μ_i strongly believes $(S_j^\infty)_{j\neq i}$. Fix $j \neq i$ and $h \in H(S_j^\infty)$. By T2, $e_j^{k_j} \cap S_j(h) \neq \emptyset$. Since μ_i strongly believes $e_j^{k_j} \subseteq S_j^\infty$, we get $\mu_i(S_j^\infty \times S_{-j,i}|h) = 1$, where $S_{-j,i} := S_{I\setminus\{i,j\}}$. ■

Proof of Proposition 7. As shown in the proof of Remark 2, by T2 every $\mu_i \in \Delta_i^e$ strongly believes $(S_j^\infty)_{j\neq i}$. Hence, for each $i \in I$, $br_i(\Delta_i^e) \cap S_i^\infty = S_{i,e}^1$.

Fix $i \in I$ and $\mu_i \in \Delta_i^e$. For each $j \neq i$ and $h \in H(S_{j,e}^1) = H(br_j(\Delta_j^e) \cap S_j^\infty)$, by T3 there is $n \leq k_j$ such that $\emptyset \neq e_j^n \cap S_j(h) \subseteq br_j(\Delta_j^e) \cap S_j^\infty = S_{j,e}^1$ (recall that $e_j^n \subseteq S_j^\infty$). Then, since μ_i strongly believes e_j^n, $1 = \mu_i(e_j^n \times S_{-j,i}|h) \leq \mu_i(S_{j,e}^1 \times S_{-j,i}|h)$. Thus, μ_i strongly believes $(S_{j,e}^1)_{j\neq i}$, besides $(S_j^\infty)_{j\neq i}$. Hence, $S_{i,e}^2 = br_i(\Delta_i^e) \cap S_i^\infty = S_{i,e}^1$ for each $i \in I$. By induction, we get (i) $S_e^1 = S_e^\infty$.

Fix $s = (s_i)_{i\in I} \in S_e^1$. For all $i \in I$, $s_i \in br_i(\mu_i)$ for some μ_i that strongly believes $(e_j^0)_{j\neq i}$ and thus e_{-i}^0. Then, for all $h \in H(e^0) \cap H(s)$, we must have $s_i(h) = s_i'(h)$ for some $s_i' \in e_i^0 \cap S_i(h)$, otherwise we would have $\zeta(s_i, s_{-i}') \notin \zeta(e^0)$ for any $s_{-i}' \in e_{-i}^0 \cap S_{-i}(h)$, violating Realization-strictness (T1). Hence, $\zeta(S_e^1) \subseteq \zeta(e^0)$. By T3, we have $e_i^0 = e_i^0 \cap S_i(h^0) \subseteq br_i(\Delta_i^e) \cap S_i^\infty = S_{i,e}^1$ for each $i \in I$, thus $e^0 \subseteq S_e^1$. So, $\zeta(S_e^1) = \zeta(e^0)$. With (i), we get (ii) $\zeta(S_e^\infty) = \zeta(e^0)$.

By $e^0 \subseteq S_e^1$ and (i), we get (iii) $e^0 = S_e^\infty \cap e^0$.

By (ii) and (iii), $\zeta(S_e^\infty) = \zeta(e^0) = \zeta(S_e^\infty \cap e^0)$: e is credible, self-enforcing and truthful. ■

Proof of Theorem 3 (Only if). Fix an implementable outcome set P and a self-enforcing agreement $e = (e_i)_{i \in I}$ that implements it. Let M be the smallest $m \geq 0$ such that $S_e^\infty = S_e^m$ (it exists by finiteness of the game).

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

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

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

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

So, $\overline{e}_i^q$ is the union of the boxes along the line that connects box $e_i^q \cap S_{i,e}^M$ with box $e_i^0 \cap S_{i,e}^{M-q}$. Starting from $\overline{e}_i^0 = e_i^0 \cap S_{i,e}^M$, every increase of q by 1 shifts the line by 1 towards south-east, until $\overline{e}_i^{k_i+M+1} = e_i^{k_i+1} \cap S_{i,e}^0 = S_i^\infty$. The boxes north-west of the line are subsets of the boxes along the line.

Without loss of generality, suppose that $\overline{e}_i^n \subsetneq \overline{e}_i^{n+1}$ for every n.[39] Then, $\overline{e} = ((\overline{e}_i^0, ..., \overline{e}_i^{k_i+M+1}))_{i \in I}$ is an agreement, and it prescribes P because

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

where the first equality is by implementation of P, the second by self-enforceability of e, and the third by construction. By $\overline{e}_i^{k_i+M+1} = S_i^\infty$ for every $i \in I$, $\overline{e}$ satisfies T2. Since e is self-enforcing, by Proposition 2, $\overline{e}^0 = e^0 \cap S_e^M$ satisfies

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

Realization-strictness, thus $\overline{e}$ satisfies T1. Finally, I show that $\overline{e}$ satisfies T3.

For each $j \in I$, $S^M_{j,e}$ is the set of all $s_j \in S^\infty_j$ such that $s_j \in br_j(\mu_j)$ for some μ_j that strongly believes $((S^q_{i,e})_{i\neq j})^M_{q=0}$ and $((e^q_i)^{k_i}_{q=0})_{i\neq j}$. I am going to show that, for each $i \neq j$, strong belief in $(\overline{e}^q_i)^{k_i+M+1}_{q=0}$ is equivalent to strong belief in $(S^q_{i,e})^M_{q=0}$ and $(e^q_i)^{k_i}_{q=0}$. Then, $S^M_{j,e} = br_j(\Delta^{\overline{e}}_j) \cap S^\infty_j$, which will be useful later.

First, I show that every μ_j that strongly believes $(S^q_{i,e})^M_{q=0}$ and $(e^q_i)^{k_i}_{q=0}$ strongly believes also $(\overline{e}^q_i)^{k_i+M+1}_{q=0}$. Since $e^{k_i+1}_i = S^\infty_i = S^0_{i,e}$, μ_j strongly believes also $e^{k_i+1}_i$. Fix $q \in \{0, ..., k_i + M + 1\}$. For each $h \in H(\overline{e}^q_i)$, by construction $h \in H(e^n_i \cap S^m_{i,e})$ for some n and m with $e^n_i \cap S^m_{i,e} \subseteq \overline{e}^q_i$. Since μ_j strongly believes e^n_i and $S^m_{i,e}$, we have $1 = \mu_j((e^n_i \cap S^m_{i,e}) \times S_{-j,i}|h) \leq \mu_j(\overline{e}^q_i \times S_{-j,i}|h)$. Hence, μ_j strongly believes $\overline{e}^q_i$.

Second, I show that every μ_j that strongly believes $(\overline{e}^q_i)^{k_i+M+1}_{q=0}$ strongly believes also $(e^q_i)^{k_i}_{q=0}$ and $(S^q_{i,e})^M_{q=0}$.

Fix $n = 0, ..., k_i$ and $h \in H(e^n_i)$. Fix the highest $m \in \{0, ..., M\}$ such that $h \in H(S^m_{i,e})$ (it exists because $S^0_{i,e} = S^\infty_i \supseteq e^n_i$). By Remark 1, there exists μ'_j that strongly believes $(S^q_{i,e})^M_{q=0}$ and $(e^q_i)^{k_i}_{q=0}$, and thus $\mu'_j(e^n_i \times S_{-j,i}|h) = \mu'_j(S^m_{i,e} \times S_{-j,i}|h) = 1$. Hence, $e^n_i \cap S^m_{i,e} \cap S_i(h) \neq \emptyset$. By construction, $e^n_i \cap S^m_{i,e} \subseteq \overline{e}^{M-m+n}_i$. So, $\overline{e}^{M-m+n}_i \cap S_i(h) \neq \emptyset$. Recall that $\overline{e}^{M-m+n}_i$ is the union of sets $e^{n'}_i \cap S^{m'}_{i,e}$ with $n' + (M - m') = n + (M - m)$. If $n' < n$, then $e^{n'}_i \subset e^n_i$, and if $n' > n$, then $m' > m$, thus $S^{m'}_{i,e} \cap S_i(h) = \emptyset$ by definition of m. So, $\overline{e}^{M-m+n}_i \cap S_i(h) \subseteq e^n_i$. Since μ_j strongly believes $\overline{e}^{M-m+n}_i$, we have $1 = \mu_j(\overline{e}^{M-m+n}_i \times S_{-j,i}|h) \leq \mu_j(e^n_i \times S_{-j,i}|h)$. So, μ_j strongly believes e^n_i.

Fix $m = 0, ..., M$ and $h \in H(S^m_{i,e})$. Fix the lowest $n \in \{0, ..., k_i + 1\}$ such that $h \in H(e^n_i)$ (it exists because $e^{k_i+1}_i = S^\infty_i \supseteq S^m_{i,e}$). By Remark 1, there exists μ'_j that strongly believes $(S^q_{i,e})^M_{q=0}$ and $(e^q_i)^{k_i}_{q=0}$, and thus $\mu'_j(e^n_i \times S_{-j,i}|h) = \mu'_j(S^m_{i,e} \times S_{-j,i}|h) = 1$. Hence, $e^n_i \cap S^m_{i,e} \cap S_i(h) \neq \emptyset$. By construction, $e^n_i \cap S^m_{i,e} \subseteq \overline{e}^{M-m+n}_i$. So, $\overline{e}^{M-m+n}_i \cap S_i(h) \neq \emptyset$. Recall that $\overline{e}^{M-m+n}_i$ is the union of sets $e^{n'}_i \cap S^{m'}_{i,e}$ with $n' + (M - m') = n + (M - m)$. If $m' > m$, then $S^{m'}_{i,e} \subseteq S^m_{i,e}$, and if $m' < m$, then $n' < n$, thus $e^{n'}_i \cap S_i(h) = \emptyset$ by definition of n. So, $\overline{e}^{M-m+n}_i \cap S_i(h) \subseteq S^m_{i,e}$. Since μ_j strongly believes $\overline{e}^{M-m+n}_i$, we have $1 = \mu_j(\overline{e}^{M-m+n}_i \times S_{-j,i}|h) \leq \mu_j(S^m_{i,e} \times S_{-j,i}|h)$. So, μ_j strongly believes $S^m_{i,e}$.

With $m = M$, the previous paragraph shows that, for each $i \in I$ and $h \in H(S_{i,e}^M)$, there is q such that $\emptyset \neq \overline{e}_i^q \cap S_i(h) \subseteq S_{i,e}^M$. So, since $S_{i,e}^M = br_i(\Delta_i^{\overline{e}}) \cap S_i^\infty$, $\overline{e}$ satisfies T3. ■

9.2 Appendix to Section 6

Here I show the properties of CURB sets and BRS's of prices used for the uniqueness result. Fix $a_1 \neq 25$ and let $a_2 = 100 - a_1$; I show the existence of $a_1' \neq a_1$ such that, for every symmetric CURB set $P \times P$ at (a_1, a_2) and every BRS $P_1 \times P_2$ at (a_1', a_2), firm 1's revenues at (a_1, a_2) against $\min P$ are not higher than at (a_1', a_2) against $\min P_2$. At (a_1, a_2) I will drop the subscript for the prices that, by symmetry, are equal for the two firms.

Let first $a_1 < 25$ and $a_2 = 100 - a_1$. Let $a_1' = 25$.

Claim: At (a_1, a_2), for every CURB set $P \times P$, $\min P \leq p^*$.
Proof: For every $p = p^* + k$ with $k > 0$, we have $p^F(p) = p^* + k/2 < p$. Then, there is a best reply to p lower than p. So, $\min P < p$. ■

Claim: At $(25, a_2)$, for every BRS $P_1 \times P_2$, $\min P_2 \geq p_2^* - 1$.[40]
Proof: Fix i such that $\max P_i - p_i^* \geq \max P_{-i} - p_{-i}^*$. For each $k > 1$, $p_i = p_i^* + k$ is never a best reply to a belief over $\left[0, p_{-i}^* + k\right] \cap \mathbb{N}_0$: for each $p_{-i} \leq p_{-i}^* + k$, we have $p_i > p_{-i} - (a_2 - 25)$, and $p_i - 1 > p_i^*$ is closer than $p_i = p_i^* + k$ to $p_i^F(p_{-i}) \leq p_i^F(p_{-i}^* + k) = p_i^* + k/2$, so either $p_i - 1$, or price 1 (if p_i brings zero demand) are better replies than p_i. Then, $\max P_i - p_i^*$ cannot be k, and thus $\max P_j \leq p_j^* + 1$ for each $j = 1, 2$.

Now fix i such that $p_i^* - \min P_i \geq p_{-i}^* - \min P_{-i}$. I show that for each $k > 1$, $p_i = p_i^* - k$ is dominated over $\widetilde{P}_{-i} := \left[p_{-i}^* - k, \max P_{-i}\right] \cap \mathbb{N}_0$. Then, $1 \geq p_i^* - \min P_i \geq p_2^* - \min P_2$, as desired.

[40] In the proof of this and of the next claim, I will implicitly use the expressions for (p_1^*, p_2^*) at $(25, a_2)$, so I report them here for reader's convenience:

$$(p_1^*, p_2^*) = \left(\frac{225 + a_2}{3}, \frac{375 - a_2}{3}\right)$$

.

If $0 < p_i \leq p^*_{-i} - (a_2 - 25)$, I show that $2p_i$ dominates p_i. We have

$$\begin{aligned} 2p_i &\leq 2p^*_{-i} - 2(a_2 - 25) < p^*_i, & (1) \\ 2p_i &\leq p^*_i - k + p^*_{-i} - (a_2 - 25) < p^*_{-i} - k + (a_2 - 25), & (2) \end{aligned}$$

where the strict inequalities are equalities for $a_2 = 75$, thus verified for $a_2 > 75$. For each $p_{-i} \in \left[p^*_{-i} - k, p_i + (a_2 - 25)\right) \cap \mathbb{N}_0$, $2p_i$ is closer than $p_i = p^*_i - k$ to $p^F_i(p_{-i}) \geq p^F_i(p^*_{-i} - k) = p^*_i - k/2$ by (1), and not larger than $p^+_i(p_{-i})$ by (2). For each $p_{-i} \geq p_i + (a_2 - 25)$, p_i brings revenues of at most $100p_i$, and $2p_i$ brings demand higher than 50, because firm 1 is closer to the center and

$$2p_i \leq p_i + p^*_{-i} - (a_2 - 25) < p_i + (a_2 - 25) \leq p_{-i}, \quad (3)$$

where the strict inequality is an equality for $a_2 = 75$, thus verified for $a_2 > 75$.

If $p_i > p^*_{-i} - (a_2 - 25)$, by $\max P_{-i} \leq p^*_{-i} + 1$ we get $p_i > p^-_i(\max P_{-i})$.

When $p_i > \max P_{-i} - (a_2 - 25)$, $p_i + 1$ dominates p_i: for each $p_{-i} \in \widetilde{P}_{-i}$,

$$p_i + 1 = p^*_i - k + 1 < p^+_i(p^*_{-i} - k) \leq p^+_i(p_{-i}),$$

and since $p_i + 1 < p^*_i$ by $k > 1$, $p_i + 1$ is closer than $p_i = p^*_i - k$ to $p^F_i(p_{-i}) \geq p^F_i(p^*_{-i} - k) = p^*_i - k/2$.

When $p_i = \max P_{-i} - (a_2 - 25)$, note preliminarly that prices $0, 1, 2$ are dominated by 4, so suppose that $p_i > 2$. I argue that p_i is dominated by $p'_i := 2(p_i - 1) > p_i$. Note that p'_i has already been shown to satisfy (1)-(2)-(3), because by $\max P_{-i} \leq p^*_{-i} + 1$, we get $p_i - 1 \leq p^*_{-i} - (a_2 - 25)$. For each $p_{-i} \in \widetilde{P}_{-i}$ with $p_{-i} < \max P_{-i}$ the argument above that uses (1) and (2) applies. For $p_{-i} = \max P_{-i}$, by (3) p'_i brings demand higher than $(25 + a_2)/2$, thus the difference in revenues between p'_i and p_i is higher than

$$2(p_i - 1)\left(\frac{25 + a_2}{2}\right) - p_i\left(a_2 + \frac{100 - a_2}{2}\right) = a_2\left(\frac{p_i}{2} - 1\right) - 25(1 + p_i) > 0,$$

where the inequality comes from $a_2 > 75$ and $p_i > p^*_{-i} - (a_2 - 25) > 16$. ■

Claim: Firm 1 has non-lower revenues against any integer $p_2 \geq p_2^* - 1$ at $(25, a_2)$ than against p_2^* at (a_1, a_2).
Proof: Firm 1's best reply to $p_2^* = 100$ at (a_1, a_2) is $p_1^* = 100$, with demand 50. With $p_1 = 100$, firm 1 gets demand at least 50 at $(25, a_2)$ against the smallest integer $p_2 \geq p_2^* - 1$, which is 99 if $a_2 = 76$, and at least $(372 - a_2)/3 > 100 - (a_2 - 75)$ otherwise. ■

The combination of the three claims yields the desired result for $a_1 < 25$.

Now let $a_1 > 25$, $a_2 = 100 - a_1$. Let $a_1' = 0$. Note preliminarly these facts.

Fact: (i) it is optimal to undercut any $p \geq \lfloor \overline{p} \rfloor + 3$ at (a_1, a_2);
(ii) it is not optimal to undercut any $p_2 \leq \lfloor \overline{p}_2 \rfloor + 1$ at $(0, a_2)$.
Proof: Note that, at any locations $(\widetilde{a}_1, \widetilde{a}_2)$,[41]

$$100 \cdot (\overline{p}_2 - (\widetilde{a}_2 - \widetilde{a}_1)) = p_1^F(\overline{p}_2) \left(\frac{1}{2} p_1^F(\overline{p}_2) \right). \tag{4}$$

Fix $p_2 = \overline{p}_2 + k$ with $k \in [-\overline{p}_2, 200]$ — for higher k is obviously optimal to undercut. Firm 1's revenues from $p_1^-(p_2)$ differ from the left-hand side of (4) by $100k - 100$. Firm 1's revenues without undercutting are bounded above by

$$\max_{p_1} p_1 D_1(p_1, p_2) = p_1^F(p_2) \left(\frac{1}{2} p_1^F(p_2) \right) = \frac{1}{2} \left(p_1^F(\overline{p}_2) + \frac{k}{2} \right)^2, \tag{5}$$

and with $p_1 = \left\lfloor p_1^F(p_2) \right\rfloor$, if $p_2 - (\widetilde{a}_2 - \widetilde{a}_1) < p_1^F(p_2) < p_2 + (\widetilde{a}_2 - \widetilde{a}_1)$, they are bounded below by

$$\left(p_1^F(p_2) - \frac{1}{2} \right) \left(\frac{1}{2} p_1^F(p_2) \right) = \frac{1}{2} \left(p_1^F(\overline{p}_2) + \frac{k}{2} \right)^2 - \frac{1}{4} \left(p_1^F(\overline{p}_2) + \frac{k}{2} \right). \tag{6}$$

At (a_1, a_2), note first that $\overline{p} < 97$ (compute it for $a_1 = 26$ and note that it is decreasing in a_1). Then, if $k \in [2, 200]$, (5) differs from the right-hand side of

[41] Indeed, $\overline{p}_{-i}$ is derived in [16] as the price p_{-i} that equalizes firm i's supremum of the revenues from undercutting and the optimal revenues without undercutting in case of an interior solution (see equation 5).

(4) by

$$\frac{1}{2}\left(kp^F(\overline{p})+\frac{1}{4}k^2\right) < \left(49.25+\frac{1}{8}k\right)k < 100k-100,$$

where the first inequality is due to $p^F(\overline{p}) < 98.5$. So, it is optimal to undercut. At $(0, a_2)$, note first that $\overline{p}_2 < 125$ (it is 125 for $a_2 = 75$ and it decreases as a_2 decreases). Then, if $k \in [-\overline{p}_2, 1]$, by $p_2 \leq \overline{p}_2 + 1 < 3a_2$ we get

$$p_1^F(p_2) = \frac{p_2 + a_2}{2} \in (p_2 - a_2, p_2 + a_2),$$

and (6) differs from the right-hand side of (4) by

$$\begin{aligned}\frac{1}{2}\left(kp_1^F(\overline{p}_2)+\frac{1}{4}k^2\right) - \frac{1}{4}\left(p_1^F(\overline{p}_2)+\frac{k}{2}\right) &= \\ \left(\frac{1}{2}p_1^F(\overline{p}_2)+\frac{1}{8}k-\frac{1}{8}\right)k - \frac{1}{4}p_1^F(\overline{p}_2) &> 100k-100,\end{aligned}$$

where the inequality is satisfied for $k = 1$, thus also for $k < 1$. So, it is optimal not to undercut. ■

Claim: At (a_1, a_2), for every symmetric CURB set $P \times P$,

$$\min P \leq \lfloor \overline{p} \rfloor + 3 - (a_2 - a_1) =: \widehat{p}.$$

Proof: Let $\overline{p}' := \lfloor \overline{p} \rfloor + 3$. Suppose by contradiction that $\min P \geq \widehat{p}+1$. Thus, $p^+(\min P) \geq \overline{p}'$. Throughout, recall that all the best replies to (conjectures over) prices in P are in P by closedness under rational behavior.

First, I show the existence of $p, p' \in P$ such that

$$p < \left\lfloor p^F(p) \right\rfloor = p' < \overline{p}' \leq p^F(p').$$

It cannot be optimal to undercut $\min P$ or a best reply to it, otherwise we would fall below $\min P$. By Fact (i), it is optimal to undercut any $p \geq \overline{p}'$. Hence, $\min P < \overline{p}'$, and since $p^+(\min P) \geq \overline{p}'$, $\left\lfloor p^F(\min P) \right\rfloor$ best replies to $\min P$ and is below $\overline{p}'$. For every $\widetilde{p} \leq \overline{p}' - 1$, we have $p^F(\widetilde{p}) = 50 + \widetilde{p}/2 \geq \widetilde{p} + 1$, where the inequality comes from $\overline{p} < 97$, thus $\widetilde{p} \leq 98$. So, if $p^F(\left\lfloor p^F(\min P) \right\rfloor) \geq$

$\overline{p}'$, $\min P$ and $\left\lfloor p^F(\min P)\right\rfloor$ are the desired p and p'; else, iterating best replies, we rise to the desired p, p', because for every $\widetilde{p} \in P \cap (\min P, \overline{p}')$, $\left\lfloor p^F(\dot{\widetilde{p}})\right\rfloor$ best replies to $\widetilde{p}$: undercutting $\widetilde{p}$ would bring below $\min P$, and $p^F(\min P) < p^+(\min P)$ implies $p^F(\dot{\widetilde{p}}) < p^+(\widetilde{p})$.

Now, consider the belief ν over $\{p, p'\}$ such that $p^F(\mathbb{E}_\nu(\cdot)) = \overline{p}'$. The expected revenues from $\widetilde{p} \in (p' - (a_2 - a_1), p + 100)$ are bounded above by

$$\widetilde{p}\left(\nu(p)\left(50 + \frac{p - \widetilde{p}}{2}\right) + \nu(p')\left(50 + \frac{p' - \widetilde{p}}{2}\right)\right) = \widetilde{p}\left(50 + \frac{\mathbb{E}_\nu(\cdot) - \widetilde{p}}{2}\right); \quad (7)$$

note that $50 + (p - \widetilde{p})/2$ is positive as long as $\widetilde{p} < p + 100$, so it does not underestimate demand against p. The maximum of (7) is at $\widetilde{p} = p^F(\mathbb{E}_\nu(\cdot)) = \overline{p}'$, and since $\overline{p}' \leq p^+(\min P) \leq p^+(p)$, it represents the true expected revenues. For each $\widetilde{p} \geq p + 100$, since $p^F(p') = p^F(\left\lfloor p^F(p)\right\rfloor) \leq 75 + p/4$, $\widetilde{p}$ cannot be optimal. Each $\widetilde{p} \leq p' - (a_2 - a_1)$ cannot best reply to ν because it is below $\min P$. Thus, $\overline{p}'$ is the best reply to ν, but by Fact (i) the best reply to $\overline{p}'$ is $p^-(\overline{p}') < \min P$, a contradiction. ■

Claim: At $(0, a_2)$, for every BRS $P_1 \times P_2$,

$$\min P_2 \geq \min\left\{\left\lfloor p_2^F(\lfloor \overline{p}_2\rfloor + 1 - a_2)\right\rfloor, \lfloor \overline{p}_2\rfloor\right\} =: \widehat{p}_2.$$

Proof: Let $p_1^m := \min P_1$ and $p_2^m := \min\left\{\left\lfloor p_2^F(p_1^m)\right\rfloor, p_2^+(p_1^m)\right\}$. Since firm 1 is at 0, firm 2 has no incentive to undercut. Then, for each $p_1 \in P_1$, firm 2's revenues are strictly increasing in p_2 up to $\min\left\{\left\lfloor p_2^F(p_1)\right\rfloor, p_2^+(p_1)\right\} \geq p_2^m$. Thus, every $p_2 < p_2^m$ is dominated over P_1 by p_2^m. Hence, $\min P_2 \geq p_2^m$.

There remains to show that $p_1^m \geq \lfloor \overline{p}_2\rfloor + 1 - a_2$. Suppose not: $p_1^m \leq \lfloor \overline{p}_2\rfloor - a_2$. Let $p_2' := p_1^m + a_2 + 1 \leq \lfloor \overline{p}_2\rfloor + 1$. I show that $P_1 \times P_2$ is not a BRS because p_1^m is dominated by $p_1' := \left\lfloor p_1^F(p_2')\right\rfloor$ over $\{p_2^m, p_2^m + 1, ...\}$. We have

$$p_1' \leq p_1^F(p_2') = \frac{1}{2}p_1^m + a_2 + \frac{1}{2} < \frac{1}{2}p_2^m + a_2 < p_1^+(p_2^m).$$

By Fact (ii), $p_1^m = p_1^-(p_2')$ does not best reply to p_2', so p_1' does. Then, p_1' does better than p_1^m against any $p_2 \geq p_2' - 1$. The same is true against each

$p_2 \in \{p_2^m, ..., p_1^m + a_2 - 1\}$ because p_1' is closer than p_1^m to $p_1^F(p_2)$:

$$\begin{aligned} p_1^F(p_2) &\geq p_1^F(p_2^m) = \frac{1}{2}a_2 + \frac{1}{2}\min\left\{\left\lfloor 100 - \frac{1}{2}a_2 + \frac{1}{2}p_1^m \right\rfloor, p_1^m + a_2 - 1\right\} \geq \\ &\geq \min\left\{50 + \frac{1}{4}a_2 + \frac{1}{4}p_1^m - \frac{1}{4}, \frac{1}{2}p_1^m + a_2 - \frac{1}{2}\right\} > \frac{1}{2}a_2 + \frac{3}{4}p_1^m + \frac{1}{4} = \\ &= \frac{p_1^F(p_2') + p_1^m}{2} \geq \frac{p_1' + p_1^m}{2}, \end{aligned}$$

where the strict inequality comes from

$$\begin{cases} 50 - \frac{1}{4}a_2 - \frac{1}{2} > \frac{1}{2}p_1^m \\ \frac{1}{2}\left(a_2 - \frac{3}{2}\right) > \frac{1}{2}\left(\frac{1}{2}p_1^m\right) \end{cases},$$

which is true because

$$a_2 - \frac{3}{2} > 50 - \frac{1}{4}a_2 - \frac{1}{2} > \frac{1}{2}\left(400 - 2a_2 - 40\sqrt{100 - a_2}\right) = \frac{1}{2}\left(\overline{p}_2 - a_2\right) \geq \frac{1}{2}p_1^m,$$

where the second inequality holds for $a_2 = 51$ and $a_2 = 75$, hence also in between by convexity of the right-hand side. ■

Claim: Firm 1 has non-lower revenues against $\widehat{p}_2$ at $(0, a_2)$ than against $\widehat{p}$ at (a_1, a_2).

Proof: Suppose first $a_1 \leq 48$. For any p_1, firm's 1 demand at (a_1, a_2) against $\widehat{p}$ is not higher than at $(0, a_2)$ against $\widehat{p}_2$, because $\widehat{p}_2$ is higher than $\widehat{p}$ by more than a_1: when $\widehat{p}_2 = \left\lfloor p_2^F(\lfloor \overline{p}_2 \rfloor + 1 - a_2)\right\rfloor$, $\widehat{p}_2 - \widehat{p}$ is greater than

$$\left(300 - \frac{3}{2}a_2 - 20\sqrt{100 - a_2} - 1\right) - \left(400 - 2a_2 - 40\sqrt{100 - a_2} + 3\right) = \\ \frac{1}{2}a_2 + 20\sqrt{100 - a_2} - 104 > 100 - a_2 = a_1,$$

where the inequality is satisfied for $a_2 = 51$ and $a_2 = 75$, thus also in between by concavity of the left-hand side; when $\widehat{p}_2 = \lfloor \overline{p}_2 \rfloor$, $\widehat{p}_2 - \widehat{p}$ is greater than

$$\left(400 - a_2 - 40\sqrt{100 - a_2} - 1\right) - \left(400 - 2a_2 - 40\sqrt{100 - a_2} + 3\right) = a_2 - 4 \geq a_1.$$

If $a_1 = 49$, at (a_1, a_2) the best reply to $\widehat{p} = 21$ is $p^-(\widehat{p}) = 18$ and it brings revenues 1800, while at $(0, a_2)$ the best reply to $\widehat{p}_2 = \lfloor \overline{p}_2 \rfloor = 69$ is $p_1 = 60$ and it brings revenues 1800 as well.[42] ■

The combination of the last three claims yields the result for $a_1 > 25$.

References

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

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

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

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

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

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

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

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

[42] It is easy to see that, since undercutting is cheap, the minimum of any symmetric CURB set at (a_1, a_2) is actually way below $\lfloor \overline{p} \rfloor$, so there is a strict incentive to move to $a_1' = 0$.

[9] Battigalli, P. and A. Prestipino, "Transparent Restrictions on Beliefs and Forward Induction Reasoning in Games with Asymmetric Information", *The B.E. Journal of Theoretical Economics*, **13(1)**, 2013, 79-130.

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

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

[12] Battigalli, P. and P. Tebaldi, "Interactive Epistemology in Simple Dynamic Games with a Continuum of Strategies," *Economic Theory*, **68**, 2019, 737-763.

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

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

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

[16] Catonini, E., "A simple solution to the Hotelling problem", working paper, 2019.

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

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

[19] D'Aspremont, C, J. J. Gabszewicz, J. F. Thisse "On Hotelling's stability in competition," *Econometrica*, **47**, 1979, 1145-1150.

[20] Dixit, A., "The Role of Investment in Entry-Deterrence", *Economic Journal*, **90**, 1980, 95-106.

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

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

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

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

[25] Harrington, J. "A Theory of Collusion with Partial Mutual Understanding", *Research in Economics*, **71(1)**, 2017, 140-158.

[26] Hotelling, H. "Stability in competition," *Economic Journal*, **39**, 1929, 41-57.

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

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

[29] Ollár, M., and A. Penta. "Full Implementation and Belief Restrictions," *American Economic Review*, **107(8)**, 2017, 2243-2277.

[30] Osborne, M. and C. Pitchik "Equilibrium in Hotelling's model of spatial competition," *Econometrica*, **55**, 1987, 911-922.

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

[32] Penta, A., "Robust Dynamic Implementation", *Journal of Economic Theory*, **160**, 2015, 280-316.

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

[34] Siniscalchi, M., "Structural Rationality in Dynamic Games", working paper, 2020.

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