Safe Pareto Improvements for Delegated Game Playing*

Caspar Oesterheld and Vincent Conitzer

February 15, 2021

Abstract

A set of players delegate playing a game to a set of representatives, one for each player. We imagine that each player trusts their respective representative's strategic abilities. Thus, we might imagine that per default, the original players would simply instruct the representatives to play the original game as best as they can. In this paper, we ask: are there safe Pareto improvements on this default way of giving instructions? That is, we imagine that the original players can coordinate to tell their representatives to only consider some subset of the available strategies and to assign utilities to outcomes differently than the original players. Then can the original players do this in such a way that the payoff is guaranteed to be weakly higher than under the default instructions for all the original players? In particular, can they Pareto-improve without probabilistic assumptions about how the representatives play games? In this paper, we give some examples of safe Pareto improvements. We prove that the notion of safe Pareto improvements is closely related to a notion of outcome correspondence between games. We also show that under some specific assumptions about how the representatives play games, finding safe Pareto improvements is NP-complete.

Keywords: program equilibrium; delegation; bargaining; Pareto efficiency; smart contracts

*This draft is work-in-progress. Please send any comments to `caspar.oesterheld@duke.edu`. A shorter version of this paper has been published in the Proceedings of AAMAS 2021; see `https://users.cs.duke.edu/~ocaspar/SPIAAMAS.pdf` for a version of the AAMAS publication with appendices.

1 Introduction

Between Aliceland and Bobbesia lies a sparsely populated desert. Until recently, neither of the two countries had any interest in the desert. However, geologists have recently discovered that it contains large oil reserves. Now, both Aliceland and Bobbesia would like to annex the desert, but they worry about a military conflict that would ensue if both countries insist on annexing.

Table 1 models this strategic situation as a normal-form game. The strategy DM (short for "Demand with Military") denotes a military invasion of the desert, demanding annexation. If both countries send their military with such an aggressive mission, the countries fight a devastating war. The strategy RM (for "Refrain with Mility") denotes yielding the territory to the other country, but building defenses to prevent an invasion of one's original territories. Alternatively, the countries can choose to not raise a military force at all, while potentially still demanding control of the desert by sending only its leader (DL, short for "Demand with Leader"). In this case, if both countries demand the desert, war does not ensue. Finally, they could neither demand nor build up a mility (RL). If one of the two countries has their military ready and the other does not, the militarized country will know and will be able to invade the other country. In game-theoretical terms, militarizing therefore strictly dominates not militarizing.

Instead of making the decision directly, the parliaments of Aliceland and Bobbesia appoint special commissions for making this strategic decision, led by Alice and Bob, respectively. The parliaments can instruct these *representatives* in various ways. They can explicitly tell them what to do – for example, Aliceland could directly tell Alice to play DM. However, we imagine that the parliaments trust the commissions' judgments more than they trust their own and hence they might prefer to give an instruction of the type, "make whatever demands you think are best for our country" (perhaps contractually guaranteeing a reward in proportion to the utility of the final outcome). They might not know what that will entail, i.e., how the commissions decide what demands to make given that instruction. However – based on their trust in their representatives – they might still believe that this leads to better outcomes than giving an explicit instruction.

We will also imagine these instructions are (or at least can be) given publicly and that the commissions are bound (as if by a contract) to follow these instructions. In particular, we imagine that the two commissions can see each other's instructions. Thus, in instructing their commissions, the countries play a game with bilateral precommitment. When instructed to play a game as best as they can, we imagine that the commissions play that game in the usual way, i.e., without further abilities to credibly commit or to instruct subcommittees and so forth.

It may seem that without having their parliaments ponder equilibrium selection, Aliceland and Bobbesia cannot do better than leave the game to their representatives. Unfortunately, in this default equilibrium, war is still a possibility. Even the brilliant strategists Alice and Bob may not always be able to resolve the difficult equilibrium selection problem to the same pure Nash equilibrium.

In the literature on commitment devices and in particular the literature on program equilibrium, important ideas have been proposed for avoiding such bad outcomes. Imagine for a moment that Alice and Bob will play a Prisoner's Dilemma (rather than the Demand Game of Table 1). Then the default of (Defect, Defect) can be Pareto-improved upon. Both original players (Aliceland and Bobbesia) can use the following instruction for their representatives: "If the opponent's instruction is equal to this instruction, Cooperate; otherwise Defect." [11, 15, 26] Then it is a Nash equilibrium for both players to use this instruction. In this equilibrium, (Cooperate, Cooperate) is played and it is thus Pareto-optimal and Pareto-better than the default.

In cases like the Demand Game, it is more difficult to apply this approach to improve upon the default of simply delegating the choice. Of course, if one could calculate the expected utility of submitting the default instructions, then one could similarly commit the representatives to follow some (joint) mix over the Pareto-optimal outcomes ((RM, DM), (DM, RM), (RM, RM), (DL, DL), etc.) that Pareto-improves on the default expected utilities. However, we will assume that the original players are unable or unwilling to form probabilistic expectations about how the representatives play the Demand Game, i.e., about what would happen with the default instructions. If this is the case, then this type of Pareto-improvement on the default is unappealing.

The goal of this paper is to show and analyze how even without forming probabilistic beliefs about the representatives, the original players can Pareto-improve on the default equilibrium. We will call such improvements *safe Pareto improvements* (SPIs). We here briefly give an example in the Demand Game.

The key idea is for the original players to instruct the representatives to select only from {DL, RL}, i.e., to not raise a military. Further, they tell them to disvalue the conflict outcome (DL, DL) as they would disvalue the original conflict outcome of war in the default equilibrium. Overall, this

		Player 2			
		DM	RM	DL	RL
Player 1	DM	$-5, -5$	$2, 0$	$5, -5$	$5, -5$
	RM	$0, 2$	$1, 1$	$5, -5$	$5, -5$
	DL	$-5, 5$	$-5, 5$	$1, 1$	$2, 0$
	RL	$-5, 5$	$-5, 5$	$0, 2$	$1, 1$

Table 1: The Demand Game

		Player 2's rep.	
		DL	RL
Player 1's rep.	DL	$-5, -5$	$2, 0$
	RL	$0, 2$	$1, 1$

Table 2: A safe Pareto improvement for the Demand Game

means telling them to play the game of Table 2. (Again, we could imagine that the instructions specify Table 2 to be how Aliceland and Bobbesia financially reward Alice and Bob.) Importantly, Aliceland's instruction to play that game must be conditional on Bobbesia also instructing their commission to play that game, and vice versa. Otherwise, one of the countries could profit from deviating by instructing their representative to always play DM or RM (or to play by the original utility function).

The game of Table 2 is isomorphic to the DM-RM part of the original Demand Game of Table 1. Of course, the original players know neither how the original Demand Game nor the game of Table 2 will be played by the representatives. However, since these games are isomorphic, one should arguably expect them to be played isomorphically. For example, one should expect that $(\mathrm{RM}, \mathrm{DM})$ would be played in the original game if and only if $(\mathrm{RL}, \mathrm{DL})$ would be played in the modified game. However, the conflict outcome $(\mathrm{DM}, \mathrm{DM})$ is replaced in the new game with the outcome $(\mathrm{DL}, \mathrm{DL})$. This outcome is harmless (Pareto-optimal) for the original players.

Contributions Our paper generalizes this idea to arbitrary normal-form games and is organized as follows. In Section 2, we introduce some notation for games and multivalued functions that we will use throughout this paper. In Section 3, we introduce the setting of delegated game playing for this paper and define and motivate the concept of safe Pareto improvements in more detail. In Section 4, we briefly review the the concepts of program games and program equilibrium and show that SPIs can be implemented as program equilibria. In Section 5, we introduce a notion of outcome correspondence between games. This relation expresses the original players' beliefs about similarities between how the representatives play different games. For example, in our example, the Demand Game of Table 1 (arguably) corresponds to the game of Table 2 in that the representatives (arguably) would play (DM, DM) in the original game if and only if they play (DL, DL) in the new game, and so forth. We also show some basic results (reflexivity, transitivity, etc.) about the outcome correspondence relation on games. In Section 6 we show that the notion of outcome correspondence is central to deriving SPIs. In particular, we show that a game Γ^s is an SPI on another game Γ if and only if there is a Pareto-improving outcome correspondence relation between Γ^s and Γ.

To derive SPIs, we need to make some assumptions about outcome correspondence, i.e., about which games are played in similar ways by representatives. We give two very weak assumptions of this type in Section 7. The first is that the representatives play isomorphic games isomorphically. The second is that the representatives' play is invariant under the removal of strictly dominated strategies. For example, we assume that in the Demand Game the representatives only play DM and RM. Moreover we assume that we could remove DL and RL from the game and the representatives would still play the same strategies as in the original Demand Game with certainty. Our SPI for the Demand Game can be proven using these assumptions. Section 8 shows that determining whether there exists an SPI based on these assumptions is NP-complete. Section 9 considers a different setting in which we allow the original players to let the representatives choose from newly constructed strategies whose corresponding outcomes map arbitrarily onto feasible payoff vectors from the original game. In this new setting, finding SPIs can be done in polynomial time. We conclude by discussing the problem of selecting between different SPIs on a given game (Section 10) and giving some ideas for directions for future work (Section 11).

2 Preliminaries

2.1 Games

We here recall some basic definitions from game theory. An *n-player game* is a tuple $(A, \mathbf{u})$ of a set $A = A_1 \times ... \times A_n$ of *(pure) strategy profiles* (or outcomes) and a function $\mathbf{u}\colon A \to \mathbb{R}^n$ that assigns to each outcome a utility for each player. Instead of, $(A, \mathbf{u})$ will also write $(A_1, ..., A_n, u_1, ..., u_n)$. We say that $a_i \in A_i$ *strictly dominates* $a_i' \in A_i$ if for all $a_{-i} \in A_{-i}$, $u_i(a_i, a_{-i}) > u_i(a_i', a_{-i})$. For any given game $\Gamma = (A, \mathbf{u})$, we will call any game $\Gamma' = (A', \mathbf{u}')$ a *subset game* of Γ if $A_i' \subseteq A_i$ for $i = 1, ..., n$. Note that a subset game may assign different utilities to outcomes than the original game. For any set of strategies S, we denote by $\Gamma - S \coloneqq ((A_1 - S) \times ...., (A_n - S), \mathbf{u})$ the game that arises from Γ by removing the strategies S for all players.

We say that some utility vector $\mathbf{y} \in \mathbb{R}^n$ is a Pareto-improvement on (or is Pareto-better than) $\mathbf{y}' \in \mathbb{R}^n$ if $y_i \geq y_i'$ for $i = 1, ..., n$. We will also denote this by $\mathbf{y} \geq \mathbf{y}'$. Note that, contrary to convention, we allow $\mathbf{y} = \mathbf{y}'$. Whenever we require one of the inequalities to be strict, we will say that $\mathbf{y}$ is a strict Pareto improvement on $\mathbf{y}'$. In a given game, we will also say that an outcome $\mathbf{a}$ is a Pareto-improvement on another outcome $\mathbf{a}'$ if $\mathbf{u}(\mathbf{a}) \geq \mathbf{u}(\mathbf{a}')$. We say that $\mathbf{y}$ is Pareto-optimal or Pareto-efficient relative to some $S \subset \mathbb{R}^n$ if there is no element of S that strictly Pareto-dominates $\mathbf{y}$.

The Demand Game of Table 1 is an example of a game that we will use throughout this paper. As noted earlier, DM and RM strictly dominate DL and RL. The game of Table 2 is a subset game of the Demand Game.

2.2 Multivalued functions

For sets M and N, a *multi-valued function* $\Phi\colon M \multimap N$ is a function which maps each element $m \in M$ to a set $\Phi(m) \subseteq N$. For a subset $Q \subseteq M$, we define $\Phi(Q) \coloneqq \bigcup_{m \in Q} \Phi(m)$. Note that $\Phi(Q) \subseteq N$ and that $\Phi(\emptyset) = \emptyset$. For any set M, we define the identity function $\mathrm{id}_M\colon M \multimap M\colon m \to \{m\}$. Also, for two sets M, N, we define $\mathrm{all}_{M,N}\colon M \multimap N\colon m \mapsto N$. We define the inverse $\Phi^{-1}\colon N \multimap M\colon n \mapsto \{m \in M \mid n \in \Phi(m)\}$. Note that $\Phi^{-1}(\emptyset) = \emptyset$ for any multi-valued function Φ. For sets M, N, Q and functions $\Phi\colon M \multimap N$, $\Psi\colon N \multimap Q$, we define the composite $\Psi \circ \Phi\colon M \multimap Q\colon m \mapsto \Psi(\Phi(m))$. As with regular functions, composition of multi-valued functions is associative. We say that $\Phi\colon M \multimap N$ is *single-valued* if $|\Phi(m)| = 1$ for all $m \in M$. Whenever a multi-valued function is single-valued, we can apply many of the terms for regular functions. For example, we will take injectivity, surjectivity, and bijectivity for single-valued functions to have the usual meaning. We will never apply these notions to non-single-valued functions.

3 Delegation and safe Pareto improvements

We consider a setting in which a given game Γ is played through what we will call *representatives*. For example, the representatives could be humans whose behavior is determined or incentivized by some contract à la the principal–agent literature [14].

We imagine that one way in which the representatives can be instructed is to in turn play a subset game $\Gamma^s = (A_1^s \subseteq A_1, ..., A_n^s \subseteq A_n, \mathbf{u}^s)$ of the original game, *without necessarily specifying a strategy or algorithm for solving such a game*. We emphasize, again, that $\mathbf{u}^s$ is allowed to be a vector of entirely different utility functions. For any subset game Γ^s, we denote by $\Pi(\Gamma^s)$ the outcome that arises if the representatives play the subset game Γ^s of Γ. Because in many games, it is not clear what the right choice is, the original players might be uncertain about $\Pi(\Gamma^s)$ for many games Γ^s. We will therefore model each $\Pi(\Gamma^s)$ as a random variable.

The original players trust their representatives to the extent that we take $\Pi(\Gamma)$ to be a default way for the game to played for any Γ. For example, in the Game of Chicken, it is not clear what the right action is. Thus, if one can simply delegate the decision to someone with more relevant expertise, that is the first option one would consider.

We are interested in whether and how the original players can jointly Pareto-improve on the default. Of course, one option is to compute the expected utilities in the default ($\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$) and then let the representatives play a distribution over outcomes whose expected utility exceeds that default expected utility. However, this is unrealistic if Γ is a complex game with multiple Nash equilibria. For one, the precise point of delegation is that the original players are unable or unwilling to properly evaluate Γ. Second, there is no widely agreed upon, universal procedure for selecting an action in the face of equilibrium selection problems.

We address this problem in a typical way. Essentially, we require of any attempted improvement that it incurs no regret in the worst-case. That is, we are interested in subset games Γ^s that are Pareto improvements *with certainty* under weak and purely qualitative assumptions about Π.

Definition 1. *Let Γ^s be a subset game of Γ. We say Γ^s is a safe Pareto improvement (SPI) on Γ if $\mathbf{u}(\Pi(\Gamma^s)) \geq \mathbf{u}(\Pi(\Gamma))$ with certainty. We say that Γ^s is a strict SPI if furthermore, there is a player i s.t. $u_i(\Pi(\Gamma^s)) > u_i(\Pi(\Gamma^s))$ with positive probability.*

4 Program equilibrium

So far, we have been vague about the details of the strategic situation that the original players face in instructing their representatives. From what set of actions can they choose? How can they jointly let the representatives play some new subset game Γ^s? Are SPIs Nash equilibria of the meta game played by the representatives? In this section, we briefly describe one way to fill this gap by discussing the concept of program games and program equilibrium [3, 6, 8, 18, 26]. This section is essential to understanding why SPIs are relevant. However, the remaining technical content of this paper does not rely on this section and the main ideas presented here are straightforward from previous work. For more detail, see Appendix A.

For any game $\Gamma = (A, \mathbf{u})$, the program equilibrium literature considers the following meta game. First, each player i chooses from a set of computer programs. Each program then receives as input a vector containing everyone else's chosen program. Each player i's program then returns an action from A_i, player i's set of actions in Γ. Together these actions then form an outcome $\mathbf{a} \in A$ of the original game. Finally, the utilities $\mathbf{u}(\mathbf{a})$ are realized according to the utility function of Γ. The meta game can be analyzed like any other game. Its Nash equilibria are called *program equilibria*. Importantly, the program equilibria can implement payoffs not implemented by any Nash equilibria of Γ itself. For example, in the Prisoner's Dilemma, both players can submit a program that says: "If the opponent's chosen computer program is equal to this computer program, Cooperate; otherwise Defect." [11, 15, 26] This is a program equilibrium which implements mutual cooperation.

In the setting for our paper, we similarly imagine that each player i can choose from a set of programs that in turn choose from A_i. However, the types of program that we have in mind here are more sophisticated than those typically considered in the program equilibrium literature. Specifically we imagine that the programs are executed by intelligent *representatives* who are themselves able to competently choose an action for player i in any given game Γ^s, without the original player having to describe how this choice is to be made. The original player may not even understand much about this program other than that it generally plays well. Thus, in addition to the elementary instructions used in a typical computer program (branches, comparisons, arithmetic operations, etc.), we allow player i to use an instruction "Play $\Pi_i(\Gamma^s)$" in the program she submits. To jointly let the representatives play, e.g., the SPI Γ^s of Table 2 on the Demand Game of Table 1, the representatives can both use an instruction that says, "If the opponent's chosen program is analogous to this one, play $\Pi_i(\Gamma^s)$; otherwise play DM". Assuming some minimal rationality requirements on the representatives (i.e., on how "play $\Pi_i(\Gamma^s)$" is implemented), this is a Nash equilibrium. More generally, we can prove the following result (see Appendix A for a proof).

Theorem 1. *Let Γ be a game and Γ^s be an SPI of Γ. Now consider a program game on Γ, where each player i can choose from a set of computer programs that output actions for Γ. In addition to the normal kind of instructions, we allow the use of the command "play $\Pi_i(\Gamma')$" for any subset game Γ' of Γ. Finally, assume that $\Pi(\Gamma)$ guarantees each player i at least that player's minimax utility (a.k.a. threat point) in the base game Γ. Then $\Pi(\Gamma^s)$ is played in a program equilibrium, i.e., in a Nash equilibrium of the program game.*

As an alternative to having the original players choose contracts separately, we could imagine the use of jointly signed contracts which only come into effect once signed by all players [cf. 12, 16]. Also compare earlier work by Sen [25] and [22], which we discuss in Appendix B.

5 Outcome correspondence between games

In this section, we introduce a notion of outcome correspondence, which we will see is essential to constructing SPIs.

Definition 2. *Consider two games $\Gamma = (A_1, ..., A_n, \mathbf{u})$ and $\Gamma' = (A'_1, ..., A'_n, \mathbf{u}')$. We write $\Gamma \sim_\Phi \Gamma'$ for $\Phi\colon A \multimap A'$ if $\Pi(\Gamma') \in \Phi(\Pi(\Gamma))$ with certainty.*

Note that $\Gamma \sim_\Phi \Gamma'$ is a statement about Π, i.e., about how the representatives choose. Whether such a statement holds generally depends on the specific representatives being used. In Section 7, we describe two general circumstances under which it seems plausible that $\Gamma \sim_\Phi \Gamma'$. For example, if two games Γ and Γ' are isomorphic, then one might expect $\Gamma \sim_\Phi \Gamma'$, where Φ is constructed from the n isomorphisms of the particular action spaces.

We now state some basic facts about the relation $\sim$, many of which we will use throughout this paper.

Lemma 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.*

For completness, we prove these in Appendix C. Items 1–3 show that $\sim$ has properties resembling those of an equivalence relation. Note, however, that since $\sim$ is not a binary relationship, $\sim$ itself cannot be an equivalence relation in the usual sense. Item 4 states that we can make an outcome correspondence claim less precise and it will still hold true. Item 5 states that in the extreme, it is always $\Gamma \sim_{\mathrm{all}_{A,A'}} \Gamma'$, where $\mathrm{all}_{A,A'}$ is the trivial, maximally imprecise outcome correspondence function that confers no information. Item 6 shows that $\sim$ can be used to express the elimination of outcomes, i.e., the belief that a particular outcome (or strategy) will never occur.

6 Safe Pareto improvements through outcome correspondence

We now show that as advertised, outcome correspondence is closely tied to SPIs. The following theorem shows not only how outcome correspondences can be used to find (and prove) SPIs. It also shows that any SPI requires an outcome correspondence relation with what we will call a *Pareto-improving* correspondence function.

Theorem 3. *Let* $\Gamma = (A, \mathbf{u})$ *be a game and* $\Gamma^s = (A^s, \mathbf{u}^s)$ *be a subset game of* Γ*. Then* Γ^s *is an SPI on* Γ *if and only if there is* Φ *such that* $\Gamma \sim_\Phi \Gamma^s$ *and for all* $\mathbf{a} \in A$ *it is for all* $\mathbf{a}^s \in \Phi(\mathbf{a})$ *the case that* $\mathbf{u}(\mathbf{a}^s) \geq \mathbf{u}(\mathbf{a})$*.*

Proof. $\Leftarrow$: By definition, $\Pi(\Gamma^s) \in \Phi(\Pi(\Gamma))$ with certainty. Hence, for $i = 1, 2$, $u_i(\Pi(\Gamma^s)) \in u_i(\Phi(\Pi(\Gamma)))$ with certainty. Hence, by assumption about Φ, with certainty, $u_i(\Pi(\Gamma^s)) \geq u_i(\Pi(\Gamma))$.

$\Rightarrow$: Assume that $u_i(\Pi(\Gamma)) \geq u_i(\Pi(\Gamma^s))$ with certainty for $i = 1, 2$. We define $\Phi \colon A \to A^s \colon \mathbf{a} \mapsto \{\mathbf{a}^s \in A^s \mid \mathbf{u}(\mathbf{a}^s) \geq \mathbf{u}(\mathbf{a})\}$. It is immediately obvious that Φ is Pareto-improving as required. Also, whenever $\Pi(\Gamma) = \mathbf{a}$ and $\Pi(\Gamma^s) = \mathbf{a}^s$ for any $\mathbf{a} \in A$ and $\mathbf{a}^s \in A^s$, it is (by assumption) with certainty $\mathbf{u}(\mathbf{a}^s) \geq \mathbf{u}(\mathbf{a})$. Thus, by definition of Φ, it holds that $\mathbf{a}^s \in \Phi(\mathbf{a})$. We conclude that $\Gamma \sim_\Phi \Gamma^s$ as claimed. □

Note that the theorem concerns weak SPIs and therefore allows the case where with certainty $\mathbf{u}(\Pi(\Gamma)) = \mathbf{u}(\Pi(\Gamma^s))$. To show that some Γ^s is a *strict* SPI, we need additional information about which outcomes occur with positive probability.

We now illustrate how outcome correspondences can be used to derive the SPI for the Demand Game from the introduction as per Theorem 3. Of course, at this point we do not have any assumptions about when games are equivalent. We will introduce some in the following section. Nevertheless, we can already sketch the argument. Let Γ be the Demand Game of Table 1. First, it seems plausible that Γ is in some sense equivalent to Γ', where $\Gamma' = \Gamma - \{\mathrm{DL}, \mathrm{RL}\}$ is the game that results from removing DL and RL for both players from Γ. Again, strict dominance could be given as an argument. We can formalize this as $\Gamma \sim_\Phi \Gamma'$, where $\Phi(a_1, a_2) = \{(a_1, a_2)\}$ if $a_1, a_2 \in \{\mathrm{DM}, \mathrm{RM}\}$ and $\Phi(a_1, a_2) = \emptyset$ otherwise. In a second step, it seems plausible that $\Gamma' \sim_\Psi \Gamma^s$, where Γ^s is the game of Table 2 and Ψ is the isomorphism between Γ' and Γ^s. Finally, we can use transitivity to obtain $\Gamma \sim_{\Psi \circ \Phi} \Gamma^s$. To see that $\Psi \circ \Phi$ is Pareto-improving for the original utility functions of Γ, notice that Φ does not change utilities at all. Ψ maps the conflict outcome $(\mathrm{DM}, \mathrm{DM})$ onto the outcome $(\mathrm{DL}, \mathrm{DL})$, which is better for both original players. Other than that, Ψ, too, does not change the utilities. Hence, $\Psi \circ \Phi$ is Pareto-improving. By Theorem 3, Γ^s is therefore an SPI on Γ.

In principle, Theorem 3 does not hinge on $\Pi(\Gamma)$ and $\Pi(\Gamma^s)$ resulting from playing games. An analogous result holds for any random variables over A and A^s. In particular, this means that Theorem 3 applies also if the representatives receive other kinds of instructions (cf. Section 4). However, it seems hard to establish non-trivial outcome correspondences between $\Pi(\Gamma)$ and other types of instructions. Still, the use of more complicated instructions can be used to derive different kinds of SPIs. For example, if there are different game SPIs, then the original players could tell their representatives to randomize between them in a coordinated way.

7 Assumptions about outcome correspondence

To make any claims about how the original players should play the meta-game, i.e., about what instructions they should submit, we have to make assumptions about how the representatives choose and (by Theorem 3) about outcome correspondence in particular. We here make two fairly weak assumptions. The first is that the representatives play two isomorphic games isomorphically.

Assumption 1. *Let* $\Gamma = (A, \mathbf{u})$ *and* $\Gamma' = (A', \mathbf{u}')$ *be two games for which there are single-valued bijections* $\Phi_i \colon A_i \to A'_i$ *for* $i = 1, ..., n$ *such that* $\mathbf{u}(a_1, ..., a_n) = \mathbf{u}'(\Phi_1(a_1), ..., \Phi_n(a_n))$ *for all* $\mathbf{a} \in A$*. Then for one tuple* Φ *of such bijections,* $\Gamma \sim_\Phi \Gamma'$*.*

Similar desiderata have been discussed in the context of equilibrium selection, e.g., by Harsanyi and Selten [10, Chapter 3.4]. In fact, they consider a generalization in which the utilities are allowed to be linear transformations of each other. Although this generalization is extremely plausible, we omit it here for simplicity.

One could criticize Assumption 1 by referring to focal points (introduced by Schelling [23, pp. 54–58]) as an example where context and labels of strategies matter. A possible response might be that in games where context plays a role, that context should be included as additional information and not be considered part of $(A, \mathbf{u})$. Assumption 1 would then either not apply to such games with (relevant) context or would require one to, in some way, translate the context along with the strategies. However, in this paper we will not formalize context, and assume that there is no decision-relevant context.

Assumption 2. *Let* $\Gamma = (A, \mathbf{u})$ *be an arbitrary* n*-player game where* $A_1, ..., A_n$ *are pairwise disjoint, and* $\tilde{a}_i \in A_i$ *be strictly dominated by some other strategy* $\hat{a}_i \in A_i$*. Then* $\Gamma \sim_\Phi \Gamma - \{\tilde{a}_i\}$*, where for all* $a_{-i} \in A_{-i}$*,* $\Phi(\tilde{a}_i, a_{-i}) = \emptyset$ *and* $\Phi(a_i, a_{-i}) = \{(a_i, a_{-i})\}$ *whenever* $a_i \neq \tilde{a}_i$*.*

Assumption 2 expresses that representatives should never play strictly dominated strategies. Moreover, it states that we can remove strictly dominated strategies from a game and the resulting game will be played in the same way by the representatives. For example, this implies that when evaluating a strategy a_i, the representatives do not take into account how many other strategies a_i strictly dominates. Assumption 2 also allows (via Transitivity of $\sim$ as per Lemma 2.3) the iterated removal of strictly dominated strategies. The notion that we can (iteratively) remove strictly dominated strategies is common in game theory [20, 13, 19, Section 2.9, Chapter 12] and has rarely been questioned. It is also implicit in the solution concept of Nash equilibrium – if a strategy is removed by iterated strict dominance, that strategy is played in no Nash equilibrium. However, like the concept of Nash equilibrium, the elimination of strictly dominated strategies becomes implausible if the game is not played in the usual way. In particular, for Assumption 2 to hold, we will in most games Γ have to assume that the representatives cannot in turn make credible precommitments (or delegate to further subrepresentatives) or play the game iteratively [2].

With Assumptions 1 and 2 we can finally state our example SPIs formally: Our proofs are in Appendix D.

Proposition (Example) 4. *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*.*

Proposition (Example) 5. *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* $P(\Pi(\Gamma)=(\text{DM}, \text{DM})) > 0$*,* Γ^s *is a strict SPI.*

8 Computing safe Pareto improvements

In this section, we ask how computationally costly it is for the original players to identify for a given game Γ a non-trivial SPI Γ^s. In particular, we ask whether a given game Γ has a non-trivial SPI that can be proved using only Assumptions 1 and 2, Transitivity (Lemma 2.3) and Theorem 3. Formally:

Definition 3. *The SPI decision problem consists in deciding for any given* Γ*, whether there is a sequence of outcome correspondences* $\Phi^1, ..., \Phi^k$ *and a sequence of subset games* $\Gamma^0 = \Gamma, \Gamma^1, ..., \Gamma^k$ *of* Γ *s.t.:*

1. *(Non-triviality:) If we fully reduce* Γ^k *and* Γ *using iterated strict dominance (Assumption 2), the two resulting games are not equal. (Of course, they are allowed to be isomorphic.)*
2. *For* $i = 1, ..., k$*,* $\Gamma^{i-1} \sim_{\Phi^i} \Gamma^i$ *is valid by a single application of either Assumption 1 or Assumption 2.*
3. *For all* $\mathbf{a} \in A$*, and whenever* $\mathbf{a}^s \in (\Phi^k \circ \Phi^{k-1} \circ ... \circ \Phi^1)(\mathbf{a})$*, it is the case that* $u(\mathbf{a}^s) \geq \mathbf{u}(\mathbf{a})$*.*

For the strict SPI decision problem, we further require:

(4.) There is a player i *and an outcome* $\mathbf{a}$ *that survives iterated elimination of strictly dominated strategies from* Γ *s.t.* $u_i((\Phi^k \circ \Phi^{k-1} \circ ... \circ \Phi^1)(\mathbf{a})) > u_i(\mathbf{a})$*.*

Many variants of this problem may be considered. For example, we might generalize it to allow imposing additional properties on the SPI. This will generally not change the computational complexity of the problem. One may also wish to compute all SPIs, or – in line with multi-criteria optimization [7, 28] – all SPIs that cannot in turn be safely improved upon. However, in general there may exist exponentially many such SPIs. To retain any hope of developing an efficient algorithm, one would therefore have to first develop a more efficient representation scheme [cf. 21, Sect. 16.4].

Theorem 6. *The (strict) SPI decision problem is NP-complete, even for 2-player games.*

Proposition 7. *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* $O(m^l)$*.*

The full proof is tedious (see Appendix E), but the main idea is simple. To find an SPI on Γ based on Assumptions 1 and 2, one has to first iteratively remove all strictly dominated actions to obtain a reduced game Γ', which the representatives would play the same as the original game. This can be done in polynomial time. One then has to map the actions Γ' onto the original Γ in such a way that each outcome in Γ' is mapped onto a weakly Pareto-better outcome in Γ. Our proof of NP-hardness works by reducing from the subgraph isomorphism problem, where the payoff matrices of Γ', Γ represent the adjacency matrices of the graphs.

Besides being about a specific set of assumptions about $\sim$, note that Theorem 6 and Proposition 7 also assume that the utility function of the game is represented explicitly in normal form as a payoff matrix. If we changed the game representation (e.g., to boolean circuits, extensive form game trees, quantified boolean formulas, or even Turing machines), this can affect the complexity of the SPI problem [cf. 9]. In fact, even reducing a game using strict dominance by pure strategies – which contributes only insignificantly to the complexity of the SPI problem for normal-form games – is difficult in some game representations [4, Section 6].

9 Safe Pareto improvements under improved coordination

In this section, we imagine that the players are able to simply invent new token strategies with new payoffs that arise from mixing existing feasible payoffs. To define this formally, we first define for any game $\Gamma = (A, \mathbf{u})$,

$$\mathcal{C}(\Gamma) := \mathbf{u}(\Delta(A)) = \left\{ \sum_{\mathbf{a}\in A} p_{\mathbf{a}} \mathbf{u}(\mathbf{a}) \;\middle|\; \forall \mathbf{a}{\in}A{:}p_{\mathbf{a}}{\in}[0,1], \sum_{\mathbf{a}\in A} p_{\mathbf{a}} = 1 \right\}$$

to be the set of feasible coordinated payoff vectors of Γ, which is exactly the convex closure of $\mathbf{u}(A)$, i.e., of the deterministically achievable utilities of the original game.

For any game Γ, we then imagine that in addition to subset games, the players can let the representatives play a *perfect-coordination token game* $(A^s, \mathbf{u}^s, \mathbf{u}^e)$, where for all i, $A^s_i \cap A_i = \emptyset$ and $u^s_i \colon A^s \to \mathbb{R}$ are arbitrary utility functions to be used by the representatives and $\mathbf{u}^e \colon A^s \to \mathcal{C}(\Gamma)$ are the utilities that the original players assign to the token strategies.

The instruction $(A^s, \mathbf{u}^s, \mathbf{u}^e)$ lets the representatives play the game $(A^s, \mathbf{u}^s)$ as usual. However, the strategies A^s are imagined to be meaningless token strategies which do not resolve the given game Γ. Once some token strategies $\mathbf{a}^s$ are selected, these are translated into some probability distribution over A, i.e., over outcomes of the original game, thus giving rise to (expected) utilities $\mathbf{u}^e(\mathbf{a}^s) \in \mathcal{C}(\Gamma)$. These distributions and thus utilities are specified by the original players. We here imagine in our definition of $\mathcal{C}(\Gamma)$ that these distributions over A could require the representatives to correlate their choices for the original game for any given $\mathbf{a}^s$.

Definition 4. *Let Γ be a game. A perfect-coordination SPI for Γ is a perfect-coordination token game $(A^s, \mathbf{u}^s, \mathbf{u}^e)$ for Γ s.t. $\mathbf{u}^e(\Pi(A^s, u^s)) \geq \mathbf{u}(\Pi(\Gamma))$ with certainty. We call $(A^s, \mathbf{u}^s, \mathbf{u}^e)$ a strict perfect-coordination SPI if there furthermore is a player i for whom $u^e_i(\Pi(A^s, u^s)) > u_i(\Pi(\Gamma))$ with positive probability.*

As an example, imagine that Γ is just the DM-RM subset game of the Demand Game of Table 1. Then, intuitively, an SPI under improved coordination could consist of the original players telling the representatives, "Play as if you were playing the DM-RM subset game of the Demand Game, but whenever you find yourself playing (DM, DM), randomize [according to some given distribution] between the other (Pareto-optimal) outcomes instead". Formally, $A^s_1 = \{\hat{D}, \hat{R}\}, A^s_2 = \{\hat{D}, \hat{R}\}$ would then consist of tokenized versions of the original strategies. The utility functions u^s_1, u^s_2 are then simply the same as in the original Demand Game except that they are applied to the token strategies. E.g., $\mathbf{u}^s(\hat{D}, \hat{R}) = (2, 0)$. The utilities for the original players remove the conflict outcome. For example, the original players might specify $\mathbf{u}^e(\hat{D}, \hat{D}) = (1, 1)$, representing that the representatives are supposed to play (RM, RM) in the $(\hat{D}, \hat{D})$ case. For all other outcomes $(\hat{a}_1, \hat{a}_2)$, it must be the case that $\mathbf{u}^e(\hat{a}_1, \hat{a}_2) = \mathbf{u}^s(\hat{a}_1, \hat{a}_2)$ because the other outcomes cannot be Pareto-improved upon. As with our earlier SPIs for the Demand Game, Assumption 1 implies that $\Gamma \sim_\Phi \Gamma^s$, where Φ maps the original conflict outcome (DM, DM) onto the Pareto-optimal $(\hat{D}, \hat{D})$.

Relative to the SPIs considered up until now, these new types of instructions put significant additional requirements on how the representatives interact. They now have to engage in a two-round process of first choosing and observing one another's token strategies and then playing the corresponding distribution over outcomes from the original game. Further, it must be the case that this additional coordination does not affect the payoffs of the original outcomes. The latter may not be the case in, e.g., the Game of Chicken. That is, we could imagine a Game of Chicken in which coordination is possible but that the rewards of the game change if the players do coordinate. After all, the underlying story in the Game of Chicken is that the positive reward (admiration from peers) is attained precisely for accepting a grave risk.

With these more powerful ways to instruct representatives, we can now replace individual outcomes of the default game *ad libitum*. For example, in the reduced Demand Game, we singled out the outcome (DM, DM) as Pareto-suboptimal and replaced it by a Pareto-optimal outcome, while keeping all other outcomes the same. This allows us to construct SPIs in many more games then before.

Definition 5. *The strict full-coordination SPI decision problem consists in deciding for any given* Γ *whether under Assumption 1 there is a perfect-coordination SPI* Γ^s *for* Γ.

Lemma 8. *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* $\mathcal{C}(\Gamma)$.

For completeness the linear program is given in Appendix F. Based on Lemma 8, Algorithm 1 decides whether there is a strict perfect-coordination SPI for a given game Γ.

Algorithm 1: An algorithm for deciding the strict perfect-coordination SPI problem.

Data: Game Γ, set $\mathrm{supp}(\Pi(\Gamma)$

for $\mathbf{a} \in \mathrm{supp}(\Pi(\Gamma))$ **do**
    if $\mathbf{u}(\mathbf{a})$ *is Pareto-suboptimal within* $\mathcal{C}(\Gamma)$ **then**
        Return True;

Return False;

It is easy to see that this algorithm runs in polynomial time (in the size of, e.g., the normal form representation of the game). It is also correct: if it returns True, simply replace the Pareto-suboptimal outcome while keeping all other outcomes the same; if it returns False, then all outcomes are Pareto-optimal within $\mathcal{C}(\Gamma)$ and so there can be no strict SPI. We summarize this result in the following proposition.

Proposition 9. *Assuming* $\mathrm{supp}(\Pi(\Gamma))$ *is known and that Assumption 1 holds, it can be decided in polynomial time whether there is a strict perfect-coordination SPI.*

From the problem of deciding whether there are strict SPIs under improved coordination at all, we move on to the question of what different perfect-coordination SPIs there are. In particular, one might ask what the cost is of only considering *safe* Pareto improvements relative to acting on a probability distribution over $\Pi(\Gamma)$ and the resulting expected utilities $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$. We start with a lemma that directly provides a characterization. So far, all the considered perfect-coordination SPIs $(A^s, \mathbf{u}^s, \mathbf{u}^e)$ for a game $(A, \mathbf{u})$ have consisted in letting the representatives play a game $(A^s, \mathbf{u}^s)$ that is isomorphic to the original game, but Pareto-improves (from the original players' perspectives, i.e., $\mathbf{u}^e$) at least one of the outcomes. It turns out that we can restrict attention to this very simple type of SPI under improved coordination.

Lemma 10. *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* $\mathcal{C}(\Gamma)$ *such that under Assumption 1 the game*

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

is also an SPI on Γ*, with*

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

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

We prove this in Appendix G. Because of this result, we will focus on these particular types of SPIs, which simply create an isomorphic game with different (Pareto-better) utilities. Note, however, that without assigning exact probabilities to the distributions of $\Pi(\Gamma), \Pi(\Gamma')$, the original players will in general not be able to *construct* a Γ^s that satisfies the expected payoff equalities. For this reason, one could still conceive of situations in which a different type of SPI would be chosen by the original players and the original players are unable to instead choose an SPI of the type described in Lemma 10.

Lemma 10 directly implies a characterization of the expected utilities that can be achieved with perfect-coordination SPIs. Of course, this characterization depends on the exact distribution of $\Pi(\Gamma)$. We omit the statement of this result. However, we state the following implication.

Corollary 11. *Under Assumption 1, the set of Pareto improvements that are safely achievable with perfect coordination*

$$\{\mathbb{E}[\mathbf{u}(\Gamma')] \mid \Gamma' \textit{ is perfect-coordination SPI on } \Gamma\}$$

is a convex polygon.

Because of this result, one can also efficiently optimize convex functions over the set of perfect-coordination SPIs. Even without referring to the distribution $\Pi(\Gamma)$, many interesting questions can be answered efficiently. For example, we can efficiently identify the perfect-coordination SPI that maximizes the minimum improvements across players and outcomes $\mathbf{a} \in A$.

In the following, we aim to use Lemma 10 and Corollary 11 to give maximally strong positive results about what Pareto improvements can be safely achieved, without referring to exact probabilities over $\Pi(\Gamma)$. To keep things simple, we will do this only for the case of two players. To state our results, we first need some notation: We use

$$\mathrm{PF}(\mathcal{C}) \coloneqq \left\{\mathbf{y} \in \mathcal{C} \;\middle|\; \nexists \mathbf{y}' {\in} \mathcal{C}, i \in \{1, ..., n\} : \mathbf{y}' \geq \mathbf{y}, y'_i > y\right\}$$

to denote the Pareto frontier of a convex polygon $\mathcal{C}$ (or more generally convex, closed set). For any real number $x \in \mathbb{R}$, we use $\pi_i(x, \mathcal{C}(\Gamma))$ to denote the $\mathbf{y}' \in \mathcal{C}(\Gamma)$ which maximizes y'_{-i} under the constraint $y'_i = x$. (Recall that we consider 2-player games, so y'_{-i} is a single real number.) Note that such a $\mathbf{y}'$ exists if and only if x is i's utility in some feasible payoff vector. We first state our result formally. Afterwards, we will give a graphical explanation of the result, which we believe is easier to understand.

Theorem 12. *Make Assumption 1. Let* Γ *be a two-player game. Let* $\mathbf{y} \in \mathbb{R}^2$ *be some potentially unsafe Pareto improvement on* $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$*. For* $i = 1, 2$*, let* $x_i^{\min/\max} = \min / \max u_i\left(\mathrm{supp}(\Pi(\Gamma))\right)$*. Then:*

A) If there is some element in $\mathcal{C}(\Gamma)$ *which Pareto-dominates all of* $\operatorname{supp}(\Pi(\Gamma))$ *and if* $\mathbf{y}$ *is Pareto-dominated by an element of at least one of the following three sets:*

- L_1 := *the line segment between* $\pi_1(x_1^{\min}, \mathrm{PF}(\mathcal{C}(\Gamma))$ *and* $\pi_1(x_1^{\max}, \mathrm{PF}(\mathcal{C}(\Gamma))$*;*
- L_2 := *the segment of the curve* $\mathrm{PF}(\mathcal{C}(\Gamma))$ *between* $\pi_1(x_1^{\max}, \mathrm{PF}(\mathcal{C}(\Gamma))))$ *and* $\pi_2(x_2^{\max}, \mathrm{PF}(\mathcal{C}(\Gamma))))$*;*
- L_3 := *the line segment between* $\pi_2(x_2^{\max}, \mathrm{PF}(\mathcal{C}(\Gamma))$ *and* $\pi_2(x_2^{\min}, \mathrm{PF}(\mathcal{C}(\Gamma))$*.*

Then there is an SPI under improved coordination Γ^s *such that* $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma^s))\right] = \mathbf{y}$*.*
B) If there is no element in $\mathcal{C}(\Gamma)$ *which Pareto-dominates all of* $\operatorname{supp}(\Pi(\Gamma))$ *and if* $\mathbf{y}$ *is Pareto-dominated by an element each of* L_1 *and* L_3 *as defined above, then there is a perfect-coordination SPI* Γ^s *such that* $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma^s))\right] = \mathbf{y}$*.*

We now illustrate the result graphically. We start with Case A, which is illustrated in Figure 1. The Pareto-frontier is the solid line in the north and east. The points marked x indicate outcomes in $\operatorname{supp}(\Pi(\Gamma))$. The point marked by a filled circle indicates the expected value of the default equilibrium $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$. For some $\mathbf{y} \in \mathbb{R}^2$ to be a Pareto-improvement, it must be to the north-east of the filled circle. The vertical dashed lines starting at the two extreme x marks illustrate the application of π_1 to project $x_1^{\min/\max}$ onto the Pareto frontier. The dotted line between these two points is L_1. Similarly, the horizontal dashed lines starting at x marks illustrate the application of π_2 to project $x_2^{\min/\max}$ onto the Pareto frontier. The line segment between these two points is L_3. In this case, this line segments lies on the Pareto frontier. The set L_2 is simply that part of the Pareto frontier, which Pareto-dominates all elements of $\operatorname{supp}(\Pi(\Gamma))$, i.e., the part of the Pareto frontier to the north-east between the two intersections with the northern horizontal dashed line and eastern vertical dashed line.

Case B of Theorem 12 is depicted in Figure 2. Note that here the two line segments L_1 and L_3 intersect. To ensure that a Pareto improvement is safely achievable, the theorem requires that it is below both of these lines.

Theorem 12 is proven by re-mapping each of the outcomes of the original game as per Lemma 10. For example, the projection of the default equilibrium $\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right]$ (i.e., the filled circle) onto L_1 is obtained as an SPI by

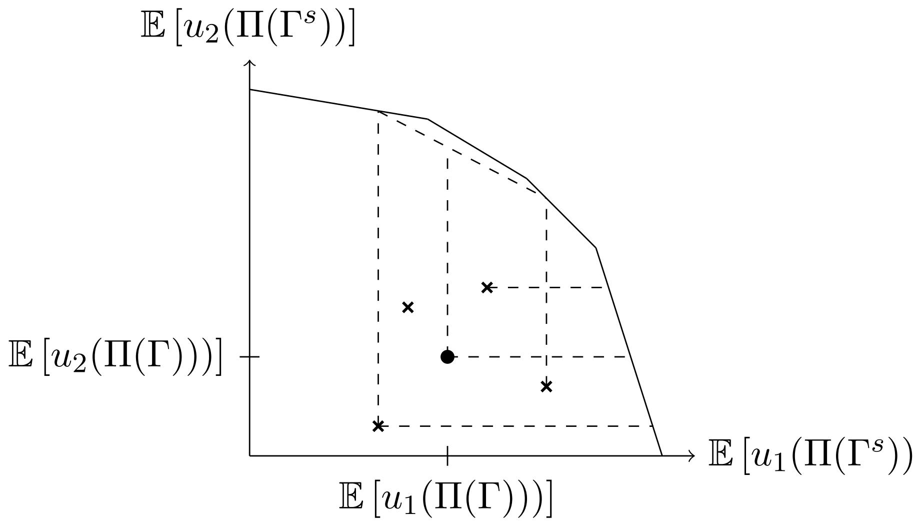

Figure 1: This figure illustrates Theorem 12, Case A.

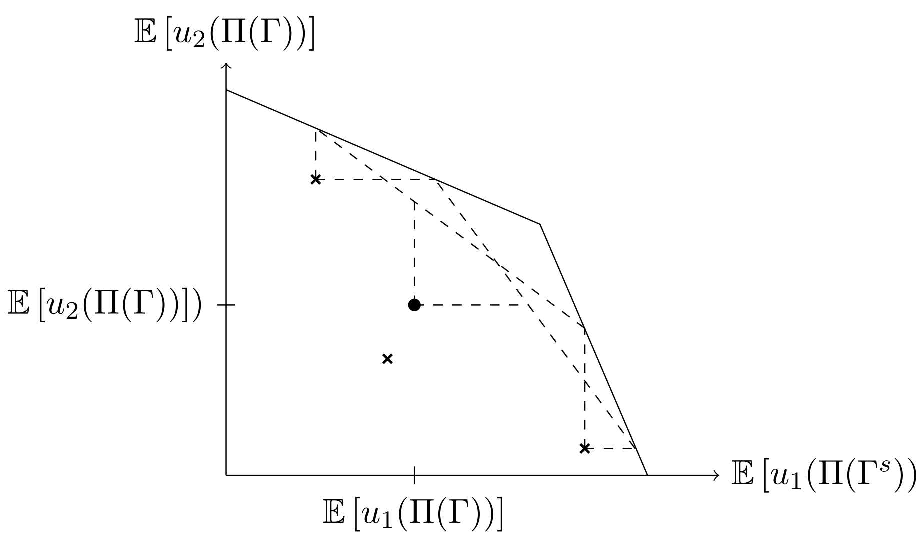

Figure 2: This figure illustrates Theorem 12, Case B.

projecting all the outcomes (i.e., all the x marks) onto L_1. In Case A, any utility vector $\mathbf{y} \in L_2$ that Pareto-improves on all outcomes of the original game can be obtained by re-mapping all outcomes onto $\mathbf{y}$. Other kinds of $\mathbf{y}$ are handled similarly. For a full proof, see Appendix H.

As a corollary of Theorem 12, we can see that all (potentially unsafe) Pareto improvements in the DM-RM subset game of the Demand Game of Table 1 are equivalent to some perfect-coordination SPI. However, this is not always the case:

Proposition 13. *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.* $u_i(\mathbf{a}) > \mathbb{E}[u_i(\Pi(\Gamma))]$ *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*,* $\mathbb{E}[u_i^e(\Pi(A^s, \mathbf{u}^s))] = u_i(\mathbf{a})$*.*

We prove this in Appendix I.

10 The SPI selection problem

In the Demand Game, there happens to be a single non-trivial SPI. However, in general (even without the type of coordination assumed in Section 9) there may be multiple incomparable SPIs that result in different payoffs for the players. If multiple SPIs are available, the original players would be left with the difficult decision of which SPI to demand in their instruction.

This difficulty of choosing what SPI to demand cannot be denied. However, we would here like to emphasize that players can profit from the use of SPIs even without addressing this SPI selection problem. To do so, a player picks an instruction that is very compliant ("dove-ish") w.r.t. what SPI is chosen, e.g., one that simply goes with whatever SPI the other players demand as long as that SPI cannot further be safely Pareto-improved upon. In many cases, all such SPIs benefit both players. For example, SPIs in bargaining scenarios like the Demand Game remove the conflict outcome, which benefits all parties. Thus, a player can expect a safe improvement even under such maximally compliant demands on the selected SPI.

11 Conclusion and future directions

Safe Pareto improvements are a promising new idea for delegating strategic decision making. To conclude this paper, we discuss some ideas for further research on SPIs.

Straightforward technical questions arise in the context of the complexity results of Section 8. First, what impact on the complexity does varying the assumptions have? Our NP-completeness proof is easy to generalize at least to some other types of assumptions. It would be interesting to give a generic version of the result. We also wonder whether there are plausible assumptions under which the complexity changes in interesting ways. Second, one could ask how the complexity changes if we use more sophisticated game representations (see the remarks at the end of that section). Third, one could impose additional restrictions on the sought SPI. For example, some of the players may be unable to have their representative maximize arbitrary utility functions. We could then ask whether there is an SPI in which only a given subset of the players adopt different utility functions and restrictions on the set of available strategies. Fourth, we could restrict the games under consideration. Are there games in which it becomes easy to decide whether there is an SPI?

It would also be interesting to see what real-world situations can already be interpreted as utilizing SPIs, or could be Pareto-improved upon using SPIs.

Acknowledgments

This work was supported by NSF under Award IIS-1814056. Some early work on this topic was conducted by Caspar Oesterheld while working at the Foundational Research Institute (now the Center on Long-Term Risk). For valuable comments and discussions, we thank our anonymous referees, Keerti Anand, Tobias Baumann, Jesse Clifton, Max Daniel, Lukas Gloor, Adrian Hutter, Anni Leskelä, Brian Tomasik and Johannes Treutlein; as well as attendees of a 2017 talk at the Future of Humanity Institute at the University of Oxford and participants of the May 2019 Effective Altruism Foundation research retreat.

References

[1] Krzysztof R. Apt. "Uniform Proofs of Order Independence for Various Strategy Elimination Procedures". In: *The B.E. Journal of Theoretical Economics* 4.1 (2004), pp. 1–48. DOI: 10.2202/1534-5971.1141.

[2] Robert Axelrod. *The Evolution of Cooperation*. New York: Basic Books, 1984.

[3] Mihaly Barasz et al. *Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic*. Jan. 2014. URL: https://arxiv.org/abs/1401.5577.

[4] Vincent Conitzer and Tuomas Sandholm. "Complexity of (Iterated) Dominance". In: *Proceedings of the 6th ACM conference on Electronic commerce*. Vancouver, Canada: Association for Computing Machinery, June 2005, pp. 88–97. DOI: 10.1145/1064009.1064019.

[5] Stephen A. Cook. "The complexity of theorem-proving procedures". In: *STOC '71: Proceedings of the third annual ACM symposium on Theory of computing*. New York: Association for Computing Machinery, May 1971, pp. 151–158. DOI: 10.1145/800157.805047.

[6] Andrew Critch. "A Parametric, Resource-Bounded Generalization of Löb's Theorem, and a Robust Cooperation Criterion for Open-Source Game Theory". In: *Journal of Symbolic Logic* 84.4 (Dec. 2019), pp. 1368–1381. DOI: 10.1017/jsl.2017.42.

[7] Matthias Ehrgott. *Multicriteria Optimization*. 2nd ed. Berlin: Springer, 2005.

[8] Lance Fortnow. *Program equilibria and discounted ComputationTime*. Tech. rep. 1473. Kellogg School of Management – Center for Mathematical Studies in Economics and Management Science, Northwestern University, 2008. URL: https://www.econstor.eu/bitstream/10419/31197/1/587667478.PDF.

[9] Joaquim Gabarró, Alina García, and Maria Serna. "The complexity of game isomorphism". In: *Theoretical Computer Science* 412.48 (Nov. 2011), pp. 6675–6695. DOI: 10.1016/j.tcs.2011.07.022.

[10] John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: The MIT Press, 1988.

[11] J. V. Howard. "Cooperation in the Prisoner's Dilemma". In: *Theory and Decision* 24 (May 1988), pp. 203–213. DOI: 10.1007/BF00148954.

[12] Adam Tauman Kalai et al. "A commitment folk theorem". In: *Games and Economic Behavior* 69 (2010), pp. 127–137. DOI: 10.1016/j.geb.2009.09.008.

[13] Elon Kohlberg and Jean-Francois Mertens. "On the Strategic Stability of Equilibria". In: *Econometrica* 54.5 (Sept. 1986), pp. 1003–1037. DOI: 10.2307/1912320.

[14] Jean-Jacques Laffont and David Martimort. *The Theory of Incentives – The Principal-Agent Model*. Princeton, NJ: Princeton University Press, 2002.

[15] R. Preston McAfee. "Effective Computability in Economic Decisions". May 1984. URL: https://www.mcafee.cc/Papers/PDF/EffectiveComputability.pdf.

[16] Dov Monderer and Moshe Tennenholtz. "Strong mediated equilibrium". In: *Artificial Intelligence* 173.1 (Jan. 2009), pp. 180–195. DOI: 10.1016/j.artint.2008.10.005.

[17] John von Neumann. "Zur Theorie der Gesellschaftsspiele". In: *Mathematische Annalen* 100 (1928), pp. 295–320. DOI: https://doi.org/10.1007/BF01448847.

[18] Caspar Oesterheld. "Robust Program Equilibrium". In: *Theory and Decision* 86.1 (Feb. 2019), pp. 143–159.

[19] Martin J. Osborne. *An Introduction to Game Theory*. New York: Oxford University Press, 2004.

[20] David G. Pearce. "Rationalizable Strategic Behavior and the Problem of Perfection". In: *Econometrica* 54.4 (July 1984), pp. 1029–1050.

[21] Guillaume Perez. "Decision diagrams: constraints and algorithms". PhD thesis. Université Côte d'Azur, 2017. URL: https://tel.archives-ouvertes.fr/tel-01677857/document.

[22] Werner Raub. "A General Game-Theoretic Model of Preference Adaptions in Problematic Social Situations". In: *Rationality and Society* 2.1 (Jan. 1990), pp. 67–93.

[23] Thomas C. Schelling. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press, 1960.

[24] Alexander Schrijver. *Theory of Linear and Integer Programming*. Chichester, UK: John Wiley & Sons, 1998.

[25] Amartya Sen. “Choice, orderings and morality”. In: *Practical Reason*. Ed. by Stephan Körner. New Haven, CT, USA: Basil Blackwell, 1974. Chap. II, pp. 54–67.

[26] Moshe Tennenholtz. “Program equilibrium”. In: *Games and Economic Behavior* 49.2 (Nov. 2004), pp. 363–373.

[27] Wiebe van der Hoek, Cees Witteveen, and Micheal Wooldridge. “Program equilibrium – a program reasoning approach”. In: *International Journal of Game Theory* 42 (3 Aug. 2013), pp. 639–671.

[28] Luc N. van Wassenhove and Ludo F. Gelders. “Solving a bicriterion scheduling problem”. In: *European Journal of Operational Research* 4 (1980), pp. 42–48.

A Proof of Theorem 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 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

$$U \colon \mathrm{PROG} \to \mathbb{R}^n \colon \mathbf{c} \mapsto \mathbb{E}\left[\mathbf{u}(exec(\mathbf{c}))\right]. \tag{1}$$

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 take as input vectors in PROG and return an action a_i. The function *exec* on input $\mathbf{c} \in \mathrm{PROG}$ then executes each player i's program c_i on $\mathbf{c}$ to assign i an action. The definition implicitly assumes that PROG only contains programs that halt when fed one another as input. A *program equilibrium* is then simply a Nash equilibrium of the program game.

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

$$v_i^\Gamma = \min_{\boldsymbol{\sigma}_{-i} \in \times_{j \neq i} \Delta(A_j)} \max_{\sigma_i \in \Delta(A_i)} u_i(\sigma_i, \boldsymbol{\sigma}_{-i}). \tag{2}$$

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,

$$(\textit{minimax}(i,j))_{j\in\{1,...,n\}-\{i\}} \in \underset{\boldsymbol{\sigma}_{-i}\in\times_{j\neq i}\Delta(A_j)}{\arg\min} \max_{\sigma_i\in\Delta(A_i)} u_i(\sigma_i,\boldsymbol{\sigma}_{-i}). \tag{3}$$

Note that for $n=2$, each player's threat point is computable in polynomial time via linear programming; and that by the minimax theorem [17], the threat point is equal to the maximin utility, i.e.,

$$v_i^\Gamma = \max_{\sigma_i\in\Delta(A_i)} \min_{\sigma_{-i}\in\Delta(A_{-i})} u_i(\sigma_i,\sigma_{-i}), \tag{4}$$

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

Tennenholtz' [26] main result on program games is the following:

Theorem 14 (Tennenholtz 2004 [26]). *Let $\Gamma=(A,\mathbf{u})$ be a game and let $\mathbf{x}\in\mathbf{u}\left(\times_{i=1}^n\Delta(A_i)\right)$ be a (feasible) payoff vector. If $x_i\geq v_i^\Gamma$ for $i=1,...,n$, then $\mathbf{x}$ is the utility of some program equilibrium of a program game on Γ.*

Throughout the rest of this section, our goal is to use similar ideas as Tennenholtz did for Theorem 14 to construct for any SPI Γ^s on Γ, a program equilibrium that results in the play of $\Pi(\Gamma^s)$. As noted in the main text, 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 safe Pareto improvement on the Demand Game, each player would only want their representative to choose from $\{\text{DL},\text{RL}\}$ rather than $\{\text{DM},\text{DM}\}$ if the other player's representative does the same. It would then seem that in a program equilibrium in which $\Pi(\Gamma^s)$ 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 [26], because it is the simplest. We could similarly use the variant proposed by Fortnow [8], techniques based on Löb's theorem [3, 6], or ϵ-grounded mutual simulation [18] or even (meta) Assurance Game preferences (see Appendix B).

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 code as this one, then play $\Pi(\Gamma^s)$. Otherwise, if there is a player j who submits a different source code, punish player j by playing her *minimax* strategy". Note that 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".

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

Data: Everybody's source code $\mathbf{c}$, my index i

1 **for** $j \in \{1, ..., n\} - \{i\}$ **do**

2   **if** $c_j \neq c_i$ **then**

3   Play $minimax(i, j)$;

4 Play $\Pi_i(\Gamma^s)$;

Proposition 15. *Let Γ be a game and let Γ^s be an SPI on Γ. 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 in expectation. Then $\mathbf{c}$ is a program equilibrium and $apply(\mathbf{c}) = \Pi(\Gamma^s)$.*

Proof. By inspection of Algorithm 2, we see that $exec(\mathbf{c}) = \Pi(\Gamma^s)$. It is left to show that $\mathbf{c}$ is a Nash equilibrium. So let i be any player and $c_i' \in \mathrm{PROG}_i - \{c_i\}$. We need to show that $\mathbb{E}\left[u_i(exec(\mathbf{c}_{-i}, c_i'))\right] \leq \mathbb{E}\left[u_i(exec(\mathbf{c}))\right]$. Again, by inspection of $\mathbf{c}$, $exec(\mathbf{c}_{-i}, c_i')$ is the threat point of Player i. Hence,

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

as required. □

Theorem 1 follows immediately.

B A discussion of work by Sen (1974) and Raub (1990) on preference adaptation games

We here discuss Raub's [22] paper in some detail, which in turn elaborates on an idea by Sen [25]. 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 program equilibrium and bilateral precommitment.

In Sections 1, 3 and 4, 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 paper are raw utility functions without any surrounding control structures, etc. 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. (Note that this payoff matrix resembles the classic Stag Hunt studied in game theory.)

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,

		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

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 is 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 4 and Appendix A, we imagine that this is taken care of by additional bilateral commitment mechanisms that are not the focus of this paper. For Raub's paper, on the other hand, ensuring mutual cooperation to be stable in the new game Γ^s is arguably the key idea. Still, we could pursue the approach of the present paper even when we limit 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 [26], Oesterheld [26], Barasz et al. [3] and van der Hoek, Witteveen, and Wooldridge [27], ordered in increasing order of similarity of the main idea of the scheme.

C Proof of Proposition 2

Lemma 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 equality 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) 4

Proposition (Example) 4. *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 2.5, we further obtain $\Gamma - \{\text{Cooperate}\} \sim_{\text{all}} \Gamma^s$, where Γ^s is as described in the proposition. Hence, by transitivity, $\Gamma \sim_{\text{all}\circ\Phi} \Gamma^s$. It is easy to verify that the function $\text{all} \circ \Phi$ is Pareto-improving. □

D.2 Proof of Proposition (Example) 5

Proposition (Example) 5. *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 $P(\Pi(\Gamma){=}(\text{DM}, \text{DM})) > 0$, Γ^s is a strict SPI.*

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 RL for both players. We can then apply Lemma 2.3 (Transitivity) to obtain $G \sim_\Phi \hat{G} = (\{\text{DM}, \text{RM}\}, \{\text{DM}, \text{RM}\}, u_1, u_2)$, where

$$\Phi(a_1, a_2) = \begin{cases} \{(a_1, a_2)\} & \text{if } a_1, a_2 \in \{\text{DM}, \text{RM}\} \\ \emptyset & \text{otherwise} \end{cases}. \tag{5}$$

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 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 6

We here prove Theorem 6.

Throughout this section, we use the following lemma.

Lemma 16. *Let Φ, Ψ be isomorphisms between Γ,Γ'. If Φ is (strictly) Pareto-improving, then so is Ψ.*

This will allow us to conclude from the existence of a Pareto-improving isomorphism Φ that there is Pareto-improving Ψ s.t. $\Gamma \sim_{\Psi} \Gamma'$ by Assumption 1, even if there are multiple isomorphisms between Γ, Γ'.

Proof. Let $\Gamma = (A, \mathbf{u})$, $\Gamma' = (A', \mathbf{u}')$. Then for all $\mathbf{a} \in A$,

$$\begin{aligned} \mathbf{u}(\mathbf{a}) &= \mathbf{u}'(\Psi(\mathbf{a})) \\ &= \mathbf{u}(\Phi^{-1}(\Psi(\mathbf{a})) \\ &\leq \mathbf{u}(\Phi(\Phi^{-1}(\Psi(\mathbf{a})))) \\ &= \mathbf{u}(\Psi(\mathbf{a})). \end{aligned}$$

Furthermore, if Φ is strictly Pareto-improving for some $\tilde{\mathbf{a}} \in A$, then by bijectivity of Φ, Ψ, there is $\mathbf{a} \in A$ s.t. $\Phi^{-1}(\Psi(\mathbf{a})) = \tilde{\mathbf{a}}$. For this $\mathbf{a}$, the inequality above is strict and therefore $\mathbf{u}(\mathbf{a}) < \mathbf{u}(\Psi(\mathbf{a}))$. □

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 3), 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

$$u_i^s \colon A^s \to \mathbb{R} \colon \mathbf{a}^s \mapsto (u_i(\Phi_1^{-1}(a_1^s), ..., \Phi_n^{-1}(a_n^s)))_{i=1,...,n}. \tag{6}$$

Otherwise, return False.

It is easy to see that this algorithm runs in non-deterministic polynomial time. Furthermore, with Lemma 16 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 17. *Let* $\Gamma \sim_{\Phi^{\text{iso}}} \hat{\Gamma}$ *by Assumption 1 and* $\hat{\Gamma} \sim_{\Phi^{\text{red}}} \tilde{\Gamma}$ *by Assumption 2. Then there are* $\Gamma', \Psi^{\text{red}}, \Psi^{\text{iso}}$ *s.t.* $\Gamma \sim_{\Psi^{\text{red}}} \Gamma'$ *by Assumption 2,* $\Gamma' \sim_{\Psi^{\text{iso}}} \tilde{\Gamma}$ *by Assumption 1 and* $\Psi^{\text{iso}} \circ \Psi^{\text{red}} = \Phi^{\text{red}} \circ \Phi^{\text{iso}}$*.*

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 $A'_i = (\Phi_i^{\text{iso}})^{-1}(\tilde{A}_i)$ and $u'_i = u_i|_{A'}$. Next, we define

$$\Psi^{\text{red}} = (\Phi^{\text{iso}})^{-1} \circ \Phi^{\text{red}} \circ \Phi^{\text{iso}} \tag{7}$$

and

$$\Psi^{\text{iso}} = \Phi^{\text{iso}}|_{A'_1 \times A'_2}. \tag{8}$$

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

First, it is

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

as claimed. Note that the second step uses the associativity of $\circ$ on multi-valued 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 $A'_i = (\Phi_i^{\text{iso}})^{-1}(\tilde{A}_i)$ 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. (Note that because $\Psi^{\text{iso}} \circ \Psi^{\text{red}} = \Phi^{\text{red}} \circ \Phi^{\text{iso}}$, it really must be $\Gamma' \sim_{\Psi^{\text{iso}}} \tilde{\Gamma}$ rather than $\Gamma' \sim_{\tilde{\Psi}^{\text{iso}}} \tilde{\Gamma}$ for some different isomorphism $\tilde{\Psi}^{\text{iso}} \neq \Psi^{\text{iso}}$.) First, it is easy to show that Ψ^{iso} decomposes into $\Psi_1^{\text{iso}}, ..., \Psi_n^{\text{iso}}$ 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

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

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 18. *Let*

$$\Gamma^1 \sim_{\Phi^1} \ldots \sim_{\Phi^{k-1}} \Gamma^k, \tag{9}$$

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

$$\Gamma^1 \sim_{\Psi^1} \Gamma'^2 \sim_{\Psi^2} \Gamma'^3 \sim_{\Psi^3} \ldots \sim_{\Psi^{m-1}} \Gamma'^m \tag{10}$$

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

$$\Phi^{k-1} \circ \Phi^{k-2} \circ \ldots \circ \Phi^1 = \Xi \circ \Psi^{m-1} \circ \ldots \circ \Psi^1. \tag{11}$$

Proof. Start with the initial sequence of line 9. We can iteratively apply Lemma 17 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 19. *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 $i = 1, \ldots, n$. 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 $\tilde{\Gamma}^b = (\Phi_1(\hat{A}^b_1), \ldots, \Phi_n(\hat{A}^b_n), \tilde{\mathbf{u}}^a)$. Note that if the correspondence function Φ is Pareto-improving, so is $\Phi|_{\hat{A}^b}$.*

Lemma 19 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 known to be path-independent [1, 20].

Proposition 20. *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 that show $\Gamma \sim_{\Phi} \Gamma^s$ for Pareto-improving Φ. Then by Lemma 18, there is Γ' such that $\Gamma \sim_{\Psi^{\text{red}}} \Gamma'$ via an arbitrary number of applications of Assumption 2 and $\Gamma' \sim_{\Psi^{\text{iso}}} \Gamma^s$ 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 full iterated removal as per Assumption 2. By Lemma 19, there is a Pareto-improving outcome correspondence $\Gamma^r \sim_{\tilde{\Psi}^{\text{iso}}} \Gamma^{sr}$ as per Assumption 2. By construction, our algorithm finds (guesses) this Pareto-improving outcome correspondence. $\square$

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 7. *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 by 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 $(n, a\colon \{1, ..., n\} \times \{1, ..., n\} \to \mathbb{B})$, where $n \in \mathbb{N}$ and $\mathbb{B} := \{0, 1\}$. 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 $\phi\colon \{1, ..., n\} \to \{1, ...n'\}$ such that for all $j \neq l$

$$a(j, l) \leq a'(\phi(j), \phi(l)). \tag{12}$$

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 from node j to node l in G, there must also be an edge in the same direction between the corresponding nodes $\phi(j), \phi(l)$ 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 [5, Theorem 2].

Lemma 21. *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 using a subset of the ideas 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 $A_1 = A_2 = \{1, ..., n\}$,$\hat{A}_1 = \hat{A}_2 = \{1, ..., \hat{n}\}$, $u_1(j, l) = u_2(j, l) = a(j, l)$ 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 Ψ is a Pareto-improving isomorphism between $\Gamma^{c,D}$ and $\tilde{\Gamma} = (\tilde{A}_1, \tilde{A}_2, \tilde{u}_1, \tilde{u}_2)$, where we define

$$\tilde{A}_1 = \tilde{A}_2 = \{C\} \times (\phi(\{1, ..., n\}) \cup \{\hat{n} + \phi(j) \mid j = 1, ..., n\}), \tag{13}$$

for $i = 1, 2$ 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 $\tilde{u}_i(\tilde{a}_1, \tilde{a}_2) = u_i(\Psi_i^{-1}(\tilde{a}_1, \tilde{a}_2))$. Again, it will then follow from Lemma 16 that in particular it is $\Gamma^{c,D} \sim_{\Psi'} \tilde{\Gamma}$ via Assumption 1 for some potentially different, but still Pareto-improving Ψ'. 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:

$$\tilde{u}_i(\Psi(a_1, a_2)) = u_i(\Psi^{-1}(\Psi(a_1, a_2))) = u_i(a_1, a_2). \tag{14}$$

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 $i = 1, 2$

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

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

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

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

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

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

$$\begin{aligned} \mathbf{u}^c((D,j),(D,l)) &= (2,2) \\ &< (3,3) \\ &= \mathbf{u}^c((C,\Phi(j)),(C,\Phi(j)+\hat{n})) \\ &= \mathbf{u}^c(\Psi_1(D,j),\Psi_2(D,l)). \end{aligned}$$

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

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

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 a strict 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 20), 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. $\Psi\left((\{D\} \times \{1,...,2n\})^2\right) \subseteq \left((\{C\} \times \{1,...,2\hat{n}\})^2\right)$: 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

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

   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. $\Psi_{-i}(D,l) \neq \Psi_i(D,j) - \hat{n}$, where we define $(D,k) - \hat{n} := (D,k-\hat{n})$. Then for that l it is

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

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

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

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,

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

   contradicting the assumption that Ψ is Pareto-improving.

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

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

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

Given these, we can define our graph isomorphism as

$$\phi\colon \{1, ..., n\} \to \{1, ..., \hat{n}\}\colon j \mapsto \pi_2(\Psi_1(D, j)), \tag{15}$$

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

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

Hence, ϕ is a subgraph isomorphism as desired. □

F Proof of Lemma 8

Lemma 8. *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* $\mathcal{C}(\Gamma)$*.*

For an introduction to linear programming, see, e.g., Schrijver [24]. In short, a linear program is a specific type of constrained optimization problem that can be solved efficiently.

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

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

□

G Proof of Lemma 10

Lemma 10. *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* $\mathcal{C}(\Gamma)$ *such that under Assumption 1 the game*

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

is also an SPI on Γ*, with*

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

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. First note that $(\hat{A}, \hat{\mathbf{u}})$ is isomorphic to Γ. Thus by Assumption 1, there is isomorphism Φ s.t. $\Gamma \sim_\Phi (\hat{A}, \hat{\mathbf{u}})$. WLOG assume that Φ simply maps $a_1^{i_1}, ..., a_n^{i_n} \mapsto \hat{a}_1^{i_1}, ..., \hat{a}_n^{i_n}$. Then define $\mathbf{u}^e$ as follows:

$$\mathbf{u}^e(\hat{a}_1^{i_1}, ..., \hat{a}_n^{i_n}) = \mathbb{E}\left[\mathbf{u}'(\Pi(\Gamma')) \mid \Pi(\Gamma) = (a_1^{i_1}, ..., a_n^{i_n})\right]. \tag{16}$$

Here $\mathbf{u}'$ describes the utilities that the original players assign to the outcomes of Γ'. Since $\mathbf{u}'$ maps onto $\mathcal{C}(\Gamma)$ and $\mathcal{C}(\Gamma)$ is convex, $\mathbf{u}^e$ as defined also maps into $\mathcal{C}(\Gamma)$ 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,

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

as required. □

H Proof of Theorem 12

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 by Corollary 11 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 $\mathbf{u}^e = \mathbf{y}$.

It is left to show that all elements of $L_{1/2}$ are safely 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 $y_1 = \mathbb{E}[u_1(\Pi(\Gamma))]$ is safely achievable. In other words, $\mathbf{y}$ is the projection northward of the default utility, or $\mathbf{y} = \pi_1(\mathbb{E}[\mathbf{u}(\Pi(\Gamma))], L_1)$. This $\mathbf{y}$ is also one of the endpoints of L_1'. To achieve this utility, we construct the equivalent game as per Lemma 10, 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 onto L_1 of the utility of the corresponding outcome (a_1, a_2) in Γ^s. That is,

$$\mathbf{u}^e(\hat{a}_1, \hat{a}_2) = \pi_1(\mathbf{u}(a_1, a_2), L_1). \tag{17}$$

Note that because $\mathcal{C}(\Gamma)$ is convex and the endpoints of the line segment L_1 are by definition in $\mathcal{C}(\Gamma)$, it is $L_1 \subseteq \mathcal{C}(\Gamma)$. Hence, all values of $\mathbf{u}^e$ as defined in Eq. 17 are feasible. Because all outcomes in the original game lie below the line L_1, π_1 is linear. Hence,

$$\begin{aligned} \mathbb{E}\left[\mathbf{u}^e(\Pi(\Gamma^s))\right] &= \mathbb{E}\left[\pi_1(\mathbf{u}(\Pi(\Gamma)), L_1)\right] && (18) \\ &= \pi_1(\mathbb{E}\left[\mathbf{u}(\Pi(\Gamma))\right], L_1) && (19) \end{aligned}$$

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 11, 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 $i = 2, j = 3$. 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 10. 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 $\text{supp}(\Pi(\Gamma))$ 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

		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.

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 11, 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 13

Proposition 13. *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.* $u_i(\mathbf{a}) > \mathbb{E}\left[u_i(\Pi(\Gamma))\right]$ *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*,* $\mathbb{E}\left[u_i^e(\Pi(A^s, \mathbf{u}^s))\right] = u_i(\mathbf{a})$*.*

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 $\mathcal{C}(\Gamma)$ that maximizes the sum of the two players' utilities. By linearity of expectation and convexity of $\mathcal{C}(\Gamma)$, 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 outcomes that still gives utilities of $(4, 0)$ and $(0, 4)$, respectively, because these are Pareto-optimal within the set of feasible payoff vectors.

$\square$