Expanding and Evaluating the Applicability of Safe Pareto Improvements

Nathaniel Sauerberg
University of Texas
Austin, United States
nsauerberg@utexas.edu

Caspar Oesterheld
Foundations of Cooperative AI Lab,
Carnegie Mellon University
Pittsburgh, United States
oesterheld@cmu.edu

ABSTRACT

Background: Safe Pareto improvements (SPIs) are commitments in games which leave all players better off with certainty.
Objectives and Research Questions: We study what SPIs can be achieved via joint commitment to token games: fictitious, cheap-talk games rendered meaningful by commitments to take payoff-relevant actions (such as strategies in the original game) as a function of the token game's outcome. Such commitment effectively allows the players to replace the original interaction with a new game whose payoffs they can design to be isomorphic to and entrywise Pareto improving on the original game.
Methods: We theoretically analyze three ways of augmenting token games and conduct computational experiments to measure the probability that different methods achieve SPIs in random games.
Results: First, we allow commitments to burn utility and show that this enables SPIs in games where commitment to strategies alone cannot. Second, we consider commitment to utility transfers and again show this enables SPIs in new classes of games. Finally, we consider allowing t he players to make monetary contracts with outside parties based on the outcome of the token game. Under very mild assumptions on what contracts the outside parties accept, this enables SPIs in all non-zero-sum games.
Conclusions: We introduce a new approach to mutually beneficial commitment. We characterize its power theoretically, and show experimentally that our methods substantially increase the frequency and magnitude of SPIs in random games.

KEYWORDS

Commitment, Bargaining, Safe Pareto Improvements, Algorithmic Game Theory, Cooperative AI, Linear Programming

ACM Reference Format:
Nathaniel Sauerberg and Caspar Oesterheld. 2026. Expanding and Evaluating the Applicability of Safe Pareto Improvements. In *Appears at the 8th Games, Agents, and Incentives Workshop (GAIW-26). Held as part of the Workshops at the 25th International Conference on Autonomous Agents and Multiagent Systems., Paphos, Cyprus, May 2026*, IFAAMAS, 17 pages.

Appears at the 8th Games, Agents, and Incentives Workshop (GAIW-26). Held as part of the Workshops at the 25th International Conference on Autonomous Agents and Multiagent Systems., Armstrong, Curry, Hosseini, Mattei, Tsang, Wąs (Chairs), May 2026, Paphos, Cyprus. . . . $ACM ISBN 978-x-xxxx-xxxx-x/YY/MM

1 INTRODUCTION

It's well-known that joint commitment in games can enable mutually beneficial agreements. For instance, players can commit to play (Cooperate, Cooperate) in a prisoners' dilemma. This is considered mutually beneficial because it's generally accepted that the default outcome would be (Defect, Defect). However, in other games, it's much less clear how to evaluate how the game would be played by default, and therefore whether a particular agreement is mutually beneficial. For instance, consider the game of Chicken. Is an agreement to randomize uniformly over (Dare, Swerve) and (Swerve, Dare) mutually beneficial? Answering this requires evaluating what equilibrium, if any, would be played by default. For example, if both players subjectively expect to Dare while the other Swerves, neither will consider such an agreement beneficial.

Much prior work has characterized the impact of various commitment frameworks [e.g. 7, 16, 18, 24, 30] with so-called *folk theorems*: any feasible payoff profile can be achieved in equilibrium so long as each player achieves at least the minimum utility they can unilaterally guarantee for themselves ("individual rationality"). The upside of such results is limited by a general version of the problem discussed above: any Pareto optimal payoff profile can be reached, but agreeing on a particular such profile might require agreeing how the game would be played by default.

In this paper, we allow players to make joint, binding commitments before playing a game, and seek commitments which guarantee a Pareto improvement regardless of how the game is played. We call such agreements safe Pareto improvements (SPIs). In particular, we seek commitments which induce a game which is isomorphic to the original, but Pareto better. This approach overcomes several difficulties. First, it may just be very unclear how the players would play the game by default (e.g. if and how they would have resolved equilibrium selection), and therefore whether agreeing to play any particular outcome is in fact an improvement on the default. Second, even if the players do have beliefs about how the game would be played by default, they might disagree about what will happen. For instance, each player might expect to achieve their preferred Nash equilibrium in Chicken, in which case no mutually beneficial agreement exists. Finally, the incentives around bargaining and strategic posturing might prevent the elicitation of the players' true beliefs. In Chicken, for instance, the players might claim to expect their preferred equilibrium regardless of their true beliefs, so as to claim that they intend to Dare and the other player should Swerve.

Our approach circumvents all three of these potential obstacles for achieving Pareto improvements. By inducing a game which is isomorphic to the original, we preserve the strategic structure of the game: the equilibria, the dominance relationships, etc. Therefore,

	Dare	Swerve	Concede
Dare	-10, -10	5, 1	8, 0
Swerve	1, 5	3, 3	8, 0

Table 1: Chicken with Concede

	Token Dare	Token Swerve
Token Dare	-2.5, -2.5	5, 3
Token Swerve	3, 5	4, 4

Table 2: A Token Game with Payments SPI on Chicken with Concede

each player should plausibly play the new game the same as the original, and have the same beliefs about how the other players will play. But because every possible outcome in the new game Pareto-dominates the corresponding outcome in the original game, such commitment leaves all players better off, regardless of their beliefs.

Our approach to achieving these goals is based on token games, first introduced in [27]. A token game is a normal-form game which is symbolic and is played over cheap talk. For instance, token game with action spaces T_i could be played by having each player i privately write down their action t_i on a piece of paper, which are then simultaneously revealed to determine the token outcome t. This is not inherently meaningful, but can be rendered meaningful by commitments to act in the original game as a function of the token outcome. The original game payoffs of these actions then determine the payoffs of the token game.

As an example, consider the game Chicken with Concede, Table 1. This is a version of chicken in which Player 2 has an additional action, Concede, which is strictly dominated by Swerve. Therefore, we assume Concede is never played and can be removed without changing how the game is played.

We'd like to construct a token game which is isomorphic to the Chicken with Concede (excluding the concede action), but Pareto better. Such a game is given in Table 2. First, observe that the payoffs of the token game can be achieved by committing to play strategies and transfer utility in the original game. For example, if (Token Dare, Token Swerve) is played, the players commit to play (Dare, Concede) in the original game, and Player 1 commits to transfer 3 utility to Player 2, resulting in payoffs of $(8-3, 0+3) = (5, 3)$. The payoffs of (Token Swerve, Token Dare) and (Token Swerve, Token Swerve) can be implemented by similar commitments. Finally, the payoffs of (Token Dare, Token Dare) can be implemented by committing to play a correlated strategy profile that randomized between (Dare, Dare) and (Swerve, Swerve) with a particular probability.

Second, observe that the token game is isomorphic to the original game (after removing Concede). For an outcome in the original game with payoff profile (v_1, v_2), the corresponding outcome in the token game has payoff profile $(0.5v_1 + 2.5, 0.5v_2 + 2.5)$. We'll therefore assume that the token game would be played isomorphically to the original game, and that the players' beliefs about how it would be played are similarly isomorphic. But each outcome in the token game Pareto-dominates the corresponding outcome in the original game. Therefore, regardless of their beliefs about how the original game would be played, both players should prefer to play the token game, and regardless of how the original game is played, the token game guarantees a Pareto improvement on the original. We've found our desired Pareto improving commitment, which we'll refer to as a Safe Pareto Improvement (SPI).

Contributions. In this work, we extend the token game framework for achieving safe Pareto improvements in three ways, allowing players to additionally commit to (1) burn utility, (2) transfer utility to each other, and (3) transfer utility to outside parties. For each extension, we characterize the existence of SPIs and give computational complexity results.

Money burning (Section 3). We show that, perhaps surprisingly, money burning strictly increases the power of token games. We characterize the existence of money-burning token game SPIs in two-player games (Theorem 1). Finally, we show that a large class of natural objectives can be optimized over money-burning token SPIs in polynomial time using linear programming (Theorem 2).

Utility transfers (Section 4). We characterize the existence of payment token game SPIs in n-player games (Theorem 3). Notably, payment token SPIs can exist even in games where no actions are eliminated and no simple SPI exists. We also give a linear program for optimizing over payment token SPIs (Theorem 4).

Outside bettors (Section 5). We introduce a scheme in which the players construct a financial security tied to the token game's outcome and sell it to an outside party. We show that token SPIs with outside bettors exist under extremely mild assumptions—essentially whenever the game is not constant-sum (Theorems 5 and 6), and how under stronger assumptions, they allow the players to achieve the maximum social welfare in the original game while preserving the original game's strategic structure and distribution of payoffs (Theorem 7).

Computational Experiments (Section 6). We present computational experiments comparing the effectiveness of these interventions on randomly generated games. We find that the extensions introduced in this paper—especially token games with payments—substantially increase the frequency and magnitude of SPIs relative to prior methods.

1.1 Related Work

As discussed in the introduction, safe Pareto improvements were first introduced by [22] and further studied by [10, 23, 27]. Besides the idea of SPIs themselves we borrow some of the conceptual features of this prior work, such as the idea of reducing games by elimination of irrational actions and using isomorphisms between games. However, we offer a simplified theoretical setup. Roughly, we directly define when we consider one game to be an SPI on another game (under the assumptions that strictly dominated actions can be removed and isomorphic games are played isomorphically). In contrast, prior papers separately introduce the assumptions and the SPI concept. Thus, our *definitions* of SPIs (Definitions 2.2 and 2.3) are characterization results in prior papers [e.g. 27, Lemma 3.1].

In this paper we primarily build on the token games of Sauerberg and Oesterheld [27]. We extend these token games by allowing players to commit to burn or transfer utility depending on the outcome and by allowing transfers with an outside expected-utility maximizing party. Our experimental results also cover the utility function SPIs of Oesterheld and Conitzer [22].

A lot of prior game-theoretic work has studied the use of commitments to money burning [20, 29] and to utility transfers [e.g. 1, 4, 8, 15, 19, 26, 28]. However, none of this prior work uses such commitment to achieve SPIs or uses token games.

Our Section 5 shares some thematic connection to a line of work, initiated by [3], on redistributing budget surplus generated by the well-known VCG mechanism [5, 14, 31] while preserving incentive compatibility. In particular, our outside bettors approach (Section 5) is reminiscent of the idea of choosing a player or set of players who cannot receive the items but are entitled to the entire budget surplus [11, 21]. However, our surplus is generated through a very different mechanism: token games rather than payments in an auction.

2 PRELIMINARIES

2.1 Basic game theory

We first introduce some game-theoretic notation and terminology. An *n-player (normal-form) game* G is a pair $(A, \mathbf{u})$, where $A = \bigtimes_i A_i$ for some nonempty set of *actions* A_i for each player i, and $\mathbf{u} : A \to \mathbb{R}^n$ is a *utility function*. The elements of A are *outcomes* or action profiles, and $u_i(a)$ is Player i's utility for outcome a. We use $\Delta(A)$ to denote the set of distributions over A, also called correlated strategy profiles. We extend $\mathbf{u}$ to distributions over outcomes by taking the expectation and define $\mathbf{u}(S) = \{\mathbf{u}(s) : s \in S\}$ for any set S of outcomes or distributions over outcomes. We index the set of players $i \in \{1, \ldots, n\} = [n]$, and use $-i$ to denote the set of players other than i. The (utilitarian) *social welfare* $\mathrm{SW}(a)$ of an outcome a is the sum of the players' payoffs $\mathrm{SW}(a) = \sum_{i \in [n]} u_i(a)$. Slightly abusing notation, we let $\mathrm{SW}(\cdot)$ apply to distributions over outcomes and payoff vectors as well. We also let $\mathrm{SW}^{max}(S)$ denote the maximum social welfare in the set S, $\max_{s \in S} \mathrm{SW}(s)$, with $\mathrm{SW}^{min}(A)$ defined analogously.

A *subgame* G' of an n-player game $G = (A, \mathbf{u})$ is a game $G' = (A', \mathbf{u}')$ where $A' = \times A'_i$ for nonempty subsets $A'_i \subseteq A_i$ and where $\mathbf{u}'$ is $\mathbf{u}$ restricted to A'. We imagine that Player i is the same entity (person, etc.) in all games and that their utility is comparable between games. For this reason, we usually consider a single utility function $\mathbf{u}$ which applies to outcomes of any game. This allows us to reason about a (safe) Pareto improvements between games.

An outcome a' *is a (weak) Pareto improvement on* a, denoted $\mathbf{u}(a') \succeq \mathbf{u}(a)$, if for all i we have that $u_i(a') \geq u_i(a)$. Furthermore, a' is a *strict* Pareto improvement on a, or $\mathbf{u}(a') \succ \mathbf{u}(a)$, if additionally there is a player i for which $u_i(a') > u_i(a)$. We say that an outcome a is *Pareto optimal* within a set S if there is no $s \in S$ with $s \succ a$. We also apply all of these terms to the payoff vectors $v = \mathbf{u}(a)$ themselves.

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

2.2 Safe Pareto improvements

We now formally introduce the concept of safe Pareto improvements (SPIs). Relative to prior work we have simplified the theoretical foundation, as discussed briefly in Section 1.1.

Strategic interactions (as represented by normal-form games) can have actions that are clearly irrational, e.g., because they are strictly dominated by another action. One of the main ingredients to the SPI framework is that we are given some apparatus for removing such irrational actions.

Definition 2.1. A *reduction function* $\mathtt{Red}$ is a function that maps any game G onto a subgame of G s.t. (1) for all games G we have that $\mathtt{Red}(\mathtt{Red}(G)) = \mathtt{Red}(G)$; and (2) for all games G and G' with isomorphism ϕ we have that ϕ (restricted to the outcome set of $\mathtt{Red}(G)$) is also an isomorphism between $\mathtt{Red}(G)$ and $\mathtt{Red}(G')$.

Various ways to reduce games have been discussed in the game-theoretic literature. Unless otherwise specified, $\mathtt{Red}$ fully reduces the game by iterated elimination of strictly dominated strategies [17, 25], which is the notion of reduction used by all prior SPI work [22, 23, 27]. We will also consider the elimination of dominated strategies in *mixed* strategies [2, 12, 25]. To avoid calling attention to the specific reduction function, we'll often denote $\bar{G} = \mathtt{Red}(G)$ and write $\bar{A}$ for the action space of $\bar{G}$.

Combining the notion of reduction and the notion of game isomorphism, we can define the notion of an SPI.

Definition 2.2. A game G' *is an isomorphism SPI on* G if $\mathtt{Red}(G')$ is isomorphic to $\mathtt{Red}(G)$ and this isomorphism is outcome-wise strictly Pareto improving.

As a simple example, let G be the game of Chicken, and let G' be a version of G where each utility is increased by, say, 1. Then G' is an SPI on G. Intuitively, if G' is an SPI on G, then everyone involved should prefer that G' rather than G is played. For instance, they should (in most cases) have the same beliefs about what will happen in G' versus G, and favor G' because the corresponding outcomes are better. For more detailed theoretical justifications of SPIs, see [22, 23, 27]. (Note also that there may be multiple isomorphisms between two given games. In Definition 2.2 we only talk about *one* of the isomorphisms. This is justified by the fact that all isomorphisms between two games must have the same effects on utilities, see Section A.)

In addition to isomorphism SPIs, we consider simple SPIs, which exist roughly whenever all outcomes of one reduced game are Pareto better than all outcomes of another.

Definition 2.3. A game G' is a simple SPI on G if all outcomes in $\mathtt{Red}(G')$ are weakly Pareto better than all outcomes in $\mathtt{Red}(G)$ and this relation is strict for at least one pair.

In this work, we're interested in token games which are SPIs on the base game G, which we call token game SPIs or token SPIs.

	Hawk	Dove	Crazy	Refrain
Hawk	0, 0	6, 2	0, 4	0, 0
Dove	2, 6	4, 4	4, 0	0, 0
Crazy	4, 0	0, 4	0, 0	0, 0
Superdove	−2, 12	−2, 12	−2, 12	5, 5

Table 3: Hawk-Dove-Crazy: A game to illustrate that burning money is sometimes useful for constructing token game SPIs

	Token Hawk	Token Dove	Token Crazy
Token Hawk	3, 0	6, 2	3, 4
Token Dove	4, 6	5, 4	5, 0
Token Crazy	5, 0	3, 4	3, 0

Table 4: A money-burning token SPI on the Hawk-Dove-Crazy game of Table 3

3 MONEY BURNING

In this section, we consider token games which are realized via joint commitments to play correlated strategy profiles in the base game and to burn utility as a function of the token outcome.

Formally, given a base game $G = (A, \mathbf{u})$, a token game with money burning on G is a game $\mathcal{T} = (T, \mathbf{u})$ for which there exists $\Psi : T \to \Delta(A)$ such that for all players i and token outcomes t, $u_i(t) \leq u_i(\Psi(t))$. The players can realize such a token game by committing that, whenever outcome t obtains, they'll play strategy profile $\Psi(t)$ and each player i will burn utility $u_i(\Psi(t)) - u_i(t) \geq 0$, leaving each with the intended utility of $u_i(t)$. Sauerberg and Oesterheld's (2026) token games instead require that $u_i(t) = u_i(\Psi(t))$.

It's perhaps unexpected that commitment to burning money could ever be helpful in achieving SPIs, especially considering that the space $\mathbf{u}(\Delta(A))$ of feasible payoffs is already convex.

Example: Consider the game in Table 3. The reduced game is a symmetric 3x3 game where each player has actions hawk, dove, and crazy. It works roughly like rock/paper/scissors: Hawk beats Dove, Dove beats Crazy, and Crazy beats Hawk, though the exact payoffs vary. Roughly, if one keeps who "wins" the RPS aspect fixed, it's generally better for both if Dove is played and Crazy is not. In addition, Player 1 has a "Superdove" action which get exploited for $(-2, 10)$ by any of Player 2's normal actions but which results in payoffs of $(5, 5)$ if Player 2 refrains from exploiting. Similar to in the trust game, refrain is strictly dominated for Player 2 and therefore both Refrain and Superdove are eliminated under iterated strict dominance.

We construct a token game which is isomorphic to the reduced Hawk-Dove-Crazy game. The isomorphism maps Player 1 payoffs of v_1 to $.5v_1 + 3$ and Player 2 payoffs by the identity. Each of Player 1's payoffs is mapped half of the way towards her maximum payoff of 6. One way to view this is that for Player 1, the SPI is equivalent in expectation to a $1/2$ probability of playing the game normally and $1/2$ probability of automatically getting her maximum payoff of 6.

The $(4, 6)$ payoff of (T Dove, T Hawk) can be realized by mixing between (Superdove, Refrain) and (Superdove, Hawk) in a 6:1 ratio for payoffs of $(6/7)(5, 5) + (1/7)(-2, 12) = (4, 6)$. The $(6, 2)$ payoff of (T Hawk, T Dove) can be realize by playing (Hawk, Dove). All other token payoffs are Pareto dominated by the $(5, 5)$ of (Superdove, Refrain) and so can be realized by commitments to play (Superdove, Refrain) and burn utility. Hence, this is a valid money burning token game SPI.

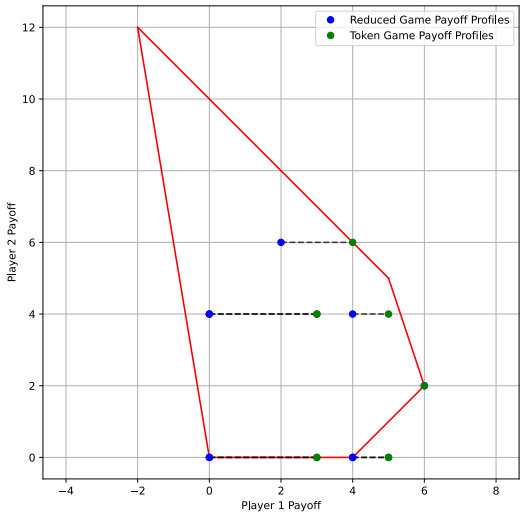

Figure 1: Geometric visualization of the money-burning token game SPI for the Hawk-Dove-Crazy game

Interestingly, the ability to commit to burn utility is necessary to implement this, or any, token SPI on Hawk-Dove-Crazy. Here, we give a relatively informal argument for this; A formal proof that money burning can be necessary for achieving token game SPIs is given in Corollary 1.

Consider the functions on the players payoffs induced by an isomorphism. The $(6, 2)$ payoff of (Hawk, Dove) is Pareto optimal in G, and hence must be a fixed point of each player's isomorphism. Hence, any SPI must improve only P1's payoffs while keeping P2's constant, and each of P1's other payoffs must be strictly improved. However, the $(4, 0)$ payoff of (Crazy, Hawk) cannot be improved for Player 1 without also increasing Player 2's payoff, and so Player 2's ability to commit to money burning is necessary for an SPI to exist.

The effect of the money burning token SPI on Hawk-Dove-Crazy is visualized geometrically in Figure 1. The red lines are the boundary of the set of feasible payoffs in Hawk-Dove-Crazy. The blue points are the payoff profiles in the reduced original game, while the green points are the payoffs of the money burning token game SPI. The bottom right blue point at $(4, 0)$ is improved in the money burning token game SPIs, but is on the boundary of the feasible region so cannot be improved without money burning. Hence, money burning is necessary for an SPI to exist.

We now turn to the questions of which games admit MB token game SPIs and of optimizing over such SPIs.

First, it is easy to see that simple SPIs exist if and only if there is $\delta \in \Delta(A)$ which is Pareto better than all of $\bar{A}$ and strictly so for at least one $a \in \bar{A}$. (Money burning makes no difference here.)

For the rest of this section, we focus on isomorphism SPIs. Let $\mathcal{T}$ be an isomorphism SPI on G. Then by definition there exists an isomorphism ϕ between $\mathtt{Red}(G)$ and $\mathtt{Red}(\mathcal{T})$. We will associate with $\mathcal{T}$ the transformation on *payoffs* induced by ϕ. This function, which we'll call a utility isomorphism function and denote $\hat{\Psi}$, has

two important properties. First, each payoff $\hat{\Psi}(\mathbf{u}(a))$ for $a \in \bar{A}$ must be attainable via commitments to play the strategy profile $\Psi(t)$ and burn the (nonnegative) utility vector $\hat{\Psi}(\mathbf{u}(t)) - \mathbf{u}(\Psi(t))$. In other words, $\hat{\Psi}$ must be into the set of feasible payoff $\mathcal{F}^{\text{MB}}(G) = \{v : \exists v' \in u(\Delta(A)), v' \succeq v\}$. Secondly, by the definition of isomorphism, $\hat{\Psi}$ must be player-wise positive affine, i.e. be of the form $\hat{\Psi}_i(v) = m_i v_i + b_i$ for some $m_i, b_i \in \mathbb{R}$ with $m_i > 0$ for each player i. In other words, $\hat{\Psi}$ must be a player-wise positive affine function $\hat{\Psi} : \mathbf{u}(\bar{A}) \to \mathcal{F}^{\text{MB}}(G)$. We'll call such $\hat{\Psi}$ "valid". Of course, for any valid $\hat{\Psi}$, we can construct a token game SPI with money burning which induces this utility isomorphism. Therefore, when characterizing the existence of and optimize over money burning token game SPIs, we'll often just reason about the space of valid $\hat{\Psi}$.

Now, we introduce our characterization of MB token game SPIs in two-player games.

THEOREM 1. *A two-player game G admits a MB-token game isomorphism SPI if and only if either (a) $\bar{A}$ contains no outcomes which are Pareto optimal in $\boldsymbol{u}(A)$ or if (b) it contains exactly one payoff v^* which is Pareto optimal in $\boldsymbol{u}(A)$, this v^* is at least one player's maximum payoff in $\bar{A}$, and that player has at least one more distinct payoff in $\bar{G}$.*

PROOF SKETCH. The proof works by reasoning about the types of constraints points in v put on $\hat{\Psi}$, and in particular any fixed points of $\hat{\Psi}$. In particular, any $v^* \in \mathbf{u}(\bar{A})$ which is Pareto optimal in G must be a fixed point of $\hat{\Psi}$ (in both dimensions): it cannot be increased in either dimension without being decreased in the other, which cannot happen because $\hat{\Psi}$ must be Pareto improving.

Only If / Nonexistence: If $\mathbf{u}(\bar{A})$ has two distinct payoffs which are Pareto optimal in G, then each $\hat{\Psi}_i$ has two fixed points and must be the identity. In addition, say for some Pareto optimal v^*, v_i^* is "intermediate" in $\mathbf{u}_i(\bar{A})$, i.e. Player i has both higher and lower payoffs in than v_i^* in $\mathbf{u}(\bar{A})$. This also implies then $\hat{\Psi}_i$ must be the identity, since this v_i^* is a intermediate fixed point and so any positive affine $\hat{\Psi}_i$ aside from the identity would either decrease Player i's payoff at most values above or below v_i^*. These together imply that, for an SPI to exist, there can be at most $v^* \in \mathbf{u}(\bar{A})$ which is Pareto optimal in G, and if it exists it must be at least one player's maximum or minimum payoff. But a Pareto optimal v^* cannot represent a player's minimum utility unless it's also a player's maximum utility, so that case is covered in the positive result.

If / Existence: In the case where $\bar{G}$ contains no payoffs which are Pareto optimal in G, We show that the $\hat{\Psi}_i(v) = v + \varepsilon(v^{max} - v)$ is feasible for some $\varepsilon > 0$. Let $v^{max} = \left(\max_{v \in \mathbf{u}(A)} v_1, \max_{v \in \mathbf{u}(A)} v_2\right)$ be the (typically infeasible) profile of each player's maximum payoff. Geometrically, this corresponds to mapping each payoff some tiny fraction of the way towards this maximum point. This is clearly feasible for any point v on the interior of $\mathcal{F}^{\text{MB}}(G)$. The boundary points which might pose a problem are those v with $v_i = v_i^{max}$ for some i, but for these points $\hat{\Psi}$ only changes them in the other dimension, which is feasible since these v are not Pareto optimal.

In the second case above, where $\bar{G}$ contains a Pareto optimal payoff v^* where (without loss of generality) v_1^* is Player 1's maximum payoff in $\bar{A}$, we show that letting $\hat{\Psi}_1(v_1) = v_1 + \varepsilon(v_1^* - v_1)$ and $\hat{\Psi}_2 = id$ is feasible for some $\varepsilon > 0$. Geometrically, this corresponds to projecting each point an ε fraction of the way towards the line defined by $v_1 = v_1^*$. This is feasible for some nonzero ε because all remaining points v in $\mathbf{u}(\bar{A})$ are Pareto dominated, so either they're Pareto dominated strictly in dimension 1 or $v_1 = v_1^{max(G)}$, in which case $v_1^{max} = v_1^*$ and the the $\hat{\Psi}$ is feasible for any ε. □

We have shown that money burning makes token games more powerful. The following result, however, restricts the additional power from money burning in two-player games. In particular, it shows that when Sauerberg and Oesterheld's [2026] token games don't yield an SPI, then a money-burning token game SPI (which may exist) will only ever benefit one of the players.

COROLLARY 1. *If a two-player game G does not admit a standard token game SPI, it cannot admit a MB token game SPI which is strict for both players.*

PROOF. Consider a game G which admits a MB token game SPI but not a standard token game SPI. If $\bar{G}$ contained no outcomes which were Pareto optimal in G, G would admit a standard token game SPI (by the charactarization of standard token SPIs from [27], Theorem 4.3). Hence, G must be as in the second condition in Theorem 1: it contains exactly one payoff v^* which is Pareto optimal in G and v^* is (wlog) Player 1's maximum payoff in $\bar{G}$.

By Theorem 4.3 of [27], for such a G not to admit a standard token SPI, this v^* must be an intermediate payoff for Player 2. But then v_2^* is an intermediate fixed point of $\hat{\Psi}_2$, which therefore must be the identity and the SPI cannot be strict for Player 2. □

Although our characterization only applies to two-player games, we now show that the existence of money-burning token game SPIs can be decided in general by linear programming (just like token SPIs without money burning [27]). Moreover, we can also use linear programming to optimize over such SPIs.

We consider optimization over a large class of natural objectives on isomorphism SPIs. It includes, for example, maximizing a player's subjective expected utility for playing the token game under some belief, i.e. probability distribution over outcomes of the token game. It also includes maximziming (weighted) sums of the players' utilities under such subjective beliefs, which are allowed to be different for each player. It also allows maximizing a player's worst-case (over outcomes) utility in the token game, or worst-case utility gain. And we can again aggregate these over players, either by taking some weighted sum over the players' objectives or by maximizing the minimum benefit across players.

More formally, we optimize over the class of objectives defined by the minimum over a (polynomially sized) set of linear functions on variables of the form $\hat{\Psi}_i(\mathbf{u}(a))$ and $\hat{\Psi}_i(\mathbf{u}(a)) - u_i(a)$, for outcomes $a \in \bar{A}$ and players i. These are, respectively, the players' payoffs for specific outcomes in the token game and their payoff gains relative to the base game under the utility correspondence. We'll call these min-linear objectives.

THEOREM 2. *It can be decided in polynomial time via linear programming whether a game G admits a MB-token game isomorphism SPI. Furthermore, min-linear objectives over such SPIs can be optimized efficiently.*

There's a subtlety regarding what it means to optimize over these SPIs. Because our utility correspondence functions must have $m_i >$

	Dare	Swerve
Dare	-10, -10	6, 1
Swerve	1, 5	3, 3

(a) Asymmetric Chicken

	Dare	Swerve
Dare	-6.8, -10	6, 1
Swerve	2, 5	3.6, 3

(b) A payments token game SPI

Table 5: Asymmetric Chicken and a payments token game SPI

0, the space of SPIs is not closed. However, we can still optimize linear objectives over SPIs in the strongest sense one could hope for given this issue: we can efficiently decide whether the instance admits an SPI, and if so, find it. If not, we can compute the supremum of the objective value over SPIs, and find a family of SPIs which approach this supremum.

4 TOKEN GAME SPIS WITH UTILITY PAYMENTS

In this section, we consider the token games which are realized by joint commitments to play correlated strategy profiles and transfer utility as a function of the token outcome. We assume players can transfer each other utility losslessly, linearly, and without any budget constraints.

Formally, given a base game $G = (A, \mathbf{u})$, a payments token game on G is a game $\mathcal{T} = (T, \mathbf{u})$ for which $\mathbf{u}(t) \in [\min_{a \in A} \mathrm{SW}(a), \max_{a \in A} \mathrm{SW}(a)]^n$. (Recall that $\mathrm{SW}(a) = \sum_i u_i(a)$ denotes the utilitarian social welfare of an outcome.) These token payoffs are exactly those which can be realized by a commitment to a correlated strategy profile and payment vector, i.e. the $\mathbf{u}(t)$ such that $\mathbf{u}(t) = \mathbf{u}(\Psi(t)) + p(t)$ for some $\Psi(t) \in \Delta(A)$ and $p(t) \in \mathbb{R}^n$ with $\sum_i p_i(t) = 0$.

Note that allowing the players to additionally commit to money burning makes no difference here: any payoff profile with social welfare within $[\min_{a \in A} \mathrm{SW}(a), \max_{a \in A} \mathrm{SW}(a)]$ can already be achieved with transfers alone. Thus, the only payoff profiles additionally rendered feasible by money burning are ones with social welfare less than $\min_{a \in A} \mathrm{SW}(a)$. However, these payoff profiles are useless for token SPIs because they'd necessarily leave at least one player worse off than in the original game.

For example, consider the game in Table 5a, Asymmetric Chicken. The only asymmetry is that Player 1's utility for (Dare, Swerve) is higher than Player 2's for (Swerve, Dare). (Dare, Swerve) is the unique social welfare maximizing outcome.

The token game Table 5b is a payments (isomorphism) token game SPI on Asymmetric Chicken. The utility correspondence is given by $\hat{\Psi}_1(v_1) = 0.8v_1 + 1.2$ and $\hat{\Psi}_2(v_2) = v_2$, so the SPI is only strict for Player 1. It's easy to verify that each token payoff has social welfare of at most 7, the maximum social welfare in Asymmetric Chicken, and is hence feasible. Note also that this token SPI is maximal: Since the social welfare of (Token, Swerve, Token Dare) is 7, it's infeasible to improve $\hat{\Psi}$ any further.

It's worth noting that Asymmetric Chicken admits a payments token game SPI even though the reduction function doesn't eliminate any actions and there is no simple SPI. Such games do not admit SPIs under most previously considered forms of commitment, including the token game SPIs of [27], the MB token games SPIs of Section 3, and the utility function SPIs of [22].

We now turn to the question of characterizing and computing payment token SPIs. As in the previous section, there's a simple characterization of *simple* token game SPIs: they exist essentially whenever a payoff profile can simultaneously give each player their maximum payoff in $\bar{G}$. The only difficulty is ensuring strictness.

LEMMA 1. *A game G admits a simple* Red *payments token SPI if and only if either its reduced game has multiple distinct payoff profiles and $\sum_i \max_{a \in \bar{A}} u_i(a) \leq \mathrm{SW}^{max}(G)$ or if its reduced game has exactly one distinct payoff profile and $\sum_i \max_{a \in \bar{A}} u_i(a) < \mathrm{SW}^{max}(G)$.*

Therefore, we'll focus on isomorphism SPIs for the rest of the section. As in Section 3, we associate each isomorphism SPI with its utility isomorphism $\hat{\Psi}$, which captures the effect the isomorphism has on the players' base game payoffs. By the definition of isomorphism, $\hat{\Psi}$ is required to be playerwise positive affine, and is required to be Pareto improving on $\mathbf{u}(\bar{A})$ (weakly everywhere and strictly on at least one point) by the definition of SPI. Finally, $\hat{\Psi}$ can be achieves by a payments token game if and only if $\mathrm{SW}(\hat{\Psi}(v)) \leq \mathrm{SW}^{max}(G)$ for all $v \in \mathbf{u}(\bar{A})$. We call such a utility isomorphism valid, and reason about payment token game SPIs by reasoning about the space of valid $\hat{\Psi}$.

THEOREM 3. *A game G admits a payments token SPI if and only if either there are no payoff profiles in $\boldsymbol{u}(\bar{A})$ which are social welfare maximizing in $\boldsymbol{u}(A)$ or all three of the following hold: (1) some Player i achieves the same payoff v_i^* in all social welfare maximizing payoff profiles in $\boldsymbol{u}(\bar{A})$, (2) this v_i^* is either Player i's maximum or minimum payoff in $\boldsymbol{u}(\bar{A})$, and (3) Player i achieves at least one other payoff in $\boldsymbol{u}(\bar{A})$.*

PROOF SKETCH. *Only if / Nonexistence:* The nonexistence results rely on the fact that any payoff profile v^* which is social welfare maximizing in A must be a fixed point of $\hat{\Psi}$ (in all dimensions), since if $\hat{\Psi}(v^*)$ is greater than v^* for one player, it would necessarily be less than v^* for another player and fail to be Pareto improving. We show that for any Player i who fails to meet any of the three conditions above, $\hat{\Psi}_i$ must be the identity. This is because If (1) fails, $\hat{\Psi}_i$ must be linear (positive affine) and have at least two distinct fixed points. If (2) fails, $\hat{\Psi}_i$ must be linear (positive affine) with a fixed point at v_i^* and while having $\hat{\Psi}_i(v_i) \geq v_i$ both above and below v_i^*. If (3) fails, then the v_i^* is a fixed point of $\hat{\Psi}_i^*$ and also the only input.

If / Existence: If no payoff profiles in $\mathbf{u}(\bar{A})$ are social welfare maximizing in $\mathbf{u}(A)$, then $\hat{\Psi}(v) = v + (\varepsilon, \varepsilon, \ldots, \varepsilon)$ is feasible for some sufficiently small ε. In the other case, we construct a valid $\hat{\Psi}$ which is the identity for all players but Player i, where $\hat{\Psi}_i$ has a fixed point at v_i^* and has $\hat{\Psi}_i(v_i) > v_i$ for all other $v_i \in \mathbf{u}_i(\bar{A})$. In the case where v_i^* is Player i's maximum payoff in $\mathbf{u}(\bar{A})$ we achieve this by $\hat{\Psi}_i(v_i) = v_i + \varepsilon(v_i^* - v_i)$ for some small $\varepsilon > 0$,and when it's Player i's minimum payoff, by $\hat{\Psi}_i(v_i) = v_i + \varepsilon(v_i - v_i^*)$. Geometrically, these can be thought of as mapping each v_i either an ε-fraction of the way towards a maximum or an ε-fraction of the way away from a minimum. □

As with the MB token SPIs, we can optimize min-linear objectives over payment token SPIs via linear programming.

THEOREM 4. *It can be decided in polynomial time via linear programming whether a game G admits a payment token game isomorphism SPI. Furthermore, min-linear objectives over such SPIs can be optimized efficiently.*

	Dare	Swerve
Dare	-10, -10	5, 1
Swerve	1, 5	3, 3

Table 6: Symmetric Chicken

5 TOKEN GAMES WITH OUTSIDE BETTORS

One can view the payoffs in the payment token games from the previous section as each realized via a strategy profile $\Psi(t)$ and a net payment vector $p(t)$ which sums to 0. Because the token payoffs $\mathbf{u}(t)$ are not all social welfare maximizing, these strategy profiles $\Psi(t)$ are also not always social welfare maximizing. The players could instead realize the token payoffs by a commitments to play some social welfare maximizing profile, make some payments to each other to realize the token payoffs, and then "give away" the excess utility by making payments to parties outside the game. However, there's an apparent problem: the players cannot give these payments to anything they value without violating the isomorphism constraint, for the same reason they cannot keep it for themselves.

In this section, we develop a way for the players to benefit from this excess utility by making contracts with outside parties to make payments depending on the outcome of the token game. We extend the transferrable utility assumption from the previous section: Utility can now be transferred between the players in the game and some outside party. We imagine that before playing the game, the players act as follows. First, they create a security or financial instrument. They then construct a token game $\mathcal{T}$ with the exact same payoffs as the original game G. Each token payoff $\mathbf{u}(t)$ is realized by playing some social-welfare-maximizing outcome a^* in G, paying the holder of the security $\text{SW}(a^*) - \text{SW}(t)$, and making payments among the players so that each receives $u_i(t) - u_i(a^*)$. The players then sell the security to an outside party before playing the game and split the proceeds among themselves (equally, say). Finally, they play the token game as usual.

It's important to note that the players' gain from selling the security is independent of the outcome of the token game, and so therefore the incentives are exactly as given by the token payoffs. Suppose the security sells for x utils, which the players split equally. One can view the overall result of the protocol as the players getting x/n utility for free and then playing a token game with identical payoffs to the original game, or equivalently as constructing a token game with utility correspondence $v \mapsto v + (x/n, ..., x/n)$.

Here's a simple example. Consider Symmetric Chicken, the game in Table 6. (It's identical to the Asymmetric Chicken of Table 5a, except that the payoffs for (Dare, Swerve) are $(5, 1)$ rather than $(6, 1)$.) In Symmetric Chicken, the social welfare of all the outcomes aside from (Dare, Dare) is 6, and hence there is no payments token game SPI. However, there is a token game SPI with outside bettors (under very mild assumptions). In this case, the security is worth $6 - (-20) = 26$ if (Dare, Dare) occurs, and 0 otherwise. Suppose the outside bettor believes there's a nonzero probability that (Dare, Dare) is played and is hence willing to buy the security for a strictly positive price. For example, if the outside bettor expects the players to play the mixed Nash equilibrium, where each player plays Dare with probability $2/13$, the security is worth $26 * (2/13)^2 \approx 0.615$. In any such case, the players make a profit from selling the security and achieve a strict SPI.

It's important to note that the overall scheme is not a safe Pareto improvement from the perspective of the outside bettor: it's not guaranteed to leave them better off. In particular, if the outcome of the token game has social welfare equal to the maximum social welfare in G, the bettor doesn't receive payments from the players and, having purchased the security for a strictly positive amount, is strictly worse off. Instead, the agreement is a more typical, non-SPI agreement for the bettor, where they make an agreement which (they believe) has positive value in expectation. For this reason, the scheme requires the outside party to be willing to assign probabilities to outcomes of the token game, at least to some extent.

We don't see this as an issue. The outside parties don't face the types of problematic incentives that the players themselves do, since they don't inherently have any stake in or control over the outcome of the game. And the outside party's role is not unusual; it's analogous to offering an insurance contract to the players. (Interestingly, the outside bettors' contract with the players is the reverse of a typical insurance contract: they pay the players a constant amount in exchange for being paid a large amount if very socially suboptimal outcomes occur in the token game.)

We now give our theoretical results about token game SPIs with outside bettors. We first show that SPIs exist under extremely mild conditions on the ways that the outside bettors form beliefs about the outcome of the token game and bid on the security given their beliefs.

THEOREM 5. *Suppose an outside bettor assigns nonzero probability to some token outcome t with $\text{SW}(t) < \text{SW}^{max}(G)$ and is willing to pay a strictly positive amount for securities with a positive expected value under their beliefs. Then there's an outside bettors token game SPI on G.*

THEOREM 6. *Suppose for any game, there is an outside bettor who assigns strictly positive probability to all outcomes which are possible under the assumptions and pays a strictly positive amount for a security if and only if it has positive expected value under their beliefs. Then a token game SPI with outside bettors fails to exist if and only if $\bar{G}$ is constant sum and all outcomes in $A \setminus \bar{A}$ have weakly lower social welfare than those in $\bar{A}$.*

The results above show that token SPIs with outside bettors exist under very mild assumptions on the game and the outside bettors. The intuition is that, even if the bettors are very risk averse and uncertain, it's reasonable that they would be willing to purchase the security for a nonzero amount. However, they don't show anything about the magnitude of the SPIs.

We can give substantially stronger results under stronger assumptions. In particular, consider the very natural setting where the bettors form a probability distribution over outcomes of the token game and are willing to pay the expected value of the security under this distribution. This might happen, for example, if there are multiple bettors with identical beliefs bidding for the security in a second price auction, and therefore the winner must pay the full expected value of the security. In this case, the scheme allows the players to achieve expected social welfare equal to their maximum social welfare in the original game, where the expectation is taken with respect to the outside bettors' belief distribution.

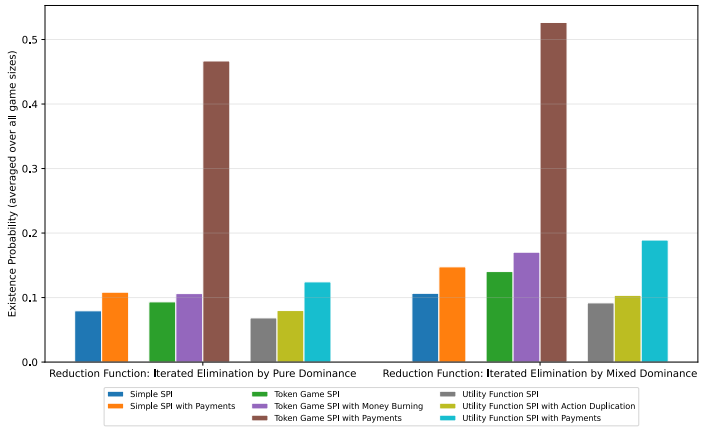

Figure 2: Empirical existence probability of different types of SPIs under a uniform distribution over payoffs.

THEOREM 7. *Consider a game G, and suppose an outside bettor with probability distribution D over outcomes of G is willing to purchase the security for its expected value under D. Then there is a token game SPI with outside bettors on G in which the players' expected social welfare (under this D) is equal to their maximum SW in the base game.*

It's interesting to note that there's an alternate version of the scheme which works even if no outside parties willing to assign probabilities to outcomes, if instead there are multiple sets of players playing disjoint games. In this case, the players could jointly agree to "swap" their securities without any other compensation, i.e. players from game 1 commit to transfer the surplus from their game to the players in game 2, and vice versa. This results in an SPI for all players, without requiring any probabilistic beliefs about the outcomes of the games.

6 EXPERIMENTAL RESULTS

The SPI framework sometimes allows us to prove that we can improve a strategic interaction by some intervention, e.g., transforming it into a specific token game. We think that token games and in particular token games with the additional features studied here (commitments to burning and transferring money, transfers with an outside party) are particularly suitable (compared to other kinds of interventions) for achieving SPIs on many strategic interactions. In the case of token games with outside bettors we were able to support this claim with strong theoretical results, see Section 5. (Compare similar results for the utility function token game SPIs of Oesterheld and Conitzer [22, Sect. 5]). But for many modes of intervention (regular token games, token games with money burning, token games with transfers, etc.), there is likely no simple characterization of how often SPIs under the given intervention exists.

In this section, we therefore compare the applicability of different modes of intervention by conducting computational experiments. Roughly, we generate a large number of games. For each, we compute SPIs under various interventions. We record how often an SPI exists, as well as how large a social welfare gain they afford.

We describe the experiment and its result in detail below. The code for our experiments can be found in the supplementary material.

6.1 Experimental setup

Generating random games: We generate 1000 random games of size m by m, for each $m \in \{3, \dots, 9\}$. Each entry of the payoff matrix is sampled i.i.d., we consider drawing from both the uniform distribution on the interval $[0, 1]$ or a standard normal distribution.

Reduction assumption: We consider two different reduction functions Red: iterated elimination of strictly dominated strategies in *pure* and in *mixed* strategies.

Modes of intervention: We compute SPIs with respect to the following interventions: Token games as per Sauerberg and Oesterheld [27], token games with money burning (Section 3), token games with payments (Section 4) and utility function SPIs as per Oesterheld and Conitzer [22]. (We do not compute SPIs via token games with outside bettors, because they exist with probability 1 in these games.) As a point of comparison, we also consider simple SPIs implemented by a single correlated strategy, as well as simple SPIs implemented by a single correlated strategy and payments. In each case, we optimize over SPIs. As the objective value we use the expected social welfare improvement under the uniform distribution over outcomes in the fully reduced default game.

Computational details: For all the different types of token SPIs we implement the linear programs of Sections 3 and 4 and Sauerberg and Oesterheld [27, Section 4]. For iterated elimination of strictly dominated strategies in mixed strategies we also use linear programs [6]. For linear programming, we use the CVXPY modeling language [9] and the CLARABEL LP solver [13] (chosen by default by CVXPY). Computing utility function SPIs is NP-complete [22]. However, using a classic depth-first search algorithm we are able to find (and optimize over) utility function SPIs in acceptable time for the given game sizes.

Measured quantities: For every game and every type of SPI we record whether an SPI exists at all. We also record the objective value of the optimal SPI – i.e., the expected social welfare gain under the uniform distribution over the reduced default game.

6.2 Results

Figure 2 shows what fraction of the time SPIs exist in games with payoffs sampled uniformly from the unit interval under different modes of intervention and different reduction functions. We find qualitatively similar results across using both uniform and normal distributions, and when looking at both existence probability and expected utility gain, see Figures 3 to 5 in the appendix.

The interventions we introduce in this paper, token games with money burning and token games with payments, are substantially more effective than both standard token game [27], and utility function SPIs [22], and the baseline of simple SPIs. We also find that the token SPI framework is generally more effective than the utility function SPI framework.

Our results show that of the modes of interventions tested in our experiment, the combination of token games and payments is especially potent. In particular, we find a superadditive effect: token games with payments are more effective than naively summing over the effectiveness of token games (without payments) and any one

payments-based intervention (simple SPIs with payments or utility function SPIs with payments) would suggest.

We also find that making a stronger reduction assumption (iterated elimination of strictly dominated strategies in mixed rather than pure strategies) tends to allow for more SPIs to exist. While this result is intuitive, it's not immediately implied by the fact that mixed dominance is stronger. This is because SPIs must be strictly Pareto improving for at least one outcome in the reduced game. For instance, suppose there's a game which reduces to a single outcome under mixed dominance, and this outcome is Pareto optimal in the full game. Such a game admits no SPI under mixed dominance, but if pure dominance is unable to reduce it to the single outcome, then it admits a (simple) SPI under pure dominance.

REFERENCES

[1] Ashton Anderson, Yoav Shoham, and Alon Altman. 2010. Internal Implementation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1* (Toronto, Canada) *(AAMAS '10)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 191–198.

[2] Krzysztof R. Apt. 2004. Uniform Proofs of Order Independence for Various Strategy Elimination Procedures. *The B.E. Journal of Theoretical Economics* 4, 1 (2004), 1–48. https://doi.org/10.2202/1534-5971.1141

[3] Ruggiero Cavallo. 2006. Optimal decision-making with minimal waste: strategyproof redistribution of VCG payments. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems* (Hakodate, Japan) *(AAMAS '06)*. Association for Computing Machinery, New York, NY, USA, 882–889. https://doi.org/10.1145/1160633.1160790

[4] Phillip JK Christoffersen, Andreas A Haupt, and Dylan Hadfield-Menell. 2023. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 448–456.

[5] Edward H. Clarke. 1971. Multipart Pricing of Public Goods. *Public Choice* 11, 1 (1971), 17–33. https://doi.org/10.1007/bf01726210

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

[7] Yuan Deng and Vincent Conitzer. 2017. Disarmament Games. *Proceedings of the AAAI Conference on Artificial Intelligence* 31, 1 (Feb. 2017). https://doi.org/10.1609/aaai.v31i1.10573

[8] Yuan Deng, Pingzhong Tang, and Shuran Zheng. 2016. Complexity and Algorithms of K-Implementation. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems* (Singapore, Singapore) *(AAMAS '16)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 5–13.

[9] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[10] Anthony DiGiovanni, Jesse Clifton, and Nicolas Macé. 2025. Safe Pareto Improvements for Expected Utility Maximizers in Program Games. In *Proceedings of the 24rd International Conference on Autonomous Agents and Multiagent Systems*.

[11] Boi Faltings. 2004. A budget-balanced, incentive-compatible scheme for social choice. In *International Workshop on Agent-Mediated Electronic Commerce*. Springer, 30–43.

[12] Itzhak Gilboa, Ehud Kalai, and Eitan Zemel. 1990. On the order of eliminating dominated strategies. *Operations Research Letters* 9, 2 (3 1990), 85–89. https://doi.org/10.1016/0167-6377(90)90046-8

[13] Paul J. Goulart and Yuwen Chen. 2024. Clarabel: An interior-point solver for conic programs with quadratic objectives. arXiv:2405.12762 [math.OC]

[14] Theodore Groves. 1973. Incentives in Teams. *Econometrica* 41, 4 (1973), 617–631. https://doi.org/10.2307/1914085

[15] Adam Kalai and Ehud Kalai. 2013. Cooperation in strategic games revisited. *The Quarterly Journal of Economics* 128, 2 (2013), 917–966.

[16] Adam Tauman Kalai, Ehud Kalai, Ehud Lehrer, and Dov Samet. 2010. A commitment folk theorem. *Games and Economic Behavior* 69, 1 (2010), 127–137.

[17] Elon Kohlberg and Jean-Francois Mertens. 1986. On the Strategic Stability of Equilibria. *Econometrica* 54, 5 (1986), 1003–1037. https://doi.org/10.2307/1912320

[18] Vojtech Kovarik, Caspar Oesterheld, and Vincent Conitzer. 2024. Recursive Joint Simulation in Games. *arXiv preprint arXiv:2402.08128* (2024).

[19] Dov Monderer and Moshe Tennenholtz. 2004. k-Implementation. *Journal of Artificial Intelligence Research* 21 (2004), 37–62.

[20] Hervé Moulin. 1976. Cooperation in mixed equilibrium. *Mathematics of Operations Research* 1, 3 (1976), 273–286.

[21] Hervé Moulin. 2009. Almost budget-balanced VCG mechanisms to assign multiple objects. *Journal of Economic theory* 144, 1 (2009), 96–119.

[22] Caspar Oesterheld and Vincent Conitzer. 2022. Safe Pareto improvements for delegated game playing. *Autonomous Agents and Multi-Agent Systems* 36, 2 (2022), 46.

[23] Caspar Oesterheld and Vincent Conitzer. 2025. Choosing what game to play with no regrets or controversies — inferring safe (Pareto) improvements in binary constraint structures. In *Proceedings of the Twentieth Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2025)*, Adam Bjorndahl (Ed.). Düsseldorf, Germany, 246–265. https://cgi.cse.unsw.edu.au/~eptcs/paper.cgi?TARK2025:36

[24] Caspar Oesterheld, Johannes Treutlein, Roger B Grosse, Vincent Conitzer, and Jakob Foerster. 2023. Similarity-based cooperative equilibrium. *Advances in Neural Information Processing Systems* 36 (2023), 24434–24465.

[25] David G. Pearce. 1984. Rationalizable Strategic Behavior and the Problem of Perfection. *Econometrica* 54, 4 (1984), 1029–1050.

[26] Nathaniel Sauerberg and Caspar Oesterheld. 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.

[27] Nathaniel Sauerberg and Caspar Oesterheld. 2026. Promises Made, Promises Kept: Safe Pareto Improvements via Ex Post Verifiable Commitments. In *Proceedings of the AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/2505.00783

[28] Eric Sodomka, Elizabeth Hilliard, Michael Littman, and Amy Greenwald. 2013. Coco-Q: Learning in Stochastic Games with Side Payments. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 1471–1479. https://proceedings.mlr.press/v28/sodomka13.html

[29] Filipe Souza and Leandro Rêgo. 2012. Mixed equilibrium: when burning money is rational. (2012).

[30] Moshe Tennenholtz. 2004. Program Equilibrium. 49 (2004), 363–373.

[31] William Vickrey. 1961. Counterspeculation, Auctions, and Competitive Sealed Tenders. *The Journal of Finance* 16, 1 (1961), 8–37. https://doi.org/10.1111/j.1540-6261.1961.tb02789.x

A ADDRESSING THE POSSIBLE MULTIPLICITY OF ISOMORPHISMS

LEMMA 2. *Let $G = (A, \mathbf{u})$ and $G' = (A', \mathbf{u}')$ be two normal-form games. Let ϕ and $\hat{\phi}$ be two isomorphisms from G to G'. Then for all a in A we have that $\mathbf{u}'(\phi(a)) = \mathbf{u}'(\hat{\phi}(a))$.*

PROOF SKETCH. Both isomorphisms must map each player's highest/lowest-utility outcomes in G to that player's highest/lowest-utility outcomes in G'. This uniquely determines the coefficients m_i and b_i associated with the isomorphism. Thus all isomorphisms are associated with the same positive affine transformation on payoffs. Since every outcome a has a unique payoff profile, any outcome a must be mapped by all isomorphisms onto outcomes with the same utility in G'. □

The following result follows immediately.

PROPOSITION 1. *Let G and G' be two normal-form games. Let ϕ and $\hat{\phi}$ be isomorphisms from G to G'. If ϕ is Pareto-improving, so is $\hat{\phi}$.*

This proposition justifies a subtlety of the definition of isomorphism SPIs (Definition 2.2). The definition only requires the existence of a single Pareto-improving isomorphism. Without Proposition 1, one might wonder what happens if there are multiple isomorphisms, some of which are and some of which aren't Pareto-improving. In such a case, Definition 2.2 would be questionable. Fortunately, Proposition 1 shows that such cases cannot occur.

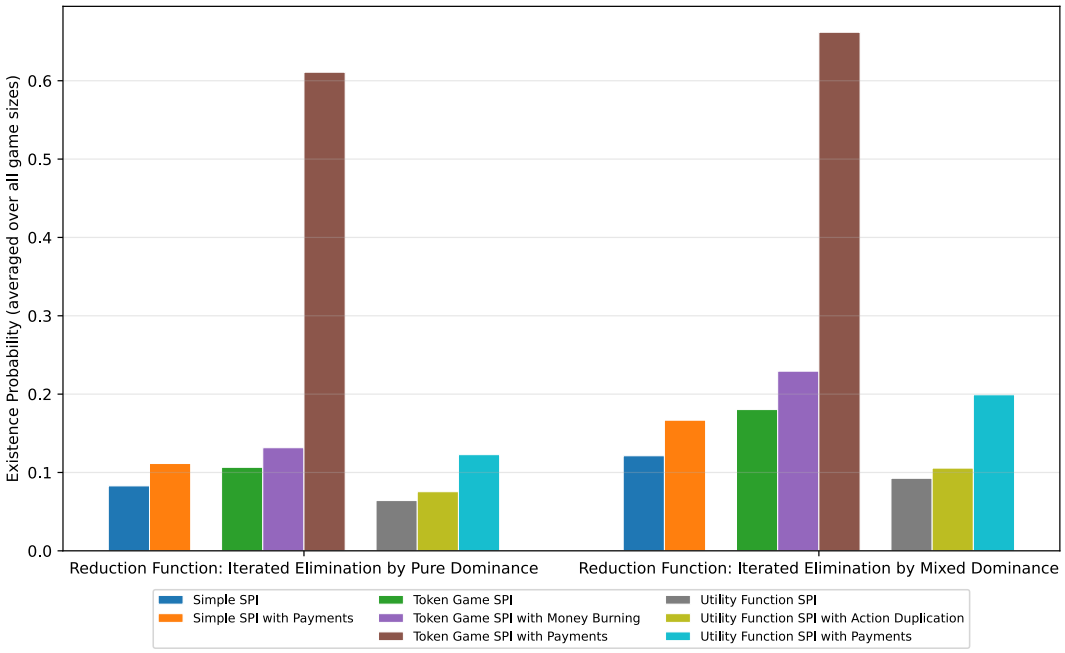

Figure 3: Empirical existence probability of SPIs under a normal distribution over payoffs

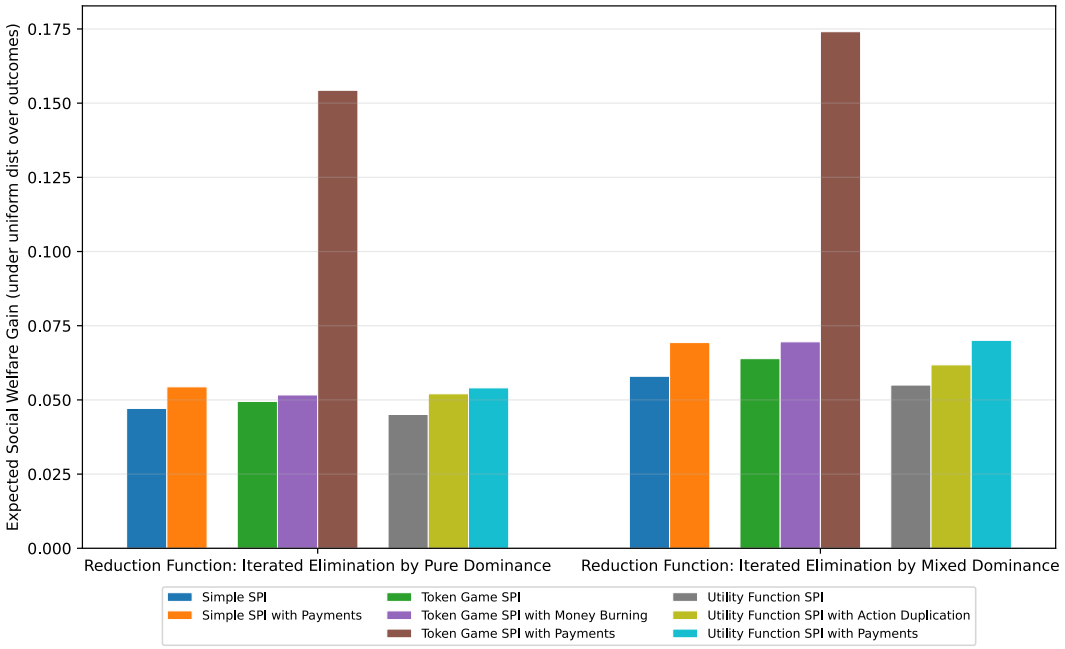

Figure 4: Empirical average social welfare gain of SPIs under a uniform distribution over payoffs

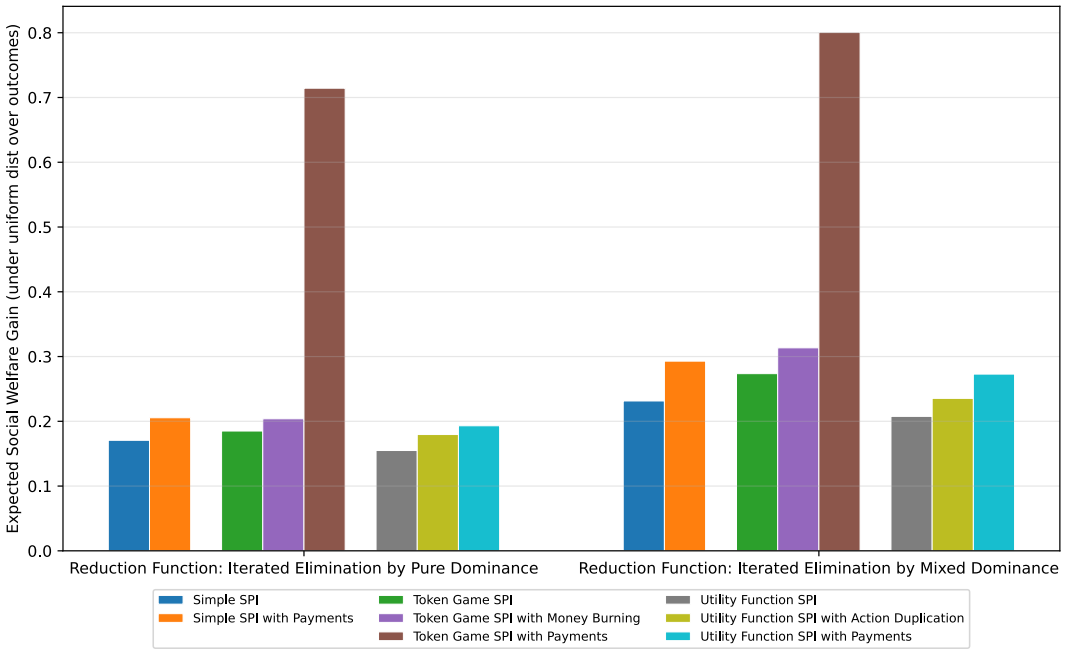

Figure 5: Empirical average social welfare gain of SPIs under a normal distribution over payoffs

A PROOFS FOR SECTION 3 (MONEY BURNING)

THEOREM 1. *A two-player game G admits a MB-token game isomorphism SPI if and only if either (a) $\bar{A}$ contains no outcomes which are Pareto optimal in $\boldsymbol{u}(A)$ or if (b) it contains exactly one payoff v^* which is Pareto optimal in $\boldsymbol{u}(A)$, this v^* is at least one player's maximum payoff in $\bar{A}$, and that player has at least one more distinct payoff in $\bar{G}$.*

PROOF. Let $V = \mathbf{u}(\bar{A})$ be the set of payoffs in the reduced game $\bar{G}$. Define $v_i^{max(G)} = \max_{v \in \mathbf{u}(A)} v_i$ to be each player's maximum payoff in G, and likewise $v_i^{max(V)} = \max_{v \in V} v_i$ to be each player's maximum payoff in the reduced game $\bar{G}$. We must decide whether there exists a valid $\hat{\Psi}$: one which is player-wise positive affine and from v into the feasible region $\mathcal{F}^{\text{MB}}(G)$.

A valid (positive affine) $\hat{\Psi}_i$ is determined by the choice of $\hat{\Psi}_i(v_i)$ for two inputs v_i: there is only one line through those two points. Much of the proof operates by reasoning about fixed points of $\hat{\Psi}$. For instance, if v is Pareto optimal in $\mathcal{F}^{\text{MB}}(G)$, it must be a fixed point of all $\hat{\Psi}_i$: it cannot be increased in any dimension without being decreased in another dimension, which would make $\hat{\Psi}$ fail to be everywhere improving for that player. In addition, a valid $\hat{\Psi}_i$ with $m_i \neq 1$ has exactly one fixed point, say $\check{v}_i$. For an $\hat{\Psi}_i$ which doesn't decrease any v_i in V, this must either be mapping towards a high value, e.g. $v_i \mapsto v_i + \varepsilon(\check{v}_i - v_i)$, or away from a low value, e.g. $v_i \mapsto v_i + \varepsilon(v_i - \check{v}_i)$.

If: Consider the first case (a). Suppose V contains no points which are Pareto optimal in $\mathcal{F}^{\text{MB}}(G)$. (Note that Pareto optimality in $\mathbf{u}(A)$ and $\mathcal{F}^{\text{MB}}(G)$ are identical.) Then for all $v \in \mathbf{u}(\bar{A})$) for which $v_i \neq v_i^{max(G)}$ for either i, the feasible region contains an ε_v ball around v. And any v with $v_1 = v_1^{max(G)}$ must have $v_2 \neq v_2^{max}$ (or else v would be Pareto optimal), and $v + (0, \varepsilon_v)$ is feasible for some sufficiently small ε_v. Of course, the case for $v_2 = v_2^{max(G)}$ is symmetric. Therefore, the $\hat{\Psi}(v) = v + \varepsilon(v^{max} - v)$ is feasible for some sufficiently small ε. This is clearly strictly Pareto improving and entrywise positive affine, as desired.

Consider the second case (b). Suppose there is a single outcome $v^* \in \bar{A}$ which is Pareto optimal in $\mathcal{F}^{\text{MB}}(G)$, and that (without loss of generality) v_1^* is Player 1's maximum payoff in $\bar{G}$. We claim that the $\hat{\Psi}$ defined by $\hat{\Psi}_2 = id$ and $\hat{\Psi}_1(v_1) = (1-\varepsilon)v_1 + \varepsilon v_1^*$ is feasible. Note that this would be strictly Pareto improving since Player 1 has multiple distinct payoffs in $\bar{G}$. We just need to show that, for all other $v' \neq v$, $v' + (\varepsilon, 0)$ is feasible for some sufficiently small ε (which can depend on v'). Because v' is not Pareto optimal in G, there is some $v'' \succ v'$ and strictly so in at least one dimension. If $v_1'' > v_1'$, we're immediately done. Consider the case where all v'' $\mathbf{u}(A)$ which Pareto dominate v' have $v_1' = v_1''$. Then in particular, the v'' which is on the Pareto frontier of G must be the point on the Pareto frontier with maximum value in the first dimension, i.e. $v_1'' = v_1^{max(G)} = v_1'$. In this case, the proposed $\hat{\Psi}$ has $\hat{\Psi}(v') = v'$, which is feasible for any ε. Hence, the proposed $\hat{\Psi}$ is feasible for some $\varepsilon > 0$ and there's an SPI, as desired.

Only If: Observe that any v^* which is Pareto optimal in $\mathcal{F}^{\text{MB}}(G)$ must be a fixed point of $\hat{\Psi}$ for both players: it cannot be increased in one dimension without being decreased in the other, which would make $\hat{\Psi}$ fail to be everywhere improving for that second player. Therefore, if v^* is intermediate for a player i, $\hat{\Psi}_i$ must be the identity, as it's the only positive affine function which is (weakly) improving for both $v_i < v_i^*$ and $v_i > v_i^*$.

First, we consider the case where there's exactly one payoff v^* in V which is Pareto optimal in G and where $v_i \neq v_i^{max(V)}$ for either player.

We proceed by cases. We know that v^* is either minimal or intermediate in V for both players. First, if v^* is intermediate for both players, we know $\hat{\Psi}$ must be the identity for both players by the above logic, so no SPI exists. If v^* is minimal in V for both players while being Pareto optimal, it would be the only payoff in V, and therefore again no SPI exists.

The only remaining case is where v^* is minimal for (without loss of generality) P1 and intermediate for P2. Since it's intermediate for P2 (in V), there must be some some $v' \in \bar{A}$ with $v_2' > v_2^*$. Since v^* is Pareto optimal, this implies that $v_1' < v_2'$. But this contradicts that v^* is minimal for P1. Hence, this case cannot occur and we no SPIs exist, as desired.

Finally, we consider the case where there are multiple distinct payoffs in V which are Pareto optimal in $\mathcal{F}^{\text{MB}}(G)$. In this case, each one must be a fixed point of each $\hat{\Psi}_i$. This immediately implies that each $\hat{\Psi}_i$ must be the identity and no SPI exists, as desired. □

THEOREM 2. *It can be decided in polynomial time via linear programming whether a game G admits a MB-token game isomorphism SPI. Furthermore, min-linear objectives over such SPIs can be optimized efficiently.*

PROOF. We construct a polynomially sized linear program which essentially directly encodes the problem of finding or optimizing over valid $\hat{\Psi}$. Our LP is adapted from and extremely similar to that of [27] for the case of token SPIs without money burning. We simply modify the feasibility constraints to allow for money burning.

Our LP has several types of variables: For each player i, we have variables m_i and b_i which correspond to the parameters of the utility isomorphism $\hat{\Psi}_i$. Our LP takes as input the payoff vectors in the reduced game $\bar{G}$, which we denote $v^1, \ldots, v^k$. For each v^j, we have a set of variables $p_a^{v^j}$ for all $a \in A$ which collectively represent the strategy profile in G, i.e. distribution over outcomes in A, which proves the feasibility of a token payoff $\hat{\Psi}(v^j)$. The utility function $\mathbf{u} : A \to \mathbb{R}$ is also an input.

$$\text{Maximize} \quad \sum_{v^j \in V} \sum_{i \in [n]} (m_i v_i^j + b_i - v_i^j)$$

Subject to:

$$m_i \geq 0 \qquad \text{for all } i \in [n]$$

$$m_i v_i^j + b_i \geq v_i^j \qquad \text{for all } i \in [n], j \in [k]$$

$$p_a^{v^j} \geq 0 \qquad \text{for all } a \in A, j \in [k]$$

$$\sum_{a \in A} p_a^{v^j} = 1 \qquad \text{for all } j \in [k]$$

$$\sum_{a \in A} p_a^{v^j} u_i(a) \geq m_i v_i^j + b_i \qquad \text{for all } i \in [n], j \in [k]$$

The first constraint ensures that the $\hat{\Psi}$ is (weakly) positive affine. The second constraint ensures that $\hat{\Psi}$ is (weakly) Pareto improving on all v^j, which represent payoff profiles in $\bar{G}$. The next two constraints collectively ensure that for each v^j, the set $\{p_a^{v^j}\}$ corresponds to valid probability distribution over outcomes in G, i.e. $\Psi(\cdot)$. And the final constraint ensures the feasibility of $\hat{\Psi}(v^j)$ for each j: it requires that for each v^j and each Player i, Player i's expected utility for the distribution $\{p_a^{v^j}\}$ is at least that specified by $\hat{\Psi}$ for the token outcome corresponding to v^j, and therefore the exact payoff specified by $\hat{\Psi}$ can be realized by playing the distribution and committing to burn the difference between the utilities.

It's easy to verify that the program is linear (recall that the sets A and vectors v^j and $u_i(a)$ are inputs to the program) and polynomially sized: it has $O(|A|^2)$ variables and $O(|A|^2)$ constraints. Therefore, the objective of the LP can be optimized in polynomial time.

Now, we claim that a game G admits a MB-token game isomorphism SPI if and only if the optimal value of the LP is strictly greater than 0. The objective value of the LP is exactly the sum over players and payoffs of $\hat{\Psi}_i(v_i^j) - v_i^j$, i.e. the social welfare gain on payoff vector v^j from the SPI. Since each of these is non-negative, the objective is strictly positive if and only if the $\hat{\Psi}$ is strictly Pareto improving.

The only remaining subtlety is that our program requires only that $m_i \geq 0$, but valid utility correspondence functions must be strictly positive affine. This can be dealt with similarly to as in [27].

First, observe that the space of valid $\hat{\Psi}$ is convex. For any pair of valid $\hat{\Psi}$, say $\hat{\Psi}^1(v) = m^1 v + b^2$ and $\hat{\Psi}^2(v) = m^2 v + b^2$, $\alpha\hat{\Psi}^1 + (1-\alpha)\hat{\Psi}^2$ is simply $(\alpha m^1 + (1-\alpha)m^2)v + \alpha b^1 + (1-\alpha)b^2$. In particular, the convex combination still has each $m_i > 0$ and is positive affine. The convex combination is also still feasible in $\mathcal{F}^{\text{MB}}(G)$ because the underlying space $\mathcal{F}^{\text{MB}}$ itself is convex.

Also, note that the identity utility correspondence is always feasible (though it is not a strict SPI). Therefore, if the LP finds a solution corresponding to a $\hat{\Psi}$ with some $m_i = 0$, we can construct a valid $\hat{\Psi}'(v)$ by $\hat{\Psi}'(v) = (1-\varepsilon)\hat{\Psi}(v) + \varepsilon v$. This $\hat{\Psi}'$ has all $m_i \geq \varepsilon$, and is still strictly Pareto improving if $\hat{\Psi}$ is. Hence, we've successfully reduced deciding the existence of a MB-token game isomorphism SPI to checking the optimal objective value of the LP, which can be done in polynomial time, as desired.

Optimization: Consider a game G and a min-linear objective defined by a set of linear functions $F = \{f^1, \ldots, f^{|F|}\}$, where each f^k is a linear function of variables of the form $\hat{\Psi}_i(\mathbf{u}(a))$ or $\hat{\Psi}_i(\mathbf{u}(a)) - u_i(a)$ for $a \in \bar{A}$ and $i \in [n]$. Let's write $w_i^k(v^j)$ to denote the sum of the coefficients on all $\hat{\Psi}_i(\mathbf{u}(a))$ in f^k for all a with $\mathbf{u}(a) = v^j$. (This is only necessary because our LP does not distinguish betwen $\hat{\Psi}_i(\mathbf{u}(a))$ for different outcomes a with the same payoff vector $\mathbf{u}(a)$. However, all such outcomes have the same image under the utility correspondence, so this is not an issue.) Similarly, let's write $\tilde{w}_i^k(v^j)$ to denote the sum of the coefficients on all $\hat{\Psi}_i(\mathbf{u}(a)) - u_i(a)$ in f^k for all a with $\mathbf{u}(a) = v^j$.

We construct an LP which is essentially the same as the one above, except that we replace the objective with a new variable α and add constraints that $\alpha \leq f^k$ for each $k \in [|F|]$.

$$\text{Maximize} \quad \alpha$$

Subject to:

$$m_i \geq 0 \qquad \text{for all } i \in [n]$$

$$m_i v_i^j + b_i \geq v_i^j \qquad \text{for all } i \in [n], j \in [k]$$

$$p_a^{v^j} \geq 0 \qquad \text{for all } a \in A, j \in [k]$$

$$\sum_{a \in A} p_a^{v^j} = 1 \qquad \text{for all } j \in [k]$$

$$\sum_{a \in A} p_a^{v^j} u_i(a) \geq m_i v_i^j + b_i \qquad \text{for all } i \in [n], j \in [k]$$

$$\alpha \leq \sum_{i \in [n]} \sum_{v^j \in V} w_i^k(v^j)(m_i v_i^j + b_i) + \tilde{w}_i^k(v^j)(m_i v_i^j + b_i - v_i^j) \qquad \text{for all } k \in [|F|]$$

Note that our LP is still linear: the $w_i^k(v^j)$ and $\tilde{w}_i^k(v^j)$ are inputs to the program, so the last set of constraints are linear in the variables m_i, b_i, and α. Our program is also still polynomially sized, size $|F|$ is polynomial. The subtleties regarding the positivity of m_i are dealt with in the same way as above. □

B PROOFS FOR SECTION 4 (TOKEN GAME SPIS WITH UTILITY PAYMENTS)

LEMMA 1. *A game G admits a simple* Red *payments token SPI if and only if either its reduced game has multiple distinct payoff profiles and $\sum_i \max_{a\in\bar{A}} u_i(a) \leq \mathrm{SW}^{max}(G)$ or if its reduced game has exactly one distinct payoff profile and $\sum_i \max_{a\in\bar{A}} u_i(a) < \mathrm{SW}^{max}(G)$.*

PROOF. *If:* First, in the case where $\bar{G}$ has exactly one distinct payoff profile, suppose $\sum_i \max_{a\in\bar{A}} u_i(a) < \sum_i \max_{a\in\bar{A}} u_i(a) + c = \mathrm{SW}^{max}(G)$ for some constant c. Then the token game which consists of a single outcome t and where $\mathbf{u}_i(t) = \max_{a\in\bar{A}} u_i(a) + c/n$ is a simple payments token game SPI on G.

Now, consider the case where $\bar{G}$ has multiple distinct payoff profiles, and $\sum_i \max_{a\in\bar{A}} u_i(a) = \mathrm{SW}^{max}(G)$. The token game which consists of a single outcome t and where $\mathbf{u}_i(t) = \max_{a\in\bar{A}} u_i(a)$ is a simple payments token game SPI on G. This outcome t is clearly weakly Pareto improving on all $a \in \bar{A}$. And the outcome is strictly Pareto improving on any $a \in \bar{A}$ in which at least one player fails to achieve their maximum utility in $\bar{A}$, which must exist because $\bar{A}$ contains at multiple distinct payoff profiles and hence at least one player who sometimes receives less than their maximum utility.

Only If: Suppose $\mathcal{T} = (T, \mathbf{u})$ is a simple payments token SPI on G. Then by definition, there is a token outcome t such that $\mathbf{u}_i(t) \geq u_i(a)$ for all $a \in \bar{A}$. and this inequality is strict for at least one outcome $a^* \in \bar{A}$ and player i.

The first condition implies that $\mathbf{u}_i(t) \geq \max_{a\in\bar{A}} u_i(a)$ for all i. Summing both sides of the inequality over all $i \in [n]$ and noting that $\mathrm{SW}^{max}(G) \geq \mathbf{u}(t)$ yields that $\mathrm{SW}^{max}(G) \geq \sum_i \max_{a\in\bar{A}} u_i(a)$, as desired. For the case where $\bar{G}$ has exactly one distinct payoff profile, the second condition says that $\mathbf{u}_i(t) > \max_{a\in\bar{A}} u_i(a)$ for some i, and therefore summing both sides of the inequality above over i yields a strict inequality. we conclude that $\mathrm{SW}^{max}(G) \geq \mathbf{u}(t) > \sum_i \max_{a\in\bar{A}} u_i(a)$, as desired. □

THEOREM 3. *A game G admits a payments token SPI if and only if either there are no payoff profiles in $\boldsymbol{u}(\bar{A})$ which are social welfare maximizing in $\boldsymbol{u}(A)$ or all three of the following hold: (1) some Player i achieves the same payoff v_i^* in all social welfare maximizing payoff profiles in $\boldsymbol{u}(\bar{A})$, (2) this v_i^* is either Player i's maximum or minimum payoff in $\boldsymbol{u}(\bar{A})$, and (3) Player i achieves at least one other payoff in $\boldsymbol{u}(\bar{A})$.*

PROOF. Let $V = \mathbf{u}(\bar{A})$ and let V^* be the subset of V which are social welfare maximizing in $\mathbf{u}(A)$.

If: First, consider the case where $|V^*| = 1$. Let $d_v = \mathrm{SW}^{max}(G) - \mathrm{SW}(v)$ for each $v \in V$, and let $d = \min_{v\in\mathbf{u}(A)} d_v$. Then $\hat{\Psi}(v) = v + (d/n, d/n, \ldots, d/n)$ is a valid utility isomorphism, and hence G admits a payments token SPI, as desired.

Now, consider the second case. Suppose $|V^*| \geq 1$, and that (without loss of generality) Player 1 has the same payoff v_1^* in all payoff profiles in V^*. For any v with $v_1 \neq v_1^*$, $\mathrm{SW}(v) < \mathrm{SW}^{max}(G)$. Let's say $\mathrm{SW}(v) = \mathrm{SW}^{max}(G) - d_v$. We claim that there's a valid $\hat{\Psi}$ where $\hat{\Psi}_i = id$ for all $i \neq 1$, v_1^* is a fixed point of $\hat{\Psi}_i$, and $\hat{\Psi}_1(v_1) > v_1$ for all other $v_1 \neq v_1^*$. In particular, if $v_1^* = \max_{v'\in V} v_1'$, then the $\hat{\Psi}_1(v_1)$ defined by $\hat{\Psi}_1(v_1) = v_1 + \varepsilon(v_1^* - v_1)$ is feasible where $\varepsilon = \min_v \frac{d_v}{v_1^* - v_1} > 0$ since this implies $\hat{\Psi}_1(v_1) \leq v_1 + d_v$ for all $v \in V$ and therefore $\mathrm{SW}(\hat{\Psi}(v)) \leq \mathrm{SW}(v) + d \leq \mathrm{SW}^{max}(G)$. Similarly, if $v_1^* = \min_{v'\in V} v_1'$, then the $\hat{\Psi}_1(v_1)$ defined by $\hat{\Psi}_1(v_1) = v_1 + \varepsilon(v_1 - v_1^*)$ is feasible for $\varepsilon = \min_v \frac{d_v}{v_1 - v_1^*} > 0$. In either case, we have a feasible, player-wise positive affine $\hat{\Psi}$ which is strictly Pareto improving because Player 1 has at least one payoff aside from v_1^* in $\mathbf{u}(A)$. Hence, we has a payemnts token SPI, as desired.

Only If: Suppose neither of the above conditions apply. That is, for all Players i, either (1) Player i achieves multiple distinct payoffs in V^*, (2) Player i achieves a payoff in V^* which is either is neither minimal nor maximal in $u_1(\bar{A})$, or (3) Player i has one payoff in V^* and this is their only payoff in $\mathbf{u}(\bar{A})$. In each case, we show that the only valid $\hat{\Psi}_i$ is the identity, and therefore G admits no strict payments SPI.

As usual, each point in v^* must be a fixed point of $\hat{\Psi}$ and hence $\hat{\Psi}_i$. In case (1), each of Player i's payoffs in V^* must be a fixed point of $\hat{\Psi}_i$ and the only positive affine $\hat{\Psi}_i$ with multiple fixed points is the identity. In case (2), v_i^* is a fixed point of $\hat{\Psi}_i$ and there are $v_i \in V$ both strictly greater than and less than v_i^*. The only positive affine $\hat{\Psi}_i$ with v_i^* as a fixed point and which doesn't have $\hat{\Psi}_i(v_i) < v_i$ on either side of v_i^* is the identity. Finally, In case (3), v_i^* must be a fixed point of $\hat{\Psi}_i$ and Player i has no other payoffs in $\mathbf{u}(\bar{A})$, so again $\hat{\Psi}_i$ is the identity on $\mathbf{u}(\bar{A})$. □

THEOREM 4. *It can be decided in polynomial time via linear programming whether a game G admits a payment token game isomorphism SPI. Furthermore, min-linear objectives over such SPIs can be optimized efficiently.*

PROOF. We construct a linear program which directly encodes the problem of finding or optimizing over valid $\hat{\Psi}$. Our LP is substantially simpler than that of Theorem 2 because feasibility of token payoffs under payments is determined entirely by a social welfare bound, so no auxiliary distribution variables are needed.

Its only variables are m_i and b_i corresponding to the parameters of each $\hat{\Psi}_i(v_i) = m_i v_i + b_i$ for each player i. The payoff profiles v^j, which we index $v^1, \ldots, v^k$, are inputs to the LP and constitute the set of payoff profiles in $\mathbf{u}(\bar{A})$.

$$
\begin{aligned}
&\text{Maximize} \quad \sum_{v^j \in V} \sum_{i \in [n]} (m_i v_i^j + b_i - v_i^j) \\
&\text{Subject to:} \\
&m_i \geq 0 && \text{for all } i \in [n] \\
&m_i v_i^j + b_i \geq v_i^j && \text{for all } i \in [n], j \in [k] \\
&\sum_{i \in [n]} (m_i v_i^j + b_i) \leq \mathrm{SW}^{max}(G) && \text{for all } j \in [k]
\end{aligned}
$$

It's easy to see that the program is linear; recall that the payoff profiles v^j are parameters of the problem / inputs to the program. The LP has $O(n)$ variables and $O(n \cdot k)$ constraints, each of which are polynomial in the instance size, so the problem can be optimized in polynomial time.

The LP essentially directly encodes the constraints and optimizes over $\hat{\Psi}$. The first set of constraints ensures that each $\hat{\Psi}_i$ is (weakly) positive affine. The second set of constraints ensures that $\hat{\Psi}$ is weakly Pareto improving on each v^j. The final set of constraints ensures that the total social welfare of the token payoff corresponding to each v^j does not exceed $\mathrm{SW}^{max}(G)$, which is necessary and sufficient for feasibility with payments. This suffices for feasibility because the constraint that the social welfare of each token payoff is at least the minimum social welfare in G is already implied by the constraint that $\hat{\Psi}$ be weakly Pareto improving on each v^j.

Now, we claim that a game G admits a payment token game isomorphism SPI if and only if the optimal value of the LP is strictly greater than 0. The objective value of the LP is exactly the sum over players and payoff profiles of $\hat{\Psi}_i(v_i^j) - v_i^j$, i.e. the total social welfare gain across all payoff profiles from the SPI. Since each term is non-negative (by the Pareto-improving constraints), the objective is strictly positive if and only if the $\hat{\Psi}$ is strictly Pareto improving.

The only remaining subtlety is that our program requires only that $m_i \geq 0$, but valid utility correspondence functions must be strictly positive affine. As in the proof of Theorem 2, the space of valid $\hat{\Psi}$ is convex, and the identity utility correspondence is always feasible (though it is not a strict SPI). Therefore, if the LP finds a solution corresponding to a $\hat{\Psi}$ with some $m_i = 0$, we can construct a valid $\hat{\Psi}'(v) = (1-\varepsilon)\hat{\Psi}(v) + \varepsilon v$ for small $\varepsilon > 0$. This $\hat{\Psi}'$ has all $m_i \geq \varepsilon$ and is still strictly Pareto improving if $\hat{\Psi}$ is. Moreover, the convex combination preserves feasibility because if $\sum_i \hat{\Psi}_i(v_i^j) \leq \mathrm{SW}^{max}(G)$ and $\sum_i v_i^j \leq \mathrm{SW}^{max}(G)$, then $\sum_i \hat{\Psi}'_i(v_i^j) \leq \mathrm{SW}^{max}(G)$. Hence, we've successfully reduced deciding the existence of a payment token game isomorphism SPI to checking the optimal objective value of the LP, which can be done in polynomial time, as desired.

Optimization: Consider a game G and a min-linear objective defined by a set of linear functions $F = \{f^1, \ldots, f^{|F|}\}$, where each f^ℓ is a linear function of variables of the form $\hat{\Psi}_i(v_i^j)$ or $\hat{\Psi}_i(v_i^j) - v_i^j$ for $v^j \in \mathbf{u}(\bar{A})$ and $i \in [n]$. Let $w_i^\ell(v^j)$ denote the coefficient on $\hat{\Psi}_i(v_i^j)$ in f^ℓ, and let $\tilde{w}_i^\ell(v^j)$ denote the coefficient on $\hat{\Psi}_i(v_i^j) - v_i^j$ in f^ℓ.

We construct an LP with the same feasibility constraints as above, but replace the objective with a new variable α and add constraints that $\alpha \leq f^\ell$ for each $\ell \in [|F|]$:

$$
\begin{aligned}
&\text{Maximize} \quad \alpha \\
&\text{Subject to:} \\
&m_i \geq 0 && \text{for all } i \in [n] \\
&m_i v_i^j + b_i \geq v_i^j && \text{for all } i \in [n], j \in [k] \\
&\sum_{i \in [n]} (m_i v_i^j + b_i) \leq \mathrm{SW}^{max}(G) && \text{for all } j \in [k] \\
&\alpha \leq \sum_{i \in [n]} \sum_{v^j \in V} w_i^\ell(v^j)(m_i v_i^j + b_i) \\
&\qquad\qquad + \tilde{w}_i^\ell(v^j)(m_i v_i^j + b_i - v_i^j) && \text{for all } \ell \in [|F|]
\end{aligned}
$$

Note that our LP is still linear: the $w_i^\ell(v^j)$ and $\tilde{w}_i^\ell(v^j)$ are inputs to the program, so the last set of constraints are linear in the variables m_i, b_i, and α. Our program is also still polynomially sized, since $|F|$ is polynomial. The subtleties regarding the positivity of m_i are dealt with in the same way as above. □

C PROOFS FOR SECTION 5 (TOKEN GAMES WITH OUTSIDE BETTORS)

THEOREM 5. *Suppose an outside bettor assigns nonzero probability to some token outcome t with $\mathrm{SW}(t) < \mathrm{SW}^{max}(G)$ and is willing to pay a strictly positive amount for securities with a positive expected value under their beliefs. Then there's an outside bettors token game SPI on G.*

PROOF. This follows immediately from the assumption that the players sell the security for strictly more than zero. □

THEOREM 6. *Suppose for any game, there is an outside bettor who assigns strictly positive probability to all outcomes which are possible under the assumptions and pays a strictly positive amount for a security if and only if it has positive expected value under their beliefs. Then a token game SPI with outside bettors fails to exist if and only if $\bar{G}$ is constant sum and all outcomes in $A \setminus \bar{A}$ have weakly lower social welfare than those in $\bar{A}$.*

PROOF. *Only If:* Suppose the conditions do not both hold. If $\bar{G}$ is not constant sum, some $a \in \bar{A}$ has $\mathrm{SW}(a) < \mathrm{SW}^{max}(G)$. If $\bar{G}$ is constant sum but some outcome in $A \setminus \bar{A}$ has strictly higher social welfare, then every $a \in \bar{A}$ satisfies $\mathrm{SW}(a) < \mathrm{SW}^{max}(G)$. In either case, an outcome with sub-maximal social welfare exists in the reduced game $\bar{A}$, and hence there is a token outcome t for which the security has positive value when t obtains. All outcomes in $\bar{A}$ are possible under our standard assumptions, since all survive iterated elimination of strictly dominated strategies. Therefore, the security has strictly positive expected value and sells for a strictly positive price, yielding an SPI.

If: Conversely, suppose $\bar{G}$ is constant sum and all outcomes in $A \setminus \bar{A}$ have weakly lower social welfare than those in $\bar{A}$. In this case, all token outcomes with positive probability under the standard assumptions achieve the maximum social welfare in G, and hence the security pays out 0 on all such outcomes. Therefore, the security has 0 expected value to a bettor who only assigns positive probability to outcomes possible under the assumptions, and no SPI exist. □

THEOREM 7. *Consider a game G, and suppose an outside bettor with probability distribution D over outcomes of G is willing to purchase the security for its expected value under D. Then there is a token game SPI with outside bettors on G in which the players' expected social welfare (under this D) is equal to their maximum SW in the base game.*

PROOF. The players' (total) expected payment to the outside bettor under D is equal to the amount they sell the security for. In other words, the expected net payments, again under D between the players' and bettor is 0. Therefore, the players' expected social welfare is equal to the (expected) social welfare of the strategy profiles used to realize the token outcomes, and by construction these strategy profiles are social welfare maximizing in the original game G. □