Supplemental Appendix

I. Variation of the example in the Introduction.

Differently from the game in the Introduction, in the following game, even if the SPE is unique, a non-SPE outcome can be implemented with a threat that does not include any of the SPE actions, while the SPE outcome requires an explicit threat to be implemented.

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

$\longrightarrow$

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

All plans are justifiable, hence they are all rationalizable. The subgame has one pure equilibrium, (D,R), and no mixed equilibrium: when Bob is indifferent between L and C, he prefers R; if he never plays C, Ann won't play M, but then R dominates L; if he never plays L, Ann won't play U, but then R dominates C. So, the game has only one SPE, inducing outcome (S,E).

Ann and Bob can implement the Pareto-superior outcome (N,W) with the following reduced agreement. Let $e_A^0 = \{N.U, N.M\}$ and $e_B^0 = \{W\}$. Note that the agreement does not include the unique SPE action D. We have $S_e^1 = \{N.U, N.M, N.D\} \times \{W\}$. Strong belief in $S_{-i,e}^1$ does not refine beliefs further with respect to the belief in the agreement. Hence, $S_e^\infty = S_e^1 = S((N,W))$. By Proposition 4, the agreement is self-enforcing.

By Theorem 2, the SPE outcome (S,E) is implemented by the reduced agreement on SPE plans $(S, E.R)$. The path agreement on (S,E), instead, is not self-enforcing. Let $e_A^0 = \{S\}$, $e_B^0 = \{E.L, E.C, E.R\}$. We have $S_e^1 = \{S, N.U, N.M\} \times \{E.L, E.C, E.R\}$. Strong belief in $S_{A,e}^1$ induces Bob to believe in U or M at history (N,E), but all his three actions best reply to some belief over $\{U,M\}$. Hence, we have $S_e^\infty = S_e^1$, thus $\zeta(S_e^\infty) \supset \zeta(e^0)$.

II. On the example of Section 2.

In Section 2, I claim that prices below 48 and above 96, with either technology, do not best reply to any conjecture about the competitor's price. Here I prove this claim, with the qualification that every $p_i \geq 48$ with technology $k = A$ best replies to $p_{-i} = 0$, which of course can be immediately ruled out.

Fix $i = 1, 2$ and a probability distribution ν over $p_{-i} \geq 0$.

Consider first $p_i < 48$ with technology $k = A$. For each $p_{-i} \geq 0$, p_i brings positive demand. But then, the profit will be negative, because technology $k = A$ entails a marginal cost $mc = 48$. So, p_i is not a best reply to ν.

Consider now $p_i < 48$ with technology $k = B$. Technology $k = B$ entails a fix cost $F = 48^2$. Then, firm i expects a positive profit only when the expected demand is higher than 48. Now, note that, for each $p_{-i} \geq 0$, increasing p_i by ε reduces demand by at most ε. So, increasing p_i improves the expected profit. Hence, p_i is not a best reply to ν.

Move now to $p_i > 96$. For each $p_{-i} \leq 48$, p_i brings zero demand.

For each $p_{-i} \geq 96$, every $\widetilde{p}_i \geq 96$ brings demand $2(120 - \widetilde{p}_i) \leq 48$, because only the consumers at distance lower than $2(120 - \widetilde{p}_i)$ from firm i have a positive utility from buying from firm i, and they have a negative utility from buying from the competitor. But then, with either technology, price $\widetilde{p}_i = 96$ brings a higher profit than p_i, because it brings $2(p_i - \widetilde{p}_i)$ additional demand, and demand remains lower than the markup.

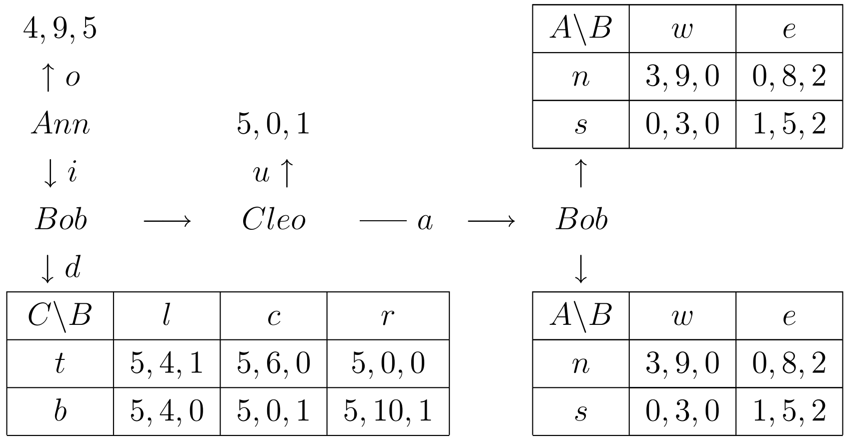
For each $p_{-i} \in (48, 96)$, $\widetilde{p}_i = 96$ brings higher profit than p_i with technology $k = A$, so a fortiori with technology $k = B$: for $k = A$, $\widetilde{p}_i = 96$ brings profit $48 \cdot (p_{-i} - 48)$, p_i brings at most[43] profit $(48 + (p_i - 96)) \cdot (p_{-i} - 48 - (p_i - 96))$, and the first is higher because $48 > p_{-i} - 48$.

Hence, if $\nu(p_{-i} > 48) \neq 0$, $\widetilde{p}_i = 96$ is a better reply than p_i. If $\nu(p_{-i} > 48) = 0$ *and* $\nu(p_{-i} = 0) < 1$, there exists $\widetilde{p}_i > 48$ that, with $k = A$, brings positive expected demand and thus positive expected profit. Thus, p_i is not a best reply to ν.

[43] The upper bound of demand $2(120 - p_i)$ may bind.

III. Another form of agreement incompleteness

Consider the following game.



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

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

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

All plans are justifiable, hence they are all rationalizable. Players want to implement outcome (o). As suggested in Section 5, we first look for the sets $S^* = S^*_A \times S^*_B \times S^*_C$ that induce (o) and satisfy Realization-strictness and Self-Justifiability. Ann's Self-Enforceability requires Bob not to play d and Cleo not to play u. Then, Bob's Self-Justifiability requires that Cleo may play t, and Cleo's Self-Justifiability requires that Bob may play e in a subgame he allows. Hence, calling S^w_B and S^e_B the binary sets of plans of Bob where the last move is w and e respectively, the desired sets S^* are those that satisfy

$$S^*_A = \{o\}\,, \quad S^*_B \subseteq S^w_B \cup S^e_B,\, S^e_B \cap S^*_B \neq \emptyset, \quad \{t.a\} \subseteq S^*_C \subseteq \{t.a, b.a\}\,.$$

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

So, we look for a tight agreement e where e^0 satisfies the conditions above. First, observe that we need $e^0_C = \{t.a\}$. If $b.a \in e^0_C$, then, regardless of e^1_A, we have $d.r \in br_B(\Delta^e_B)$, but $d.l \notin br_B(\Delta^e_B)$. So, for Bob, T3 imposes $d.l \notin e^m_B \cap S_B((i,d)) \neq \emptyset$ for some m, but then $t.a \notin br_C(\Delta^e_C)$, a violation of T3 for Cleo. Still, without restrictions on e^1_A, we have $d.c \in br_B(\Delta^e_B)$, so again

$d.l \notin e_B^m \cap S_B((i,d)) \neq \emptyset$ for some m and $t.a \notin br_C(\Delta_C^e)$. Hence, we must obtain $d.c \notin br_B(\Delta_B^e)$. So, we must impose $i.s.s \notin e_A^1$. If Ann guarantees to play n in a specific subgame, then we have $br_i(\Delta_B^e) \subseteq S_B^w$; hence, T3 imposes $e_B^0 \subseteq S_B^w$, a contradiction of the conditions on e_B^0. So, the only remaining option is $e_A^1 = \{i.n.n, i.n.s, i.s.n\}$. Then, on the one hand there is $\mu_B \in \Delta_B^e$ with $\mu_B(i.n.s|(i)) = \mu_B(i.s.n|(i)) = 1/2$ and $br_B(\mu_B) = S_B^e$; on the other hand, for every $\mu_B \in \Delta_B^e$, there is $s_B \in br_B(\mu_B) \cap (S_B^w \cup S_B^e)$ that gives to Bob an expected payoff of at least 6.5, so $d.c \notin br_B(\Delta_B^e) \cap S_B((i,d)) = \emptyset$. So, the following is a tight agreement:

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

The vagueness of Ann about in which subgame she is going to play n is a kind of agreement incompleteness that, like here, can be necessary to implement an outcome. It can be interpreted as Ann doing the following speech: “I guarantee that I will be prepared to play n in at least one contingency, but I cannot guarantee that I will be prepared to play n in both.”

This kind of strategic uncertainty also arises naturally from strategic reasoning. The example on page 50 in Battigalli [6], provided by Gul and Reny, shows that the set of justifiable plans of a player is not a Cartesian product of sets of actions at different information sets. This is the reason why (agreement-)rationalizability is defined as an elimination procedure of plans and not of actions at different information sets, and agreements are defined in terms of plans as well.

IV. Equivalence between Agreement-rationalizability and Selective Rationalizability.

In this section, I show that Agreement-rationalizability is equivalent to the original notion of Selective Rationalizability, provided and characterized epistemically in [17], for the analysis of agreements. Selective Rationalizability is defined as follows.

Definition 15 *Let $((S_i^m)_{i\in I})_{m=0}^{\infty}$ denote Rationalizability. Consider the following procedure.*

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

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

S1 $s_i \in br_i(\mu_i)$;

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

S3 μ_i strongly believes S_j^q for all $j \neq i$ and $q \in \mathbb{N}$.

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

Selective Rationalizability differs from Agreement-rationalizability because of the stronger requirement S3 in place of the requirement that $s_i \in S_i^{\infty}$. Here I argue that the two procedures are equivalent for the analysis of agreements. In a nutshell, the reason is that the two procedures are equivalent when the agreement does not restrict behavior off the rationalizable paths $\zeta(S^{\infty})$, and restricting behavior off the rationalizable paths is unneeded to induce the desired outcomes, because players do not expect to leave the rationalizable paths anyway.

To show this, one has to make sure that the desired restrictions along the rationalizable paths do not force restrictions off the rationalizable paths, because, in general, the set of rationalizable plans of a player can feature "cross restrictions" at unordered histories (see the previous section). The following lemma ensures that this issue does not arise between histories on and off the rationalizable paths. The intuition is that a deviation from the rationalizable paths always comes as a surprise, thus players have to revise their beliefs, and this disentangles their behavior after the deviation from what they would have done absent the deviation. Let $p(h)$ denote the immediate predecessor of a history $h \in H\backslash \{h^0\}$.

Lemma 1 *For every $i \in I$, $s_i \in S_i^\infty$, $\overline{h} \in H(s_i)\backslash H(S^\infty)$ with $p(\overline{h}) \in H(S^\infty)$, and $s_i' \in S_i^\infty \cap S_i(\overline{h})$, there is $s_i'' \in S_i^\infty$ such that, for each $h \in H(s_i'')$, $s_i''(h) = s_i(h)$ if $h \not\succeq \overline{h}$, $s_i''(h) = s_i'(h)$ if $h \succeq \overline{h}$.*

Proof. Fix μ_i and μ_i' that strongly believe $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $s_i \in br_i(\mu_i)$ and $s_i' \in br_i(\mu_i')$. Since $\overline{h} \in H(S_i^\infty)$, $\overline{h} \notin H(S_{-i}^\infty)$. Then, since μ_i strongly believes S_{-i}^∞ and $p(\overline{h}) \in H(S_{-i}^\infty)$, we have $\mu_i(S_{-i}(\overline{h})|p(\overline{h})) = 0$. Hence, I can construct μ_i'' that strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ as $\mu_i''(\cdot|h) = \mu_i(\cdot|h)$ for each $h \not\succeq \overline{h}$, and $\mu_i''(\cdot|h) = \mu_i'(\cdot|h)$ for each $h \succeq \overline{h}$. Clearly, $s_i'' \in br_i(\mu_i'') \subseteq S_i^\infty$. ■

Next, I formalize the elimination or absence of restrictions off the rationalizable paths. Fix $i \in I$ and a subset of rationalizable plans $\widetilde{S}_i \subseteq S_i^\infty$. I call the **canonical form of** $\widetilde{S}_i$ the set

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

The canonical form of $\widetilde{S}_i$ contains $\widetilde{S}_i$. I say $\widetilde{S}_i$ is **canonical** when it coincides with its canonical form. Analogously, given an agreement $e = ((e_i^0, ..., e_i^{k_i}))_{i\in I}$, I call the **canonical form of** e the agreement $\overline{e} = ((\overline{e}_i^0, ..., \overline{e}_i^{k_i}))_{i\in I}$ such that, for each $i \in I$ and $n = 0, ..., k_i$, $\overline{e}_i^n$ is the canonical form of e_i^n, and I say that e is canonical when it coincides with its canonical form.

Now I show that Agreement-rationalizability and Selective Rationalizability are equivalent for a canonical agreement. The proof will follow a simple inductive argument based on the next technical lemma.

Lemma 2 *For each $i \in I$, fix a collection of K_i canonical sets of rationalizable plans $(\widetilde{S}_i^k)_{k=0}^{K_i}$. Let 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 $((\widetilde{S}_j^k)_{k=0}^{K_j})_{j \neq i}$ and $(S_j^\infty)_{j \neq i}$. Let $\widehat{S}_i^*$ 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 $((\widetilde{S}_j^k)_{k=0}^{K_j})_{j \neq i}$ and $((S_j^q)_{q=0}^\infty)_{j \neq i}$. Then, S_i^* and $\widehat{S}_i^*$ are identical and canonical.*

Proof. Fix $s_i' \in S_i^*$ and μ_i' that strongly believes $((\widetilde{S}_j^k)_{k=0}^{K_j})_{j \neq i}$ and $(S_j^\infty)_{j \neq i}$ such that $s_i' \in br_i(\mu_i')$. Fix $s_i \in S_i^\infty$ and μ_i that strongly believes $((S_j^q)_{q=0}^\infty)_{j \neq i}$ such that $s_i(h) = s_i'(h)$ for all $h \in H(S^\infty) \cap H(s_i)$, and $s_i \in br(\mu_i)$. For each $\overline{h} \notin H(S^\infty)$ with $p(\overline{h}) \in H(S^\infty)$, $j \neq i$, $s_j' \in S_j^\infty \cap S_j(\overline{h})$, $k = 0, ..., K_j$, and $s_j \in \widetilde{S}_j^k \cap S_j(\overline{h})$, by Lemma 1 there is $s_j'' \in S_j^\infty$ such that, for each $h \in H(s_j'')$, $s_j''(h) = s_j'(h)$ if $h \succeq \overline{h}$, $s_j''(h) = s_j(h)$ if $h \not\succeq \overline{h}$. Hence, $s_j'' \in \widetilde{S}_j^k$. Endowed with all such s_j'''s, since μ_i' strongly believes S_{-i}^∞, I can construct μ_i'' that strongly believes $((\widetilde{S}_j^k)_{k=0}^{K_j})_{j \neq i}$ and $((S_j^q)_{q=0}^\infty)_{j \neq i}$ such that $\mu_i''(\cdot|h) = \mu_i'(\cdot|h)$ for all $h \in H(S_{-i}^\infty)$, and $\mu_i''(S_{-i}(z)|h) = \mu_i(S_{-i}(z)|h)$ for all $h \notin H(S_{-i}^\infty)$ and $z \succeq h$. Clearly, s_i is a continuation best reply to $\mu_i''(\cdot|h)$ for all $h \in H(s_i) \backslash H(S_{-i}^\infty)$. Moreover, s_i is a continuation best reply to $\mu_i''(\cdot|h)$ for all $h \in H(s_i) \cap H(S_{-i}^\infty) \subseteq H(S^\infty)$ because it induces the same outcome distribution as s_i'. So, $s_i \in br(\mu_i'') \subseteq \widehat{S}_i^* \subseteq S_i^*$. This shows that S_i^* is canonical, included in $\widehat{S}_i^*$, and thus identical to $\widehat{S}_i^*$. ■

Proposition 8 *For every canonical agreement $e = (e_i)_{i \in I}$, $S_e^\infty = \widehat{S}_e^\infty$.*

Proof. For each $i \in I$ and $n > 0$, $S_{i,e}^n$ is 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 $((e_j^k)_{k=0}^{k_j})_{j \neq i}$ and $((S_{j,e}^q)_{q=0}^{n-1})_{j \neq i}$, while $\widehat{S}_{i,e}^1$ is 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 $((e_j^k)_{k=0}^{k_j})_{j \neq i}$, $((\widehat{S}_{j,e}^q)_{q=0}^{n-1})_{j \neq i}$, and $((S_j^q)_{q=0}^\infty)_{j \neq i}$. For $n = 1$ and each $j \neq i$, $\widehat{S}_{j,e}^{n-1} = S_{j,e}^{n-1} = S_j^\infty$, so, by Lemma 2, S_e^1 and $\widehat{S}_e^1$ are identical and canonical. Then, the argument can be repeated for $n > 1$ by induction. ■

Corollary 5 *A canonical agreement is self-enforcing under Agreement-rationalizability if and only if it is self-enforcing under Selective Rationalizability.*

How do an agreement and its canonical form compare in terms of behavioral implications? For the first step of reasoning, both under Agreement-rationalizability and Selective Rationalizability, the canonical form yields the same plans as the original agreement on the rationalizable paths, and combines them with all rationalizable plans off the rationalizable paths.

Lemma 3 *Fix an agreement $e = (e_i)_{i \in I}$ and its canonical form $\overline{e} = (\overline{e}_i)_{i \in I}$. For each $i \in I$, $S^1_{i,\overline{e}}$ and $\widehat{S}^1_{i,\overline{e}}$ are the canonical forms of $S^1_{i,e}$ and $\widehat{S}^1_{i,e}$.*

Proof. The proof is identical under both procedures, so let $\widetilde{S}^1_{i,\overline{e}}, \widetilde{S}^1_{i,e}$ denote either of the two. I have just proven for Proposition 8 that $\widetilde{S}^1_{i,\overline{e}}$ is canonical. So, there only remains to show that for each $s_i \in \widetilde{S}^1_{i,\overline{e}}$, there is $s_i' \in \widetilde{S}^1_{i,e}$ such that $s_i(h) = s_i'(h)$ for all $h \in H(s_i) \cap H(S^\infty)$, and vice versa.

Fix $s_i \in \widetilde{S}^1_{i,\overline{e}}$. By definition of $\widetilde{S}^1_{i,\overline{e}}$, $s_i \in S^\infty_i$, and there is $\mu_i \in \Delta^{\overline{e}}_i$ that strongly believes $(S^\infty_j)_{j \neq i}$ such that $s_i \in br_i(\mu_i)$. For each $j \neq i$, $n = 0, ..., k_j$, and $s_j \in \overline{e}^n_j$, there is $s_j' \in e^n_j$ such that $s_j'(h) = s_j(h)$ for all $h \in H(S^\infty) \cap H(s_j')$. Endowed with all such s_j''s, I can construct $\mu_i' \in \Delta^e_i$ that strongly believes $((S^q_j)_{j \neq i})^\infty_{q=0}$ such that $\mu_i'(S_{-i}(z)|h) = \mu_i(S_{-i}(z)|h)$ for all $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$. Since μ_i' and μ_i strongly believe S^∞_{-i}, for each $h \in H(S^\infty)$, each $\widetilde{s}_i \in S^\infty_i$ yields the same outcome distribution against $\mu_i'(\cdot|h)$ and $\mu_i(\cdot|h)$. Then, since $br_i(\mu_i') \subseteq S^\infty_i$, there is $s_i' \in br_i(\mu_i') \subseteq S^1_{i,e}$ such that for $s_i'(h) = s_i(h)$ for all $h \in H(s_i') \cap H(S^\infty)$.

The proof of the vice versa is identical.[44] ■

How relevant is the class of canonical agreements? Under both Agreement-rationalizability and Selective Rationalizability, it suffices to implement all implementable outcomes. To show this, given Lemma 3, it is enough to find, for each implementable outcome set, a self-enforcing agreement that requires only one step of reasoning. For Agreement-rationalizability, this agreement is the tight agreement. I will further show that the canonical form is a tight agreement as well. (For Selective Rationalizability, an analogous characterization can be provided.)

Proposition 9 *Under Agreement-rationalizability, the canonical form of a tight agreement is a tight agreement.*

[44] The vice versa can also be proven in a simpler way: by $e^n_j \subseteq \overline{e}^n_j$ and $H(e^n_j) \cap H(S^\infty) = H(\overline{e}^n_j) \cap H(S^\infty)$, there is $\mu_i' \in \Delta^{\overline{e}}_i$ that strongly believes $((S^q_j)_{j \neq i})^\infty_{q=0}$ such that $\mu_i'(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(S^\infty)$.

Proof. Let $e = (e_i)_{i \in I}$ be a tight agreement and $\overline{e} = (\overline{e}_i)_{i \in I}$ be its canonical form, which clearly inherits T2 from the tight agreement.

For T1, fix $i \in I$ and μ_i that strongly believes $\overline{e}^0_{-i}$. Fix $\widetilde{\mu}_i$ that strongly believes $\overline{e}^0_{-i}$ and $((S^q_j)_{j \neq i})^\infty_{q=0}$ such that $\widetilde{\mu}_i(\cdot|h) = \mu_i(\cdot|h)$ for each $h \in H(\overline{e}^0_{-i})$. Thus, $\zeta(br(\mu_i) \times \overline{e}^0_{-i}) = \zeta(br(\widetilde{\mu}_i) \times \overline{e}^0_{-i})$. Since each $\overline{e}^0_j$ is the canonical form of e^0_j, I can construct μ'_i that strongly believes e^0_{-i} and $((S^q_j)_{j \neq i})^\infty_{q=0}$ such that $\mu'_i(S_{-i}(z)|h) = \widetilde{\mu}_i(S_{-i}(z)|h)$ for all $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$. Since $br(\widetilde{\mu}_i), br(\mu'_i) \subseteq S^\infty_i$ and each $s_i \in S^\infty_i$ induces the same outcome distribution with $\mu'_i(\cdot|h)$ and $\widetilde{\mu}_i(\cdot|h)$ at every $h \in H(S^\infty) \cap H(s_i)$, we obtain $\zeta(br(\widetilde{\mu}_i) \times S^\infty_{-i}) = \zeta(br(\mu'_i) \times S^\infty_{-i}) \subseteq \zeta(S^\infty)$. Then, $\zeta(br(\widetilde{\mu}_i) \times \overline{e}^0_{-i}) = \zeta(br(\mu'_i) \times \overline{e}^0_{-i})$. By $\zeta(S^\infty_i \times \overline{e}^0_{-i}) = \zeta(S^\infty_i \times e^0_{-i})$, we get $\zeta(br(\mu'_i) \times \overline{e}^0_{-i}) = \zeta(br(\mu'_i) \times e^0_{-i})$. By T1, $\zeta(br(\mu'_i) \times e^0_{-i}) \subseteq \zeta(e^0)$. By definition of canonical form, $\zeta(e^0) = \zeta(\overline{e}^0)$. Altogether, $\zeta(br(\mu_i) \times \overline{e}^0_{-i}) \subseteq \zeta(\overline{e}^0)$.

For T3, observe first that, by T2, every $\mu_i \in \Delta^e_i \cup \Delta^{\overline{e}}_i$ strongly believes $(S^\infty_j)_{j \neq i}$. Then, $S^1_{i,e} = br_i(\Delta^e_i) \cap S^\infty_i$ and $S^1_{i,\overline{e}} = br_i(\Delta^{\overline{e}}_i) \cap S^\infty_i$. With this, I am going to show that for each $h \in H(S^1_{i,\overline{e}})$, there is n such that $\emptyset \neq \overline{e}^n_i \cap S_i(h) \subseteq S^1_{i,\overline{e}}$, which gives T3 for $\overline{e}$.

Suppose first that either $h = h^0$ or $p(h) \in H(S^\infty)$. By Lemma 3, $S^1_{i,\overline{e}}$ is the canonical form of $S^1_{i,e}$. Hence, $h \in H(S^1_{i,e})$. So, by T3, there is n such that $\emptyset \neq e^n_i \cap S_i(h) \subseteq br_i(\Delta^e_i) \cap S^\infty_i = S^1_{i,e} \subseteq S^1_{i,\overline{e}}$. Suppose now that $p(h) \notin H(S^\infty)$. Fix the unique $h' \prec h$ such that $h' \notin H(S^\infty)$ and $p(h') \in H(S^\infty)$. Since $h' \in H(S^1_{i,\overline{e}})$, as just shown $\emptyset \neq \overline{e}^n_i \cap S_i(h') \subseteq S^1_{i,\overline{e}}$ for some n. Since $S_i(h) \subseteq S_i(h')$, there only remains to show that $\overline{e}^n_i \cap S_i(h) \neq \emptyset$. Since $h \in H(S^\infty_i)$, this follows from construction of $\overline{e}^n_i$ and Lemma 1. ■

Given this, I will call the canonical form of a tight agreement "canonical tight agreement".

Corollary 6 *Under Agreement-rationalizability, an outcome set is implementable if and only if it is prescribed by a canonical tight agreement.*

Now I move to Selective Rationalizability.

Lemma 4 *Every implementable outcome set under Selective Rationalizability is implemented by a canonical agreement.*

Proof. Fix a self-enforcing agreement $\widehat{e} = (\widehat{e}_i)_{i\in I}$ under Selective Rationalizability. Following the proof of Theorem 3 (only if), I can construct an agreement $e = (e_i)_{i\in I}$ such that $e^0 = \widehat{e}^0 \cap \widehat{S}^\infty_{\widehat{e}}$, and for each $i \in I$, Δ^e_i is exactly the set of all CPS's $\mu_i \in \Delta^{\widehat{e}}_i$ that strongly believe $((\widehat{S}^q_{j,\widehat{e}})_{j\neq i})^\infty_{q=0}$. Then, the set of plans $s_i \in S^\infty_i$ such that $s_i \in br_i(\mu_i)$ for some $\mu_i \in \Delta^e_i$ that strongly believes $((S^q_j)_{j\neq i})^\infty_{q=0}$ coincides with both $\widehat{S}^1_{i,e}$ and $\widehat{S}^\infty_{i,\widehat{e}}$, thus $\widehat{S}^1_e = \widehat{S}^\infty_{\widehat{e}}$. Since every $\mu_i \in \Delta^e_i$ already strongly believes $(\widehat{S}^\infty_{j,\widehat{e}})_{j\neq i}$, it already strongly believes $(\widehat{S}^1_{j,e})_{j\neq i}$, hence $\widehat{S}^1_e = \widehat{S}^2_e = \widehat{S}^\infty_e$. So we have $\widehat{S}^\infty_e = \widehat{S}^\infty_{\widehat{e}}$. With $e^0 = \widehat{e}^0 \cap \widehat{S}^\infty_{\widehat{e}}$, we get $e^0 \subseteq \widehat{S}^\infty_e$. Self-enforceability of $\widehat{e}$ gives $\zeta(\widehat{S}^\infty_{\widehat{e}}) = \zeta(\widehat{e}^0 \cap \widehat{S}^\infty_{\widehat{e}})$. Altogether,

$$\zeta(\widehat{S}^\infty_e) = \zeta(\widehat{S}^\infty_{\widehat{e}}) = \zeta(\widehat{e}^0 \cap \widehat{S}^\infty_{\widehat{e}}) = \zeta(e^0) = \zeta(e^0 \cap \widehat{S}^\infty_e),$$

so e is self-enforcing and induces $\zeta(\widehat{S}^\infty_{\widehat{e}})$.

Consider now the canonical form $\overline{e}$ of e. By Lemma 3, each $\widehat{S}^1_{i,\overline{e}}$ is the canonical form of $\widehat{S}^1_{i,e}$. Hence, $\zeta(\widehat{S}^1_{\overline{e}}) = \zeta(\widehat{S}^1_e)$. The intersection of the canonical forms of two sets contains the canonical form of the intersection of the sets. This implies $\zeta(e^0 \cap \widehat{S}^1_e) \subseteq \zeta(\overline{e}^0 \cap \widehat{S}^1_{\overline{e}})$. I will show that $\widehat{S}^1_{\overline{e}} = \widehat{S}^\infty_{\overline{e}}$. Then,

$$\zeta(\widehat{S}^\infty_{\overline{e}}) = \zeta(\widehat{S}^1_{\overline{e}}) = \zeta(\widehat{S}^1_e) = \zeta(\widehat{S}^\infty_e) = \zeta(e^0 \cap \widehat{S}^\infty_e) \subseteq \zeta(\overline{e}^0 \cap \widehat{S}^\infty_{\overline{e}}),$$

so with $\zeta(\widehat{S}^\infty_{\overline{e}}) \supseteq \zeta(\overline{e}^0 \cap \widehat{S}^\infty_{\overline{e}})$, $\overline{e}$ is self-enforcing and induces $\zeta(\widehat{S}^\infty_e) = \zeta(\widehat{S}^\infty_{\widehat{e}})$.

To show that $\widehat{S}^1_{\overline{e}} = \widehat{S}^\infty_{\overline{e}}$, I prove that, for each $i \in I$, every $\overline{\mu}_i \in \Delta^{\overline{e}}_i$ already strongly believes $(\widehat{S}^1_{j,\overline{e}})_{j\neq i}$. Fix $j \neq i$ and $h \in H(\widehat{S}^1_{j,\overline{e}})$; I show that, for some n, $\emptyset \neq \overline{e}^n_j \cap S_j(h) \subseteq \widehat{S}^1_{j,\overline{e}}$. Suppose first that either $h = h^0$ or $p(h) \in H(S^\infty)$. By Lemma 3, $\widehat{S}^1_{j,\overline{e}}$ is the canonical form of $\widehat{S}^1_{j,e}$. Hence, $h \in H(\widehat{S}^1_{j,e})$. As proven above, every $\mu_i \in \Delta^e_i$ strongly believes $\widehat{S}^1_{j,e}$. Therefore, there is n such that $\emptyset \neq e^n_j \cap S_j(h) \subseteq \widehat{S}^1_{j,e} \subseteq \widehat{S}^1_{j,\overline{e}}$. Note that a canonical set contains the canonical form of all its subsets. Then, since $\overline{e}^n_j$ is the canonical form of e^n_j and $\widehat{S}^1_{j,\overline{e}}$ is canonical, we also have $\emptyset \neq \overline{e}^n_j \cap S_j(h) \subseteq \widehat{S}^1_{j,\overline{e}}$. Suppose now that $p(h) \notin H(S^\infty)$. Fix the unique $h' \prec h$ such that $h' \notin H(S^\infty)$ and $p(h') \in H(S^\infty)$. Since $h' \in H(\widehat{S}^1_{j,\overline{e}})$, as just shown $\emptyset \neq \overline{e}^n_j \cap S_j(h') \subseteq \widehat{S}^1_{j,\overline{e}}$ for some n. Since $S_j(h) \subseteq S_j(h')$, there only remains to show that $\overline{e}^n_j \cap S_j(h) \neq \emptyset$. Since $h \in H(S^\infty_j)$, this follows from construction of $\overline{e}^n_j$ and Lemma 1. ■

The final proposition states the equivalence between Agreement-rationalizability and Selective Rationalizability for the analysis of implementability across all agreements.

Proposition 10 *An outcome set is implementable under Agreement-rationalizability if and only if it is implementable under Selective Rationalizability.*

Proof. Suppose an outcome set is implementable under Agreement-rationalizability. Then, by Corollary 6, it is implemented by a canonical tight agreement $\overline{e} = (\overline{e}_i)_{i\in I}$, and by Proposition 8, $S^{\infty}_{\overline{e}} = \widehat{S}^{\infty}_{\overline{e}}$, thus $\overline{e}^0 \cap S^{\infty}_{\overline{e}} = \overline{e}^0 \cap \widehat{S}^{\infty}_{\overline{e}}$ and implementation under Selective Rationalizability obtains. Suppose now an outcome set is implementable under Selective Rationalizability. Then, by Lemma 4, it is implemented by a canonical agreement $\overline{e} = (\overline{e}_i)_{i\in I}$, and by Proposition 8 $S^{\infty}_{\overline{e}} = \widehat{S}^{\infty}_{\overline{e}}$, thus $\overline{e}^0 \cap S^{\infty}_{\overline{e}} = \overline{e}^0 \cap \widehat{S}^{\infty}_{\overline{e}}$ and implementation under Agreement-rationalizability obtains. ■

V. On different epistemic priority orderings

Epistemic priority to the agreement

In this section, I formalize the claim of Section 7.1 that the analysis of Sections 4 and 5 can be replicated under epistemic priority to the agreement by using Strong-Δ-Rationalizability in place of Agreement-rationalizability, and I prove that this expands the collection of implemental outcome sets.

Strong-Δ-Rationalizability can be defined like Agreement-rationalizability with S_i^0 in place of S_i^∞ as initialization. Let $((S_{i,\Delta^e}^q)_{i\in I})_{q=0}^\infty$ denote Strong-Δ-Rationalizability under belief in the agreement. Allow agreements to feature non-rationalizable plans. Then, the following holds.

Remark 3 *Under priority to the agreement, the results of Sections 4 and 5 hold through verbatim after substituting everywhere:*

1. *agreement-rationalizable (S_e^∞) with strongly-Δ-rationalizable plans ($S_{\Delta^e}^\infty$);*
2. *rationalizable plans (S^∞) with justifiable plans (S^1) in Corollary 2, Proposition 6, and Theorem 2, and with all plans (S) elsewhere.*

To verify Remark 3, one can follow the proofs in the main appendix, i.e., under priority to rationality, with the following substitutions: replace $(S_i^\infty)_{i\in I}$ with $(S_i)_{i\in I}$,[45] and $((S_{j,e}^q)_{j\in I})_{q=0}^\infty$ with $((S_{j,\Delta^e}^q)_{j\in I})_{q=0}^\infty$.

A credible agreement under priority to rationality needs not be credible under priority to the agreement: as shown in the companion paper [17], Selective Rationalizability does not refine Strong-Δ-Rationalizability for given first-order belief restrictions. (I prove in [2] that the two procedures are outcome-equivalent for path agreements.) Despite of this, across all agreements, more outcome sets can be implemented under priority to the agreement.

[45] Except in the proof of Proposition 6, where S_i^∞ must be substituted by S_i^1 (but S_j^∞ with S_j as usual). Moreover, in the proof of Proposition 2, the CPS μ_i' constructed to show Realization-strictness shall not strongly believe $((S_j^q)_{j\neq i})_{q=0}^\infty$.

Of course most statements become trivial/superfluous with S_i in place of S_i^∞ , but the substitution allows to keep the same phrasing.

Proposition 11 *If an outcome set is implementable under priority to rationality, then it is implementable under priority to the agreement.*

Proof. Fix an implementable outcome set $P \subset Z$ under priority to rationality. By Corollary 6, it is implemented by a canonical tight agreement $\overline{e} = ((\overline{e}_i^0, ..., \overline{e}_i^{k_i}))_{i \in I}$. Then, by Corollary 5 and Proposition 8, $\overline{e}$ implements P also under Selective Rationalizability. Let M be the smallest m such that $S^m = S^{m+1}$. For each $i \in I$ and $n = k_i + 1, ..., k_i + M$, let $\overline{e}_i^n = S_i^{M+1-(n-k_i)}$. Without loss of generality, suppose that, for each $n = k_i + 1, ..., k_i + M$, $\overline{e}_i^n \neq \overline{e}_i^{n-1}$ (if not, $\overline{e}_i^n$ can simply be eliminated from the list); then, $\overline{e}^* := ((\overline{e}_i^0, ..., \overline{e}_i^{k_i+M}))_{i \in I}$ is an agreement. This agreement incorporates requirement S3 of Selective Rationalizability. Hence, Strong-Δ-Rationalizability under $\overline{e}^*$ is identical to Selective Rationalizability under $\overline{e}$. Therefore, $\overline{e}^*$ implements P under priority to the agreement. ■

For instance, by Remark 3.2, under priority to the agreement any realization-strict Nash equilibrium in justifiable plans of a two-player game is a self-enforcing agreement, also when incompatible with strong belief in rationality.[46] An example is the entry game of Section 2, where the incumbent can deter entry also in Case 1 by threatening a low justifiable price; then, entry would be considered a sign of the entrant's irrationality, and the incumbent could have any belief about the entrant's price. In the Hotelling model, almost all location pairs would be implementable under priority to the agreement (see [16]). In all other examples of this paper, all plans are rationalizable; then, Agreement-rationalizability and Strong-Δ-Rationalizability coincide and the insights are robust to the inversion of epistemic priority.

[46] As shown by the introductory example of the companion paper, Strong-Δ-Rationalizability can yield a non-subgame perfect equilibrium outcome even in a perfect information game without relevant ties, where the unique backward induction outcome is also the only extensive-form-rationalizable one (Battigalli [6]).

Epistemic priority to the path

In this section, I revise my methodology for the study of self-enforcing agreements under priority to the path. Recall that giving priority to the path means that players, whenever possible, interpret deviations from the agreed-upon path(s) under the view that the deviator did believe that the co-players would have complied with the agreement on-path.

For simplicity of exposition, I will focus on agreements that prescribe a single path z. To analyze them under priority to the path, I specialize the elimination procedure for general epistemic priority orderings that I construct in the companion paper ([17]). Let $((S^q_{j,z})_{j\in I})^\infty_{q=0}$ denote Agreement-rationalizability under the *path* agreement on z, which I will call *z-rationalizability*. Fix an agreement $e = (e_i)_{i\in I}$ with $e^0 \subseteq S(z)$ and $\times_{i\in I} e^{k_i}_i \subseteq S^\infty_z$.

Definition 16 *Let $S^0_{e^z} = S^\infty_z$. Fix $n > 0$ and suppose to have defined $((S^q_{j,e^z})_{j\in I})^{n-1}_{q=0}$. For each $i \in I$ and $s_i \in S^\infty_{i,z}$, let $s_i \in S^n_{i,e^z}$ if $s_i \in br_i(\mu_i)$ for some $\mu_i \in \Delta^e_i$ that strongly believes $((S^q_{j,e^z})_{j\neq i})^{n-1}_{q=0}$.*

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

Definition 16 captures the following reasoning scheme. First, each order of belief in rationality is maintained as long as compatible with the observed behavior. Second, each order of belief in the path is maintained as long as compatible with all orders of belief in rationality. Third, each order of belief in the *whole* agreement is maintained as long as compatible with all the aforementioned beliefs. The first two levels of epistemic priority are captured by Agreement-rationalizability under the path agreement on z. Thus, the credibility of the path agreement is a preliminary test for the self-enforceability of an agreement that prescribes z under priority to the path. Then, the z-rationalizable plans $(S^\infty_{i,z})_{i\in I}$ are refined using the belief in the whole agreement. So, the agreement must be compatible with strategic reasoning around the path.

The analysis of Sections 4 and 5 can be replicated under priority to the path. Allow agreements (including SES's) to prescribe only one path z and feature only z-rationalizable plans. Then, the following holds.

Remark 4 *Under priority to the path, the results of Sections 4 and 5 hold through verbatim after substituting everywhere:*

1. *outcome sets P with single outcomes z;*
2. *selectively-rationalizable (S_e) with z-agreement-rationalizable plans (S_{e^z});*
3. *rationalizable plans (S^∞) with z-rationalizable plans (S^∞_z).*

To verify Remark 4, one can follow the proofs in the main appendix with the following substitutions: replace $P \subseteq Z$ with $z \in Z$, $(S^\infty_i)_{i\in I}$ with $(S^\infty_{i,z})_{i\in I}$, $((S^q_{j,e})_{j\in I})^\infty_{q=0}$ with $((S^q_{j,e^z})_{j\in I})^\infty_{q=0}$, and $((S^q_j)_{j\in I})^\infty_{q=0}$ with $((S^q_{j,z})_{j\in I})^\infty_{q=0}$. Although a self-enforcing agreement under priority to the path needs not be self-enforcing under priority to rationality, the following holds.

Proposition 12 *If an outcome is implementable under priority to the path, then it is implementable under priority to rationality.*

Proof. The analysis of Section IV can be replicated with z-agreement-rationalizability in place of Agreement-rationalizability, z-rationalizability in place of Rationalizability, and "z-selective rationalizability" in place of Selective Rationalizability, where S^∞_z replaces S^∞ as initialization and $((S^q_{z,i})_{i\in I})^\infty_{q=0}$ replaces $((S^q_i)_{i\in I})^\infty_{q=0}$ in requirement S3. Then, the proof of this proposition is identical to the proof of Proposition 11. ■

Now I formalize the solution of the example of Section 7.2. Ann and Bob would like to agree on the SPE with path $z = (FR.W, W.FR)$. To check whether the agreement is self-enforcing under priority to the path, we must first carry out Agreement-rationalizability under the path agreement on z.

All plans are justifiable, hence they are all rationalizable. Then, z-rationalizability goes as follows (the second action of a plan s_i refers to the history $(s_i(h^0), W)$):

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

At the first step of reasoning, Bob already concludes that he has the incentive to comply with the path, because when $\mu_B(S_A(z)|h^0) = 1$, a deviation in the first period reduces his first-stage payoff, and his second-stage payoff cannot be higher than on path. For Ann, instead, it is optimal to deviate to W in the first period if $\mu_A(s_B|h^0) \geq 2/3$ for $s_B \in S_B(z)$ such that $s_B((W, W)) = W$: the deviation reduces her first-stage payoff by 1 but is expected to lead to history (W, W), where her second-stage expected payoff is higher than 2, whereas on path it is just 1. Note that under this belief Ann free-rides at history (W, W).

At the second step of reasoning, for Ann nothing changes because she already expected Bob to comply. For Bob, strong belief in $S^1_{A,z}$ entails

$$\mu_B(\{W.FR.FR, W.FR.W\} \,|(W, W)) = 1,$$

which together with $\mu_B(S_A(z)|h^0) = 1$ implies that the only optimal plan of Bob is $W.W.FR$. Then, strong belief in $S^2_{B,z}$ entails that Ann has the incentive to work in the first period and free ride in the second period if Bob has worked. So, $S^3_{A,z}$ is disjoint from $S_A(z)$, and then Bob cannot give probability 1 to both, so we obtain $S^4_{B,z} = \emptyset$. This means that no agreement that prescribes z is credible under priority to the path. Instead, under priority to rationality, the agreement on the SPE plans is self-enforcing.

When z-rationalizability is non-empty, it is not necessarily the case that z can be implemented under priority to the path. In the companion paper ([17]), I provide an example of a SPE whose path constitutes a credible path agreement but is not implementable under priority to the path (because two players cannot agree on any effective threat to deter the deviation of a third player).

In all the examples in the main body that prescribe a specific path (including the SES in Hotelling), the conclusions do not change under priority to the path. Hence, the insights from the examples are robust to this finer epistemic priority ordering. In the example of Section 2, there is no agreed-upon path, but suppose that entry was costless and E was player 2's payoff from a path that firm 1 and firm 2 agree to follow if firm 2 does not enter. Then, the analysis of Section 2 would be capturing strategic reasoning under priority to this path.

The path of the example above resembles a "*path that can be upset by a convincing deviation*", a notion proposed by Osborne [4] for repeated coordination games.[47] Osborne proves that such paths are not stable, in the sense of Kohlberg and Mertens [27]. I will prove that these outcomes cannot be implemented under priority to the path, because the corresponding path agreements are not credible. Battigalli and Siniscalchi [11] have shown an analogous relationship between the iterated intuitive criterion (Cho and Kreps [18]) and Strong-Δ-Rationalizability with belief restrictions on the equilibrium outcome distribution. Sobel et al. [5] provide similar arguments for divine equilibrium (Banks and Sobel [3]). In the aforementioned example of the companion paper, stability and implementability under priority to the path rule out the same SPE. The common message of these works is that strategic stability and related refinements capture instances of forward induction reasoning based on the belief in the equilibrium path. Still, these refinements focus on sequential equilibrium. But the logics of subgame perfection clash precisely with this particular way of rationalizing deviations: if the deviator believed in the equilibrium path, she certainly does not believe in the threat.[48] The example of Supplemental Appendix I is a case in point: if Ann and Bob reach an

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

[48] Interestingly, Man [3] finds that also the invariance argument, used to motivate the notions of forward induction of Kohlberg and Mertens [27] and Govindan and Wilson [22], does not imply sequential equilibrium.

agreement that prescribes the SPE outcome (S, E), but Ann deviates, if Bob believes that Ann believed that he would have played E, he must conclude that she does not believe in the SPE threat R, and then she may continue with any action except precisely her SPE action D.

To conclude, I prove that the paths that can be upset by a convincing deviation are not credible. Fix a two-player (i and j) static game G with action sets A_i and A_j and payoff function $v_k : A_i \times A_j \to \mathbb{R}$, $k = i, j$. Let b^k and c^k be the first- and second-ranked stage-outcomes of G for player $k = i, j$. A path $\overline{z} = (\overline{a}^1, .., \overline{a}^T)$ of Nash equilibria of the T-fold repetition of G *can be upset by a convincing deviation* if there exist $\tau \in \{1, ..., T-1\}$ and $\widehat{a}_i \neq \overline{a}_i^\tau$ such that, letting $\overline{T} := T - \tau$,

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

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

Condition I says that player i benefits from a unilateral deviation at τ only if followed by her preferred subpath.[49] Condition J says that player j cannot benefit from a unilateral deviation from that subpath even if followed by her preferred subpath.[50]

Proposition 13 *Let $\overline{z}$ be a path that can be upset by a convincing deviation. The path agreement on $\overline{z}$ is not credible.*

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

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

[50] This implies that i's preferred stage-outcome is Nash, reason why Osborne refers to coordination games.

Fix $n \in \mathbb{N}$ and suppose that $S_i^{n-1}(z) \neq \emptyset$. Fix $s_j \in S_j$ with $\overline{\mu}_i(s_j|h^0) \neq 0$. Since $\overline{\mu}_i$ strongly believes S_j^∞ and $S_j(\overline{z})$, $s_j \in S_j^\infty(\overline{z})$. Fix μ_j that strongly believes $(S_i^q)_{q=0}^\infty$ with $s_j \in br_j(\mu_j)$. Let $p(h) \in H$ be the immediate predecessor of h. Since $\overline{\mu}_j$ strongly believes $S_i(\overline{z})$, for each $h \notin H(S_i(\overline{z}))$ with $p(h) \prec \overline{z}$, $\overline{\mu}_j(S_i(h)|p(h)) = 0$. Thus, there exists μ_j' that strongly believes $(S_i^q)_{q=0}^{n-1}$ such that (i) $\mu_j'(\cdot|h^0) = \overline{\mu}_j(\cdot|h^0)$, (ii), $\mu_j'(S_i(z)|\widehat{h}) = 1$, and (iii) $\mu_j'(\cdot|h) = \mu_j(\cdot|h)$ for all $h \in H(S_j(\overline{z}))$ with $h \nprec \overline{z}$ and $h \nsucceq \widehat{h}$. Then, there exists $s_j' \in br_j(\mu_j') \subseteq S_j^n$ such that: by $br_j(\overline{\mu}_j) \cap S_j(\overline{z}) \neq \emptyset$, $\overline{\mu}_j(S_i(z)|h^0) = 1$, and (i), $s_j' \in S_j(\overline{z}) \subseteq S_j(\widehat{h})$; by (ii) and (J), $s_j' \in S_j(z)$; by (iii) and $s_j, s_j' \in S_j(\overline{z})$, $s_j'(h) = s_j(h)$ for all $h \in H(S_j(\overline{z}))$ with $h \nsucceq \widehat{h}$. With these s_j''s, I can construct μ_i that strongly believes $(S_j^q)_{q=0}^n$ such that $\mu_i(S_j(z)|h^0) = 1$, and $\mu_i(S_j(\widetilde{z})|h^0) = \overline{\mu}_i(S_j(\widetilde{z})|h^0)$ for all $\widetilde{z} \nsucc \widehat{h}$. Thus, by $br_i(\overline{\mu}_i) \cap S_i(\overline{z}) \neq \emptyset$, $\overline{\mu}_i(S_j(\overline{z})|h^0) = 1$, and (I), $\emptyset \neq br_i(\mu_i) \cap S_i(z) \subseteq S_i^{n+1}(z)$. So, by induction, there exists μ_i that strongly believes $(S_j^q)_{q=0}^\infty$ and $S_j(\overline{z})$ such that $\emptyset \neq br_i(\mu_i) \cap S_i(z) \subseteq S_{i,e}^1(z)$. On the other hand, for every μ_i that strongly believes $S_j(\overline{z})$, by (I) $br_i(\mu_i) \cap S_i(\widehat{h}) \subseteq S_i(z)$, so $S_{i,e}^1(\widehat{h}) \subseteq S_i(z)$. The two things combined imply that for every μ_j that strongly believes $S_{i,e}^1$ and $S_i(\overline{z})$, $\mu_j(S_i(z)|\widehat{h}) = 1$. So, by (J), $S_{j,e}^2(\widehat{h}) \subseteq S_j(z)$. Since $S_j(\overline{z}) \subseteq S_j(\widehat{h})$, for every μ_i that strongly believes $S_{j,e}^2$ and $S_j(\overline{z})$, $\mu_i(S_j(z)|h^0) = 1$, so by (I) $br_i(\mu_i)(\overline{z}) = \emptyset$. Hence $S_{i,e}^3(\overline{z}) = \emptyset$. So, $S_{j,e}^4 = \emptyset$. ■

References

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

[2] Catonini, E. "On non-monotonic strategic reasoning," *Games and Economic Behavior*, **120**, 2020, 209-224.

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

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

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