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

Nathaniel Sauerberg* and Caspar Oesterheld*
University of Texas and Carnegie Mellon University
nsauerberg@utexas.edu
*denotes equal contribution

Setting and Motivation

- Suppose before playing a NFG G, players can make joint "EPV" commitments
  - EPV = ex post verifiable: can be verified after the game
  - e.g., can commit to play some a_i, or to not play a_i
  - cannot commit e.g. to "fairly access whether your deserve a raise"
- This "solves" some games
  - Agree to (S, S) in stag hunt or (C, C) in prisoner's dilemma
- But what about more complicated games, even e.g. chicken?
  - Multiple Nash equilibria, different preferences over them
  - Players might just not know what will happen, or might disagree on it
  - Or, might have incentive to (pretend to) disagree,
    - e.g. each claims they think they will Dare, other will Swerve
- **Question:** Can we still reach beneficial agreements in such games?

	Coop.	Defect
Coop.	3, 3	-1, 4
Defect	4, -1	0, 0

	Swerve	Dare
Swerve	2, 2	0, 5
Dare	5, 0	-10, -10

	Stag	Hare
Stag	2, 2	0, 1
Hare	1, 0	1, 1

Safe Pareto Improvements

Safe Pareto Improvements (SPIs) [Oesterheld and Conitzer, 2022]
- Interventions (e.g. agreements) which necessarily leave all players better off
- Regardless of how individual games are played
- Under mild assumptions on relationships between how different games are played

Def: a is **Pareto** better than a' if all players prefer weakly prefer it: $u_i(a) \geq u_i(a')$ for all i
- And *strictly* Pareto better if one player strictly prefers it: $u_i(a) > u_i(a')$

Dominance Assumption:
- An action a_1 *strictly dominates* a_1' if playing a_1 always gives P1 strictly higher utility, regardless of what the other players do
- **Assumption:** removing strictly dominated actions doesn't change a game's outcome
- Iteratively applying this leads to a unique game Red(G)

Isomorphism Assumption:
- An isomorphism from game G to G' says the games are "strategically equivalent", i.e. it multiplies each player's utility by a constant and then adds a constant
- For each player i, there is a bijection on actions $\beta_i : A_i \to A_i'$ and a function on utilities $v_i \mapsto m_i v_i + b_i$ such that, for all i and outcomes a, $u_i(\beta(a)) = m_i u_i(a) + b_i$
- **Assumption:** Isomorphic games are played isomorphically
  - i.e., if the outcome of G would be a, then the outcome of G' would be $\beta(a)$

Under these assumptions, we show there are two types of SPIs:
- G' is a ***simple SPI*** on G if all outcomes in Red(G') are Pareto better than all outcomes in Red(G), and at least one is strictly* Pareto better (*for at least one player)
  - e.g. (stag, stag) in stag hunt; (cooperate, cooperate) in prisoner's dilemma
- G' is an ***isomorphism SPI*** on G if Red(G') is isomorphic to Red(G) via a strictly Pareto improving isomorphism β, i.e. $u_i(\beta(a)) \geq u_i(a)$ for all

We generally focus on isomorphism SPIs.
They can help even when, as above, players cannot agree to play a single strategy profile

Disarmament SPIs

Disarmament: (joint) commitment not to play certain actions

Example

	C Swerve	C Dare	D Swerve	D Dare
C Swerve	3, 4	1, 7	-10, 10	-10, 10
C Dare	6, 2	-9, -8	-10, 10	-10, 10
D Swerve	10, -10	10, -10	2, 2	0, 5
D Dare	10, -10	10, -10	5, 0	-10, -10

- The full game G reduces to the (D, D) subgame, call it Red(G)
  - By strict dominance: C actions are strictly dominated
- **Proposal:** Agree not to play D actions
- The resulting game is isomorphic to but Pareto better than red(G)
  - Players utilities are (+1, +2) entrywise along the isomorphism
- This an **SPI:** *"We don't know what would have happened in G, but whatever is it we'll get that (+1, +2)"*

Theorem: In general games, finding disarmament SPIs is NP-Complete. (Deciding for a given disarmament is GI-complete.)

Token Game SPIs

Token Games: High Level Idea
- Create a "token" game, where each player has action space T_i
- Play the token game by writing actions t_i down on paper, simultaneously flipping them over
- Commit to play a correlated strategy profile in G as a fn of this token outcome t
- This lets us construct a new game from arbitrary feasible payoffs in G

Example

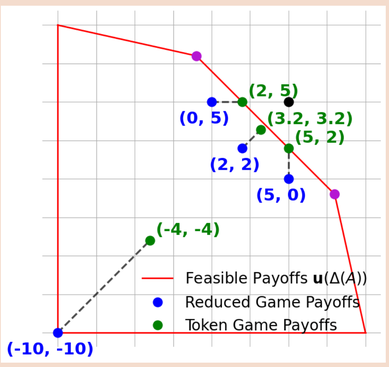
Original Game G

	C Swerve	C Dare	D Swerve	D Dare
C Swerve	2, 2	-1, 8	-10, 10	-10, 10
C Dare	8, -1	-50, -50	-10, 10	-10, 10
D Swerve	10, -10	10, -10	2, 2	0, 5
D Dare	10, -10	10, -10	5, 0	-10, -10

Token Game

	Token Swerve	Token Dare
Token Swerve	3.2, 3.2	2, 5
Token Dare	5, 2	-4, -4

$v_i \mapsto .6v_i + 2$



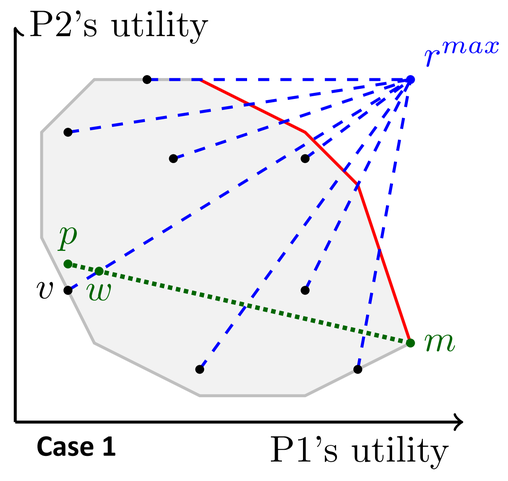
- The original game reduces to red(G) = the (D, D) subgame
  - C actions are strictly dominated
- The (C, C) subgame is a higher upside but much riskier chicken
  - Disarmament is no longer and SPI
- **Token SPI:** Play the token game instead
- Payoffs can be realized (in expectation) by committing to strategy profiles in G
  - e.g. (Token Swerve, Token Dare) = (2,5) = (2/3) (CS, CD) + (1/3) (CD, CS) = (2/3) (-1, 8) + (1/3) (8, -1)
  - Token game is isomorphic to red(G), but Pareto better
  - For each player payoffs of v_i are mapped to $.6v_i + 2$

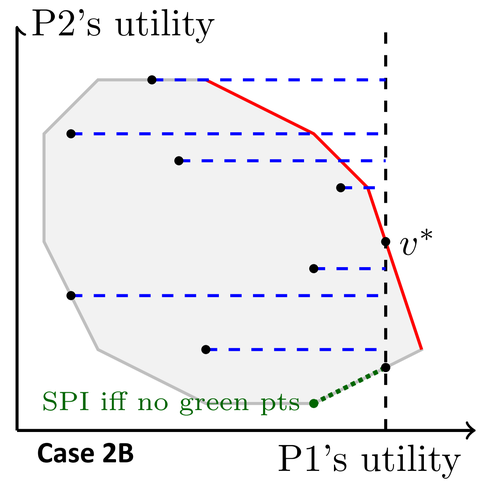
Theorem: Characterization of Token Game SPIs
- Given game G
- Let V be the set of payoffs in the red(G)
- Let $V^* \subseteq V$ be the subset of V which is Pareto optimal in $\mathbf{u}(A)$

1. If $|V^*| = 0$, there is an SPI
2. If $|V^*| = 1$, say $V^* = \{v^*\}$
   A. If $v_i^* \in \{v_i^{min}, v_i^{max}\}$ for both players, there is an SPI
   B. If $v_i^* \in \{v_i^{min}, v_i^{max}\}$ for one player, there is sometimes* an SPI
   C. If $v_i^* \notin \{v_i^{min}, v_i^{max}\}$, for either player, there is no SPI
3. If $|V^*| = 2$, there is no SPI

Proof Sketch
- Token payoffs must be isomorphic to G
- Consider the positive affine fns induced on the payoffs, say $f_i(v_i) \mapsto m_i v_i + b_i$
- Any $v^* \in V^*$ must be fixed points of the f_i

Case 3: each f_i has two fixed points, so must be the identity
Case 2C: each f_i has an *intermediate* fixed point v_i^*, To have $f_i(v_i) \geq v_i$ both above and below v_i^*, it must be the identity
Case 1: Let f_i map each pt towards Player i's best payoff, i.e. $f_i(v_i) = v_i + \varepsilon(r_i^{max} - v_i)$
Case 2B: Say v^* is v_i^{max}. We need $f_2 = \texttt{id}$ (like 2c), but $f_1(v_1) = v_1 + \varepsilon(r_1^{max} - v_1)$ might* work
- It definitely works if (**Case 2A**) $v_2^* = v_2^{min}$





Theorem: In general n-player games, we can decide the existence token game SPIs in polynomial time w/ linear programming

More Stuff!

More Settings in the Paper

"Pure" Token Games:
- Token payoffs be from single outcomes in the original game, not distributions over outcomes
- We give a quasi-polynomial time algorithm and a hardness result

"Default-Remapping Commitment:
- Players can credibly reveal what action they would have play the base game
  - e.g., maybe a player intends to copy what a public figure will do, or ask an LLM what to do
- This enables SPIs in almost all games

Future Work
- Stronger forms of commitment for token game, e.g. utility transfers
- Bayesian Games: What is the right notion isomorphism?
- New assumptions on relationships between how games are played
- AI agents + SPIs:
  - Can AIs use existing forms of SPI?
  - Can we design AIs to make new/ stronger forms of SPIs possible?
  - Can we design SPIs that leverage features of AI agents, e.g. default-remapping with LLMs

Presenter: Nathaniel Sauerberg

Paper: