A PROOF OF THEOREM 4.1 – PROGRAM EQUILIBRIUM IMPLEMENTATIONS OF SAFE PARETO IMPROVEMENTS

This paper considers the meta-game of delegation. SPIs are a proposed way of playing these games. However, throughout most of this paper, we do not analyze the meta-game directly as a game using the typical tools of game theory. We here fill that gap and in particular prove Theorem 4.1, which shows that SPIs are played in Nash equilibria of the meta game, assuming sufficiently strong contracting abilities. As noted, this result is essential. However, since it is mostly an application of existing ideas from the literature on program equilibrium, we left a detailed treatment out of the main text.

A *program game* for $\Gamma = (A, \mathbf{u})$ is defined via a set $\mathrm{PROG} = \mathrm{PROG}_1 \times ... \times \mathrm{PROG}_n$ and a non-deterministic mapping $exec\colon \mathrm{PROG}_1 \times ... \times \mathrm{PROG}_n \rightsquigarrow A$. We obtain a new game with action sets PROG and utility function

$$U\colon \mathrm{PROG} \longrightarrow \mathbb{R}^n\colon \mathbf{c} \mapsto \mathbb{E}\left[\mathbf{u}(exec(\mathbf{c}))\right]. \tag{4}$$

Though this definition is generic, one generally imagines in the program equilibrium literature that for all i, PROG_i consists of computer programs in some programming language, such as Lisp, that takes as input vectors in PROG and returns an action a_i. The actions $exec(\mathbf{c})$ as follows. For Player i's function, we execute program c_i on input $\mathbf{c}$.[2] The definition implicitly assumes that PROG only contains programs that halt when fed one another as input. For the present paper, we add the following feature to the underlying programming language. A program can call a "black box subroutine" $\Pi_i(\Gamma')$ for any subset game Γ' of Γ, where $\Pi_i(\Gamma')$ is a random variable over A_i' and $\Pi(\Gamma') = (\Pi_1(\Gamma'), ..., \Pi_n(\Gamma'))$.

We need one more definition. For any game Γ and player i, we define Player i's *threat point* (a.k.a. minimax utility) v_i^Γ as

$$v_i^\Gamma = \min_{\boldsymbol{\sigma}_{-i} \in \bigtimes_{j \neq i} \Delta(A_j)} \; \max_{\sigma_i \in \Delta(A_i)} u_i(\sigma_i, \boldsymbol{\sigma}_{-i}). \tag{5}$$

In words, v_i^Γ is the minimum utility that the players other than i can force onto i, under the assumption that i reacts optimally to their strategy. We further will use $minimax(i, j) \in \Delta(A_j)$ to denote the strategy for Player j that is played in the minimizer σ_{-i} of the above. Of course, in general, there might be multiple minimizers σ_{-i}. In the following, we will assume that the function *minimax* breaks such ties in some consistent way, such that for all i,

$$(minimax(i, j))_{j \in \{1,...,n\} - \{i\}} \in \operatorname*{arg\,min}_{\boldsymbol{\sigma}_{-i} \in \bigtimes_{j \neq i} \Delta(A_j)} \; \max_{\sigma_i \in \Delta(A_i)} u_i(\sigma_i, \boldsymbol{\sigma}_{-i}). \tag{6}$$

Note also that each player's threat point is computable in polynomial time via linear programming. Note also that by the minimax theorem [33], the threat point is equal to the maximin utility, i.e.,

$$v_i^\Gamma = \max_{\sigma_i \in \Delta(A_i)} \; \min_{\boldsymbol{\sigma}_{-i} \in \bigtimes_{j \neq i} \Delta(A_j)} u_i(\sigma_i, \boldsymbol{\sigma}_{-i}), \tag{7}$$

so v_i^Γ is also the minimum utility that Player i can guarantee for herself under the assumption that the opponents see her mixed strategy and react in order to minimize Player i's utility.

The role of the threat points $v_1^\Gamma, ..., v_n^\Gamma$ in characterizations of equilibria of games derived from Γ is immediately intuitive. In particular, if some payoff vector $\mathbf{x} \in \mathbb{R}^n$ has some i such that $x_i < v_i^\Gamma$, then a feasible payoff vector $\mathbf{x}$ can't be a payoff vector of any equilibrium, since Player i would be better of by playing their maximin utility for a guaranteed expected utility of at least v_i^Γ. Tennenholtz [31] showa that for program games, the converse is also true. I.e., if $x_i > v_i^\Gamma$ for all i, then x_i is Player i's payoff in some program equilibrium for Γ. As has been known by game theorists since the 50s, the same is also true for equilibria of repeated games [20, Chapter 15].

We are now ready to construct for any SPI Γ^s on Γ, a program equilibrium that results in the play of $\Pi(\Gamma^s)$. As noted in Section 3, the Player i's instruction to her representative to play the game Γ^s will usually be conditional on the other player telling her representative to also play her part of Γ^s and *and vice versa.* After all, if Player i simply tells her representative to maximize u_i^s from A_i^s regardless of Player $-i$'s instruction, then Player $-i$ will often be able to profit from deviating from the Γ^s instruction. For example, in the bilateral safe Pareto improvement of Proposition (Example) 2, each player would only want their representative to choose from $\{\mathrm{DL}, \mathrm{RL}\}$ rather than $\{\mathrm{DM}, \mathrm{DM}\}$ if Player 2's representative does the same. It would then seem that in an equilibrium in which Γ^s is played, each program c_i would have to contain a condition of the type, "if the opponent code plays as in $\Pi(\Gamma^s)$ against me, I also play as I would in $\Pi(\Gamma^s)$." But in a naive implementation of this, each of the programs would have to call the other, leading to an infinite recursion.

In the literature on program equilibrium, various solutions to this problem have been discovered. We here use the general scheme proposed by Tennenholtz [31], because it is the simplest. We would similarly use the variant proposed by Fortnow [8], techniques based on Löb's theorem [3, 7], or ϵ-grounded mutual simulation [19] or even (meta) Assurance Game preferences (see Appendix B). These alternative approaches have

In our equilibrium, we let each player submit code as sketched in Algorithm 2. Roughly, each player uses a program that says, "if everyone else submitted the same source as this one, then play $\Pi(\Gamma^s)$. Otherwise, if there is a player j who submits a different source code, punish player j with her threat point". Before showing that this is the desired equilibrium, a few remarks. For convenience, Algorithm 2 receives the player number i as input. This way, every player can use the exact same source code. Otherwise the original players would have to provide slightly different programs and in line 2 of the algorithm, we would have to use a more complicated comparison, roughly: "if $c_j \neq c_i$ are the same, except for the player index used".

Proposition A.1. *Let* $\mathbf{c}$ *be the program profile consisting only of Algorithm 2 for each player. Assume that* $\Pi(\Gamma)$ *guarantees each player at least threat point utility. Then* $\mathbf{c}$ *is a program equilibrium and* $apply(\mathbf{c}) = \Pi(\Gamma^s)$.

Proof. By inspection of Algorithm 2, we see that $apply(\mathbf{c}) = \Pi(\Gamma^s)$. It is left to show that $\mathbf{c}$ is a Nash equilibrium. So let i

[2] Note that we here assume that each program receives not only everyone else's program source codes but also its own source code as input. This convenient for us (as Algorithm 2 references its own source code). It is also realistic: computer programs can open their source code as a file on the computer on which they are running and contracts can and frequently refer to themselves. However, by the diagonal lemma, we could also have written a version of Algorithm 2 which reproduces its source code into c_i and thus needs no explicit access to ability to reference its source code. (This is also known as quining.)

Algorithm 2: A program equilibrium implementation of an SPI Γ^s of Γ.

```
Data: Everybody's source code $\mathbf{c}$, my index $i$
for $j \in \{1, ..., n\} - \{i\}$ do
    if $c_j \neq c_i$ then
        Play player $i$'s strategy in player $j$'s threat point;
Play $\Pi_i(\Gamma^s)$;
```

be any player and $c_i' \in \mathrm{PROG}_i - \{c_i\}$. We need to show that $\mathbb{E}\left[u_i(apply(\mathbf{c}_{-i}, c_i'))\right] \leq \mathbb{E}\left[u_i(apply(\mathbf{c}))\right]$. Again, by inspection $apply(\mathbf{c}_{-i}, c_i')$, is the threat point of Player i. Hence,

$$\begin{aligned} \mathbb{E}\left[u_i(apply(\mathbf{c}_{-i}, c_i'))\right] &= v_i && (8)\\ &\leq \mathbb{E}\left[u_i(\Pi(\Gamma))\right] && (9)\\ &\leq \mathbb{E}\left[u_i(\Pi(\Gamma^s))\right] && (10)\\ &\leq \mathbb{E}\left[u_i(apply(\mathbf{c}))\right] && (11) \end{aligned}$$

□

Theorem 4.1 follows immediately.

B A DISCUSSION OF WORK BY SEN (1974) AND RAUB (1990) ON PREFERENCE ADAPTION GAMES

We here discuss Raub's [25] paper in some detail, which in turn elaborates on an idea by Sen [28]. Superficially, Raub's setting seems somewhat similar to ours, but we here argue that it should be thought of as closer to the work on bilateral precommitment.

In Sections 1 and 3, we briefly discuss multilateral commitment games, which have been discussed before in various forms in the game-theoretic literature. Our paper extends this setting by allowing instructions that let the representatives play a game without specifying an algorithm for solving that game. On first sight, it appears that Raub pursues a very similar idea. Translated to our setting, Raub allows that as an instruction, each player i chooses a new utility function $u_i^s \colon A \to \mathbb{R}$, where A is the set of outcomes of the original game Γ. Given instructions $u_1^s, ..., u_n^s$, the representatives then play the game $(A, \mathbf{u}^s)$. In particular, each representative can see what utility functions all the other representatives have been instructed to maximize. However, what utility function representative i maximizes is not conditional on any of the instructions by other players. In other words, the instructions in Raub's are raw utility functions, not the more complicated type of instructions that motivate our study of SPIs. Raub then asks for equilibria $\mathbf{u}^s$ of the meta-game that Pareto-improve on the default outcome.

To better understand how Raub's approach relates to ours, we here give an example of the kind of instructions Raub has in mind. (Raub uses the same example in his paper.) As the underlying game Γ, we take the Prisoner's Dilemma. Now the main idea of his paper is that the original players can instruct their representatives to adopt so-called *Assurance Game* preferences. In the Prisoner's Dilemma, this means that the representatives prefer to cooperate if the other representative cooperates, and prefer to defect if the other player defects. Further, they prefer mutual cooperation over mutual defection. An example of such Assurance Game preferences is given in Table 3.

		Player 2	
		Cooperate	Defect
Player 1	Cooperate	4, 4	1, 3
	Defect	3, 1	2, 2

Table 3: Assurance Game preferences for the Prisoner's Dilemma

The Assurance Game preferences have two important properties.

(1) If both players tell their representatives to adopt Assurance Game preferences, (Cooperate, Cooperate) is a Nash equilibrium. (Defect, Defect) is a Nash equilibrium as well. However, since (Cooperate, Cooperate) is Pareto-better than (Defect, Defect), the original players could reasonably expect that the representatives play (Cooperate, Cooperate).

(2) Under reasonable assumptions about the rationality of the representatives, it is a Nash equilibrium of the meta-game for both players to adopt Assurance Game preferences. If Player 1 tells her representative to adopt Assurance Game preferences, then Player 2 maximizes his utility by telling his representative to also maximize Assurance Game preferences. After all, representative 1 prefers defecting if representative 2 defects. Hence, if Player 2 instructs his representative to adopt preferences that suggest defecting, then he should expect representative to defect as well.

The first important difference between Raub's approach and ours related to item 2. We have ignored the issue of making SPIs Γ^s Nash equilibria of our meta game. As we have explained in Section 3 and Appendix A, we imagine that this is taken care of by additional bilateral commitment mechanisms. For Raub's paper, on the other hand, ensuring mutual cooperation to be stable in the meta game is arguably the key idea. Still, we could pursue the approach of the present paper even when we limited assumptions to those that consist only of a utility function.

The second difference is even more important. Raub assumes that – as in the PD – the default outcome of the game ($\Pi(\Gamma)$ in the formalism of this paper) is known. (Less significantly, he also assumes that it is known how the representatives play under assurance game preferences.) Of course, the key feature of the setting of this paper is that the underlying game Γ might be difficult (through equilibrium selection problems) and thus that the original players might be unable to predict $\Pi(\Gamma)$.

These are the reasons why we cite Raub in our section on bilateral commitment mechanisms. Arguably, Raub's paper could be seen as very early work on program equilibrium, except that he uses utility functions as a programming language for representative. In this sense, Raub's Assurance Game preferences are analogous to the program equilibrium schemes of Tennenholtz [31], Oesterheld [19], Barasz et al. [3] and van der Hoek et al. [32], ordered in increasing order of similarity of the main idea of the scheme.

C PROOF OF PROPOSITION 5.2

Lemma 5.2. *Let* $\Gamma = (A, \mathbf{u}), \Gamma' = (A', \mathbf{u}'), \hat{\Gamma} = (\hat{A}, \hat{\mathbf{u}})$ *and* $\Phi, \Xi \colon A \multimap A', \Psi \colon A' \multimap \hat{A}$.

1. *Reflexivity:* $\Gamma \sim_{\mathrm{id}_A} \Gamma$*, where* $\mathrm{id}_A \colon A \multimap A \colon \mathbf{a} \mapsto \{\mathbf{a}\}$.
2. *Symmetry: If* $\Gamma \sim_\Phi \Gamma'$*, then* $\Gamma' \sim_{\Phi^{-1}} \Gamma$.
3. *Transitivity: If* $\Gamma \sim_\Phi \Gamma'$ *and* $\Gamma' \sim_\Psi \hat{\Gamma}$*, then* $\Gamma \sim_{\Psi \circ \Phi} \hat{\Gamma}$.
4. *If* $\Gamma \sim_\Phi \Gamma'$ *and* $\Phi(\mathbf{a}) \subseteq \Xi(\mathbf{a})$ *for all* $\mathbf{a} \in A$*, then* $\Gamma \sim_\Xi \Gamma'$.
5. $\Gamma \sim_{\mathrm{all}_{A,A'}} \Gamma'$*, where* $\mathrm{all}_{A,A'} \colon A \multimap A' \colon \mathbf{a} \mapsto A'$.
6. *If* $\Gamma \sim_\Phi \Gamma'$ *and* $\Phi(\mathbf{a}) = \emptyset$*, then* $\Pi(\Gamma) \neq \mathbf{a}$ *with certainty.*

Proof. 1. By reflexivity of equality, $\Pi(\Gamma) = \Pi(\Gamma)$ with certainty. Hence, $\Pi(\Gamma) \in \mathrm{id}_A(\Pi(\Gamma))$ by definition of id_A. Therefore, $\Gamma \sim_{\mathrm{id}_A} \Gamma$ by definition of $\sim$, as claimed.

2. $\Gamma \sim_\Phi \Gamma'$ means that $\Pi(\Gamma') \in \Phi(\Pi(\Gamma))$ with certainty. Thus,
$$\Pi(\Gamma) \in \{\mathbf{a} {\in} A \mid \Pi(\Gamma') {\in} \Phi(\mathbf{a})\} = \Phi^{-1}(\Pi(\Gamma')),$$
where the equation is by the definition of the inverse of multi-valued functions. We conclude (by definition of $\sim$) that $\Gamma' \sim_{\Phi^{-1}} \Gamma$ as claimed.

3. If $\Gamma \sim_\Phi \Gamma', \Gamma' \sim_\Psi \hat{\Gamma}$, then by definition of $\sim$, (i) $\Pi(\Gamma') \in \Phi(\Pi(\Gamma))$ and (ii) $\Pi(\hat{\Gamma}) \in \Psi(\Pi(\Gamma'))$, both with certainty. The former (i) implies $\{\Pi(\Gamma')\} \subseteq \Phi(\Pi(\Gamma))$. Hence,
$$\Psi(\Pi(\Gamma')) = \Psi(\{\Pi(\Gamma')\}) \subseteq \Psi(\Phi(\Pi(\Gamma))).$$
With ii, it follows that $\Pi(\hat{\Gamma}) \in \Psi(\Phi(\Pi(\Gamma)))$ with certainty. By definition, $\Gamma \sim_{\Psi \circ \Phi} \hat{\Gamma}$ as claimed.

4. It is
$$\Pi(\Gamma') \in \Phi(\Pi(\Gamma)) \subseteq \Xi(\Pi(\Gamma))$$
with certainty. Thus, by definition $\Gamma \sim_\Xi \Gamma'$.

5. By definition of Π, it is $\Pi(\Gamma') \in A'$ with certainty. By definition of $\mathrm{all}_{A,A'}$, it is $\mathrm{all}_{A,A'}(\Pi(\Gamma)) = A'$ with certainty. Hence, $\Pi(\Gamma') \in \mathrm{all}_{A,A'}(\Pi(\Gamma))$ with certainty. We conclude that $\Gamma \sim_{\mathrm{all}_{A,A'}} \Gamma'$ as claimed.

6. With certainty, $\Pi(\Gamma') \in \Phi(\Pi(\Gamma))$ (by assumption). Also, with certainty $\Pi(\Gamma') \notin \emptyset$. Hence, $\Phi(\Pi(\Gamma)) \neq \emptyset$ with certainty. We conclude that $\Pi(\Gamma) \neq \mathbf{a}$ with certainty.

□

D EXAMPLES

D.1 Proof of Proposition (Example) 1

Proposition (Example) 1. *Let* Γ *be the Prisoner's Dilemma and* $\Gamma^s = (A_1^s, A_2^s, u_1^s, u_2^s)$ *be any subset game of* Γ *with* $A_1^s = A_2^s = \{\text{Cooperate}\}$*. Then under Assumption 2,* Γ^s *is a strict SPI on* Γ.

Proof. By applying Assumption 2 twice and Transitivity once, $\Gamma \sim_\Phi \Gamma - \{\text{Cooperate}\}$, where $\Phi(\text{Defect}, \text{Defect}) = \{(\text{Defect}, \text{Defect})\}$ and $\Phi(a_1, a_2) = \emptyset$ for all $(a_1, a_2) \neq (\text{Defect}, \text{Defect})$. By Lemma 5.2.5, we further obtain $\Gamma - \{\text{Cooperate}\} \sim_{\mathrm{all}} \Gamma^s$, where Γ^s is as described in the proposition. Hence, by transitivity, $\Gamma \sim_{\mathrm{all} \circ \Phi} \Gamma^s$. It is easy to verify that the function $\mathrm{all} \circ \Phi$ is Pareto-improving. □

D.2 Proof of Proposition (Example) 2

Proposition (Example) 2. *Let* Γ *be the Demand Game of Table 1 and* Γ^s *be the subset game described in Table 2. Under Assumptions 1 and 2,* Γ^s *is an SPI on* Γ*. Further, if* $P(\Pi(\Gamma){=}(DM, DM)) > 0$*,* Γ^s *is a strict safe Pareto improvement.*

Proof. Let $(A_1, A_2, u_1, u_2) = \Gamma$. We can repeatedly apply Assumption 2 to eliminate from Γ the strategies DL and DR for both players. We can then apply Lemma 5.2.3 (Transitivity) to obtain $G \sim_\Phi \hat{G} = (\{\text{DM}, \text{RM}\}, \{\text{DM}, \text{RM}\}, u_1, u_2)$, where
$$\Phi(a_1, a_2) = \begin{cases} \{(a_1, a_2)\} & \text{if } a_1, a_2 \in \{\text{DM}, \text{RM}\} \\ \emptyset & \text{otherwise} \end{cases}. \tag{12}$$
Next, by Assumption 1, $\hat{\Gamma} \sim_\Psi \Gamma^s$, where $\Psi_i(\text{DM}) = \text{DL}$ and $\Psi_i(\text{RM}) = \text{RL}$ for $i = 1, 2$. We can then apply Lemma 5.2.3 (Transitivity) again, to infer $\Gamma \sim_{\Psi \circ \Phi} \Gamma^s$. It is easy to verify that for all $(a_1, a_2) \in A_1 \times A_2$, it is for all $(a_1^s, a_2^s) \in \Psi(\Phi(\Gamma^s))$ the case that $\mathbf{u}(a_1^s, a_2^s) \geq \mathbf{u}(a_1, a_2)$. □

E PROOF OF THEOREM 8.2

We here prove Theorem 8.2. In our proof we focus in particular on a version with a strictness requirement which requires a few additional ideas.

We start by showing that the SPI problem is in NP at all. The following algorithm can be used to determine whether there is a safe Pareto improvement: Reduce the given game Γ until it can be reduced no further to obtain some subset game $\Gamma' = (A', \mathbf{u})$. Then non-deterministically select injections $\Phi_i \colon A_i' \to A_i$. If $\Phi = (\Phi_1, ..., \Phi_n)$ is (strictly) Pareto-improving (as required in Theorem 6.1), return True with the solution Γ^s defined as follows: The set of action profiles is defined as $A^s = \bigtimes_i \Phi_i(A_i')$. The utility functions are
$$u_i^s \colon A^s \to \mathbb{R} \colon \mathbf{a}^s \mapsto (u_i(\Phi_1^{-1}(a_1^s), ..., \Phi_n^{-1}(a_n^s)))_{i=1,...,n}. \tag{13}$$
Otherwise, return False.

It is easy to see that this algorithm runs in non-deterministic polynomial time. Furthermore, it is easy to see that if this algorithm finds a solution Γ^s, that solution is indeed a safe Pareto improvement. It is left to show that if there is a safe Pareto improvement via a sequence of Assumption 1 and 2 outcome correspondences, then the algorithm indeed finds a safe Pareto improvement. To prove this fact, we prove a few simple lemmata.

First, one might worry that the algorithm only ever finds sequences of outcome correspondences that start with a number of reductions and end with a single isomorphism step. Perhaps some safe Pareto improvements can only be found by considering very different sequences? The following two lemmata show that this is not an issue, i.e., that it is sufficient to consider sequences that start with a number of reductions and end in a single isomorphism step.

Lemma E.1. *Let* $\Gamma \sim_{\Phi^{\mathrm{iso}}} \hat{\Gamma}$ *by Assumption 1 and* $\hat{\Gamma} \sim_{\Phi^{\mathrm{red}}} \tilde{\Gamma}$ *by Assumption 2. Then there is a* Γ' *s.t.* $\Gamma \sim_{\Psi^{\mathrm{red}}} \Gamma'$ *by Assumption 2,* $\Gamma' \sim_{\Psi^{\mathrm{iso}}} \tilde{\Gamma}$ *by Assumption 1 and* $\Psi^{\mathrm{iso}} \circ \Psi^{\mathrm{red}} = \Phi^{\mathrm{red}} \circ \Phi^{\mathrm{iso}}$.

Intuitively, this means that isomorphism steps as per Assumption 1 and reduction steps as per Assumption 2 commute. Instead of first applying Assumption 1 and then Assumption 2 to a game, we can also apply Assumption 2 first and then Assumption 1 to obtain the same game $\tilde{\Gamma}$ in both cases.

Proof. We construct Γ', Ψ^{red}, Ψ^{iso} as follows. First, $\Gamma' = (A_1', ..., A_n', \mathbf{u}')$, where $A_i' = (\Phi_i^{\mathrm{iso}})^{-1}(\tilde{A}_i)$ and $u_i' = u_i|_{A'}$. Next, we define
$$\Psi^{\mathrm{red}} = (\Phi^{\mathrm{iso}})^{-1} \circ \Phi^{\mathrm{red}} \circ \Phi^{\mathrm{iso}} \tag{14}$$

and

$$\Psi^{\text{iso}} = \Phi^{\text{iso}}|_{A'_1 \times A'_2}. \tag{15}$$

We now need to show that these satisfy the consequents of the lemma.

First, it is

$$\begin{aligned} \Psi^b \circ \Psi^{\text{red}} &= \Phi^{\text{iso}}|_{A'_1 \times A'_2} \circ \left((\Phi^{\text{iso}})^{-1} \circ \Phi^{\text{red}} \circ \Phi^{\text{iso}} \right) && (16) \\ &= \left(\Phi^{\text{iso}}|_{A'_1 \times A'_2} \circ (\Phi^{\text{iso}})^{-1} \right) \circ \Phi^{\text{red}} \circ \Phi^{\text{iso}} && (17) \\ &= \Phi^{\text{red}} \circ \Phi^{\text{iso}} && (18) \end{aligned}$$

as claimed. Note that the second step uses the associativity of $\circ$ on multivalued functions. The third step uses the fact that Φ^{iso} is a single-valued bijection, which means that $\Phi^{\text{iso}} \circ \left(\Phi^{\text{iso}}\right)^{-1} = \text{id}$. Of course, Φ^{iso} is here restricted to $A'_1 \times A'_2$, but this is not a problem, because $A'_i = (\Phi^{\text{iso}}_i)^{-1}(\tilde{A}_i)$ and $\tilde{A}$ is the codomain of $\Phi^{\text{red}} \circ \Phi^{\text{iso}}$. Hence, the restriction is inconsequential.

Second, we need to show that it is indeed $\Gamma' \sim_{\Psi^{\text{iso}}} \tilde{\Gamma}$ by Assumption 1. First, it is easy to show that Ψ^{iso} decomposes into $\Psi^{\text{iso}}_1, ..., \Psi^{\text{iso}}_n$ as required because Φ^{iso} decomposes. Further, Ψ^{iso} is a single-valued injection because Φ^{iso} is a single-valued bijection. It is surjective because its codomain is defined as the image of its domain.

Now let $(a'_1, a'_2) \in A'_1 \times A'_2$. It is

$$\begin{aligned} \mathbf{u}'(\mathbf{a}') &= \mathbf{u}(\mathbf{a}') && (19) \\ &= \hat{\mathbf{u}}(\Phi^{\text{iso}}(\mathbf{a}')) && (20) \\ &= \tilde{\mathbf{u}}(\Phi^{\text{iso}}(\mathbf{a}')) && (21) \\ &= \tilde{\mathbf{u}}(\Psi^b(\mathbf{a}')), && (22) \end{aligned}$$

as required.

Finally, we have to show that $\Gamma \sim_{\Psi^{\text{red}}} \Gamma'$ by Assumption 2. We leave this as an exercise to the reader. □

Lemma E.2. *Let*

$$\Gamma^1 \sim_{\Phi^1} ... \sim_{\Phi^{k-1}} \Gamma^k, \tag{23}$$

where each outcome correspondence is due to a single application of Assumption 1 or 2. Then there is a sequence $\Gamma'^2, ..., \Gamma'^m$ with $m \leq k-1$ such that

$$\Gamma^1 \sim_{\Psi^1} \Gamma'^2 \sim_{\Psi^2} \Gamma'^3 \sim_{\Psi^3} ... \sim_{\Psi^{m-1}} \Gamma'^m \tag{24}$$

all by single applications of Assumption 2, $\Gamma'^m \sim_{\Xi} \Gamma^k$ by a single application of Assumption 1, and

$$\Phi^{k-1} \circ \Phi^{k-2} \circ ... \circ \Phi^1 = \Xi \circ \Psi^{m-1} \circ ... \circ \Psi^1. \tag{25}$$

Proof. Start with the initial sequence of line 23. We can iteratively apply Lemma E.1 to obtain a new sequence of the same length in which one first applies only Assumption 2 and then only Assumption 1 while obtaining the same composite outcome correspondence function. We can summarize all the applications of Assumption 1 into a single step applying that assumption. □

A second potential worry about our algorithm is that it reduces the game completely and only then looks for a Pareto-improving isomorphism step. Perhaps in some cases one has to only *partially* reduce and then look for a Pareto-improving isomorphism step? The next lemma shows that the answer to this is no and that one can restrict oneself to sequences that fully reduce.

Lemma E.3. *Let $\Gamma = (A, \mathbf{u})$, $\hat{\Gamma}^a = (\hat{A}^a, \mathbf{u})$, $\hat{\Gamma}^b = (\hat{A}^b, \mathbf{u})$ such that $\hat{A}^b_i \subseteq \hat{A}^a_i \subseteq A_i$ for $i = 1, ..., n$. If there is a subset game $\tilde{\Gamma}^a = (\tilde{A}^a, \tilde{\mathbf{u}}^a)$ of Γ such that $\hat{\Gamma}^a \sim_{\Phi} \tilde{\Gamma}^a$ by Assumption 1, then $\hat{\Gamma}^b \sim_{\Phi|_{\hat{A}^b}} \tilde{\Gamma}^b$, where $\tilde{\Gamma}^b = (\Phi_1(\hat{A}^b_1), ..., \Phi_n(\hat{A}^b_n), \tilde{\mathbf{u}}^a)$. Note that if the correspondence function Φ is Pareto-improving, so is $\Phi|_{\hat{A}^b}$.*

Lemma E.3 shows that it is enough to consider isomorphism steps from fully reduced versions of Γ. A third worry might be that even so, elimination via Assumption 2 might be path-dependent and therefore we have to consider the resulting games from multiple paths. However, iterated elimination of strictly dominated strategies is path-independent [1, 22].

Proposition E.4. *If there is a safe Pareto improvement for a given game, then the above algorithm applied to that game returns True.*

Proof. Let us say there is a sequence of outcome correspondences as per Assumptions 1 and 2. Then by Lemma E.2, there is Γ' such that $\Gamma \sim_{\Psi} \Gamma'$ via an arbitrary number of applications of Assumption 2 and $\Gamma' \sim_{\Phi} \Gamma^s$ via a single application of Assumption 1. Because of the path-independence of iterated removal of strictly dominated strategies, Γ' contains (as a subset game with equal utility functions) the unique Γ^r arising from iterated removal as per Assumption 2. By Lemma E.3, there is a Pareto-improving outcome correspondence $\Gamma^r \sim_{\Phi'} \Gamma^{sr}$ as per Assumption 2. □

Overall, we have now shown that our non-deterministic polynomial-time algorithm is correct and therefore that the SPI problem is in NP. Note that the correctness of other algorithms can be proven using very similar ideas. For example, instead of first reducing and then finding an isomorphism, one could first find an isomorphism, then reduce and then (only after reducing) test whether the overall outcome correspondence function is Pareto-improving. One advantage of reducing first is that there are fewer isomorphisms to test if the game is smaller. In particular, the number of possible isomorphisms is exponential in the number of strategies in the reduced game Γ' but polynomial in everything else. Hence, by implementing our algorithm deterministically, we obtain the following positive result.

Proposition 8.3. *For games Γ with $|A_1| + ... + |A_n| = m$ that can be reduced (via iterative application of Assumption 2) to a game Γ' with $|A'_1| + ... + |A'_n| = l$, the (strict) SPI decision problem can be solved in $O(m^l)$.*

We now proceed to showing that the safe Pareto improvement problem is NP-hard. We will do this reducing the subgraph isomorphism problem to the (two-player) safe Pareto improvement problem. We start by briefly describing one version of that problem here.

A *(simple, directed) graph* is a tuple $(n, a\colon \{1, ..., n\} \times \{1, ..., n\} \rightarrow \mathbb{B})$, where $n \in \mathbb{N}$ and $\mathbb{B} \coloneqq \{0, 1\}$. We call a the adjacency function of the graph. Since the graph is supposed to be simple and therefore free of self-loops (edges from one vertex to itself), we take the values $a(j, j)$ for $j \in \{1, ..., n\}$ to be meaningless.

For given graphs $G = (n, a)$,$G' = (n', a')$ a subgraph isomorphism from G to G' is an injection $\phi\colon \{1, ..., n\} \rightarrow \{1, ...n'\}$ such that for all $j \neq l$

$$a(j,l) \leq a'(\phi(j),\phi(l)). \tag{26}$$

In words, a subgraph isomorphism from G to G' identifies for each node in G a node in G' s.t. if there is an edge between nodes j and l in G, there must also be an edge (in the same direction) between the corresponding nodes in G'. Another way to say this is that we can remove some set of $(n' - n)$ nodes and some edges from G' to get a graph that is just a relabeled (isomorphic) version of G.

Given two graphs G, G', the subgraph isomorphism problem consists in deciding whether there is a subgraph isomorphism ϕ between G, G'. The problem is well-known to be NP-complete [6, Theorem 2].

Lemma E.5. *The subgraph isomorphism problem is reducible in linear time with linear increase in problem instance size to the safe Pareto improvement problem. As a consequence, the safe Pareto improvement problem is NP-hard.*

Proof. We conduct our proof only for the strict safe Pareto improvement problem. Reducing to the non-strict safe Pareto improvement problem is a little easier and can be done with the same ideas as in this proof.

So take graphs $G = (n, a)$ and $\hat{G} = (\hat{n}, \hat{a})$. We will transform these step-wise into a single game.

First, we define the games $\Gamma^a = (A_1, A_2, u_1, u_2)$ and $\hat{\Gamma}^a = (\hat{A}_1, \hat{A}_2, \hat{u}_1, \hat{u}_2)$, where $A_1 = A_2 = \{1, ..., n\}$,$\hat{A}_1 = \hat{A}_2 = \{1, ..., \hat{n}\}$, $\mathbf{u}(j,l) = a(j,l)$ for all $j,l \in \{1, ..., n\}$ with $j \neq l$ and $u_1(j,j) = u_2(j,j) = 2$. We analogously define $\hat{u}_1 = \hat{u}_2$ based on $\hat{a}$. Setting the utility functions is the main idea of the entire proof, of course, and will become clearer below. Setting the utilities $u_1(j,j) = u_2(j,j) = 2$ is to ensure that Pareto-improving mappings Φ between Γ^a and $\hat{\Gamma}^a$ satisfy $\Phi_1(j) = \Phi_2(j)$ for all j, and thus directly relate to subgraph isomorphisms.

Next, we add dummy strategies to Γ^a,$\hat{\Gamma}^a$, to obtain two purposes. We want to remove exact equivalences and allow only *strict Pareto improvements*; and we want to remove the possibility of reducing either of these games via Assumption 2. In particular, we consider Γ^b as follows: $A_i^b = \{1, ..., 2n\}$; $\mathbf{u}^b(j,l) = \mathbf{u}^a(j,l)$ if $j,l \in \{1, ..., n\}$, $\mathbf{u}^b(j, n{+}j) = \mathbf{u}^b(n{+}j, n) = (3,3)$ for $j \in \{1, .., n\}$, $\mathbf{u}^b(j,l) = (-1,-1)$ otherwise. We define $\hat{\Gamma}^b$ analogously, except that utilities of $(3,3)$ are to be replaced by $(4,4)$.

Finally, we construct from $\Gamma^b, \hat{\Gamma}^b$ a single game Γ^c. Roughly, the idea is for Γ^c to contain as subset games both Γ^b and $\hat{\Gamma}^b$, but to reduce to Γ^b via Assumption 2. We construct Γ^c thus: $A_i^c = (\{D\} \times A_i^b) \cup (\{C\} \times \hat{A}_i^b)$ and

$$\begin{aligned}
\mathbf{u}^c((C,\hat{a}_1),(C,\hat{a}_2)) &= \hat{\mathbf{u}}^b(\hat{a}_1,\hat{a}_2) \text{ for all } \hat{a}_1 \in \hat{A}_1^b, \hat{a}_2 \in \hat{A}_2^b\\
\mathbf{u}^c((D,a_1),(D,a_2)) &= \mathbf{u}^b(a_1,a_2) \text{ for all } a_1 \in A_1^b, a_2 \in A_2^b\\
u_i^c((D,a_i),(C,\hat{a}_{-i})) &= 5 \text{ for all } a_i \in A_i^b, \hat{a}_{-i} \in \hat{A}_2^b\\
u_{-i}^c((D,a_i),(C,\hat{a}_{-i})) &= -5 \text{ for all } a_i \in A_i^b, \hat{a}_{-i} \in \hat{A}_2^b.
\end{aligned}$$

It is easy to show that this reduction can be computed in linear time and that it also increases the problem instance size only linearly. It is left to prove the correctness of the reduction.

We start by showing that if there is a subgraph isomorphism from G to $\hat{G}$, then there is also a safe (strict) Pareto improvement via a sequence of outcome correspondence as per Assumptions 1 and 2. So let ϕ be that subgraph isomorphism. Then we need to construct a series of outcome correspondences as per Assumptions 1 and 2.

First notice that $\Gamma^c \sim_{\Xi} \Gamma^{c,D}$ by Assumption 2, where $\Gamma^{c,D}$ is the subset game of Γ^c that contains only the strategies of type (D,j) for both players. We now show that $\Gamma^{c,D} \sim_{\Psi} \tilde{\Gamma}$ via Assumption 1, where $\tilde{\Gamma} = (\tilde{A}_1, \tilde{A}_2, \tilde{u}_1, \tilde{u}_2)$ and Ψ are defined as follows:

$$\tilde{A}_1 = \tilde{A}_2 = \{C\} \times (\phi(\{1, ..., n\}) \cup \{\hat{n} + \phi(j) \mid j = 1, ..., n\}); \tag{27}$$

Ψ_i for $i = 1,2$ is defined as $\Psi_i(D,j) = (C, \phi(j))$ if $j \in \{1, ..., n\}$ and $\Psi_i(D,j) = (C, \hat{n} + \phi(j-n))$ otherwise; and the utility function is simply defined as $\tilde{u}_i(\tilde{a}_1, \tilde{a}_2) = u_i(\Psi_i^{-1}(\tilde{a}_1, \tilde{a}_2))$. It is easy to see that Ψ_i is surjective for $i = 1,2$. From the fact that ϕ is injective and that ϕ never returns values greater than $\hat{n}$, it follows that Ψ_i is also injective for $i = 1,2$. Finally, Ψ maintains utilities by definition:

$$\tilde{u}_i(\Psi(a_1,a_2)) = u_i(\Psi^{-1}(\Psi(a_1,a_2))) = u_i(a_1,a_2). \tag{28}$$

With Transitivity, it is left to show that Ψ is Pareto-improving for the original players, i.e., for $\mathbf{u}^c$. For this we distinguish a number of different cases. If $j,l \in \{1, ..., n\}$ and $j \neq l$, then for $i = 1,2$

$$\begin{aligned}
u_i^c((D,j),(D,l)) &= a(j,l) && (29)\\
&\leq \hat{a}(\phi(j),\phi(l)) && (30)\\
&= \hat{u}_i^c((C,\phi(j)),(C,\phi(l))) && (31)\\
&= \hat{u}_i^c(\Psi_1(D,j),\Psi_2(D,l)). && (32)
\end{aligned}$$

For $j \in \{1, ..., n\}$, it is

$$\begin{aligned}
\mathbf{u}^c((D,j),(D,j)) &= (2,2) && (33)\\
&= \hat{\mathbf{u}}^c((C,\phi(j)),(C,\phi(j))) && (34)\\
&= \hat{\mathbf{u}}^c(\Psi_1(D,j),\Psi_2(D,j)). && (35)
\end{aligned}$$

For $j,l \in \{n+1, ..., 2n\}$, it is

$$\begin{aligned}
\mathbf{u}^c((D,j),(D,l)) &= \mathbf{u}^b(j,l) && (36)\\
&= (-1,-1) && (37)\\
&= \hat{\mathbf{u}}^b(\hat{n} + \phi(j-n), \hat{n} + \phi(l-n)) && (38)\\
&= \mathbf{u}^c(\Psi_1(D,j),\Psi_2(D,l)). && (39)
\end{aligned}$$

If $(j,l) \in \{1, ..., n\} \times \{n+1, ..., 2n\}$ and $l = j + n$, then

$$\begin{aligned}
\mathbf{u}^c((D,j),(D,l)) &= (2,2) && (40)\\
&< (3,3) && (41)\\
&= \mathbf{u}^c((C,\Phi(j)),(C,\Phi(j)+\hat{n})) && (42)\\
&= \mathbf{u}^c(\Psi_1(D,j),\Psi(C,l)). && (43)
\end{aligned}$$

If $l \neq j + n$, then

$$\begin{aligned}
\mathbf{u}^c((D,j),(D,l)) &= \mathbf{u}^b(j,l) && (44)\\
&= (-1,-1) && (45)\\
&= \hat{\mathbf{u}}^b(\phi(j), \hat{n} + \phi(l-n)) && (46)\\
&= \mathbf{u}^c(\Psi_1(D,j),\Psi_2(D,l)). && (47)
\end{aligned}$$

The cases $(j,l) \in \{n+1, ..., 2n\} \times \{1, ..., n\}$ work analogously.

This concludes our proof that if a graph isomorphism exists, there also exists an SPI as per Assumptions 1 and 2.

It is left to prove that if there is a safe Pareto improvement for Γ^c, then there also exists a graph isomorphism. So let $\Gamma^c \sim_{\Xi} \tilde{\Gamma}$ for some $\tilde{\Gamma}$, via some Pareto-improving outcome correspondence function

Ξ. By our earlier results (Proposition E.4), this means that there is a sequence of outcome correspondences that first fully reduces Γ to $\Gamma^{c,D}$ and then applies Assumption 1 to get $\Gamma^{c,D} \sim_{\Psi} \tilde{\Gamma}$ via some Pareto-improving Ψ.

To construct a subgraph isomorphism, we must now realize some facts about the structure of Ψ that all follow from Ψ being utility-increasing:

(1) $\Psi\left((\{D\} \times \{1, ..., 2n\}) \times (\{D\} \times \{1, ..., 2n\})\right) \subseteq ((\{C\} \times \{1, ..., 2\hat{n}\}) \times (\{C\} \times \{1, ..., 2\hat{n}\}))$: For Ψ to be strict, there has to be some overlap, i.e., there has to be (D, j) s.t. $\Psi_i(D, j) \in \{C\} \times \{1, ..., 2\hat{n}\}$. But for Ψ to be Pareto-improving for i, it has to be for all (D, l) with $l = 1, ..., 2n$ the case that $\Psi_{-i}(D, l) \in \{C\} \times \{1, ..., 2\hat{n}\}$ since otherwise it would be

$$\begin{aligned} u_i(\Psi_i(D, j), \Psi_{-i}(D, l)) &= -5 & (48)\\ &< u_i((D, j), (D, l)). & (49)\end{aligned}$$

By an analogous argument it can further be shown that $\Psi_i(D, j) \in \{C\} \times \{1, ..., 2\hat{n}\}$ for all $j = 1, ..., 2n$.

(2) For every $j \in \{1, ..., n\}$, it is $\Psi_i(D, j) \in \{C\} \times \{1, ..., \hat{n}\}$ for $i = 1, 2$. That is, Ψ_i maps the "non-dummy" strategies (which correspond to nodes in the original graph) onto "non-dummy" strategies. We will show this by showing the contrapositive, i.e., that if this were not the case, then Ψ would not be Pareto-improving.
So assume there is $j \in \{1, ..., n\}$ and $i \in \{1, 2\}$ with $\Psi_i(D, j) \in \{C\} \times \{\hat{n} + 1, ..., 2\hat{n}\}$. Because Ψ_{-i} is injective, there is an $l \in \{1, ..., n\}$ s.t. $\Psi_{-i}(D, l) \neq \Psi(j) - \hat{n}$, where we define $(D, k) - \hat{n} := (D, k - \hat{n})$. Then for that l it is

$$\begin{aligned} \mathbf{u}(\Psi_i(D, j), \Psi_{-i}(D, l)) &= (-1, -1) & (50)\\ &< (0, 0) & (51)\\ &\leq \mathbf{u}^c((D, j), (D, l)). & (52)\end{aligned}$$

(3) For all $j \in \{1, ..., n\}$ and $i \in \{1, ..., n\}$ it is $\Psi_i(D, j + n) = \Psi_{-i}(D, j) + \hat{n}$, where addition is defined to operate only on the second entry like the subtraction defined above. We again prove the contrapositive, i.e., if this is not true then Ψ is not Pareto-improving. So let us assume that there is $j \in \{1, ..., n\}$ and $i \in \{1, ..., n\}$ s.t. $\Psi_i(D, j + n) \neq \Psi_{-i}(D, j)$. Then it would be

$$\begin{aligned} \mathbf{u}^c(\Psi_i(D, j + n), \Psi_{-i}(D, j)) &\leq (2, 2) & (53)\\ &< (3, 3) & (54)\\ &= \mathbf{u}^c((C, j + n), (C, j)). & (55)\end{aligned}$$

(4) Finally, we prove that $\Psi_1 = \Psi_2$. We first show that for $j \in \{1, ..., n\}$ it is $\Psi_1(D, j) = \Psi_2(D, j)$. We do this again by showing the contrapositive. So assume $\Psi_1(D, j) \neq \Psi_2(D, j)$. Recall that by item 2, it is $\Psi_1(D, j), \Psi_2(D, j) \in \{C\} \times \{1, ..., \hat{n}\}$. Hence,

$$\begin{aligned} \mathbf{u}(\Psi_1(D, j), \Psi_2(D, j)) &\leq (1, 1) & (56)\\ &< (3, 3) & (57)\\ &= \mathbf{u}((D, j), (D, j)), & (58)\end{aligned}$$

contradicting the assumption that Ψ is Pareto-improving.

Finally, for $j \in \{1, ..., n\}$ it is

$$\begin{aligned} \Psi_1(D, n + j) &\underset{\text{Item 3}}{=} \hat{n} + \Psi_2(D, j) & (59)\\ &= \hat{n} + \Psi_1(D, j) & (60)\\ &\underset{\text{Item 3}}{=} \Psi_2(D, n + j), & (61)\end{aligned}$$

where the middle equality is due to the equality we have already proven.

Given these, we can define our graph isomorphism as

$$\phi\colon \{1, ..., n\} \to \{1, ..., \hat{n}\} \colon j \mapsto \pi_2(\Psi_1(D, j)), \qquad (62)$$

where π_2 just maps pairs (C, j) onto the second entry j. This is well-defined because of item 2. Note that because Ψ is an injection, so is ϕ.

For all $j, l \in \{1, ..., n\}$ with $j \neq l$ it is

$$\begin{aligned} \hat{a}(\phi(j), \phi(l)) &= \hat{u}_1^a(\phi(j), \phi(l)) & (63)\\ &= \hat{u}_1^c((C, \phi(j)), (C, (\phi(l)))) & (64)\\ &\underset{\text{Item 1}}{=} \hat{u}_1^c(\Psi_1(D, j), \Psi_1(D, l)) & (65)\\ &\underset{\text{Item 4}}{=} \hat{u}_1^c(\Psi_1(D, j), \Psi_2(D, l)) & (66)\\ &\geq u_1^c((D, j), (D, l)) & (67)\\ &= u_1^a(j, l) & (68)\\ &= a(j, l). & (69)\end{aligned}$$

Hence, ϕ is a subgraph isomorphism as desired. □

F PROOF OF LEMMA 9.3

Lemma 9.3. *For a given n-player game Γ and payoff vector $\mathbf{y} \in \mathbb{R}^n$, it can be decided by linear programming and thus in polynomial time whether $\mathbf{y}$ is Pareto-optimal in $C(\Gamma)$.*

For an introduction to linear programming, see, e.g., Schrijver [27]. In short, a linear program is a specific type of constrained optimization problem that can be solved efficiently.

Finding a Pareto-improvement on a given $\mathbf{x} \in \mathbb{R}^n$ can be formulated as the following linear program:

$$\begin{aligned} \text{Variables:}\quad & p_{\mathbf{a}} \in [0, 1] \text{ for all } \mathbf{a} \in A\\ \text{Maximize}\quad & \sum_{i=1}^{n} \left(\sum_{\mathbf{a} \in A} p_{\mathbf{a}} u_i(\mathbf{s}) \right) - x_i\\ \text{s.t.}\quad & \sum_{\mathbf{a} \in A} p_{\mathbf{a}} = 1\\ & \sum_{\mathbf{a} \in A} p_{\mathbf{a}} u_i(\mathbf{a}) \geq x_i \text{ for } i = 1, ..., n\end{aligned}$$

G PROOF OF LEMMA 9.5

Lemma 9.5. *Let $\Gamma = (\{a_1^1, ..., a_1^{l_1}\}, ..., \{a_n^1, ..., a_n^{l_n}\}, \mathbf{u})$ be any game. Let Γ' be a perfect-coordination SPI on Γ. Then we can define $\mathbf{u}^e$ with values in $C(\Gamma)$ such that under Assumption 1 the game*

$$\Gamma^s = \Big(\hat{A}_1 := \{\hat{a}_1^1, ..., \hat{a}_1^{l_1}\}, ..., \hat{A}_n := \{\hat{a}_n^1, ..., \hat{a}_n^{l_n}\},$$
$$\hat{\mathbf{u}}\colon (\hat{a}_1^{i_1}, ..., \hat{a}_n^{i_n}) \mapsto \mathbf{u}(a_1^{i_1}, ..., a_n^{i_n}), \mathbf{u}^e \Big) \qquad (2)$$

is also an SPI on Γ*, with*

$$\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma^s)) \mid \Pi(\Gamma)=\mathbf{a}\right] = \mathbb{E}\left[\mathbf{u}(\Pi(\Gamma')) \mid \Pi(\Gamma)=\mathbf{a}\right] \tag{3}$$

for all $\mathbf{a} \in A$ *and consequently* $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma^s))\right] = \mathbb{E}\left[\mathbf{u}(\Pi(\Gamma'))\right]$.

Proof. For any SPI Γ', construct Γ^s as in Equation 2 with

$$\mathbf{u}^e(\hat{a}_1^{i_1},...,\hat{a}_n^{i_n}) = \mathbb{E}\left[\mathbf{u}'(\Pi(\Gamma')) \mid \Pi(\Gamma) = (a_1^{i_1},...,a_n^{i_n})\right]. \tag{70}$$

Here $\mathbf{u}'$ describes the utilities that the original players assign to the outcomes of Γ'. Since $\mathbf{u}'$ maps onto $C(\Gamma)$ and $C(\Gamma)$ is convex, $\mathbf{u}^e$ as defined also maps into $C(\Gamma)$ as required. Note that for all $a_1^{i_1},...,a_n^{i_n}$ it is by assumption $\mathbf{u}'(\Pi(\Gamma')) \geq \mathbf{u}(a_1^{i_1},...,a_n^{i_n})$ with certainty. Hence,

$$\begin{aligned} u^e(\hat{a}_1^{i_1},...,\hat{a}_n^{i_n})) &= \mathbb{E}\left[\mathbf{u}'(\Pi(\Gamma')) \mid \Pi(\Gamma) = (a_1^{i_1},...,a_n^{i_n})\right] \qquad (71) \\ &\geq \mathbf{u}(a_1^{i_1},...,a_n^{i_n}). \qquad (72) \end{aligned}$$

By Assumption 1, Γ^s constructed in this way satisfies Equation 3. □

H PROOF OF THEOREM 9.7

Proof. We will give the proof based on the graphs as well, without giving all formal details. Further we assume in the following that neither L_1 nor L_3 consist of just a single point, since these cases are easy.

Case A: Note first that if $\mathbf{y}$ Pareto-improves on $\mathbf{y}'$ which in turn Pareto-improves on $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$ and $\mathbf{y}$ can be made safe, then so can $\mathbf{y}'$. Thus, it is enough to show that if $\mathbf{y}$ is in any of the listed sets L_1, L_2, L_3, it can be made safe.

It's easy to see that all payoff vectors on the curve segment of the Pareto frontier L_2 are safely achievable. After all, all payoff vectors in this set Pareto-improve on all outcomes in $\mathrm{supp}(\Pi(\Gamma))$. Hence, for each $\mathbf{y}$ on the line segment, one could select the Γ^s where $\mathbf{u}^e = \mathbf{u}$. It is left to show that all elements of $L_{1/2}$ are safely achievable.

Next, it is to be argued that all Pareto improvements on the other two line segments are achievable. Remember that not all payoff vectors on the line segments are Pareto improvements, only those that are to the north-east of (Pareto-better than) the default utility. In the following, we will use L_1' and L_3' to denote those elements of L_1 and L_3, respectively, that are Pareto-improvements on the default.

We now argue that the Pareto improvement $\mathbf{y}$ on the line L_1 for which $y_1 = \mathbb{E}\left[u_1(\Pi(\Gamma))\right]$ are safely achievable. In other words, $\mathbf{y}$ is the projection northward of the default utility, or $\mathbf{y} = \pi_1(\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right], L_1)$. This $\mathbf{y}$ is also one of the endpoints of L_1. To achieve this utility, we construct the equivalent game as per Lemma 9.5, where the utility to the original players of each outcome $(\hat{a}_1, \hat{a}_2)$ of the new game Γ^s is similarly the projection northward of the utility of the corresponding outcome (a_1, a_2) in Γ^s. That is,

$$\mathbf{u}^e(\hat{a}_1, \hat{a}_2) = \pi_1(\mathbf{u}(a_1, a_2), L_1). \tag{73}$$

Note that because $C(\Gamma)$ is convex and the endpoints of the line segment L_1 are by definition in $C(\Gamma)$, it is $L_1 \subseteq C(\Gamma)$). Hence, all values of $\mathbf{u}^e$ as defined in Eq. 73 are feasible. Because all outcomes in the original game lie below the line L_1, π_1 is linear. Hence,

$$\begin{aligned} \mathbb{E}\left[\mathbf{u}^e(\Pi(\Gamma^s))\right] &= \mathbb{E}\left[\pi_1(\mathbf{u}(\Pi(\Gamma)), L_1)\right] \qquad (74) \\ &= \pi_1(\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right], L_1) \qquad (75) \end{aligned}$$

as required.

We have now shown that one of the endpoints of L_1' is safely achievable. Since the other endpoint of L_1' is in L_2, it is also safely achievable. By Corollary 9.6, this implies that all of L_1' is safely achievable.

By an analogous line of reasoning, we can also show that all elements of L_3' are safely achievable.

Case B: Define L_1', L_3' as before as those elements of L_1, L_3 respectively that Pareto improve on the default $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$. By a similar argument as before, one can show that the utilities $\pi_i(\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right], L_j')$ is safely achievable both for $i = 1, j = 1$ and for $i = 1, j = 3$. Call these points E_1 and E_3, respectively.

We now proceed in two steps. First, we will show that there is a third safely achievable utility point E_2, which is above both L_1 and L_3. Then we will show the claim using that point.

To construct E_2, we again construct an SPI Γ^s as per Lemma 9.5. For each $(a_1, a_2) \in A_1 \times A_2$ we will set the utility $u^e(\hat{a}_1, \hat{a}_2)$ of the corresponding $(\hat{a}_1, \hat{a}_2) \in \hat{A}_1 \times \hat{A}_2$ to be above or on both L_1 and L_3, i.e., on or above a set which we will refer to as $\max(L_1, L_3)$. Formally, $\max(L_1, L_3)$ is the set of outcomes in $L_1 \cup L_3$ that are not strictly Pareto dominated by some other element of $L_1' \cup L_3'$. Note that by definition every outcome in $\mathrm{supp}(\Pi(\Gamma))$ is Pareto-dominated by some outcome in either L_1 or L_3. Hence, by transitivity of Pareto dominance, each outcome is Pareto-dominated by some outcome in $\max(L_1, L_3)$. Hence, the described $\mathbf{u}^e$ is indeed feasible.

Now note that the set of feasible payoffs of Γ is convex. Further, the curve $\max(L_1, L_3)$ is concave. Because the area above a concave curve is convex and because the intersection of convex sets is convex, the set of feasible payoffs on or above $\max(L_1, L_3)$ is also convex. By definition of convexity, $E_2 = \mathbb{E}\left[\mathbf{u}^e(\Pi(\Gamma^s))\right]$ is therefore also in the set of feasible payoffs on or above $\max(L_1, L_3)$ and therefore above both L_1 and L_3 as desired.

In our second step, we now use E_1, E_2, E_3 to prove the claim. Because of convexity of the set of safely achievable payoff vectors as per Corollary 9.6, all utilities below the curve consisting of the line segments from E_1 to E_2 and from E_2 to E_3 are safely achievable. The line that goes through E_1, E_2 intersects the line that contains L_1 at E_1, by definition. Since non-parallel lines intersect each other exactly once and parallel lines that intersect each other are equal and because E_2 is above or on L_1, the line segment from E_1 to E_2 lies entirely on or above L_1. Similarly, it can be shown that the line segment from E_2 to E_3 lies entirely on or above L_3. It follows that the $E_1 - E_2 - E_3$ curve lies entirely above or on $\min(L_1, L_3)$. Now take any Pareto improvement that lies below both L_1' and L_3'. Then this Pareto improvement lies below $\min(L_1', L_3')$ and therefore below the $E_1 - E_2 - E_3$ curve. Hence, it is safely achievable. □

I PROOF OF PROPOSITION 9.8

Proposition 9.8. *There is a game* $\Gamma = (A, \mathbf{u})$*, representatives* Π *that satisfy Assumptions 1 and 2, and an outcome* $\mathbf{a} \in A$ *s.t.* $u_i(\mathbf{a}) > u_i(\Pi(\Gamma))$ *for all players i, but there is no perfect-coordination SPI* $(A^s, \mathbf{u}^s, \mathbf{u}^e)$ *s.t. for all players i,* $\mathbb{E}\left[u_i^e(\Pi(A^s, \mathbf{u}^s)\right] = u_i(\mathbf{a})$.

		Player 2		
		a	b	c
Player 1	a	$-5, -5$	$4, 0$	$10, -100$
	b	$0, 4$	$1, 1$	$10, -100$
	c	$-100, 10$	$-100, 10$	$3, 3$

Table 4: An example of a game in which – depending on Π – a Pareto improvement may not be safely achievable.

Proof. Consider the game in Table 4. Strategy c can be eliminated by strict dominance (Assumption 2) for both players, leaving a typical Chicken-like payoff structure with two pure Nash equilibria ((a, b) and (b, a)), as well as a mixed Nash equilibrium $(3/8 * a + 5/8 * b, 3/8 * a + 5/8 * b)$.

Now let us say that in the resulting game $P(\Pi(\Gamma){=}(a, b)) = p = P(\Pi(\Gamma){=}(b, a))$ for some p with $0 < p \leq 1/2$. Then one (unsafe) Pareto-improvement would be to simply always have the representatives play (c, c) for a certain payoff of $(3, 3)$. Unfortunately, there is no *safe* Pareto improvement with the same expected payoff. Notice that $(3, 3)$ is the unique element of $C(\Gamma)$ that maximizes the sum of the two players' utility. By linearity of expectation and convexity of $C(\Gamma)$, if for any Γ^s it is $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma^s))\right] = (3, 3)$, it must be $\mathbf{u}(\Pi(\Gamma^s)) = (3, 3)$ with certainty. Unfortunately, in any safe Pareto improvement the outcomes (a, b) and (b, a) must corresponds to outcomest that still gives utilities of $(4, 0)$ and $(0, 4)$, respectively, because these are Pareto-optimal within the set of feasible payoff vectors.

□