8 Supplemental Appendix

8.1 On SPE and self-enforcing agreements

Consider the following game.

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

$\longrightarrow$

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

All plans are justifiable, hence they are all rationalizable. The subgame has one pure equilibrium, (D,R), and no mixed equilibrium: for Ann to be indifferent between U and M, Bob must randomize over $\{L,C\}$, but when he is indifferent between them, he prefers R; for Ann to be indifferent between U and D or M and D, Bob must randomize over, respectively, $\{L,R\}$ and $\{C,R\}$, but R dominates L over $\{U,D\}$ and C over $\{M,D\}$. So, the game has only one SPE, inducing outcome (S,E).

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

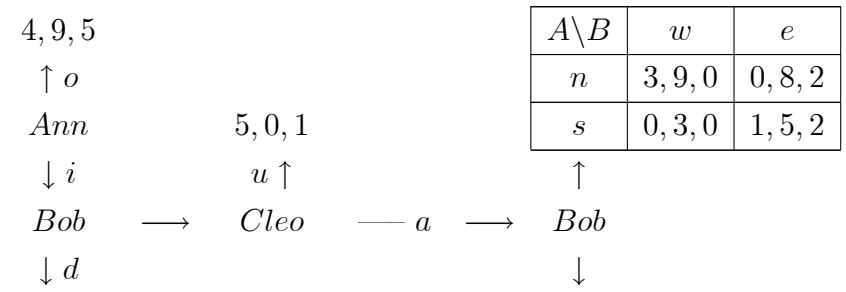
Note moreover that if Ann believes in the SPE path, it is not rational for her to deviate and then play D. Thus, if Bob interprets the deviation of Ann as an attempt to increase her payoff with respect to the equilibrium (as implicitly assumed by strategic stability and related refinements, see Section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

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

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

8.2 Another form of agreement incompleteness

Consider the following game.



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

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

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

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

satisfy

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

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

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

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

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

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

This kind of strategic uncertainty also arises naturally from strategic reasoning. The example on page 50 in Battigalli [6] (provided by Gul and Reny) shows that already the set of justifiable plans of a player is not a Cartesian product of sets of actions at different information sets. This is the reason why (selective) 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.

8.3 Proofs for Sections 5 and 6

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

Proof of Proposition 8

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

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

for each $n = k_i + 1, ..., k_i + M + 1$, let $\overline{e}_i^n = S_i^{k_i + M + 1 - n}$. Assume without loss of generality that $\overline{e}_i^n \subsetneq \overline{e}_i^{n+1}$ for each $n = 0, ..., k_i + M$,[47] so that $\overline{e} = (\overline{e}_i)_{i \in I}$ is an agreement. I am going to show that $\overline{e}$ is tight under priority to the agreement. Indicate with T1^a, T2^a and T3^a the conditions of tightness under priority to the agreement (i.e., with S_i in place of S_i^∞).

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

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

believes e^0_{-i} such that (ii) $\mu_i(S_{-i}(z)|h) = \mu'_i(S_{-i}(z)|h)$ for all $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$. Note that $\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}), \zeta(\rho_i(\mu'_i) \times \overline{e}^0_{-i}) \subseteq \zeta(S^\infty)$.[48] Then: by (ii), $\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}) = \zeta(\rho_i(\mu'_i) \times \overline{e}^0_{-i})$; by (i), $\zeta(\rho_i(\mu'_i) \times e^0_{-i}) = \zeta(\rho_i(\mu'_i) \times \overline{e}^0_{-i})$. So, we obtain

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

where the inclusion holds by T1 and the last equality by construction.

Moreover, $\overline{e}$ satisfies T2^a by $\overline{e}_i^{k_i+M+1} = S_i$. It remains to show that $\overline{e}$ satisfies T3^a. I will show later that[49]

$$\rho_i(\Delta_i^{\overline{e}}) \cap S_i = \left\{ s_i \in S_i^\infty : \exists s'_i \in \rho_i(\Delta_i^e) \cap S_i^\infty, \forall h \in H(s'_i) \cap H(S^\infty), s'_i(h) = s_i(h) \right\}. \tag{2}$$

Now, fix $h \in H(\rho_i(\Delta_i^{\overline{e}}) \cap S_i)$. Suppose first that either $h = h^0$ or $p(h) \in H(S^\infty)$. Then, by (2), $h \in H(\rho_i(\Delta_i^e) \cap S_i^\infty)$. By T3, there is n such that $\emptyset \neq e_i^n \cap S_i(h) \subseteq \rho_i(\Delta_i^e) \cap S_i^\infty$. Then, by definition of $\overline{e}_i^n$, $\overline{e}_i^n \cap S_i(h) \neq \emptyset$, and for each $s_i \in \overline{e}_i^n \cap S_i(h) \subset S_i^\infty$, there is $s'_i \in e_i^n$ such that $s'_i(h) = s_i(h)$ for all $h \in H(s'_i) \cap H(S^\infty)$, thus $s'_i \in e_i^n \cap S_i(h)$. So, $s'_i \in \rho_i(\Delta_i^e) \cap S_i^\infty$. But then, by (2), $s_i \in \rho_i(\Delta_i^{\overline{e}}) \cap S_i$.

Suppose now that $p(h) \notin H(S^\infty)$. Fix the unique $h' \prec h$ such that $h' \notin H(S^\infty)$ but $p(h') \in H(S^\infty)$. As shown, there is n such that $\emptyset \neq \overline{e}_i^n \cap S_i(h') \subseteq \rho_i(\Delta_i^{\overline{e}}) \cap S_i$. So, it suffices to show that $\overline{e}_i^n \cap S_i(h) \neq \emptyset$. Fix $s_i \in \rho_i(\Delta_i^{\overline{e}}) \cap S_i \cap S_i(h)$ and $s'_i \in \overline{e}_i^n \cap S_i(h') \subseteq \rho_i(\Delta_i^{\overline{e}}) \cap S_i$. By (2), $s_i, s'_i \in S_i^\infty$. Fix μ_i, μ'_i that strongly believe $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that $s_i \in \rho_i(\mu_i)$ and $s_i \in \rho_i(\mu'_i)$. Since $h' \notin H(S^\infty)$, $p(h') \in H(S^\infty)$, and $h' \in H(S_i^\infty)$, we have $h' \notin H(S_{-i}^\infty)$, so $\mu'_i(S_{-i}(h')|p(h)) = 0$. Then, I can construct μ''_i that strongly believes $((S_j^q)_{j \neq i})_{q=0}^\infty$ such that $\mu''_i(\cdot|\widetilde{h}) = \mu'_i(\cdot|\widetilde{h})$ for each $\widetilde{h} \not\succeq h'$ and $\mu''_i(\cdot|\widetilde{h}) = \mu_i(\cdot|\widetilde{h})$ for each $\widetilde{h} \succeq h'$. So, there is $s''_i \in \rho_i(\mu''_i) \subseteq S_i^\infty$ such that $s''_i(\widetilde{h}) = s'_i(\widetilde{h})$ for each $\widetilde{h} \not\succeq h'$ with $\widetilde{h} \in H(s_i)$ and $s''_i(\widetilde{h}) = s_i(\widetilde{h})$ for each $\widetilde{h} \succeq h'$ with $\widetilde{h} \in H(s'_i)$. Hence, $s''_i \in \overline{e}_i^n$

[48] To see this, fix $\widetilde{\mu}_i$ that strongly believes $((S_j^q)_{j \neq i})_{q=0}^\infty$ with $\widetilde{\mu}_i(\cdot|h) = \mu_i(\cdot|h)$ for all $h \in H(\overline{e}^0_{-i})$. So, $\zeta(\rho_i(\widetilde{\mu}_i) \times \overline{e}^0_{-i}) = \zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i})$. By $\rho_i(\widetilde{\mu}_i) \subseteq S_i^\infty$ and $\overline{e}^0_{-i} \subseteq S_{-i}^\infty$, $\zeta(\rho_i(\widetilde{\mu}_i) \times \overline{e}^0_{-i}) \subseteq \zeta(S^\infty)$. Hence, $\zeta(\rho_i(\mu_i) \times \overline{e}^0_{-i}) \subseteq \zeta(S^\infty)$ as well.

[49] Of course, the intersection with S_i is superfluous here. It will be substituted with S_i^∞ in the next proof.

and $s_i'' \in S_i(h)$.

Finally I prove (2). First I prove "$\subseteq$". For each $j \neq i$, $n = 0, ..., k_j$, and $s_j \in \overline{e}_j^n$, there is $s_j' \in e_j^n$ such that $s_j'(h) = s_j(h)$ for all $h \in H(S^\infty) \cap H(s_j')$. Moreover, by T2, $H(e_j^{k_j}) \supseteq H(S_j^\infty) \supseteq H(S^\infty)$. Hence, for each $\mu_i \in \Delta_i^{\overline{e}}$ (which strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ by construction of $\overline{e}$) I can construct $\mu_i' \in \Delta_i^e$ that strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ (which is possible because $e_j^{k_j} \subseteq S_j^\infty$) such that $\mu_i'(S_{-i}(z)|h) = \mu_i(S_{-i}(z)|h)$ for all $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$. Then, for every $s_i \in \rho_i(\mu_i) \cap S_i \subset S_i^\infty$, there is $s_i' \in \rho_i(\mu_i') \subset S_i^\infty$ such that $s_i'(h) = s_i(h)$ for all $h \in H(s_i') \cap H(S^\infty)$.

Now I prove "$\supseteq$". Fix $s_i \in S_i^\infty$, $\mu_i' \in \Delta_i^e$, and $s_i' \in \rho_i(\mu_i') \cap S_i^\infty$ with $s_i'(h) = s_i(h)$ for all $h \in H(s_i') \cap H(S^\infty)$. Fix μ_i'' that strongly believes $((S_j^q)_{j\neq i})_{q=0}^\infty$ such that $s_i \in \rho_i(\mu_i'')$. For each $j \neq i$, $n = 0, ..., k_j$, $s_j \in e_j^n \subseteq S_j^\infty$, $h \in H(s_j) \backslash H(S^\infty)$ with $p(h) \in H(S^\infty)$, and $s_j' \in S_j^\infty \cap S_j(h)$, fix μ_j, μ_j' that strongly believe $((S_k^q)_{k\neq j})_{q=0}^\infty$ such that $s_j \in \rho_j(\mu_j)$ and $s_j' \in \rho_j(\mu_j')$. By $\mu_j(S_{-j}(h)|p(h)) = 0$, I can construct μ_j'' that strongly believes $((S_k^q)_{k\neq j})_{q=0}^\infty$ such that $\mu_j''(\cdot|\widetilde{h}) = \mu_j(\cdot|\widetilde{h})$ for each $\widetilde{h} \not\succeq h$, and $\mu_j''(\cdot|\widetilde{h}) = \mu_j'(\cdot|\widetilde{h})$ for each $\widetilde{h} \succeq h$. So, there is $s_j'' \in \rho_j(\mu_j'') \subseteq S_j^\infty$ such that $s_j''(\widetilde{h}) = s_j(\widetilde{h})$ for each $\widetilde{h} \not\succeq h$ with $\widetilde{h} \in H(s_j)$ and $s_j''(\widetilde{h}) = s_j'(\widetilde{h})$ for each $\widetilde{h} \succeq h$ with $\widetilde{h} \in H(s_j')$. Hence, $s_j'' \in \overline{e}_j^n$. With all such s_j'''s, I can construct $\mu_i \in \Delta_i^{\overline{e}}$ such that $\mu_i(S_{-i}(z)|h) = \mu_i'(S_{-i}(z)|h)$ for all $h \in H(S^\infty)$ and $z \in \zeta(S^\infty)$, and $\mu_i(S_{-i}(z)|h) = \mu_i''(S_{-i}(z)|h)$ for all $h' \in H(S_i^\infty) \backslash H(S^\infty)$ with $p(h') \in H(S^\infty)$, $h \succeq h'$, and $z \succeq h$. Hence, $s_i \in \rho_i(\Delta_i^{\overline{e}})$. ■

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

Proposition 10 *Let $\overline{z} = (\overline{a}^1, ..., \overline{a}^T)$ be a path that can be upset by a convincing deviation. The path agreement on $\overline{z}$ is not credible.*

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

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

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

Condition I says that player i benefits from a unilateral deviation at τ only if followed by her preferred subpath.[50] Condition J says that player j cannot benefit from a unilateral deviation from that subpath even if followed by her preferred subpath.[51]

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

Fix $n \in \mathbb{N}$ and suppose that $S_i^{n-1}(z) \neq \emptyset$. Fix $s_j \in S_j$ with $\overline{\mu}_i(s_j|h^0) \neq 0$. Since $\overline{\mu}_i$ strongly believes S_j^∞ 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 $s_j \in \rho_j(\mu_j)$. Since $\overline{\mu}_j$ strongly believes $S_i(\overline{z})$, for each $h \notin H(S_i(\overline{z}))$ with $p(h) \prec \overline{z}$, $\overline{\mu}_j(S_i(h)|p(h)) = 0$. Thus, there exists 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 $h \not\prec \overline{z}$ and $h \not\succeq \widehat{h}$. Then, there exists $s_j' \in \rho_j(\mu_j') \subseteq S_j^n$ such that: by $\rho_j(\overline{\mu}_j) \cap S_j(\overline{z}) \neq \emptyset$, $\overline{\mu}_j(S_i(z)|h^0) = 1$, and (i), $s_j' \in S_j(\overline{z}) \subseteq S_j(\widehat{h})$; by (ii) and (J), $s_j' \in S_j(z)$; by (iii) and

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

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

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

8.4 On the definition of Selective Rationalizability.

Consider the following, alternative definition of Selective Rationalizability.

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

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

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

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

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

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

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

This is the definition of Selective Rationalizability provided and characterized epistemically in [18]. It differs from the definition used in this paper

because of requirement S3 in place of the requirement that $s_i \in S_i^\infty$. Here I argue that the two definitions are equivalent for the analysis of agreements.

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

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

Proof. By induction.

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

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

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

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

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

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

(ii), $\widehat{\mu}_i$ strongly believes $((S^m_{j,e})_{j\neq i})^n_{m=0}$ and, by (a), also $((S^m_j)_{j\neq i})^\infty_{m=0}$. By (i), $s_i \in \rho_i(\widehat{\mu}_i)$.

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

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

come up with new beliefs), this indirect effect can also be obtained directly by only restricting her behavior at the rationalizable histories. This can be proven formally with the same arguments of the proof of Proposition 8.

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

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

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

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

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

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

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

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

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