Promises Made, Promises Kept: Safe Pareto Improvements via Ex Post Verifiable Commitments

Nathaniel Sauerberg[*1], **Caspar Oesterheld**[*2]

[1]University of Texas at Austin
[2]Foundations of Cooperative AI Lab, Carnegie Mellon University
nsauerberg@utexas.edu, oesterheld@cmu.edu

Abstract

A safe Pareto improvement (SPI) is a modification of a game that leaves all players better off with certainty. SPIs are typically proven under qualitative assumptions about the way different games are played. For example, we assume that strictly dominated strategies can be iteratively removed and that isomorphic games are played isomorphically. In this work, we study SPIs achieved through three types of *ex post* verifiable commitments – promises about player behavior from which deviations can be detected by observing the game. First, we consider disarmament – commitments not to play certain actions. Next, we consider SPIs based on *token games*. A token game is a game played by simply announcing an action (via cheap talk). As such, its outcome is intrinsically meaningless. However, we assume the players commit in advance to play specific (pure or correlated) strategy profiles in the original game as a function of the token game outcome. Under such commitments, the token game becomes a new, meaningful normal-form game. Finally, we consider default-conditional commitment: SPIs in settings where the players' default ways of playing the original game can be credibly revealed and hence the players can commit to act as a function of this default. We characterize the complexity of deciding whether SPIs exist in all three settings, giving a mixture of characterizations and efficient algorithms and NP- and GRAPH ISOMORPHISM-hardness results.

Extended version — https://arxiv.org/abs/2505.00783

1 Introduction

Among the most important applications of game theory is guiding decisions that shape a downstream strategic interaction. To make this tractable, it's common to reduce games to a single value by assuming that the games will resolve in a particular way. For example, the literature on Stackelberg games generally makes an explicit assumption that the followers will play the best or worst (e.g. Coniglio, Gatti, and Marchesi 2020; Sauerberg and Oesterheld 2024; Basilico, Coniglio, and Gatti 2016) Nash equilibrium for the leader, while mechanism design typically assumes that the truthful equilibrium will be played. Similarly, notions like price of anarchy/stability (Koutsoupias and Papadimitriou 1999), value of mediation (Ashlagi, Monderer, and Tennenholtz 2008), and value of recall (Berker et al. 2025) evaluate the importance of a particular affordance (centralized control, mediation, or recall) by bounding the ratio of some welfare objective between the games with and without the affordance, assuming a particular (best- or worst-case) equilibrium will be played in each game. Work on game theory with simulation asks when allowing one player to pay to learn the other's strategy introduces a new Nash equilibrium which is Pareto-better than all existing equilibria (Kovařík, Oesterheld, and Conitzer 2023; Kovařík et al. 2025).

However, compressing complex games down to a single outcome is not without loss. In particular, equilibrium selection is an unsolved and arguably unsolvable problem (Norde et al. 1996), so it may not be safe to assume that players play any particular equilibrium, or indeed play an equilibrium at all. Safe Pareto improvements (SPIs) (Oesterheld and Conitzer 2022) offer a more general framework for analyzing interventions–one that doesn't require assumptions about how individual games are resolved. SPIs are interventions that improve *all* possible outcomes of the game, given explicit, usually mild assumptions on the relationships between the ways different games are played. For example, we might assume that isomorphic games are played isomorphically and that removing strictly dominated strategies doesn't change how a game is played. In this paper, we introduce and study SPIs via *ex post* verifiable commitments–commitments that the other players or an outside observer can verify to have been followed by observing the game.

Consider the following strategic interaction, modeled by Table 1. Two countries are disputing the control of a seaway passing that's newly navigable due to melting sea ice. Part of the seaway lies within each country's established territory, but

	FA	FN	PA	PN
FA	$2, 2$	$-1, 8$	$-10, 10$	$-10, 10$
FN	$8, -1$	$-50, -50$	$-10, 10$	$-10, 10$
PA	$10, -10$	$10, -10$	$2, 2$	$0, 5$
PN	$10, -10$	$10, -10$	$5, 0$	$-10, -10$

Table 1: Seaway Dispute. The actions stand for claim Full/-Partial seaway via Navy/Announcement.

[*]These authors contributed equally.

	Token PA	Token PN
Token PA	$3.2, 3.2$	$2, 5$
Token PN	$5, 2$	$-4, -4$

Table 2: Token Game for Seaway Dispute

most had previously been unclaimed sea ice. The countries choose both (1) whether to claim the full seaway (F) or only the disputed part (P), and (2) whether to assert their claim via naval occupation (N) or diplomatic announcement (A). If either both countries claim the full seaway (F) or both claim only the disputed portion (P), the game has typical "chicken" dynamics.

- (N, A) or (A, N): If one country claims with their Navy and the other via Announcement, the country that claims with its Navy wins control of the disputed territory.
- (A, A): If both countries claim via Announcement, they will eventually agree to joint control through diplomacy.
- (N, N): If both countries claim with their Navies, it results in costly warfare. This is especially costly if the countries claim the full seaway and so invade each other.

If only one country attempts to claim the other's territory, the international outrage results in an outcome favorable to the other country. For this reason, PA strictly dominates FA and FN. Therefore, we assume that by default the players will play PN or PA, so the game is equivalent to its bottom right quadrant. There are increasing returns to controlling more of the seaway, so the sum of the payoffs for the different outcomes follows $SW(\mathrm{FN}, \mathrm{FA}) > SW(\mathrm{PN}, \mathrm{PA}) > SW(\mathrm{PA}, \mathrm{PA}) = SW(\mathrm{FA}, \mathrm{FA})$.

Suppose both countries believe they're likely to achieve a payoff of 5 from $(\mathrm{PN}, \mathrm{PA})$ or $(\mathrm{PA}, \mathrm{PN})$ in the default game (or they want to represent that belief for strategic reasons). No strategy profile gives both players a utility of at least 5 (or even more than 3.5), so no agreement to play a single strategy profile is possible. However, the players can still guarantee a Pareto improvement if they commit to resolve the game by playing the "token game" in Table 2. They act in the token game simply by (privately) writing their actions on pieces of paper and then flipping them over to observe the outcome. Once the token outcome t is observed, the players are required to play a (predetermined) correlated strategy profile in the original game with expected utility $\mathbf{u}(t)$. Any jointly observable source of randomness, e.g. physical or cryptographic (Blum 1983) coin flips, can be used to draw an outcome from this correlated strategy profile. Given such a source, it's *ex post* verifiable that the players play according to the prescribed outcome, so we assume they do.

For this to work, of course, all payoffs in the token game must be realizable (in expectation) by some correlated strategy profile in the original game. Here, the $(5, 2)$ payoff can be achieved by $(2/3)(\mathrm{FN}, \mathrm{FA}) + (1/3)(\mathrm{FA}, \mathrm{FN}) = (2/3)(8, -1) + (1/3)(-1, 8)$, and the $(2, 5)$ payoff symmetrically. The $(3.2, 3.2)$ payoff can be achieved by playing $.2(\mathrm{PA}, \mathrm{PA}) + .4(\mathrm{FN}, \mathrm{FA}) + .4(\mathrm{FA}, \mathrm{FN})$, and the $(-4, -4)$ payoff can be achieved by playing $(1/2)(\mathrm{PA}, \mathrm{PA}) + (1/2)(\mathrm{PN}, \mathrm{PN})$.

Also, observe that the token game is isomorphic to the bottom right quadrant of the original game: a payoff of v_i in the original game corresponds to a payoff of $.6v_i + 2$ in the token game. Therefore, we can reasonably assume that the players will play them isomorphically. E.g., if a player would have played PA in the original game, they'd play Token PA in the token game, and so on. Since each token outcome Pareto improves on its counterpart, the token game is a guaranteed Pareto improvement on the original game, an SPI.

In this paper, we consider SPIs achieved by three different types of *ex post* verifiable commitment: the token game SPIs exemplified by the previous example and two others. Situations where *ex post* verifiable commitments can be made credible are frequent. In particular, they could be enforced by reputation costs or by external authorities through legal contracts. Moreover, *ex post* verifiability seems close to necessary for any type of external enforcement of commitment.

Existing schemes for achieving SPIs, such as the delegation game SPIs proposed in (Oesterheld and Conitzer 2022), require forms of commitment that seem more difficult to achieve. In the delegation game setting, the original players delegate the game to representatives, assigning them a utility function but otherwise leaving how to play up to the representatives. By default, they instruct the representatives with their true utility functions, but SPIs can sometimes be achieved by the players instead making joint agreements to assign alternate utility functions.

Making it credible that one's representative indeed plays according to an alternate utility function seems to require a high degree of transparency into the representative's decision making process. Though such transparency is sometimes achievable, it seems unattainable in cases where the decisions are being made in the minds of the people or groups of people with stake in the outcome of the game. Importantly, the SPIs achieved in the present paper don't require any assumptions on the process that decides how to play the games. In particular, our schemes allow the original, self-interested players to simply play the modified game themselves.

Contributions. In Section 3, we consider SPIs that can be achieved by players committing not to take, i.e., *disarm*, particular actions. We find that deciding the existence of disarmament SPIs is NP-complete. We further find that deciding whether a *given* disarmament is an SPI is polynomial-time equivalent to the graph isomorphism problem (which is believed to be NP-intermediate, i.e., in NP, but neither in P nor NP-hard).

Next, in Section 4, we study token game SPIs like the one for the Seaway Dispute described above. We distinguish two types of token SPIs. In the first type, the token outcomes can represent distributions over outcomes of the default game (as in the Seaway Dispute example). In this case, the existence of SPIs can be decided in polynomial time, and we obtain an explicit characterization of the existence of SPIs in the two-player case. In the second type, token outcomes can only represent a single outcome of the default game. There, we give an algorithm for finding SPIs that runs in polynomial time in games with a constant number of players and quasi-polynomial time in general. We also show that the problem becomes NP-complete in more succinct game representations

(e.g. payoff tables that only store non-zero entries).

Finally, in Section 5, we study SPIs achieved by in a setting where players can credibly reveal their default. We show that finding unilateral default-remapping SPIs, where a player commits to act according to some function of their default strategy, is NP-hard. However, in the omnilateral default-remapping setting, where all players can credibly reveal their default action and jointly commit to play a strategy profile as a function of the default outcome, we show that SPIs exist whenever the original game contains Pareto-suboptimal outcomes after dominated strategies are iteratively removed.

Throughout the paper, we give only proof sketches of most results. The extended version of this paper, which includes an appendix with full proofs and additional technical content, is available at https://arxiv.org/abs/2505.00783.

1.1 Related Work

Most closely related to our paper is the prior work on safe Pareto improvements (Oesterheld and Conitzer 2022, 2025; DiGiovanni, Clifton, and Macé 2025). We use the framework from Oesterheld and Conitzer (2022, 2025), see our Section 2. We consider different interventions on strategic interaction than Oesterheld and Conitzer (see above). The token games studied in Section 4 resemble the token games studied by Oesterheld and Conitzer (2022, Sect. 5). The main difference is that they allow giving the agents arbitrary utility functions over the token outcomes. In contrast, we assume that the utility of a token outcome is simply the utility of the distribution over base game outcomes associated with the token outcome. Consequently, the results are different. Our study of default-conditional SPIs (Section 5.2) is inspired by DiGiovanni, Clifton, and Macé (2025), but they focus on the implementation of and incentives for commitment to default-remapping in a program game setting and don't study the unilateral default remapping case.

Many forms of *ex post* verifiable commitment (e.g., Stackelberg games) have been studied. Most closely related is the prior work on disarmament games (Deng and Conitzer 2017, 2018; see also Renou 2009; Bade, Haeringer, and Renou 2009; Collina, Derr, and Roth 2024), which studies similar forms of commitment to our Section 3, though not in an SPI framework. The literature on program games and open source game theory (Tennenholtz 2004; Critch, Dennis, and Russell 2022; c.f. Kalai et al. 2010) can also be viewed as a studying form of *ex post* verifiable commitment: players commit to play the output of a computer program which can read the other players' programs. In addition to being the setting of (DiGiovanni, Clifton, and Macé 2025), this bears some similarity to our token games: In both, the players commit to condition their final actions on a prior interaction (the token game or the program game). We discuss more distantly related work on *ex post* verifiability in the appendix.

2 Preliminaries

Game theory. We here introduce some game-theoretic notation and terminology. An *n-player (normal-form) game (NFG)* G is a pair $(A, \mathbf{u})$, where $A = \times_i A_i$ for some a nonempty set of *actions* A_i for each player i, and $\mathbf{u} : A \to \mathbb{R}^n$ is a *utility function*, with $u_i(a)$ Player i's utility if the players play a. We assume $|A_i| \geq 2$ for all Players i unless otherwise stated. We call the elements of A *outcomes* or action profiles. We use $\Delta(A)$ to denote the set of *correlated strategy profiles*, i.e., distributions over A. We extend $\mathbf{u}$ to strategy profiles by taking the expectation: $\mathbf{u}(c) \coloneqq \sum_{a \in A} c(a)\mathbf{u}(a)$, where $c(a)$ is the probability assigned to outcome a by the correlated strategy profile c. We define $\mathbf{u}(A) = \{\mathbf{u}(a) : a \in A\}$, and define $\mathbf{u}(\Delta(A))$ similarly. We use $-i$ to denote the set of players other than i.

For any n-player game $G = (A, \mathbf{u})$ and nonempty sets $\hat{A}_1 \subseteq A_1, ..., \hat{A}_n \subseteq A_n$ and letting $\hat{A} = \hat{A}_1 \times ... \times A_n$, note that $(\hat{A}, \mathbf{u}_{|\hat{A}})$ is a new game, where $u_{|\hat{A}}$ denotes the restriction of $\mathbf{u}$ to $\hat{A}$. We call this a subgame of G. We typically just write $(\hat{A}, \mathbf{u})$, omitting that $\mathbf{u}$ is restricted to the new action sets. We will often obtain a subgame by removing some set of actions A'_i. We then use $G - A'_i$ as shorthand for $(A_1 \times ... \times A_{i-1} \times (A_i - A'_i) \times A_{i+1} \times ... \times A_n, \mathbf{u})$, the subgame of G obtained by removing A'_i from G.

Given utility functions $\mathbf{u}$, we say that some outcome a' *is a (weak) Pareto improvement on* a if for all i we have that $u_i(a') \geq u_i(a)$. We then write $\mathbf{u}(a') \succeq \mathbf{u}(a)$. We say that a' is a *strict* Pareto improvement on a, or $\mathbf{u}(a') \succ \mathbf{u}(a)$, if additionally there is a player i s.t. $u_i(a') > u_i(a)$. We say that an outcome a is *Pareto optimal* within some set if there is no strict Pareto improvement on a in that set.

Let G be a game and let $a_i, a'_i \in A_i$ be actions for Player i. We say that a'_i *strictly dominates* a_i if for all $a_{-i} \in A_{-i}$ we have that $u_i(a'_i, a_{-i}) > u_i(a_i, a_{-i})$.

A function $\phi\colon A \to A'$ defined by bijections $\phi_i : A_i \to A'_i$ is a *(game) isomorphism* from $(A, \mathbf{u})$ to $(A', \mathbf{u}')$ if there exist some $m, b \in \mathbb{R}^n$ with all $m_i > 0$ such that $u'_i(\phi_1(a_1), ..., \phi_n(a_n)) = m_i u_i(a) + b_i$ for all $a \in A$ and all players i. An isomorphism is Pareto improving if $u'_i(\phi(a)) \geq u_i(a)$ for all players i and all $a \in A$ and strictly Pareto improving if this inequality is strict for at least one player and outcome.

We take Player i to be the same person (or company, etc.) across games. For this reason, we don't allow isomorphisms to permute players (as done by e.g. Harsanyi and Selten 1988; Gabarró, García, and Serna 2011). We also assume the utilities of a player are comparable between games and generally consider a single utility function $\mathbf{u}$ which operates on outcomes of all games. This renders the notion of Pareto improvements between games meaningful.

Outcome correspondences. Following Oesterheld and Conitzer (2022), we reason about safe Pareto improvements by reasoning about outcome correspondence relationships. We imagine that the players' strategies across all games can be represented by an (unknown) policy function Π which maps arbitrary games to their outcome. An outcome correspondence is a claim relating the results of playing two different games G, G', i.e. a claim about the relationship between $\Pi(G)$ and $\Pi(G')$. For example, one possible outcome correspondence is the claim: if playing $G = (A, \mathbf{u})$ would result in some $a \in A$, then playing $G' = (A', \mathbf{u})$ would result in a' or a'' (for some $a', a'' \in A'$). This is a claim about Π: If $\Pi(G) = a$, then $\Pi(G') \in \{a', a''\}$. More generally, let

$G = (A, \mathbf{u})$ and $G' = (A', \mathbf{u}')$ be two games and let Φ be a multivalued function from A to A'. Then $G \sim_\Phi G'$ denotes the (outcome correspondence) claim that, whatever outcome $\Pi(G)$ is, the outcome $\Pi(G')$ will satisfy $\Pi(G') \in \Phi(\Pi(G))$.

We typically have to make assumptions about what kinds of outcome correspondences hold between games. We make essentially the same assumptions as Oesterheld and Conitzer (2022, 2025). The first is that we can remove strictly dominated actions and the resulting game will be played in the same way (cf. Pearce 1984; Kohlberg and Mertens 1986).

Assumption A. *Let $G = (A_1, ..., A_n, \mathbf{u})$ be a game. Let $\hat{a}_i$ be an action for Player i that is strictly dominated in G. Then $G \sim_\Xi (A_i - \{\hat{a}_i\}, A_{-i}, \mathbf{u})$, where $\Xi(\mathbf{a}) = \emptyset$ if $a_i = \hat{a}_i$ and $\Xi(\mathbf{a}) = \{\mathbf{a}\}$ otherwise. In other words, $\Pi_i(G) \neq \hat{a}_i$ and $\Pi_i(G) = \Pi_i(G - \{\hat{a}_i\})$.*

We often consider the subgame obtained by iteratively removing all strictly dominated actions from G, which we denote $\bar{G} = (\bar{A}, \mathbf{u})$ and refer to as the *reduced game*. This game is well-known to be unique (Pearce 1984; Gilboa, Kalai, and Zemel 1990; Apt 2004).

Our second assumption is, roughly, that isomorphic games are played isomorphically.

Assumption B. *Let G and G' be isomorphic games without strictly dominated actions. Then let Φ be the union of the isomorphisms from G to G', i.e., for every outcome $\mathbf{a}$ of G, we let $\Phi(\mathbf{a}) = \{\phi(\mathbf{a}) \mid \phi \text{ isomorphism from } G \text{ to } G'\}$. Then $G \sim_\Phi G'$. In other words, there must be an isomorphism ϕ from G to G' such that $\phi(\Pi(G)) = \Pi(G')$.*

Further, we use the following *transitivity* rule: if $G \sim_\Phi G'$ and $G' \sim_\Xi G''$, then $G \sim_{\Xi \circ \Phi} G''$. Finally, between any two games G and G' the following trivial outcome correspondence holds: $(A, \mathbf{u}) \sim_{\mathrm{all}_{A,A'}} (A', \mathbf{u}')$, where $\mathrm{all}_{A,A'}(a) := A'$ for all $a \in A$. That is, whatever outcome occurs in G, some outcome in A' must obtain if G' were to be played.

We call G' a *safe Pareto improvement (SPI)* on G if there is a Φ s.t. 1. $G \sim_\Phi G'$, 2. for all a and $a' \in \Phi(a)$ we have that $a' \succeq a$, and 3. there exists some realization of Π (satisfying any assumptions made) such that $\Pi(G') \succ \Pi(G)$. In other words, G' is an SPI on G if there is a strictly Pareto-improving outcome correspondence from G to G', where strictly Pareto improving means that at least one outcome (which is possible under the assumptions) is guaranteed to be strictly Pareto improved.

3 Disarmament SPIs

Perhaps the simplest form of ex-post-verifiable commitment is commitment against taking particular actions. Most straightforwardly, one of the players could commit unilaterally. That is, before playing a game G, Player 1 could announce that they won't take any action from some set $\tilde{A}_1 \subset A_1$. Following Deng and Conitzer (2017, 2018), we call such a commitment a *disarmament of* $\tilde{A}_1$. If Player 1's announcement is credible, the game $G - \tilde{A}_1$ (the game obtained from G by removing the actions in $\tilde{A}_1$) is played instead of G. We also consider multilateral disarmament. That is, Players 1 and 2 jointly agree not to play $\tilde{A}_1$ and $\tilde{A}_2$, respectively. Such bilateral disarmament is still *ex post* verifiable.

	l, s	m, s	h, s	l, c	m, c	h, c
l, nf	2, 4	2, 5	1, 7	4, 4	4, 5	3, 7
m, nf	3, 4	3, 5	0, 2	5, 4	5, 5	2, 2
h, nf	5, 3	0, 2	0, 2	7, 3	2, 2	2, 2
l, f	3, 3	3, 4	2, 6	5, 2	6, 2	6, 2
m, f	4, 3	4, 4	1, 1	6, 2	7, 2	3, 1
h, f	6, 2	1, 1	1, 1	8, 1	3, 1	3, 1

Table 3: Negotiation Game.

As an example, imagine that Alice and Bob are negotiating a contract. For simplicity, imagine that each player can make just three levels of demands: (h(igh), m(edium), and l(ow)). The outcome is determined by the combination of demands. Roughly, higher demands result in better outcomes for a player, but if aggregate demands are too high (at least one high and one medium), no favorable agreement can be reached. Additionally, let's imagine that Alice – since she is a contract lawyer and Bob is not – can incorporate "loopholes" into the fine print of proposed contracts (f) or not (nf). Adding these loopholes is generally good for Alice and bad for Bob. Meanwhile, Bob can insist on a simple contract (s) to minimize the impact of loopholes. Unfortunately, simple contracts also reduce flexibility and thus decrease payoffs relative to complicated contracts (c). A version of this game is visualized in Table 3.

By default, Alice will incorporate loopholes into the contract by dominance reasoning. Anticipating this, dominance reasoning suggests that Bob will insist on a simple contract. That is, the game reduces to its lower-left quadrant.

Now imagine that Alice can credibly commit against the use of loopholes. For instance, we could imagine that she can publicly promise Alice and Bob's social circle that that she won't include loopholes; and we might imagine that the social opprobrium from breaking such a promise would outweigh the gains from including the loopholes. Then there's no reason for Bob to insist on a simple contract. Thus, after such disarmament, the game reduces to the upper-right quadrant, which is an SPI on the original game.

We now consider the computational question of whether for a given game G, there is a disarmament s.t. the game resulting from the disarmament is an SPI on G, as well as the question of whether a *given* disarmament induces an SPI.

To get started, we prove a general result characterizing SPIs induced by Assumptions A and B. Roughly, to assess whether G' is an SPI on G we only need to consider the reduced versions of the two games, $\bar{G}'$ and $\bar{G}$. For G' to be an SPI on G, we need either either $\bar{G}$ to be isomorphic to $\bar{G}'$ via a Pareto-improving isomorphism or every outcome of $\bar{G}'$ to Pareto-dominate every outcome of $\bar{G}$.

Lemma 3.1. *Consider two games G and G'. Then G' is an SPI on G under Assumptions A and B if and only if at least one of the following two conditions holds:*

1. *There is a strictly Pareto improving isomorphism between $\bar{G}$ and $\bar{G}'$, i.e., an isomorphism ϕ from $\bar{G}$ to $\bar{G}'$ where $\phi(a) \succeq a$ for all $a \in A$ and $\phi(a) \succ a$ for at least one*

$a \in A$.

2. $\boldsymbol{u}(a') \succeq \boldsymbol{u}(a)$ *for all outcomes* $a \in \bar{A}$ *and* $a' \in \bar{A}'$ *and, for at least one outcome* $a \in \bar{A}$, $\boldsymbol{u}(a') \succ \boldsymbol{u}(a)$ *for all* $a' \in \bar{A}'$.

We call an SPI *simple* if it can be proven using only condition 2. That is, G' is a *simple* SPI on G under Assumption A if, for all outcomes $a' \in \bar{A}'$ and all outcomes $a \in \bar{A}$, $\mathbf{u}(a') \succeq \mathbf{u}(a)$. Note that simple SPIs can be proved without Assumption B. Similarly, we refer to SPIs based on condition 1 as *isomorphism* SPIs. That is, a game G' is an *isomorphism* SPI on a game G under Assumptions A and B if there exists a Pareto-improving isomorphism between $\bar{G}$ and $\bar{G}'$.

We now consider the problem of deciding whether a *given* disarmament is a safe Pareto improvement. We show that even under strong restrictions this problem is graph-isomorphism-complete (GI-complete). The graph isomorphism problem is commonly believed to be NP-intermediate, i.e., in NP, not solvable in polynomial time, but not NP-hard. (For discussions of GI, see Mathon 1979; Zemlyachenko, Korneenko, and Tyshkevich 1985; Köbler, Schöning, and Torán 1993, Grohe and Schweitzer 2020.)

Theorem 3.2. *The following problem is* GI*-complete: Given a game* $G = (A_1, ..., A_n, \mathbf{u})$ *and sets of actions* $(\tilde{A}_i)_i$ *for each player, decide whether the game* $G' = (A_1 - \tilde{A}_1, ..., A_n - \tilde{A}_n, \mathbf{u})$ *is an SPI on* G *under Assumptions A and B. The problem remains* GI*-complete if we restrict attention to* $n = 2$, $|\tilde{A}_1| = 1$ *and* $\tilde{A}_2 = \emptyset$.

Proof sketch. Whether G' is a simple SPI on G can be decided in polynomial time, so we focus on isomorphism SPIs. The first central idea behind the proof is that deciding various questions about whether a given pair of games are isomorphic is GI-complete, see appendix. GI-membership is then easy to prove. For hardness, we reduce from the problem of deciding whether two games are isomorphic. To do this, we construct for any pair of games G, G', a new game with two properties. First, it reduces to G plus some gadget when no actions are disarmed. Second, under a particular unilateral disarmament with $|\tilde{A}_1| = 1$ and $\tilde{A}_2 = \emptyset$, it reduces to $G' + (\epsilon, \epsilon)$, i.e., the game arising from G' by giving each player an extra ϵ in each outcome, plus an isomorphic gadget. Whether the disarmament is an SPI then becomes equivalent to the question whether G and G' are isomorphic. □

Next we consider the problem of deciding whether a given game has *any* disarmament SPI (rather than evaluating a specific candidate). This problem is NP-complete, even if we restrict attention to unilateral SPIs.

Theorem 3.3. *The following problem is* NP*-complete: Given a game* G*, decide whether there exist sets* $A'_1, ..., A'_n$ *s.t.* $(A_1 - A'_1, ..., A_n - A'_n, \mathbf{u})$ *is a SPI on* G *under Assumptions A and B. The problem remains* NP*-complete if we restrict attention to two-player games and* $\bar{A}_2 = \emptyset$.

Proof sketch. The difficult part is proving hardness. Similar to Theorem 3.2, the first central idea is to use the NP-hardness of determining whether one game G can be isomorphically mapped into a subgame of another game G', where the isomorphism keeps utilities constant (see appendix). The main challenge of the proof then is to construct a game that (without disarmament) reduces to G (plus some gadget) and that by (unilateral disarmament) can be made to reduce to any subgame of G' (plus an isomorphic gadget) with an extra utility of ϵ for all players. Then there is a (strict) (unilateral) disarmament SPI if and only if G' has a subgame that is isomorphic to G. □

Note that if we bound the number of actions that can be disarmed, the problem returns to being GI-complete, since there are only polynomially many disarmaments to try.

Throughout this paper we consider safe *Pareto* improvements. For *unilateral* disarmaments in particular it is also natural to consider "safe u_1 improvements", i.e., unilateral disarmaments that are guaranteed (by Assumptions A and B) to be better for Player 1. Our proofs of Theorems 3.2 and 3.3 apply not just to SPIs but also to safe u_1 improvements.

4 Token Game SPIs

In this section, we consider SPIs achieved by commitments to resolve a game by playing a token game. Token games are the same type of mathematical object as "normal" games, and we'll typically denote them $\mathcal{T} = (T, \mathbf{u})$. All of our assumptions apply to token games in the same way as to normal games. However, token games are intrinsically meaningless; their actions and payoffs don't represent anything in the real world. Instead, their payoffs must be realized by playing actions in the original game. We imagine this works as follows. Suppose that instead of playing G directly, the players agree to resolve it by playing the token game $\mathcal{T} = (T, \mathbf{u})$. To do so, they simultaneously declare their token actions $t_i \in T_i$, perhaps by writing them down on pieces of paper and then flipping them over. This results in a token outcome $t \in T$. The players then realize the token payoffs $\mathbf{u}(t)$ by playing some strategy profile in G with that (expected) payoff.

We say a token game $\mathcal{T}$ can be realized in a game G if there exists a *realization function* $\Psi : T \to \mathcal{F}(A)$ such that $\mathbf{u}(t) = \mathbf{u}(\Psi(t))$ for all $t \in T$. We'll consider two cases for $\mathcal{F}(A)$: the set of correlated strategy profiles $\Delta(A)$ and the set of pure strategy profiles A, and refer to the constructed games as correlated and pure token games, respectively. Formally, a pure/correlated token SPI on a game G is a pure/correlated token game $\mathcal{T}$ which is realizable in G and where $G \sim_\Phi \mathcal{T}$ via a Pareto-improving outcome correspondence Φ.

Commitments to play as prescribed by a token game can easily be made *ex post* verifiable. The players need to ensure that each player chooses their token action before learning the others'. This can be done through cryptographic commitment (Brassard, Chaum, and Crépeau 1988; Goldreich 2004) or using physical assumptions, e.g. by privately writing the actions on pieces of paper. For correlated token games, the players also need to (*ex post* verifiably) correlate their strategies. This is easy to do given a shared source of randomness: the randomness selects an outcome/strategy profile which the players are then required to play. This randomness could be provided by physical or cryptographic coin flipping (Blum 1983), which can even be done non-interactively (c.f. Canetti, Fiat, and Gonczarowski 2025), or by having the players submit strings of (purportedly random) bits and taking

their XOR (c.f. Walsh 2024). Note that, compared to cryptographic protocols for implementing correlated equilibria (Dodis, Halevi, and Rabin 2000), our solutions are much simpler because the players don't need distinct signals. Indeed making the players' signals distinct would generally break *ex post* verifiability.

We do not consider mixed token games, as they offer little benefit over correlated token games and come with substantial drawbacks. In contrast to the other sections, we do not consider a unilateral version of token SPIs. Token SPIs are commitments to play a strategy determined by a token game, which doesn't make much sense if the other players don't participate. For further discussion, see the appendix.

Theorem 3.1 shows that there are two types of SPIs: simple SPIs and isomorphism SPIs. We first show that simple token SPIs can be found in polynomial time, before moving on to isomorphism token SPIs, which will be our primary focus.

Simple Token SPIs. Applying the definition of simple SPIs to the present setting, a token game $\mathcal{T}$ is a simple SPI on a game G under Assumption A if, (a) $\mathbf{u}(t) \succeq \mathbf{u}(a)$ for all outcomes t in the reduced token game $\bar{T}$ and all outcomes $a \in \bar{A}$ and (b) there exists an outcome $a \in \bar{A}$ such that $\mathbf{u}(t) \succ \mathbf{u}(a)$ for all $t \in \bar{T}$. As one might expect, there's a simple characterization of simple token SPIs in both the pure and correlated cases.

Theorem 4.1. *A game G admits a simple token SPI realizable in $\mathcal{F}(A)$ if and only if there exists a payoff in $\boldsymbol{u}(\mathcal{F}(A))$ which weakly Pareto dominates all of $\boldsymbol{u}(\bar{A})$ and strictly Pareto dominates at least one payoff in $\boldsymbol{u}(\bar{A})$. For both pure and correlated token SPIs, it can be decided in polynomial time whether a simple token SPI exists.*

Isomorphism Token SPIs. We'll focus on isomorphism token SPIs for the rest of the section. By definition, a token game $\mathcal{T}$ is an isomorphism SPI on a game G under Assumptions A and B if there exists a Pareto improving isomorphism between $\bar{G}$ and $\bar{T}$. We begin by making some simplifying observations. When constructing a token game, there's no reason to include any token strategies that can be eliminated by Assumption A, so we'll only consider token games that contain no strictly dominated actions. In addition, since the SPI requires an isomorphism from T to $\bar{A}$, we can also consider only token games with $|T_i| = |\bar{A}_i|$ for all i.

A token SPI $\mathcal{T}$ on G requires a Pareto-improving isomorphism from $\bar{G}$ to $\bar{\mathcal{T}}$ and a realization function $\Psi : \mathbf{u}(T) \to \mathbf{u}(\mathcal{F}(A))$. The following technical lemma shows that, rather than needing to think about these two functions and their composition, we can consider a single function $\hat{\Psi} : \mathbf{u}(\bar{A}) \to \mathbf{u}(\mathcal{F}(A))$, which we call a utility remapping function. We'll call a utility remapping function *valid* if it is entrywise positive affine and strictly Pareto improving on $\mathbf{u}(\bar{A})$. That is, (1) For all outcomes $a \in \bar{A}$, $\hat{\Psi}(\mathbf{u}(a)) \succeq \mathbf{u}(a)$, (2) For some outcome $a \in \bar{A}$, $\hat{\Psi}(\mathbf{u}(a)) \succ \mathbf{u}(a)$, and (3) For all players i, there exist $m_i, b_i \in \mathbb{R}$ with $m_i > 0$ such that $\hat{\Psi}_i(v) = m_i v_i + b_i$ for all $v \in \mathbf{u}(\bar{A})$. The lemma shows a correspondence between isomorphism token SPIs realized in $\mathcal{F}(A)$ on G and valid utility remapping functions $\hat{\Psi} : \mathbf{u}(\bar{A}) \to \mathbf{u}(\mathcal{F}(A))$. Roughly, for any isomorphism token SPI, there's a valid utility remapping function $\hat{\Psi} : \mathbf{u}(\bar{A}) \to \mathbf{u}(\mathcal{F}(A))$ which characterizes the SPI's effect on payoffs, and for any valid $\hat{\Psi}$, there's an isomorphism token SPI with the effect on payoffs described by $\hat{\Psi}$.

Lemma 4.2. *Let G be a game and $\mathcal{T}$ be an isomorphism token SPI on G under Assumptions A and B that can be realized in $\mathcal{F}(A)$. Then there exists a valid utility remapping function $\hat{\Psi} : \boldsymbol{u}(\bar{A}) \to \boldsymbol{u}(\mathcal{F}(A))$ such that, for all $a \in \bar{A}$ and any isomorphism ϕ from G to $\mathcal{T}$, $\boldsymbol{u}(\phi(a)) = \hat{\Psi}(\boldsymbol{u}(a))$. Conversely, let $\hat{\Psi} : \boldsymbol{u}(\bar{A}) \to \boldsymbol{u}(\mathcal{F}(A))$ be a valid utility remapping function on the game G. Then there exists an isomorphism token SPI $\mathcal{T}$ under Assumptions A and B that can be realized in $\mathcal{F}(A)$ and for which, for all $a \in \bar{A}$ and all isomorphisms ϕ from G to $\mathcal{T}$, $\hat{\Psi}(\boldsymbol{u}(a)) = \boldsymbol{u}(\phi(a))$. In particular, there exists an isomorphism token SPI realizable in $\mathcal{F}(A)$ if and only if there exists a valid utility remapping function into $\boldsymbol{u}(\mathcal{F}(A))$.*

We will now apply Theorem 4.2 to pure and correlated token SPIs, beginning with the latter. We show that correlated token SPIs can be found efficiently. In addition, we characterize the existence of correlated token SPIs in two-player games.

Theorem 4.3 (Characterization of isomorphism correlated token SPIs)**.** *It can be decided in polynomial time whether a game G admits an isomorphism correlated token SPI. Furthermore, if G has exactly two players, we have the following characterization of when isomorphism correlated token SPIs exist. Let $V = \boldsymbol{u}(\bar{A})$, v_i^{min} and v_i^{max} be the minimum and maximum values of v_i in V, and $V^* \subseteq V$ be the set of points in V which cannot be strictly Pareto improved in $\boldsymbol{u}(\Delta(A))$. Assume $|V| \geq 2$, as otherwise isomorphism token SPIs are equivalent to simple SPIs and there's an SPI iff the unique point in V is not Pareto optimal in $\Delta(A)$.*

1. *If $|V^*| = 0$, G admits the desired SPI.*
2. *If $|V^*| = 1$, call that point v^*. Then*
   (a) *If $v_i^* \in \{v_i^{min}, v_i^{max}\}$ for both i, G admits the desired SPI.*
   (b) *If only one player i has $v_i^* \in \{v_i^{min}, v_i^{max}\}$, G admits the desired SPI if and only if, for all v in V with $v_i \neq v_i^*$, $(v_i + \varepsilon_v, v_{-i}) \in u(\Delta(A))$ for some $\varepsilon_v > 0$.*
   (c) *If for both i, $v_i^* \notin \{v_i^{min}, v_i^{max}\}$, G does not admit the desired SPI.*
3. *If $|V^*| \geq 2$, G does not admit the desired SPI.*

Proof sketch. To prove the first part of the theorem, we reduce the decision problem to checking the optimal value of a linear program. For the characterization in the 2-player case, we use Theorem 4.2, demonstrating a valid $\hat{\Psi}$ for the positive results and showing none exists for the negative results.

A key observation for the negative results is that, if a value v_i cannot be Pareto improved within $\mathbf{u}(\Delta(A))$, strictly for Player i, then it must be a fixed point of $\hat{\Psi}_i$. In case 3, where $|V^*| \geq 2$, each of these Pareto optimal values must be a fixed point of $\hat{\Psi}_i$ in every dimension i. Hence, each $\hat{\Psi}_i$ has at least two fixed points and must be the identity by linearity, so there can be no SPI. In case 2(c), where $|V^*| = 1$, the Pareto

optimal value v^* is a fixed point of each $\hat{\Psi}_i$ at an intermediate value $v_i \notin \{v_i^{min}, v_i^{max}\}$. This implies that each $\hat{\Psi}_i$ must be the identity; otherwise it would fail to be improving on either the values less than v_i or those greater than v_i.

For case 1, when $|V^*| = 0$, we show that the utility remapping function $\hat{\Psi}(v) = (1-\varepsilon)v + \varepsilon r^{max}$, where $r^{max} = (\max_{r \in R} r_1, \max_{r \in R} r_2)$ is Pareto improving and feasible for some $\varepsilon > 0$. Geometrically, this corresponds to mapping each value v some ε of the way towards r^{max} on the line segment between v and r^{max}.

For case 2(a), where $|V^*| = 1$ and this point v^* satisfies $v_i^* \in \{v_i^{min}, v_i^{max}\}$ for both i, we have two subcases. If v^* is maximal in both dimensions, $\hat{\Psi}(v) = (1-\varepsilon)v + \varepsilon v^{max}$ is feasible by convexity. If v^* is maximal in dimension i and minimal in dimension j, we show that the $\hat{\Psi}$ defined by $\hat{\Psi}_i(v) = (1-\varepsilon)v_i + \varepsilon v_i^*$ and $\Psi_j(v) = v$ is Pareto improving and feasible for some $\varepsilon > 0$. (Note that v^* can't be minimal in both dimensions since then we would have $|V| = 1$.)

For case 2(b), where v^* satisfies $v_i^* \in \{v_i^{min}, v_i^{max}\}$ for only one i, v^* must be maximal in dimension i and hence v_i^* is a fixed point of $\hat{\Psi}_i$. Since v_j^* is an intermediate fixed point of $\hat{\Psi}_j$, $\hat{\Psi}_j$ must be the identity. Thus, the only potential Pareto-improving $\hat{\Psi}$ has the form $\hat{\Psi}_i(v) = (1-\varepsilon)v_i + \varepsilon v_i^*$, as in case 2(a). This is feasible if and only if all points v with $v_i \neq v_i^*$ can be improved in the i dimension, as desired. □

Now, we turn our focus to isomorphism *pure* token SPIs, i.e., those whose payoffs must be realized over *pure* strategy profiles. The following theorem gives an algorithm for finding such SPIs. It runs in time $|A|^{O(n)}$, scaling polynomially in the number of outcomes but exponentially in the number of players. Since we generally assume each player has at least two actions, $n \in O(\log(|A|))$ and the overall runtime is $|A|^{O(\log|A|)}$. Such runtimes are sometimes called quasipolynomial time.

Theorem 4.4. *It can be decided in time $|A|^{O(n)} \in |A|^{O(\log|A|)}$, i.e. quasipolynomial time, whether a game G admits a pure isomorphism token SPI. For any fixed number of players n, this is polynomial time.*

Proof sketch. By Theorem 4.2, the desired SPI exists if and only if there exists a valid utility remapping function $\hat{\Psi} : \mathbf{u}(\bar{A}) \to \mathbf{u}(A)$, i.e. one which is a strictly Pareto improving, entrywise positive affine function on $\mathbf{u}(\bar{A})$. We give an algorithm that decides whether such a valid $\hat{\Psi}$ exists. Let $V = \mathbf{u}(\bar{A})$ be the set of payoffs in the reduced game. First, we efficiently find a set $V' \subseteq V$ of at most $n+1$ payoffs that contains at least two distinct payoffs for every player whose utility is not constant in V.

Any choice of $\hat{\Psi}' : V' \to \mathbf{u}(A)$ for such a V' determines the parameters $(m_i, b_i)_{i \in [n]}$ of any entrywise positive affine extension of $\hat{\Psi}$ to V: If V' contains two distinct values v' and v'', then $\hat{\Psi}'(v')$ and $\hat{\Psi}'(v'')$ determine the positive affine function in dimension i. If V' is constant in dimension i, so is V and thus any extension must map all v_i to the same value $\hat{\Psi}'_i(v')$. Hence, we can check whether a given $\hat{\Psi}'$ has a valid extension to V by checking whether $mv + b \in \mathbf{u}(A)$ for each $v \in V - V'$ and whether it is strictly Pareto improving on V. This can be done in $O(|A|^2)$.

Therefore, we can simply try each of the $O(|A|^{n+1})$ possible $\hat{\Psi}'$. This is polynomial for any constant number of players and quasipolynomial in general, since all players have at least two actions and therefore $n \in O(\log_2(|A|))$. Our algorithm enumerates all valid $\hat{\Psi}$, so we can also optimize arbitrary (quasi-)polynomial-time computable objectives over SPIs by simply computing the objective value of each valid $\hat{\Psi}$. □

The quasipolynomial efficiency of finding pure token isomorphism SPIs is in some sense an artifact of the fact that the representation size of normal-form games scales exponentially in the number of players. In particular, an abstracted version of the underlying pure token SPI problem is NP-complete. In this problem, rather than a game, we're given a set of input payoff vectors and a set of target payoff vectors. These correspond to the sets of payoffs in the reduced game and full game, respectively. The problem asks whether there is a strictly Pareto-improving, entry-wise positive affine mapping from the input set to target set. Of course, such a mapping exists if and only if there's a pure isomorphism token SPI in games with these reduced and full game payoffs.

Theorem 4.5. *The following problem is NP-complete. Given a set S of input vectors and a set of target vectors T in $\mathbb{R}^n$, decide whether there exists a strictly Pareto improving, entrywise positive affine mapping from S to $S \cup T$. That is, a function $\Psi : S \to S \cup T$ such that*

1. *$\Psi(v) \succeq v$ for all $v \in S$,*
2. *$\Psi(v) \succ v$ for some $v \in S$, and*
3. *For all players i, there exist $m_i, b_i \in \mathbb{R}$ with $m_i > 0$ such that $\Psi_i(v) = m_i v_i + b_i$ for all $v \in S$.*

Proof sketch. We reduce from the problem of graph 3-coloring. Given a graph (V, E), we construct a remapping instance which has a satisfying remapping if and only if the graph admits a 3-coloring. Our vectors have one dimension corresponding to each vertex. S consists of the $\binom{n}{2}$ vectors with value 1 in two dimensions and 0 in all others. We construct T so that, in each dimension i, we must have $\Psi_i(0) = .5$ and $\Psi_i(1) \in \{1, 2, 3\}$. $\Psi_i(1)$ corresponds to the color of vertex i. Specifically, T consists of vectors which have value .5 in all but two dimensions, and in those two dimensions can be some subset of $\{1, 2, 3\}$. For each $(i, j) \notin E$, T contains all of the vectors v with $(v_i, v_j) \in \{1, 2, 3\} \times \{1, 2, 3\}$, so that the colors of i and j do not constrain each other. For each $(i, j) \in E$, T does not contain vectors v with $v_i = v_j \in \{1, 2, 3\}$, encoding the constraint that vertices i and j cannot be the same color. □

This also shows that the pure token SPI problem becomes NP-hard if the game is represented in a more succinct form, e.g. as a dictionary which only stores payoffs that aren't uniformly zero.

Optimization. In addition to deciding whether SPIs exist in various settings, we'll also consider optimizing over SPIs. We define our objectives on valid utility remapping functions, which Theorem 4.2 shows specify the effect of an SPI on the

players' payoffs. We define a class of *linear* objective functions, which includes (weighted) utilitarian social welfare under some beliefs about the outcome of the reduced game. It generally does not include either Nash social welfare or maximizing one player's benefit subject to a minimum on the other player's benefit (under beliefs about the outcome of the reduced game). Roughly, we show that linear objectives over correlated token SPIs can be efficiently optimized. In contrast, *arbitrary* objectives over pure token SPIs can be optimized in quasipolynomial time, the same time complexity as our algorithm for the decision problem.

There's an important complication regarding what it means to optimize over correlated token SPIs. Valid $\hat{\Psi}$ must have strict inequalities $m_i > 0$, so the space of valid $\hat{\Psi}$ is not closed. Thus, there may not be an optimal SPI: It might be possible to get arbitrarily close to a particular objective value but not to achieve it. We show that we can optimize linear objectives over correlated token SPIs in the strongest sense one could hope for given this issue. For details, see the appendix.

5 Default-Remapping SPIs

In this section, we consider what SPIs can be achieved if the players can credibly reveal their default strategy $\Pi_i(G)$ and thus *ex post* verifiably commit to play according to some function Ψ of this default policy. We call this default-remapping commitment and refer to Ψ as a (default-) remapping function. Unilateral default-remapping commitment involves committing to some $\Psi_i : A_i \to \mathcal{F}(A_i)$. In the omnilateral case, when all players can credibly reveal their default, the players can choose a function $\Psi : A \to \mathcal{F}(A)$ and commit to play $\Psi(a)$ whenever the default policy $\Pi(G)$ results in outcome a. We'll consider the unilateral and omnilateral versions of these commitments, as well as the pure and correlated versions, where $\mathcal{F}(A)$ is either A or $\Delta(A)$.

Our reuse of the Ψ notation highlights the relationship of omnilateral default-remapping to the token game SPIs of the previous section. In the token game setting, the players commit to play the strategy profile $\Psi(t)$, where t is the outcome of a token game $\mathcal{T}$. In the omnilateral default-remapping setting, players commit to play the strategy profile $\Psi(a)$, where a is the outcome the players *would have reached* had they played the original game G as usual.

Though the ability to credibly reveal one's default policy is a strong assumption, it applies in some scenarios. For example, a player might intend to play a future game by copying the strategy of some public figure or taking the recommendation of a forthcoming paper. If this fact is common knowledge, the player could unilaterally commit to an *ex post* verifiable remapping of their default action.

The complexity of finding default-remapping SPIs depends on whether or not all players' default actions can be credibly revealed. As such, we'll consider these two cases separately.

5.1 Unilateral Default-Remapping SPIs

We first consider the case where only a strict subset of the players can commit to a strategy remapping. For notational simplicity, we specifically assume that only Player 1 can commit to a strategy remapping $\Psi_1(\Pi_1(\tilde{G}))$. Call the resulting interaction $G^{\Psi_1 \circ \Pi_1(\tilde{G})}$. For SPI purposes, unilateral default remapping is similar to unilateral utility function commitments (although we will see some differences below and especially in the appendix). Therefore, we illustrate it with the "Complicated Temptation Game", the same example that Oesterheld and Conitzer (2022, Table 4) use to illustrate unilateral utility function SPIs, see our Table 4.

		Player 2			
		C_1	C_2	F_1	F_2
Player 1	T_1	$4,2$	$1,1$	$6,0$	$6,0$
	T_2	$1,1$	$2,4$	$6,0$	$6,0$
	R_1	$0,0$	$0,0$	$5,3$	$3,2$
	R_2	$0,0$	$0,0$	$2,2$	$3,5$

Table 4: Complicated Temptation Game

By default, this game reduces to its top-left quadrant, where Player 1 chooses between T_1, T_2 and Player 2 chooses between C_1, C_2. Player 1 can unilaterally Pareto-improve by committing to choose R_1/R_2 had she chosen T_1/T_2 in the default.

To allow for a formal analysis of unilateral default-remapping commitments, we need assumptions about outcome correspondence for interactions of the form $G^{\Psi_1 \circ \Pi_1(G)}$. Specifically, we make three assumptions. The first two parallel the elimination of dominated strategies: Actions not in the image of Ψ_1 can be removed; and elimination of dominated actions for Players $-i$ works as before. We also need an analog of the isomorphism assumption (Assumption B). We also show (Proposition 5.1) how the assumptions can be used to prove the SPI for the example in Table 4.

First, we need two elimination assumptions. The first is that we can eliminate dominated actions for players *other* than i.

Assumption C. *If in G some action $\bar{a}_i$ of some Player $i \neq 1$ is strictly dominated, then*

$$\Pi(G^{\Psi_1 \circ \Pi_1(G)}) \sim_{(\mathrm{id}, \Xi_i)} \Pi((G - \{\bar{a}_i\})^{\Psi_1 \circ \Pi_1(G)})$$

where Ξ_i is the identity function except that it maps $\bar{a}_i$ to the empty set, i.e., $\Xi_i(a_i) = \{a_i\}$ whenever $a_i \neq \bar{a}_i$ and $\Xi_i(\bar{a}_i) = \emptyset$.

The second assumption is an elimination assumption for Player 1. It says that if Ψ_1 never maps to some action $\bar{a}_1$, then we can remove $\bar{a}_1$. This is important primarily by allowing us to apply Assumption C more often.

Assumption D. *Let $G, \hat{G}$ be games and let $\Psi^{-1}(\hat{a}_1) = \emptyset$, i.e., let a_1 be an action that is not in the image of Ψ_1. Then*

$$\Pi(\hat{G}^{\Psi_1 \circ \Pi_1(G)}) \sim_{(\Xi_1, \mathrm{id})} \Pi((\hat{G} - \hat{a}_1)^{\Psi_1 \circ \Pi_1(G)})$$

where $\Xi_1(a_1) = \{a_1\}$ for all $a_1 \neq \hat{a}_1$ and $\Xi(\hat{a}_1) = \emptyset$.

Third, we need a sort of isomorphism assumption to connect interactions of the form $\Pi(G^{\Psi_1 \circ \Pi_1(G')})$ to interactions

that are just normal form games. Roughly, the following assumption states: If P1 announces that she'll play like $\Pi(G')$ but mapped into $\hat{G}$ and moreover G' and $\hat{G}$ are isomorphic under $\Psi_1, \phi_2, ..., \phi_n$ in terms of the other players' utilities (for some $\phi_2, ..., \phi_n$), then $\Pi(\hat{G}^{\Psi_1 \circ \Pi_1(G')})$ will be played isomorphically to G'.

Assumption E. *Let G' be a fully reduced game. Let $\hat{G}$ be a game in which Players -1 have no strictly dominated strategies. Let $\Psi_1 \colon A'_1 \to \hat{A}_1$. Let $\phi_i \colon A'_i \to \hat{A}_i$ s.t. $(\Psi_1, \phi_2, ..., \phi_n)$ is an isomorphism in terms of the other players' utilities (i.e., for each $i \neq 1$, ϕ_i is a bijection and there are $m_i \in \mathbb{R}_+, b_i \in \mathbb{R}$ s.t. $u_i \circ (\Psi_1, \phi_2, ..., \phi_n) = m_i u_i + b_i$). Then $G' \sim_\Phi \Pi(\hat{G}^{\Psi_1 \circ \Pi_1(G')})$, where Φ is the union of all isomorphisms $(\Psi_1, \phi_2, ..., \phi_n)$ of the form above, i.e., $\Phi(a') = \{(\Psi_1, \phi_2, ..., \phi_n)(a') \mid (\Psi_1, \phi_2, ..., \phi_n)\}$.*

Using these assumptions, we can now formally prove that the unilateral default-conditional SPI for the Complicated Temptation Game is indeed an SPI.

Proposition 5.1. *Let G be the game of Table 4. Let $\bar{G}$ be the top-left quadrant of G. Let $\Psi_1 \colon T_1 \mapsto R_1, T_2 \mapsto R_2$. From Assumptions A and C to E, it follows that $\Pi(G^{\Psi_1 \circ \Pi_1(\bar{G})})$ is an SPI on G.*

Proof. By repeated application of the dominance assumption, we get that $G \sim_\Xi \bar{G}$, where Ξ maps outcomes including F or R actions to $\emptyset$ and otherwise maps outcomes onto themselves.

Now let $\hat{G}$ be the bottom-right game. Consider $\phi_2 \colon C_1 \mapsto F_1, C_2 \mapsto F_2$. Note that (Ψ_1, ϕ_2) is isomorphism for Player 2's utility (with coefficients $a = 1, b = 1$). Note further that ϕ_2 thus defined is the only such function. Thus, we have that $\bar{G} \sim_{(\Psi_1, \phi_2)} \Pi(\hat{G}^{\Psi_1 \circ \Pi_1(\bar{G})})$.

By Assumption D, we have

$$\Pi(G^{\Psi_1 \circ \Pi_1(\bar{G})}) \sim \Pi((G - \{T_1, T_2\})^{\Psi_1 \circ \Pi_1(\bar{G})}).$$

By Assumption C,

$$\Pi((G - \{T_1, T_2\})^{\Psi_1 \circ \Pi_1(\bar{G})}) \sim \Pi(\hat{G}^{\Psi_1 \circ \Pi_1(\bar{G})}).$$

Putting it all together using the transitivity rule, we get that $G \sim \Pi(G^{\Psi_1 \circ \Pi(\bar{G})})$. It is easy to verify that the resulting outcome correspondence is Pareto improving. □

One can characterize unilateral default-remapping SPIs in a way that's analogous to the characterization in Theorem 3.1. That is, if Ψ_1 is a default-remapping SPI, then we can see this by first reducing G and $G^{\Psi_1 \circ \Pi_1(G)}$ (i.e., the game under remapping Ψ_1). We can then have two types of SPIs (simple and isomorphism SPIs): The first is that every possible outcome of the prospective SPI $G^{\Psi_1 \circ \Pi_1(G)}$ is better than every outcome of the default G. The second is that the reduced games are isomorphic. An important difference between Theorem 3.1 and the present result is that we need the condition to state that *all* the isomorphisms are Pareto-improving. It does not suffice to consider one of the isomorphisms. We explain this in detail in the appendix.

We can now prove our main result about the complexity of finding unilateral default-remapping SPIs.

Theorem 5.1. *Deciding whether a game admits a unilateral default-remapping action remapping SPI (for Player 1) under Assumptions A and C to E is* NP*-hard, even for two players.*

As usual, the key difficulty is finding some part of the full game that is isomorphic to the original game. However, contrary to the other proof, only Player 2's utilities are relevant. Thus, (in the two-player case) we cannot straightforwardly reduce from any of the subgame isomorphism problems that we consider in our other proofs (see the appendix). Instead, we will directly reduce from the subgraph isomorphism problem. Nonetheless, the proof is a straightforward adaption of earlier proofs. In fact, Oesterheld and Conitzer's (2022) construction in the proof of their NP-hardness result can be used to show Theorem 5.1. To be complete and self-contained, we show how the proof of our Theorem 3.3 can be adapted to prove Theorem 5.1.

5.2 Omnilateral Default-Remapping SPIs

Now, we consider the case where all players can commit to strategies as a function of the default outcome of the game. In this case, the players' default-remapping commitment Ψ, along with the default policy, fully determines the outcome of the game. That is, $G \sim_\Psi G^\Psi$. Because of this, we don't need to reason about outcome correspondence, or strategic dynamics in general, when proving SPIs; SPIs occur whenever the remapping function is Pareto improving. Consequently, finding default-remapping SPIs requires only finding an outcome in the reduced game which can be feasibly Pareto improved.

Theorem 5.2. *Suppose the players can make omnilateral commitments to remap outcomes of the default policy to any feasible $\mathcal{F}(A)$ strategy profile. A default-remapping SPI exists under Assumption A if and only if there exists an outcome in $\bar{A}$ which is Pareto sub-optimal in $\mathcal{F}(A)$. For both pure and correlated default-remapping, it can be decided in polynomial time whether an SPI exists.*

The setting and result of correlated default-remapping part of the theorem above is related to that of "SPIs under improved coordination" from (Oesterheld and Conitzer 2022). Our setting is more restrictive on the ability of the players to commit, but allows the same class of SPIs to be achieved. (In our language, they essentially allow the players to make an omnilateral default-remapping commitment to remap the default outcome of an arbitrary game into $\Delta(A)$, while we only allow remapping the default of the original game.)

As in Section 4, we additionally consider *optimizing* over omni-lateral default-remapping SPIs. For details, see the appendix.

6 Conclusion

In this paper, we've studied safe Pareto improvements (SPIs) achieved through disarmament, commitment to token games, and default-remapping commitment. In each setting, we've characterized the computational complexity of finding and optimizing over SPIs. By considering forms of commitment which are *ex post* verifiable and thus easier to make credible and enforce, we hope to work towards SPIs that can be more readily applied in practice.

Acknowledgments

Nathaniel Sauerberg was supported by a grant from the CLR fund and funding from the Cooperative AI Foundation (CAIF). Some of this work was carried out as part of the ML Alignment & Theory Scholars (MATS) program. Caspar Oesterheld was supported by an FLI PhD Fellowship.

References

Apt, K. R. 2004. Uniform Proofs of Order Independence for Various Strategy Elimination Procedures. *The B.E. Journal of Theoretical Economics*, 4(1): 1–48.

Ashlagi, I.; Monderer, D.; and Tennenholtz, M. 2008. On the value of correlation. *Journal of Artificial Intelligence Research*, 33: 575–613.

Bade, S.; Haeringer, G.; and Renou, L. 2009. Bilateral commitment. *Journal of Economic Theory*, 144(4): 1817–1831.

Basilico, N.; Coniglio, S.; and Gatti, N. 2016. Methods for Finding Leader-Follower Equilibria with Multiple Followers. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 1363–1364.

Berker, R. E.; Tewolde, E.; Anagnostides, I.; Sandholm, T.; and Conitzer, V. 2025. The Value of Recall in Extensive-Form Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 13631–13640.

Blum, M. 1983. Coin flipping by telephone a protocol for solving impossible problems. *ACM SIGACT News*, 15(1): 23–27.

Brassard, G.; Chaum, D.; and Crépeau, C. 1988. Minimum disclosure proofs of knowledge. *Journal of computer and system sciences*, 37(2): 156–189.

Canetti, R.; Fiat, A.; and Gonczarowski, Y. A. 2025. Zero-Knowledge Mechanisms. In *Proceedings of the 26th ACM Conference on Economics and Computation*, 338–339.

Collina, N.; Derr, R.; and Roth, A. 2024. The Value of Ambiguous Commitments in Multi-Follower Games. *arXiv preprint arXiv:2409.05608*.

Coniglio, S.; Gatti, N.; and Marchesi, A. 2020. Computing a pessimistic Stackelberg equilibrium with multiple followers: The mixed-pure case. *Algorithmica*, 82(5): 1189–1238.

Critch, A.; Dennis, M.; and Russell, S. 2022. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *arXiv preprint arXiv:2208.07006*.

Deng, Y.; and Conitzer, V. 2017. Disarmament Games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Deng, Y.; and Conitzer, V. 2018. Disarmament games with resources. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. ISBN 978-1-57735-800-8.

DiGiovanni, A.; Clifton, J.; and Macé, N. 2025. Safe Pareto Improvements for Expected Utility Maximizers in Program Games. In *Proceedings of the 24rd International Conference on Autonomous Agents and Multiagent Systems*.

Dodis, Y.; Halevi, S.; and Rabin, T. 2000. A cryptographic solution to a game theoretic problem. In *Annual International Cryptology Conference*, 112–130. Springer.

Gabarró, J.; García, A.; and Serna, M. 2011. The complexity of game isomorphism. *Theoretical Computer Science*, 412(48): 6675–6695.

Gilboa, I.; Kalai, E.; and Zemel, E. 1990. On the order of eliminating dominated strategies. *Operations Research Letters*, 9(2): 85–89.

Goldreich, O. 2004. *Foundations of Cryptography, Volume 2*. Cambridge: Cambridge University Press.

Grohe, M.; and Schweitzer, P. 2020. The graph isomorphism problem. *Communications of the ACM*, 63(11): 128–134.

Harsanyi, J. C.; and Selten, R. 1988. *A general theory of equilibrium selection in games*. The MIT Press.

Kalai, A. T.; Kalai, E.; Lehrer, E.; and Samet, D. 2010. A commitment folk theorem. *Games and Economic Behavior*, 69(1): 127–137.

Köbler, J.; Schöning, U.; and Torán, J. 1993. *The Graph Isomorphism Problem: Its Structural Complexity*. Progress in Theoretical Computer Science. Springer Science+Business Media, LLC.

Kohlberg, E.; and Mertens, J.-F. 1986. On the Strategic Stability of Equilibria. *Econometrica*, 54(5): 1003–1037.

Koutsoupias, E.; and Papadimitriou, C. 1999. Worst-case equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science*, 404–413. Springer.

Kovařík, V.; Oesterheld, C.; and Conitzer, V. 2023. Game theory with simulation of other players. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2800–2807.

Kovařík, V.; Sauerberg, N.; Hammond, L.; and Conitzer, V. 2025. Game Theory with Simulation in the Presence of Unpredictable Randomisation. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, 1191–1199.

Mathon, R. 1979. A note on the graph isomorphism counting problem. *Information Processing Letters*, 8(3): 131–136.

Norde, H.; Potters, J.; Reijnierse, H.; and Vermeulen, D. 1996. Equilibrium selection and consistency. *Games and Economic Behavior*, 12(2): 219–225.

Oesterheld, C.; and Conitzer, V. 2022. Safe Pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems*, 36(2): 46.

Oesterheld, C.; and Conitzer, V. 2025. Choosing what game to play with no regrets or controversies — inferring safe (Pareto) improvements in binary constraint structures. In Bjorndahl, A., ed., *Proceedings of the Twentieth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2025)*, 246–265. Düsseldorf, Germany.

Pearce, D. G. 1984. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica*, 54(4): 1029–1050.

Renou, L. 2009. Commitment games. *Games Econ. Behav.*, 66(1): 488–505.

Sauerberg, N.; and Oesterheld, C. 2024. Computing Optimal Commitments to Strategies and Outcome-Conditional Utility Transfers. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 1654–1663.

Tennenholtz, M. 2004. Program equilibrium. *Games and Economic Behavior*, 49(2): 363–373.

Walsh, T. 2024. Mechanisms That Play a Game, Not Toss a Coin. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 3005–3013. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zemlyachenko, V. N.; Korneenko, N. M.; and Tyshkevich, R. I. 1985. Graph isomorphism problem. *Journal of Soviet Mathematics*, 29: 1426–1481.