Choosing what game to play with no regrets or controversies – results on inferring safe (Pareto) improvements in binary constraint structures*

Caspar Oesterheld Vincent Conitzer

May 3, 2024

Abstract

We consider a setting in which we get to choose which game is played by a group of players. We would like to choose a game that favors ourselves, one of the players or the social preferences of the players. Unfortunately, given the potentially multiplicity of equilibria, it is in general (conceptually) hard to tell which of two games is better. Oesterheld and Conitzer (2022) propose that we use assumptions about *outcome correspondence* – i.e., about how the outcomes of different games relate – to allow comparisons in some cases. For example, it seems reasonable to assume that isomorphic games are played isomorphically. From such assumptions we can sometimes deduce that the outcome of one game Γ is guaranteed to be better than the outcome of another game Γ', even if we don't have beliefs about how each of Γ and Γ^s will be played individually. Following Oesterheld and Conitzer, we then call Γ^s a *safe improvement* on Γ.

In this paper, we study how to derive safe improvement relations. We first show that if we are given a set of games and arbitrary assumptions about outcome correspondence between these games, deriving safe improvement relations is co-NP-complete. We then study the (in)completeness of a natural set of inference rules for outcome correspondence. We show that in general the inference rules are incomplete. However, we also show that under natural, generally applicable assumptions about outcome correspondence the rules are complete.

*This is a very early draft. Currently the appendix is removed.

1 Introduction

Imagine you have to choose between two (say, normal-form) games, Γ_1 and Γ_2 for Alice and Bob to play tomorrow. You have some stake in the outcome of the game, e.g., because of externalities or because you are Alice's and/or Bob's friend. Which of Γ_1 and Γ_1 should you choose? (Relatedly, what if by default, Alice and Bob play some game Γ, but now they are offered to play some alternate game Γ^s instead? When would they both consent?)

In general, assessments of what game to choose are conceptually difficult, because each of the games may have multiple solutions (e.g., multiple Nash equilibria). Whether we prefer one game over another hinges on how this multiplicity of solutions is resolved. For example, imagine that you can decide whether Alice and Bob play a Game of Chicken or a trivial game in which they both receive \$5. Then (depending on your preferences over the outcomes) the decision hinges on which – if any – equilibrium of Chicken Alice and Bob will play.

In this paper, we build on the safe Pareto improvement approach of Oesterheld and Conitzer (2022) to address this problem. Roughly, we imagine that we consider some (potentially infinite) set of games. We make some (qualitative, non-probabilistic) assumptions about how the outcomes of pairs of these games, as played by a set of *agents* (Alice and Bob in the above example), relate. For example, we might assume that if two games are isomorphic, then Alice and Bob would play them isomorphically (ibid., Assumption 2). Following Oesterheld and Conitzer, we call such statements about the relationships of games *outcome correspondences*. From this set of assumptions about outcome correspondence we can infer new facts about outcome correspondence. For instance, if we know an outcome relation between Γ_a and Γ_b and an outcome relation between Γ_b and Γ_c, then we can also infer using some sort of transitivity rule (Lemma 2.1 of the present paper; Lemma 2.3 of Oesterheld and Conitzer 2022) an outcome relation between Γ_a and Γ_c. In some cases, we will be able to infer outcome correspondences that show that the outcome of one game Γ^s is always better than the outcome of another game Γ, even without resolving the multiplicity of solutions in the underlying games. We will then say that Γ^s is a *safe improvement* on Γ. If in particular we have that the outcome of Γ^s is better than the outcome Γ under a *Pareto* (partial) order over outcomes, we call Γ^s a *safe Pareto improvement* in line with Oesterheld and Conitzer.

As a simple example, consider the three games in Figure 1. (Note that all three games are essentially versions of the Game of Chicken and each have exactly three equilibria, two pure and one mixed.) We make the following assumptions about outcome correspondence, illustrated by lines in the figure. In Γ^a, the pure strategy C' is strictly dominated by mixing appropriately over D and C. Now note that Γ_b is obtained from Γ_a by removing C'. It seems plausible that Γ_a and Γ_b will therefore be played the same way. That is, if Alice and Bob play (D, D) in Γ_a, then they will also play (D, D) in Γ_b, and so on. Meanwhile, we can see that Γ_b and Γ_c are isomorphic (with payoffs transformed by the positive affine function $x \mapsto 2x + 4$). Again, it therefore seems plausible

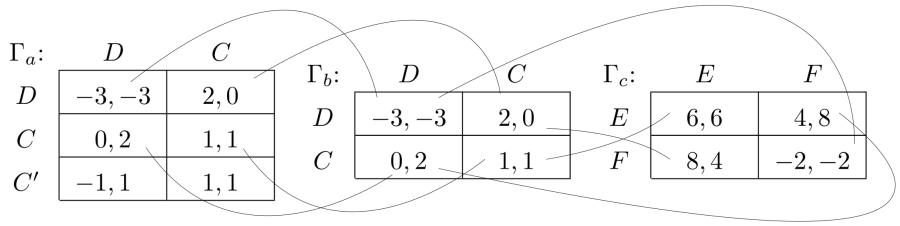

Figure 1: An example of three games with plausible outcome correspondences between them

that Γ^b and Γ^c will be played in the same way – in particular that they will be played isomorphically.

Now imagine that under these assumptions about outcome correspondence, we need to decide whether to let Alice and Bob play Γ_a or Γ_c. Imagine that our goal for the purpose of this choice is to optimize for Alice and/or Bob's preferences. Now notice that from the given outcome correspondences we can infer the aforementioned transitivity rule an outcome correspondence between Γ^a and Γ_b: if Alice and Bob play (D, D) in Γ^a, then they play (F, F) in Γ_c; if Alice and Bob play (D, C) in Γ^a, then they play (F, E) in Γ_c; and so on. By considering each outcome individually, we can see that the outcome of Γ^c is guaranteed to be strictly better for both Alice and Bob. Thus, Γ_c is a safe (Pareto) improvement on Γ_c; when given the choice we should let Alice and Bob play Γ_c instead of Γ_a. Importantly, we don't need to assume anything about how Alice and Bob resolve the equilibrium selection posed by each of $\Gamma_a, \Gamma_b, \Gamma_c$.

The concept of safe (Pareto) improvements is widely applicable to strategic settings. Oesterheld and Conitzer (2022) focus on two players who each instruct their agent by shaping their utility functions (cf. Baumann, 2017; Clifton, 2020, Sect. 4.2). Oesterheld and Sauerberg (2024) show how the concept of safe Pareto improvements can be used to assess *ex post* verifiable commitments. However, the broad conceptualization of safe (Pareto) improvements of the present paper points at many new fields of applications. For example, mechanism design studies the question of how to shape a game played by Alice and Bob. We believe that in some settings – specifically settings where mechanisms induce a game with multiple Nash equilibria – the concept of safe (Pareto) improvements can be used to assess and compare different possible mechanisms. As such, the concept of safe improvements can be viewed as complementing concepts like best-Nash (cf. price of stability) and worst-Nash (cf. price of anarchy) that are widely used for comparing different mechanisms.

Contributions: After introducing some notation and background (Section 2), we introduce a new framework for reasoning about safe (Pareto) improvement framework (Section 3). In doing so, we generalize the setting of Oesterheld and Conitzer (2022): We define safe improvements relative to arbitrary preferences over outcomes rather than specifically Pareto preferences.

Furthermore, we define safe improvements on arbitrary binary constraint structures.

Throughout this paper, we address the following question: Given some set of games and given some assumptions about outcome correspondence between these games, can we conclude that one game is a safe (Pareto) improvement over another?

In Section 4, we analyze the computational complexity of deciding this problem if we are given a finite set of games and explicitly represented outcome correspondence assumptions between these games. We prove that this problem is co-NP-complete, even under various restrictions on the outcome correspondence assumptions. Roughly speaking, we can view the assumptions as defining a binary constraint structure a la the binary constraint satisfaction problem (CSP). Binary CSP is NP-complete. However, in the context of SIs, we are not interested in finding a satisfying assignment (a way in which the agents might play the different games that satisfies all the assumptions) but in whether specific facts hold about *all* satisfying assignments. In general, this problem is *co*-NP-complete, because it is the complement of an NP-complete problem. We strengthen the result in various ways to show that it applies even if we restrict attention to problem instances that satisfy various plausibility assumptions.

In Section 5, we consider the problem of inferring safe Pareto improvements under restrictions on what type of assumptions we make. In particular, we ask whether under such restrictions local inference (such as the above application of the transitivity rule) are complete. In particular, we note that most natural generic assumptions about outcome correspondence (such as: isomorphic games are played isomorphically; dominated actions can be removed; duplicate actions can be removed) satisfy a condition called *max-closedness* known from the CSP literature. We show that under this condition, the local rules of inference are complete (Theorem 4). From that we immediately obtain an efficient algorithm for deriving S(P)Is. We show that completeness holds even if we consider countably many assumptions and games. We then demonstrate the usefulness of the completeness results for proving complexity results.

2 Preliminaries

Normal-form games We here briefly introduce some game-theoretic notation. An *n-player (normal-form) game* is a tuple $\Gamma = (A_1, ..., A_n, u_1, ..., u_n)$ of sets of *(pure) strategies* or *actions* $A_1, ..., A_n$ and *utility functions* $u_i \colon A_1 \times ... \times A_n \to \mathbb{R}$ mapping *(pure) strategy profiles* onto utilities. We call a pure strategy profile $\mathbf{a}$ a Nash equilibrium if we have for every player i that $u_i(\mathbf{a}_{-i}, a_i) \leq u_i(\mathbf{a})$. We say that a strategy a_i strictly dominates another strategy a'_i if for all $\mathbf{a}_{-i}$ we have that $u_i(a_i, \mathbf{a}_{-i}) > u_i(a'_i, \mathbf{a}_{-i})$. We say that $(\Phi_i : A_i \to A'_i)_{i=1,...,n}$ is an *isomorphism* between two n-player games Γ and Γ' if there are $\lambda_i > 0, b_i$ such that for all $a_1, ..., a_n$ we have $u_i(a_1, ..., a_n) = \lambda_i u'_i(a_1, ..., a_n) + b_i$.

Set-valued functions We use set-valued functions to express outcome correspondences between games. A set-valued function $\Phi\colon X \multimap Y$ is a binary relation between X and Y, i.e., Φ is a subset of $X \times Y$. For any $x \in X$, we write $\Phi(x) \coloneqq \{y \in Y \mid (x, y) \in \Phi\}$. Note that by specifying $\Phi(x)$ for each $x \in X$ we fully specify Φ. For any $M \subseteq X$, we write $\Phi(M) \subseteq \bigcup_{x \in M} \Phi(x)$. For any $\Phi\colon X \multimap Y$ we define $\Phi^{-1}\colon Y \multimap X$ by $\Phi^{-1} \coloneqq \{(y, x) \mid (x, y) \in \Phi\}$. For two multi-valued functions $\Phi\colon X \multimap Y$ and $\Psi\colon Y \times Z$, we define their composition $\Psi \circ \Phi$. Finally, for $\Phi, \Psi\colon X \multimap Y$ we define $\Phi \cap \Psi\colon X \multimap Y$ to be the argument-wise intersection of Φ and Ψ, i.e., the set-valued function defined by $(\Phi \cap \Psi)(x) = \Phi(x) \cap \Psi(x)$ for all $x \in X$.

Binary constraint structures A *binary constraint structure* is a triplet $(\mathcal{X}, \mathcal{D}, \mathcal{A})$, where:

- $\mathcal{X}$ is a set of variables.
- $\mathcal{D} = (D_X)_{X \in \mathcal{X}}$ is a family of finite domains, one for each variable.
- $\mathcal{A}$ is a set of binary constraints. Each binary constraint is a triplet (X, Y, Φ) where X and Y are variables and $\Phi \subseteq D_X \times D_Y$ is a subset of the domains corresponding to X and Y. Will generally write constraints as $X \sim_\Phi Y$.

An *assignment* for $(\mathcal{X}, \mathcal{D}, \mathcal{A})$ associates with each variable X a value v_X in D_X. We say that an assignment *satisfies* a binary constraint structure if for each constraint $(X, Y, \Phi) \in \mathcal{A}$ we have that $(v_X, v_Y) \in A$. We call $(\mathcal{X}, \mathcal{D}, \mathcal{A})$ finite if $\mathcal{X}$, $\mathcal{D}$ and $\mathcal{A}$ are all finite.

The binary constraint satisfaction problem (BCSP) consists in deciding whether there exists a satisfying assignment for a given binary constraint structure. The problem is well-known to be NP-complete.

Theorem 1. BCSP *is NP-complete, even if we restrict the domains to have size at most three. If we restrict the domains to have size two, then the problem is solvable in polynomial time.*

Given a binary constraint structure, we can often tighten the constraints by drawing inferences from the existing constraints. In the literature on BCSP, this is known as constraint propagation. In the following we in particular introduce the rules necessary for imposing *path consistency* Montanari, 1974. We follow the notation of Oesterheld and Conitzer (2022).

Lemma 2. *Let Γ, Γ', $\hat{\Gamma}$ be games and $\Phi, \Xi\colon A \multimap A'$, $\Psi\colon A' \multimap \hat{A}$ be outcome correspondence functions. Then the following hold:*

1. *Transitivity: If $\Gamma \sim_\Phi \Gamma'$ and $\Gamma' \sim_\Psi \hat{\Gamma}$, then $\Gamma \sim_{\Psi \circ \Phi} \hat{\Gamma}$.*
2. *Intersection: If $\Gamma \sim_\Phi \Gamma'$ and $\Gamma \sim_\Psi \Gamma'$, then $X \sim_{\Psi \cap \Xi} Y$, where $\Psi \cap \Xi$ refers to the element-wise intersection of the multi-valued functions Ψ and Ξ.*
3. *Reflexivity: $\Gamma \sim_{\mathrm{id}_A} \Gamma$, where $\mathrm{id}_A\colon A \multimap A\colon \mathbf{a} \mapsto \{\mathbf{a}\}$.*
4. *Symmetry: If $\Gamma \sim_\Phi \Gamma'$, then $\Gamma' \sim_{\Phi^{-1}} \Gamma$.*
5. *$\Gamma \sim_{\mathrm{all}_{A,A'}} \Gamma'$, where $\mathrm{all}_{A,A'}\colon A \multimap A'\colon \mathbf{a} \mapsto A'$.*

The proof can be found in Appendix E.

3 Safe (Pareto) improvements and outcome correspondence

A generic definition of safe (Pareto) improvements We now introduce a formalism for safe (Pareto) improvements, which is slightly generalized and adapted from Oesterheld and Conitzer (2022). We defer to Oesterheld and Conitzer for a more detailed introduction and motivation of the concepts in this section. We compare the high-level formalisms in detail in Appendix H.

Imagine that we have the choice between some set of variables $X_0, X_1, ..., X_n$. Each variable can result in one of a number of possible values or *outcomes*. We are uncertain which outcome any given variable might result in.

We have some preferences $\succeq$ over the outcomes of the different variables. These preferences may be incomplete, however. That is, there are pairs of outcomes of which neither is preferred to the other. For instance, you might imagine that we are choosing as a group of people with diverging preferences and our group preferences are only definite in cases of consensus.

The most standard approach to choosing between the variables involves assigning probabilistic beliefs to the outcomes. But in some cases such probabilistic beliefs are controversial or very difficult to assign. Let's imagine that instead of probabilistic beliefs, we are only willing or able to act on qualitative beliefs about how the outcomes of different variables relate. For instance, we might strongly believe that if X_i were to result in outcome o, then X_j would result in outcome o'.

Then we might still be able to adjudicate between some pairs of variables. In particular, it might be that from the outcome correspondence assumptions we can infer that one variable will always yield an outcome that is to be preferred over the outcome of another variable.

By interpreting the variables, outcomes and outcome correspondence assumptions as a binary constraint structure, we can formalize this as follows.

Definition 1. *Let $(\mathcal{X}, \mathcal{D}, \mathcal{A})$ be a binary constraint structure. Let $\succeq$ be a partial order on $\cup\mathcal{D}$ (i.e., on the union of all the outcomes of $\mathcal{A}$). Let $X, Y \in \mathcal{X}$ be two variables. We say that Y is a safe improvement on X w.r.t. $\succeq$ if each satisfying assignment $(v_Z)_{Z\in\mathcal{X}}$ of $(\mathcal{X}, \mathcal{D}, \mathcal{A})$ satisfies $v_X \leq v_Y$.*

One perspective on the safe improvement relation is that it is an application of decision-theoretic dominance with preferences $\succeq$, with uncertainty over the satisfying assignments.

If $\succeq$ is a Pareto ordering in particular, then we say that Y is a *safe Pareto improvement* on X. Note that Pareto orderings are partial orders.

Safe (Pareto) improvements in strategic settings Throughout this paper, we will typically consider safe (Pareto) improvements in strategic settings. That is, we imagine that each variable in $\mathcal{X}$ corresponds to a strategic interaction between a group of *agents*. For simplicity, we assume specifically that the agents play a normal-form game. The domains $\mathcal{D}$ are the possible outcomes

of the normal-form games. We imagine that a *principal* preferences $\succeq$ chooses which game is played by the agents. Note that $\succeq$ must compare outcomes across different normal-form games and $\succeq$ isn't necessarily (though in this paper often will be) related to the agents' utilities in the underlying game.

In game-theoretic settings we typically imagine that the principal's uncertainty is due to the multiplicity of solutions (e.g., multiplicity of Nash equilibria, rationalizable strategies). This uncertainty seems especially intractable. Therefore, it seems especially natural to be unwilling to assign probabilistic beliefs. Moreover, if we imagine that the principal is a group of decision makers, it seems plausible that they disagree about how the multiplicity of solutions is resolved.

Meanwhile, in the game-theoretic context it is natural to make some qualitative assumptions, especially about how the outcomes of different games related.

We now give two specific assumptions about outcome correspondence along the lines of Oesterheld and Conitzer (2022) and illustrated in Figure 1. These assumptions are plausible in many settings. That said, we will not always make these assumptions – we will indicate explicitly when we make any of these assumptions.

The first assumption is that we can remove strictly dominated strategies from any given game and obtain a game that will be played equivalently by the representatives.

Assumption 1. *Let $\Gamma = (A_1, ..., A_n, \mathbf{u})$ be a game. Let $a_i, a_i' \in A_i$ s.t. a_i strictly dominates a_i'. Then $\Gamma \sim_\Phi (A_1, ..., A_{i-1}, A_i - \{a_i'\}, A_{i+1}, ..., A_n, \mathbf{u})$, where Φ is defined by $\Phi(\mathbf{a}) = \emptyset$ if $a_i = a_i'$ and $\Phi(\mathbf{a}) = \{\mathbf{a}\}$ otherwise.*

The second assumption is that isomorphic games are played isomorphically.

Assumption 2. *Let $\Gamma = (A_1, ..., A_n, \mathbf{u}), \Gamma' = (A_1', ..., A_n', \mathbf{u}')$ be two isomorphic games that do not contain any strictly dominated strategies. Let $\mathcal{I}$ be the set of isomorphisms between Γ and Γ'. Then $\Gamma \sim_I \Gamma'$, where I is defined by $I(o) = \{\Phi(o) \mid \Phi \in \mathcal{I}\}$.*

We restrict the assumption to games without dominated actions, because the dominated actions might break some of the isomorphism and thus narrow down the set $\mathcal{I}$. This seems undesirable.

We give some further plausible assumptions about outcome correspondence in Section 5.1 and Appendix G.2.

4 Complexity results under general outcome correspondence assumptions

Throughout this section, we study the following computational problem: Given a (satisfiable) finite binary constraint structure, some partial order $\succeq$ over outcomes, decide whether a particular (type of) or any safe improvement claim follows from the assumptions. We will show that this problem is co-NP-complete, even under some further restrictions.

At a very high-level, the argument is based on the close connection between SIs and inference on binary constraint structures. We give a very rough sketch here: Consider the problem of finding a valid assignment of $(\mathcal{G}, \mathcal{D}, \mathcal{A})$. This problem is essentially a binary constraint satisfaction problem (CSP) and therefore NP-complete. The complement of CSP is the problem of deciding whether there is no valid assignment. Since this problem is the complement of an NP-complete problem, this problem is co-NP-complete. The problem of deciding whether $\mathcal{A}$ implies that Y is an SI on Y is essentially the problem of deciding whether $\mathcal{A}$ implies that (X, Y, Φ), where Φ is maps each outcome of X onto an at least as good outcome of Y. (Since $\succeq$ can be anything in general, Φ can be any function.) This problem in turn is the same as the problem of deciding whether there is no assignment that satisfies $\mathcal{A}$, but not (X, Y, Ξ), where Ξ is the complement of Φ, i.e., $\Xi = D_X \times D_Y - \Phi$. Therefore, deciding whether $\mathcal{A}$ implies that (X, Y, Φ) is co-NP-complete.

We would like to refine our co-NP-completeness result a little further, however. One natural question is whether the correspondence between (the complement of) binary CSP and reasoning about outcome correspondence persists if we restrict attention to assumptions about outcome correspondences that one might plausibly believe. For example, what happens if we exclude unsatisfiable sets of assumptions $\mathcal{A}$ or assumptions $\mathcal{A}$ that imply irrational behavior (such as mutual cooperation in the one-shot Prisoner's Dilemma)? It turns out that co-NP-completeness holds under some such restrictions on $\mathcal{A}$ (though cf. Section 5 for conditions under which co-NP-completeness ceases to hold). What if we restrict $\succeq$ to be aligned

We will include some restrictions on $\mathcal{A}$ below. However, since we can't expect to give a full account of what a plausible set $\mathcal{A}$ looks like, we here give a brief account of what types of games we use for our reduction from (the complement of) binary CSP. These games are sketched in Figure 2. Essentially, these are games with n pure equilibria along the diagonal. We assume that the agents successfully coordinate on a equilibrium (and thus play rationally in some sense) but we don't know which. Our assumption give us information about how the agents select equilibria across games. From considering only the payoff matrices, it is unclear how one would ever arrive at such judgments. However, we might imagine further contextual information (à la Schelling or focal points Schelling, 1960, pp. 54–58) under which these outcome correspondences would be plausible.

To (partially) enforce plausibility of the assumptions in our formal result, we define the following. Let $\mathcal{G}$ be a finite set of games. Let $\mathcal{A}_{1,2}(\mathcal{G})$ be the set of OCs that can be inferred from Assumptions 1 and 2 about $\mathcal{G}$, including by reasoning about games outside of $\mathcal{G}$.

Theorem 3. *The following decision problem is co-NP-complete: Given a finite set of games $\mathcal{G}$ and some finite satisfiable set $\mathcal{A}$ of outcome correspondences that includes $\mathcal{A}_{1,2}(\mathcal{G})$, two games $\Gamma, \Gamma' \in \mathcal{G}$, and a preference ordering $\succeq$, decide whether $\mathcal{A}$ implies that Γ' is an SI on Γ. The same holds under the following we restrict $\succeq$ to express one of the player's utilities or Pareto preferences, and/or if we consider strict and very strict S(P)I. The problem remains co-NP-hard if*

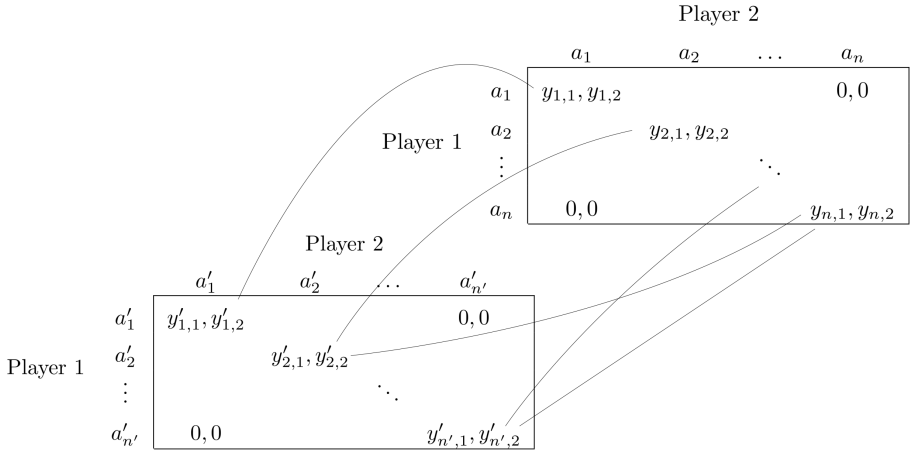

Figure 2: A type of outcome correspondence to which we can reduce binary constraint satisfaction problems without producing wildly implausible outcome correspondences.

we bound the sizes of of the games in $\mathcal{G}$.

The proof can be found in Appendix J.

The co-NP-completeness results hinge on increasing the number of games (variables) in $\mathcal{G}$. For any fixed size of $\mathcal{G}$, the problem is solvable in polynomial time.

The problem of deciding whether there exists within $\mathcal{G}$ a pair of games s.t. the first is an improvement on the second is also co-NP-complete, and so is the problem of deciding whether for a given $\Gamma \in \mathcal{G}$ there exists another game Γ' that is an SI on Γ. For details, see Proposition 25 in the appendix.

One interesting implication of the complexity results is that (assuming P$\neq$ NP) repeated application of inference rules like those in Lemma 2 are in general insufficient for deriving all relevant SPIs. This is because repeated application of such rules can be done in polynomial time. In Appendix K.3, we show this formally. We also provide an example where the rules of Lemma 2 are incomplete.

5 Completeness under natural assumptions about assumptions about outcome correspondence

So far we have considered the problem of deriving SPIs under arbitrary assumptions. We have seen that the problem of inferring S(P)Is is co-NP-complete in this case. In this section, we provide conditions under which inference as per Lemma 2 is complete. We thus also obtain that under some condition S(P)Is can be found in polynomial time.

5.1 Some further natural assumptions about outcome correspondence

To guide our results, we first give some additional natural assumptions about outcome correspondence. We focus here on the structurally most interesting examples of further assumptions. We give some further examples of assumptions in Appendix G.

Decomposition/factorization Imagine the agents play a normal-form game that consists of playing in parallel the game Stag Hunt and Chicken, i.e., they each choose from the set of actions $\{\text{Stag}, \text{Hare}\} \times \{\text{Dare}, \text{Swerve}\}$ and the reward is given by the sum of the rewards of the respective games. Then we might expect that they play an action from $\{\text{Stag}\} \times \{\text{Dare}, \text{Swerve}\}$ if and only if they play Stag in the Stag Hunt.

To formally state this, we first define product–sum games as follows. Let $\Gamma = (A_1, ..., A_n, \mathbf{u})$ and $\Gamma' = (A'_1, ..., A'_n, \mathbf{u}')$ be n-player (normal-form) game. Then define $\Gamma * \Gamma'$ to be the game

$$(A_1 \times A'_1, ..., A_n \times A'_n, ((a_1, a'_1), ..., (a_n, a'_n)) \mapsto \mathbf{u}(a_1, ..., a_n) + \mathbf{u}'(a'_1, ..., a'_n)) .$$

Assumption 3 (Decomposition into sum–product games)**.** *Let Γ, Γ' be n-player games. Let $\mathcal{I}$ be the set of isomorphisms on Γ. Then $\Gamma * \Gamma' \sim_\Phi \Gamma$, where $\Phi\colon ((a_1, a'_1), ..., (a_n, a'_n)) \mapsto \bigcup_{\Xi \in \mathcal{I}} \Xi(a_1, ..., a_n)$.*

As a sanity check, note that the Nash equilibria of $\Gamma * \Gamma'$ are simply the combinations of Nash equilibria of Γ and Γ'. Although we find Assumption 3 plausible in many cases, it may not universally apply. For instance, imagine that both Γ and Γ' are about distributing one indivisible item between the agents. It then seems plausible that in $\Gamma * \Gamma'$, one agent receives the Γ item and the other receives the Γ' item. This option is, of course, not available if only Γ is played. It's also worth noting that various on first sight plausible-seeming generalizations of Assumption 3 are problematic. For instance, if we play two *repeated* games simultaneously (and we can use our observations in one of the games to inform our choices in the other game), then the Nash equilibria of the resulting game are different from the Nash equilibria of each of the individual games.

Ordered outcomes and order-preserving correspondences Finally, we give an example of a plausible outcome correspondence assumption that is closely related to some ordering of the outcomes or pure strategies of the game. In contrast to the above, we only give this assumption for a single game, because it's unclear how this assumption generalizes.

Consider the two normal-form games in Table 1. Roughly, both games model the following interaction: \$4 are on the table and Players 1 and 2 can choose whether to demand \$3, \$2, or \$1. If the demands are compatible (i.e., sum up to at most \$4), then they both receive their demands. Otherwise, they receive some punishment. The difference between the two versions is that Player 1

	3	2	1
3	$3,1$	$-1,-1$	$-1,-1$
2	$2,1$	$2,2$	$-1,-1$
1	$1,1$	$1,2$	$1,3$

	3	2	1
3	$3,1$	$-2,-1$	$-2,-1$
2	$2,1$	$2,2$	$-2,-1$
1	$1,1$	$1,2$	$1,3$

Table 1: An example of a game with a plausible order-preserving outcome correspondence.

receives a harsher punishment in the "no-agreement" case in the right-hand game, compared to the left-hand game.

Now it seems plausible that being more harshly punished for a failure to agree on a split, should lead Player 1 to demand less aggressively, or at least not more aggressively. That is, if Player 1 demands 2 in the left-hand game, we would expect here to demand 1 or 2 in the right-hand game. Anticipating that Player 1 will be less aggressive, Player 2 can be more aggressive (or least equally aggressive) in the right-hand game.[1]

Ad Hoc Assumption 1. *Let* Γ *and* Γ' *be the left-hand and right-hand game in Table Table 1. Let* $\Phi_1 \colon \{1,2,3\} \to \{1,2,3\} \colon 1 \mapsto \{1\}, 2 \mapsto \{1,2\}, 3 \mapsto \{1,2,3\}$ *and* $\Phi_2 \colon \{1,2,3\} \to \{1,2,3\} \colon 1 \mapsto \{1,2,3\}, 2 \mapsto \{2,3\}, 3 \mapsto \{3\}$. *Then* $\Gamma \sim_{(\Phi_1,\Phi_2)} \Gamma'$.

Note that combined with Assumption 4, we get $\Gamma \sim_{\Phi} \Gamma'$, where $\Phi \colon (1,3) \mapsto \{(1,3)\}, (2,2) \mapsto \{(1,3),(2,2)\}, (3,1) \mapsto \{(1,3),(2,2),(3,1)\}$.

5.2 Inference is complete if the outcome correspondence assumptions are max-closed

We now give a sufficient condition under which inference is complete.

Definition 2 (P. G. Jeavons and Cooper 1995, Def. 2.5)**.** *Let* $\mathcal{G}$ *be a set of games and* $\mathcal{A}$ *be a set of assumptions over* $\mathcal{G}$*. We call* $\mathcal{A}$ *max-closed if there exists a family of total orders* $(\geq_G)_{G\in\mathcal{G}}$ *s.t. Each outcome correspondence* Φ *in* $\mathcal{A}$ *has the following property: If* $(x_1, y, 1), (x_2, y_2) \in \Phi$*, then* $(\max(x_1, x_2), \max(y_1, y_2)) \in \Phi$.

P. G. Jeavons and Cooper (1995) prove that under max-closedness, *satisfiability* of binary constraint structures is decidable in polynomial time. In particular, their proof shows that iterated application of the rules of Lemma 2 will infer an empty outcome correspondence if and only if the binary constraint structure is unsatisfiable. We generalize this result to show that iterated application of

[1] It's worth noting that this intuition isn't reflected in the Nash equilibria of the game. The pure Nash equilibria are the same between the two games. In the mixed equilibria, the players have to make each other indifferent. Of course, Player 2's strategies for making Player 1 indifferent are the same between the two games. Meanwhile, Player 1 is actually *less* "aggressive" in the mixed equilibria! Smaller probabilities of higher demands are now needed to make Player 2 indifferent between low and high demands.

the rules of Lemma 2 is also complete for inferring outcome correspondences, i.e., for deciding whether all satisfying assignments satisfy some OC, including SPIs.

Theorem 4. *Let $\mathcal{G}$ be a finite set of games and S be a set of OC assumptions about $\mathcal{G}$. Let S be max-closed. Then inference as per Lemma 2 is complete for $\mathcal{G}, S$. Consequently, inference on max-closed sets of assumptions S is solved in polynomial time.*

The proof can be found in Appendix D.

P. Jeavons, Cohen, and Gyssens (1995) provide a generalization of Definition 2 to partial orders under which satisfiability can still be decided in polynomial time using repeated application of Lemma 2. It follows that inference of OCs can also still be done in polynomial time under this condition. However, repeated application of the rules of Lemma 2 is insufficient in this case. We discuss this in detail in Appendix B.

To argue for the plausibility of the max-closedness condition being true about natural assumptions about outcome correspondence, we first show that all but one of our jointly satisfy the simple-overlap condition.

Proposition 5. *Assumptions 1, 2 and 4 to 6 are each individually max-closed. Any union of any subset of the four is also max-closed. Assumption 3 is not max-closed. Ad Hoc Assumption 1 is not max-closed. Ad Hoc Assumption 1 intersected with Assumption 4 is max-closed.*

5.3 Adding sum–product games retains completeness

Our completeness result (Proposition 5) so far excludes Assumption 3 and it also excludes Ad Hoc Assumption 1 on its own. The problem is that if we have two variables with linearly ordered domains and the constraints on those variables are max-closed, then constraints on the product variable of the two need not be max-closed. To address this, we will prove an additional result that says that adding such product variables or replacing variables with their product does not interfere with completeness.

Lemma 6. *Let $(\mathcal{X}, \mathcal{D}, \mathcal{A})$ be any binary constraint structure. Let $X_a, X_b \in \mathcal{X}$ with domains D_a and D_b. Then repeated inference as per Lemma 2 is complete for the following binary constraint structures*

a) $(\mathcal{X}', \mathcal{D}', \mathcal{A}')$ defined as follows: $\mathcal{X}' = \mathcal{X} \cup \{X_{ab}\}$; the domains $\mathcal{D}'$ are as in $\mathcal{D}$ and the domain D_{ab} for X_{ab} is $D_a \times D_b$;

$$\mathcal{A}' = \mathcal{A} \cup \{(X_a, X_{ab}, \{(x, (x, y)) \mid x \in D_a, y \in D_b\}), \\ (X_b, X_{ab}, \{(y, (x, y)) \mid x \in D_a, y \in D_b\})\}$$

b) $(\hat{\mathcal{X}}, \hat{\mathcal{D}}, \hat{\mathcal{A}})$ defined as follows: $\hat{\mathcal{X}} = \mathcal{X} \cup \{X_{ab}\} - \{X_a, X_b\}$; the domain for X_{ab} is $D_a \times D_b$. The constraints $\hat{\mathcal{A}}$ are constructed from $\mathcal{A}$ as follows.

First we remove the constraints involving X_a, X_b and keep the remaining constraints. Then for each constraint (X_a, Y, Φ) we add a constraint $(X_{ab}, Y, \{((x_a, x_b), y) \mid (x_a, y) \in \Phi, x_b \in D_b\})$. We analogously adapt all other constraints involving X_a and/or X_b.

With this result in hand, we can prove completeness of under all of our assumptions.

Proposition 7. *Let $\mathcal{G}$ be a finite set of games. Let S be any set of assumptions induced by some subset of Assumptions 1 to 3 and 4 to 6 and Ad Hoc Assumption 1. Then inference as per Lemma 2 on S is complete.*

5.4 Generalizing completeness to countably infinite sets of assumptions about outcome correspondence

In some cases, the set of games under considerations $\mathcal{G}$ is infinite. For instance, Oesterheld and Conitzer (2022) consider a setting in which the principals can specify the utility functions of the agents. Assuming, for instance, arbitrary fractions or decimal numbers can be used to specify the utility functions, there are then (countably) infinitely many possible games to have the agents play. Of course, it's hard to say anything positive about the complexity of inferring SPIs on general infinite binary constraint structures. But it's useful to note that the *completeness* of rules like those in Lemma 2 generalizes to infinite binary constraint structures.

Theorem 8. *Let $\mathcal{A}$ be a countable set of assumptions on some countable set of games $\mathcal{G}$. Let $\Gamma, \Gamma' \in \mathcal{G}$ and Φ be such that S implies $\Gamma \sim_\Phi \Gamma'$. Then there exists a finite set of games $\mathcal{G}^f$ containing Γ, Γ' and finite set of assumptions S^f over $\mathcal{G}^f$ such that S^f implies $\Gamma \sim_\Phi \Gamma'$.*

Corollary 9. *Let S be a countable set of outcome correspondence assumptions. If the rules of Lemma 2 are complete for every finite subset of S, then they are also complete for S.*

5.5 Strengthening the complexity result of Oesterheld and Conitzer's (2022)

Oesterheld and Conitzer (2022) prove a hardness result for SPIs under a specific set $\mathcal{G}$ and Assumptions 1 and 2. However, their result is about deciding whether an SPI can be inferred with Lemma 2. Using our completeness results, we can show that in this case inference as per Lemma 2 is equivalent to inference period and so in particular we can prove the hardness of the regular inference problem. To keep it brief, we don't include any of the strict versions of the problems in the below.

Theorem 10. *The following decision problem is NP-hard: Given a game Γ, decide whether there exists a subset game Γ^s of Γ s.t.:*

1. *(Non-triviality:) If we fully reduce* Γ *and* Γ^s *using iterated strict dominance, then* Γ *and* Γ^s *are not equal. (They are allowed to be isomorphic.)*
2. $\Gamma \sim_\Phi \Gamma^s$ *is implied by Assumptions 1 and 5.*
3. *For all* $\mathbf{a}$ *in* Γ*, we have that* $\mathbf{u}(\Phi(\mathbf{a})) \geq \mathbf{u}(\mathbf{a})$.

6 Related work

Constraint satisfaction problem Technically, our work is most closely related to the literature on binary constraint satisfaction problems and in particular constraint propagation. We have used some ideas and results from that literature, such as the NP-completeness of CSP and the max-closedness condition (P. G. Jeavons and Cooper, 1995; P. Jeavons, Cohen, and Gyssens, 1995). Generally, our perspective differs in two ways from that in the CSP literature. First, we are interested in inference problems on binary constraint structures, as opposed to finding satisfying assignments. (More specifically, we are interested in inferring SI relations.) This inference problem is, of course, somewhat different from the problem of finding satisfiable assignments, sometimes in subtle ways (e.g., see Appendix B). Second, because we work in a specific domain (strategic interactions), we are interested in specific types of constraints.

Prior work on safe Pareto improvements Conceptually, our work is of course most closely related to prior work on SPIs, especially Oesterheld and Conitzer (2022). Our basic setup and definition of safe (Pareto) improvement generalizes the setup of Oesterheld and Conitzer by allowing considering arbitrary assumptions about outcome correspondence, arbitrary interventions on how the agents play the game and arbitrary preferences over outcomes. See Appendix H for a detailed comparison of the setups.

Our complexity results are different from those of Oesterheld and Conitzer (2022) and Oesterheld and Sauerberg (2024). For instance, the NP-hardness result of Oesterheld and Conitzer (Theorem 9) hinges on $\mathcal{G}$ being infinitely large (or at least more than polynomially large as a function of the problem input size). In contrast, our co-NP-hardness results consider finite, explicitly represented sets of games. The graph-isomorphism-hardness results of Oesterheld and Sauerberg (2024), meanwhile, are due to the graph-isomorphism hardness of deciding whether Assumption 2 applies to any given pair of games.

Prior work on SPIs has not considered the question of completeness of inference as per Lemma 2 at all.

References

Baumann, Tobias (2017). *Using surrogate goals to deflect threats*. URL: https://s-risks.org/using-surrogate-goals-to-deflect-threats/.

Clifton, Jesse (2020). *Cooperation, Conflict, and Transformative Artificial Intelligence: A Research Agenda*. URL: https://longtermrisk.org/files/Cooperation-Conflict-and-Transformative-Artificial-Intelligence-A-Research-Agenda.pdf.

Dechter, Rina (2003). *Constraint Processing*. Morgan Kaufmann.

Garey, M.R., D.S. Johnson, and L. Stockmeyer (1976). "Some simplified NP-complete graph problems". In: *Theoretical Computer Science* 1.3, pp. 237–267. ISSN: 0304-3975. DOI: https://doi.org/10.1016/0304-3975(76)90059-1. URL: https://www.sciencedirect.com/science/article/pii/0304397576900591.

Jeavons, Peter, David Cohen, and Marc Gyssens (1995). "A unifying framework for tractable constraints". In: *Principles and Practice of Constraint Programming—CP'95: First International Conference, CP'95 Cassis, France, September 19–22, 1995 Proceedings 1*. Springer, pp. 276–291.

Jeavons, Peter G and Martin C Cooper (1995). "Tractable constraints on ordered domains". In: *Artificial Intelligence* 79.2, pp. 327–339.

Krom, M. R. (1967). "The Decision Problem for a Class of First-Order Formulas in Which all Disjunctions are Binary". In: *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 13.1-2, pp. 15–20. DOI: https://doi.org/10.1002/malq.19670130104. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/malq.19670130104. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/malq.19670130104.

Montanari, Ugo (1974). "Networks of constraints: Fundamental properties and applications to picture processing". In: *Information sciences* 7, pp. 95–132.

Oesterheld, Caspar and Vincent Conitzer (2022). "Safe Pareto improvements for delegated game playing". In: *Autonomous Agents and Multi-Agent Systems* 36.2, p. 46.

Oesterheld, Caspar and Nathaniel Sauerberg (2024). "Safe Pareto improvements by ex post verifiable commitments".

Schelling, Thomas C. (1960). *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.