Game Theory with Simulation of Other Players

Vojtěch Kovařík, **Caspar Oesterheld** and **Vincent Conitzer**
Foundations of Cooperative AI Lab (FOCAL), Carnegie Mellon University
vojta.kovarik@gmail.com, oesterheld@cmu.edu, conitzer@cs.cmu.edu

Abstract

Game-theoretic interactions with AI agents could differ from traditional human-human interactions in various ways. One such difference is that it may be possible to simulate an AI agent (for example because its source code is known), which allows others to accurately predict the agent's actions. This could lower the bar for trust and cooperation. In this paper, we formalize games in which one player can simulate another at a cost. We first derive some basic properties of such games and then prove a number of results for them, including: (1) introducing simulation into generic-payoff normal-form games makes them easier to solve; (2) if the only obstacle to cooperation is a lack of trust in the possibly-simulated agent, simulation enables equilibria that improve the outcome for both agents; and however (3) there are settings where introducing simulation results in strictly worse outcomes for both players.

1 Introduction

Game theory is in principle agnostic as to the nature of the players: besides individual human beings, they can be households, firms, countries, and indeed AI agents. Nevertheless, throughout most of the development of the field, game theorists have had in mind players that were either humans or entities whose decisions were taken by humans; and as with any theory, the examples one has in mind while developing that theory are likely to affect its focus. If we try to re-develop game theory specifically with AI agents in mind, how might the theory turn out different? Of course, theorems in traditional game theory will not suddenly become false just because of the change in focus. Instead, we would expect any difference to consist in the kinds of settings and phenomena for which we develop models, analysis, and computational tools.

In this paper, we focus on one specific phenomenon that is more pertinent in the context of AI agents: agents being able to *simulate* each other. If an agent's source code is available, another agent can simulate what the former agent will do, which intuitively appears to significantly change the game strategically. We consider settings in which one agent can simulate another, and if they do so, they learn what the other agent will do in the actual game; however, simulating comes at a cost to the simulator, and therefore it is not immediately clear whether and when simulation will actually be used in equilibrium. In particular, we are interested in understanding whether and when the availability of such simulation results in play that is more cooperative. For example, in settings where *trust* is necessary for cooperative behavior [Berg *et al.*, 1995], one may expect that the ability to simulate the other player can help to establish this trust. But does this in fact happen in equilibrium? And if so, does the ability to simulate foster cooperation in all games, or are there games where it backfires? Are we even able to compute equilibria of games with the ability to simulate?

In terms of related work, our setting is similar to the one of credible commitment [von Stackelberg, 1934], except that one needs to decide whether to pay for allowing the *other* player to commit. Another perspective is that we study program equilibria [Tennenholtz, 2004], except that only *one* player's program can read the other's source code, and has to pay a cost to do so. For further discussion and references, see Section 7.

In the remainder of this introduction, we describe a specific example of a trust game and use it to overview the technical results presented later. We also give several examples that illustrate how simulation can lead to different results when moving beyond trust games. ***For a quick overview, the key takeaways are in Section 1.1, highlighted in italics.***

1.1 Overview and Illustrative Examples

Trust Game As a motivation, consider the following Trust Game (cf. Figure 1; our TG is a variation on the traditional one from [Berg *et al.*, 1995]). Alice has \$100k in savings, which are currently sitting in her bank account at a 0% interest rate. She is considering hiring an AI assistant from the company Bobble to manage her savings instead. If Bobble and its AI *cooperate* with her, the collaboration generates a profit of \$50k, to be split evenly between her and Bobble. However, Alice is reluctant to *trust* Bobble, which might have instructed the AI to *defect* on Alice by pretending to malfunction, while siphoning off all of the \$150k. In fact, the only Nash equilibria of this scenario are ones where Bobble defects on Alice with high probability, and Alice, expecting this, *walks out* on Bobble.

Adding simulation Dismayed by their inability to make a profit, Bobble decides to share with Alice a portion of the AI's source code. This gives Alice the ability to spend \$7k

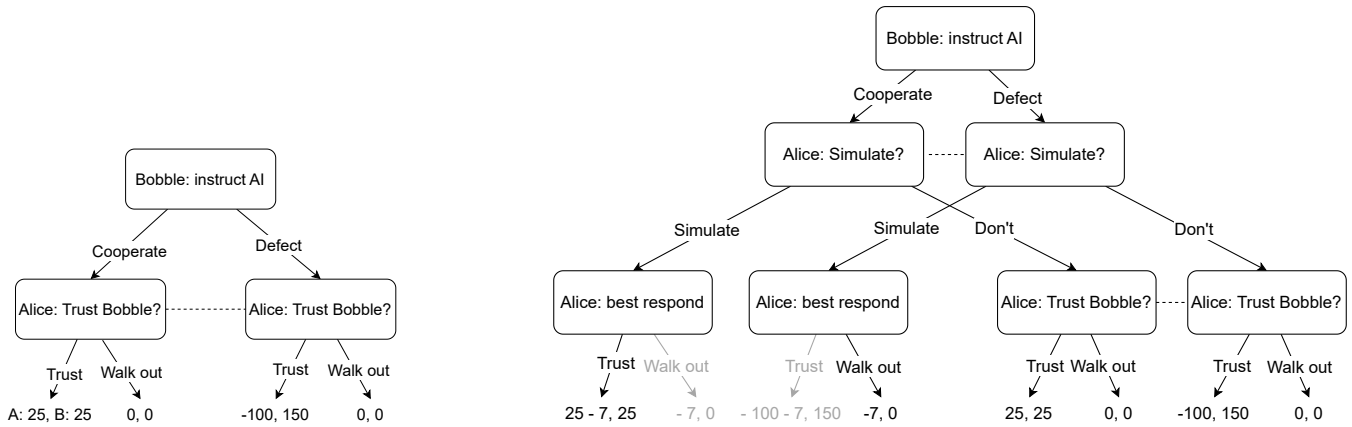


Figure 1: The underlying trust game TG (left) and the corresponding simulation game TG_{sim} (right).

on hiring a programmer, to *simulate* the AI in a sandbox and learn whether it is going to cooperate or defect. Crucially, we assume that the AI either does not learn whether it has been simulated or is unable to react to this fact. We might hope that this will ensure that Alice and Bobble can reliably cooperate. However, perhaps Alice will try to save on the *simulation cost* and trust Bobble blindly instead — and perhaps Bobble will bet on this scenario and instruct their AI to defect.

To analyze this modified game TG_{sim}, note that when Alice simulates, the only sensible followup is to trust Bobble if and only if the simulation reveals they instructed the AI to cooperate. As a result, the normal-form representation of TG_{sim} is equivalent to the normal form of the original game TG with a single added action for Alice (Fig. 2). Analyzing TG_{sim} reveals that it has two types of Nash equilibria. In one, Bobble defects with high probability and Alice, expecting this, walks out without bothering to simulate. In the other, Bobble still sometimes defects ($\pi_{\text{B}}(\text{D}) = {}^{7}/_{100}$), but not enough to stop Alice from cooperating altogether. In response, Alice simulates often enough to stop Bobble from outright defection ($\pi_{\text{A}}(\text{S}) = 1 - {}^{25}/_{150} = {}^{5}/_{6}$), but also sometimes trusts Bobble blindly ($\pi_{\text{A}}(\text{T}) = {}^{25}/_{150} = {}^{1}/_{6}$). In expectation, this makes Alice and Bobble better off by \$16.25k, resp. \$25k relative to the (*defect, walk-out*) equilibrium.

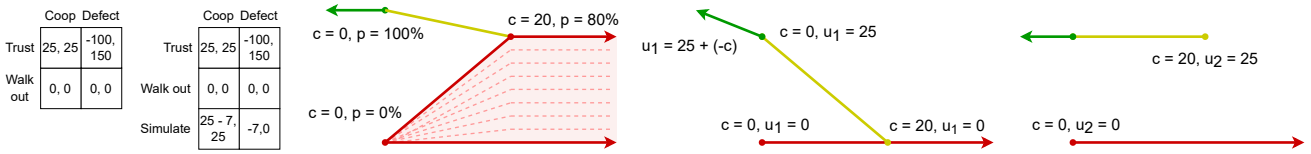
More generally, we can also consider $\text{TG}^{\text{c}}_{\text{sim}}$, a parametrization of TG_{sim} where simulation costs some $\text{c} \in \mathbb{R}$. As shown in Figure 2, the equilibria of $\text{TG}^{\text{c}}_{\text{sim}}$ are similar to the special case $\text{c} = 7$ for a wide range of c.

Generalizable properties of the trust game The analysis of Figure 2 illustrates several trends that hold more generally: First, *when simulation is subsidized, the simulation game turns into a "pure commitment game" where the simulated player is the Stackelberg leader* (Prop. 7 (i)). Conversely, *when simulation is prohibitively costly, the simulation game is equivalent to the original game* (Prop. 7 (ii)). Third, the simulation game has a finite number of breakpoints between which individual equilibria change continuously — more specifically, the simulator's strategy does not change at all while the simulated player's strategy changes linearly in c (Prop. 11). Informally speaking, *simulation games have piecewise constant/linear equilibrium trajectories*. A corollary of this observation is that *it is* not *the case that as simulation gets cheaper, the simulator must use it more and more often* (Fig. 2). Fourth, the indifference principle implies that when the simulator simulates with a nontrivial probability (i.e., neither 0 nor 1), *the value of information of simulating must be precisely equal to the simulation cost*. This also implies that *any pure NE of the original game is also a NE of the simulation game* for any $\text{c} \geq 0$ (Prop. 12). Finally, we saw that *at* $\text{c} = 0$*, the* outcome *of the simulation game becomes deterministic despite the* strategy *of the simulator being stochastic*. (For example, in the NE where Bobble always cooperates, Alice always ends up trusting him — either directly or after first simulating.) In Section 5, we show that this result holds quite generally but not always: By Theorem 2, *the equilibria of generic normal-form games with cheap simulation can be found in linear time*.

Different effects of simulation There are games where simulation behaves similarly to the Trust Game above. Indeed, in Theorem 4, we prove that *simulation leads to a strict Pareto improvement in generalized trust games* with generic payoffs (defined in Section 6). However, simulation can also affect games quite differently from what we saw so far. For example, *simulation can benefit either of the players at the cost of the other, or even be harmful to both of them*. Indeed, simulation benefits only the simulator in zero-sum games (Prop. 21), benefits only the simulated player in the Commitment Game (Fig. 3), and harms both if cooperation is predicated upon the simulated player's ability to maintain privacy (Ex. 23). In fact, there are even *cases where the Pareto optimal outcome requires simulation to be neither free nor prohibitively expensive* (Ex. 27). Finally, with multiple, incompatible ways to cooperate, a *game might admit multiple simulation equilibria* (i.e., multiple NE with $\pi_1(\text{S}) > 0$; cf. Fig. 3).

1.2 Outline

The remainder of the paper is structured as follows. First, we recap the necessary background (Section 2). In Section 3, we formally define simulation games and describe their basic properties. In Section 4, we prove several structural properties of simulation games; while these are instrumental for the



simulation cost	NE of $\mathrm{TG}^{c}_{\mathrm{sim}}$, given as the probs. of (*trust, walk-out, simulate*) and (*cooperate, defect*)		
$(-\infty, 0)$	(*simulate, cooperate*)		
$(0, 20)$	$(25/150, 0, 1 - 25/150)$ and $(1 - c/100, c/100)$	*walk-out* and $(c/25, 1 - c/25)$	(*walk-out, defect*)
$(20, \infty)$		*walk-out* and $(100/125, 25/125)$	(*walk-out, defect*)

Figure 2: Top left: The normal-form representation of the trust game from Figure 1, before and after adding simulation.
Bottom: The extremal equilibria of $\mathrm{TG}^{c}_{\mathrm{sim}}$. The non-extremal NE are precisely the convex combinations of the last two columns.
Top right: The cooperation probability and utilities under each of these NE. The non-extremal NE are light red, the dashed lines illustrate the NE trajectories from Proposition 11. Note that all the red NE (i.e., with $\pi_1(\mathrm{WO}) = 1$) yield $u_1 = u_2 = 0$.

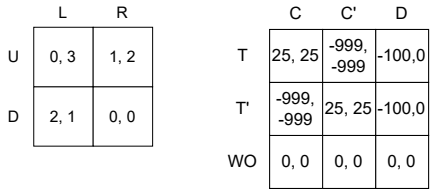


Figure 3: Left: Commitment game, where the row player prefers to *not* be able to simulate. For details, see Example 22.
Right: A variant of Trust Game with multiple simulation NE.

subsequent results, we also find them interesting in their own right. Afterwards, we analyze the computational complexity of solving simulation games (Section 5) and the effects of simulation on the players' welfare (Section 6). Finally, we review the most relevant existing work (Section 7), summarize our results, and discuss future work (Section 8). The detailed proofs are presented in the appendix (which is only available in the arXiv version of the paper).

2 Background

A two-player **normal-form game** (NFG) is a pair $\mathcal{G} = (\mathcal{A}, u)$ where $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \neq \emptyset$ is a finite set of **actions** and $u = (u_1, u_2) : \mathcal{A} \to \mathbb{R}^2$ is the **utility function**. We use P1 and P2 as shorthands for "player one" and "player two". For finite X, $\Delta(X)$ denotes the set of all probability distributions over X. A **strategy** (or policy) **profile** is a pair $\pi = (\pi_1, \pi_2)$ of **strategies** $\pi_i \in \Delta(\mathcal{A}_i)$. We denote the set of all strategies as $\Pi = \Pi_1 \times \Pi_2$. π_i is **pure** if it has support $\mathrm{supp}(\pi_i)$ of size 1. We identify such strategy with the corresponding action.

For $\pi \in \Pi$, $u_i(\pi) := \sum_{(a,b)\in\mathcal{A}} \pi_1(a)\pi_2(b)u_i(a, b)$ is the **expected utility** of π. π_1 is said to be a **best response** to π_2 if $\pi_1 \in \arg\max_{\pi_1' \in \Pi_1} u_1(\pi_1', \pi_2)$; $\mathrm{br}(\pi_2)$ denotes the set of all pure best responses to π_2. Since the **best-response utility** "$u_1(\mathrm{br}, \cdot)$" is uniquely determined by π_2, we denote it as $u_1(\mathrm{br}, \pi_2) := \max_{a\in\mathcal{A}_1} u_1(a, \pi_2)$. (The analogous definitions apply for P2 and π_1.) A **Nash equilibrium** (NE) is a strategy profile (π_1, π_2) under which each player's strategy is a best response to the strategy of the other player. We use $\mathrm{NE}(\mathcal{G})$ to denote the set of all Nash equilibria of $\mathcal{G}$.

A **pure-commitment equilibrium** (cf. [von Stackelberg, 1934]) is, informally, a subgame-perfect equilibrium of the game in which the leader first commits to a pure action, after which the follower sees the commitment and best-responds, possibly stochastically. Since our formalism will assume that P1 is the simulator, we naturally encounter situations where P2 acts as the leader. Formally, we will use $\mathrm{SE}^{\mathrm{P2}}_{\mathrm{pure}}(\mathcal{G})$ to denote all pairs (ψ_{br}, b) where the **optimal commitment** $b \in \mathcal{A}_2$ and P1's best-response policy $\psi_{\mathrm{br}} : b' \in \mathcal{A}_2 \mapsto \psi_{\mathrm{br}}(b') \in \Delta(\mathrm{br}(b')) \subseteq \Delta(\mathcal{A}_1)$ satisfy $b \in \arg\max_{b'\in\mathcal{A}_2} \mathbf{E}_{a\sim\psi_{\mathrm{br}}(b')} u_2(a, b')$.

Below, we sometimes restrict the analysis to particular classes of NFGs. To motivate the first one, recall that a property is said to be generic (typical) if it holds for almost all elements of a set [Rudin, 1987, 1.35]

Definition 1 (Generic games)**.** *We say that a statement P holds **for games with generic payoffs** if, among games whose payoffs are sampled i.i.d. from the uniform distribution over $[0, 1]$, P holds with probability 1.*

Since different joint actions in a generic-payoff NFG necessarily yield different payoffs, these games are a special case of the following more general class:

Definition 2 (No best-response utility tiebreaking)**.** *An NFG $\mathcal{G}$ is said to admit **no best-response utility tiebreaking** by* P1 *if for every pure strategy b of* P2*, any two pure best-responses $a, a' \in \mathrm{br}(b)$ give* P2 *the same utility, i.e. $u_2(a, b) = u_2(a', b) =: u_2(\mathrm{br}, b)$.*

Note that if $\mathcal{G}$ satisfies Def. 2, any pure-commitment equilibrium $\pi \in \mathrm{SE}^{\mathrm{P2}}_{\mathrm{pure}}(\mathcal{G})$ can be identified with a joint action (a, b) s.t. $a \in \mathrm{br}(b)$ and $b \in \arg\max_{b\in\mathcal{A}_2} u_2(\mathrm{br}, b)$.

3 Simulation Games

In this section, we formally define simulation games and describe their basic properties. To streamline this initial investigation of simulation games, we assume that when the simulator learns the other agent's action, they always best-respond to it — that is, they will not execute non-credible threats [Shoham and Leyton-Brown, 2008]. (However, this assumption somewhat limits the applicability of the results, and we consider moving beyond it a worthwhile future direction.)

Notation 3. For a two-player NFG $\mathcal{G}$, $\Psi_{\text{br}} := \{\psi_{\text{br}} : \mathcal{A}_2 \to \Delta(\mathcal{A}_1) \mid \forall b \in \mathcal{A}_2 : \psi_{\text{br}}(b) \in \Delta(\text{br}(b))\}$, resp. $\Psi^{\text{det}}_{\text{br}} \subset \Psi_{\text{br}}$, is the set of all stochastic, resp. pure **best-response policies**.

Definition 4 (Simulation game)**.** *(1) For a **simulation cost** $c \in \mathbb{R}$, the **simulation game** $\mathcal{G}^{c,\text{all}}_{\text{sim}}$ is defined as the NFG that is identical to $\mathcal{G}$, except that* P1 *additionally has access to "simulation" actions $s_{\psi_{\text{br}}}$, $\psi_{\text{br}} \in \Psi^{\text{det}}_{\text{br}}$, s.t. $u_1(s_{\psi_{\text{br}}}, b) := u_1(\text{br}, b) - c$, $u_2(s_{\psi_{\text{br}}}, b) := u_2(\psi_{\text{br}}(b), b)$.*

(2) For a fixed $\psi_{\text{br}} \in \Psi_{\text{br}}$, $\mathcal{G}^{c}_{\text{sim}} := \mathcal{G}^{c,\psi_{\text{br}}}_{\text{sim}}$ denotes the game where P1 *has a single additional action $s := s_{\psi_{\text{br}}}$ with $u_1(s, b) := u_1(\text{br}, b) - c$, $u_2(s, b) := \mathbf{E}_{a\sim\psi_{\text{br}}(b)} u_2(a, b)$.*

We refer to P1 as the **simulator** and to P2 as the **simulated player**. When the exact value of c is unspecified or unimportant, we write $\mathcal{G}_{\text{sim}}$ instead of $\mathcal{G}^{c}_{\text{sim}}$.

3.1 Basic Properties

In the remainder of this paper, we will only study simulation games in the context of a fixed best-response policy. To justify this decision, note that the variants (1) and (2) of Definition 4 are equivalent for most games:

Lemma 5. *If (and only if) $\mathcal{G}$ admits no best-response utility tiebreaking by* P1*, $\mathcal{G}^{c,\text{all}}_{\text{sim}}$ and $\mathcal{G}^{c}_{\text{sim}}$ are identical up to the existence of duplicate actions.*

Moreover, the problem of solving general simulation games can be reduced to the problem of solving simulation games for a fixed best-response policy:

Lemma 6. $\text{NE}(\mathcal{G}^{c,\text{all}}_{\text{sim}}) \setminus \text{NE}(\mathcal{G})$ *(i.e., the new NE introduced by adding simulation) can be written as a disjoint union* $\dot{\bigcup}_{\psi_{\text{br}}\in\Psi_{\text{br}}} \text{NE}(\mathcal{G}^{c,\psi_{\text{br}}}_{\text{sim}}) \setminus \text{NE}(\mathcal{G})$.

The next observation we make (Proposition 7) is that if simulation is too costly, then it is never used and the simulation game $\mathcal{G}_{\text{sim}}$ becomes strategically equivalent to the original game $\mathcal{G}$. Conversely, if simulation is subsidized (i.e., a negative simulation cost), then P1 will always use it, which effectively turns $\mathcal{G}_{\text{sim}}$ into a pure-commitment game with P2 moving first. (The situation is similar when simulation is free but not subsidized, except that this allows for additional equilibria where the simulation probability is less than 1.)

Proposition 7 (Equilibria for extreme simulation costs)**.** *In any simulation game $\mathcal{G}_{\text{sim}}$, we have:*

(i) For $c < 0$, simulating is a strongly dominant action. In particular, $\text{NE}(\mathcal{G}^{c}_{\text{sim}}) \subseteq \text{SE}^{\text{P2}}_{\text{pure}}(\mathcal{G})$.[1]

(ii) For $c > \max_{a\in\mathcal{A}_1, b\in\mathcal{A}_2} u_1(a, b) - \max_{\pi_1\in\Pi_1}\min_{\pi_2\in\Pi_2} u_1(\pi_1, \pi_2)$, s is a strictly dominated action. In particular, $\text{NE}(\mathcal{G}^{c}_{\text{sim}}) = \text{NE}(\mathcal{G}_{\text{sim}})$.

3.2 Information-Value of Simulation

The following definition measures the extra utility that the simulator can gain by using the knowledge of the other player's strategy:

Definition 8 (Value of information of simulation)**.** *The **value of information of simulation** for $\pi_2 \in \Pi_2$ is*

$$\mathit{VoI}_s(\pi_2) := \left(\sum\nolimits_{b\in\mathcal{A}_2} \pi_2(b) \max_{a\in\mathcal{A}_1} u_1(a, b)\right) - \max_{\pi_1\in\Pi_1} u_1(\pi_1, \pi_2).$$

Lemma 9. $\forall \pi_2 : u_1(s, \pi_2) = u_1(\text{br}, \pi_2) + \mathit{VoI}_s(\pi_2) - c$.

Lemma 9 implies that $\text{VoI}_s(\pi_2)$ always lies between 0 and the difference between maximum possible u_1 and P1's maxmin value. Moreover, to make P1 simulate with a non-trivial probability, P2 needs to pick a strategy whose value of information is equal to the simulation cost:

Lemma 10 (VoI_s is equal to simulation cost)**.** *(1) For any $\pi \in \text{NE}(\mathcal{G}^{c}_{\text{sim}})$, we have $\pi_1(s) \in (0, 1) \implies \mathit{VoI}_s(\pi_2) = c$. (2) Moreover, unless $\mathcal{G}$ admits multiple optimal commitments of* P2 *that do not have a common best-response, any $\pi \in \text{NE}(\mathcal{G}^{0}_{\text{sim}})$ has $\mathit{VoI}_s(\pi_2) = 0$.*

(Where, in (2), a set of actions having a common best-response means that $\bigcap_{b\in B} \text{br}(b) \neq \emptyset$.)

4 Structural Properties

We now review several structural properties that appear in simulation games because of the special nature of the simulation action.These results will prove instrumental when determining the complexity of simulation games (Sec. 5) and predicting the impact of simulation on the players' welfare (Sec. 6). Moreover, we find these results interesting in their own right.

The first of these properties is that a change of the simulation cost *typically* results in a very particular change in a Nash equilibrium of the corresponding game: The strategy of the simulating player (P1) doesn't change at all, while the simulated player's strategy changes linearly. However, to be technically accurate, we need to make two disclaimers. First, there is a finite number of "atypical" values of c, called breakpoints, where the nature of the NE strategies changes discontinuously.[2] Second, there can be multiple equilibria, which complicates the formal description of the result.

Proposition 11 (Trajectories of simulation NE are piecewise constant/linear)**.** *For every $\mathcal{G}$, there is a finite set of simulation-cost breakpoint values $-\infty = e_{-1} < 0 = e_0 < e_1 < \cdots < e_k < e_{k+1} = \infty$ such that the following holds: For every $c_0 \in (e_l, e_{l+1})$ and every $\pi^{c_0} \in \text{NE}(\mathcal{G}^{c_0}_{\text{sim}})$, there is a linear mapping $t_2 : c \in [e_l, e_{l+1}] \mapsto \pi^{c}_2 \in \Pi_2$ such that $t_2(c_0) = \pi^{c_0}_2$ and $(\pi^{c_0}_1, t_2(c)) \in \text{NE}(\mathcal{G}^{c}_{\text{sim}})$ for every $c \in [e_l, e_{l+1}]$.*

Since we were not able to find any existing result that would immediately imply this proposition, we provide our own proof in the appendix. However, a related result in the context of parameterized linear programming appears in [Adler and Monteiro, 1992, Prop. 2.3]. As an intuition for why this result holds, recall that in an equilibrium, each player uses a strategy that makes the other player indifferent between the actions in their support. Since P2's payoffs are not affected by c, P1 should keep their strategy constant to keep P2 indifferent, even when

[1] If we allowed P1 to consider all possible best-response policies, $\text{NE}(\mathcal{G}^{c}_{\text{sim}}) \subseteq \text{SE}^{\text{P2}}_{\text{pure}}(\mathcal{G})$ would turn into equality.

[2] While all of the non-breakpoint equilibria extend to the corresponding breakpoints as limits (Definition 13), the breakpoints might also admit additional non-limit equilibria, typically convex combinations of the limits (cf. Figure 2).

$\subset$ changes. Similarly, increasing $\subset$ linearly decreases P1's payoff for the simulate action, so P2 needs to linearly adjust their strategy to bring P1's payoffs back into equilibrium.

A particular corollary of Proposition 11 is that while one might perhaps expect simulation will gradually get used more and more as it gets more affordable, this is in fact not what happens — instead, the simulation rate is dictated by the need to balance the unchanging tradeoffs of the other player.

The second structural property of simulation games is the following refinement of Proposition 7:

Proposition 12 (Gradually recovering the NE of $\mathcal{G}$)**.** *Let π be a NE of $\mathcal{G}$. Then π, as a strategy in $\mathcal{G}^{\subset}_{\text{sim}}$ with $\pi_1(S) := 0$, is a NE precisely when $\subset \geq VoI_S(\pi_2)$.*

In particular, $VoI_S(\pi_2)$ is a breakpoint of $\mathcal{G}$.

Together, these two results imply that with $\subset = 0$, $\mathcal{G}^0_{\text{sim}}$ may have no NE in common with $\mathcal{G}$. As we increase $\subset$, the NE of $\mathcal{G}$ gradually appear in $\mathcal{G}^{\subset}_{\text{sim}}$ as well, while the simulation equilibria of $\mathcal{G}_{\text{sim}}$ (i.e., those with $\pi_1(\mathtt{S}) > 0$) gradually disappear, until eventually $\text{NE}(\mathcal{G}^{\subset}_{\text{sim}}) = \text{NE}(\mathcal{G})$.

4.1 Equilibria for Cheap Simulation

By combining the concept of value of information with the piecewise constancy/linearity of simulation equilibria, we are now in a position to give a more detailed description of Nash equilibria of games where simulation is cheap. First, we identify the equilibria of $\mathcal{G}_{\text{sim}}$ with $\subset = 0$ that might be connected to the equilibria for $\subset > 0$:

Definition 13 (Limit equilibrium of $\mathcal{G}_{\text{sim}}$)**.** *A policy profile π^0 is a **limit equilibrium** (at $\subset = 0$) of $\mathcal{G}_{\text{sim}}$ if it is a limit of some $\pi^{\subset_n} \in \text{NE}(\mathcal{G}^{\subset_n}_{\text{sim}})$ where $\subset_n \to 0_+$.*

As witnessed by the Trust Game (and Table 2 in particular), not every NE of $\mathcal{G}^0_{\text{sim}}$ is a limit equilibrium. Note that this definition automatically implies a stronger condition:

Lemma 14. *For any limit equilibrium π^0 of $\mathcal{G}_{\text{sim}}$, there is some $e > 0$ and π^e_2 such that for every $\subset \in [0, e]$, $(\pi^0_1, (1 - \frac{\subset}{e})\pi^0_2 + \frac{\subset}{e}\pi^e_2)$ is a NE of $\mathcal{G}^{\subset}_{\text{sim}}$.*

The following result shows that cheap-simulation equilibria have a very particular structure. Informally, every such NE corresponds to a "baseline" limit equilibrium π^{B} and P2's "deviation policy" π^{D}_2. As the simulation cost increases, P2 gradually deviates away from their baseline, which forces P1 to randomize between their baseline and simulating. While the technical formulation can seem daunting, all of the conditions in fact have quite intuitive interpretations that can be used for locating the simulation equilibria of small games by hand.

Lemma 15 (Structure of cheap-simulation equilibria)**.** *Let $\subset_0 \in (0, e_1)$ and suppose that $\mathcal{G}$ admits no best-response utility tiebreaking by* P1*. Then any $\pi \in \text{NE}(\mathcal{G}^{\subset_0}_{\text{sim}})$ with $\pi_1(S) \in (0, 1)$ is of the form $\pi = (\pi_1, \pi^{\subset_0}_2)$, where*

$$\begin{aligned} \pi_1 &= (1 - \pi_1(S)) \cdot \pi^B_1 + \pi_1(S) \cdot S \\ \pi^{\subset}_2 &= (1 - \alpha\subset) \cdot \pi^B_2 + \alpha\subset \cdot \pi^D_2, \quad \alpha > 0, \end{aligned}$$

and the following holds:

(i) For every $\subset \in [0, e_1]$, $(\pi_1, \pi^{\subset}_2) \in \text{NE}(\mathcal{G}^{\subset}_{\text{sim}})$.

*(ii) $\pi^B \in \Pi$ is some **baseline policy** that satisfies:*

(B1) every action in the support of π^B_1 is a best-response to every action from $\text{supp}(\pi^B_2)$*;*

(B2) every action in the support of π^B_2 is an optimal commitment by P2 *conditional on* P2 *only using strategies that satisfy (B1).*

*(iii) $\pi^D_2 \in \Pi_2$ is some **deviation policy** that satisfies:*

(D1) No $a \in \text{supp}(\pi^B_1)$ lies in $\text{br}(d)$ *for all $d \in \text{supp}(\pi^D_2)$.*

(D2) Every $d \in \text{supp}(\pi^D_2)$ satisfies one of

$$\begin{aligned} u_2(\pi^B_1, d) &> u_2(\pi^B) > u_2(\text{br}, d) && (D^{>}_2) \\ u_2(\pi^B_1, d) &= u_2(\pi^B) = u_2(\text{br}, d) && (D^{=}_2) \\ u_2(\pi^B_1, d) &< u_2(\pi^B) < u_2(\text{br}, d). && (D^{<}_2) \end{aligned}$$

(D3) If $d \in \text{supp}(\pi^D_2)$ satisfies $(D^{>}_2)$, resp. $(D^{<}_2)$, it maximizes the attractiveness ratio *r_d, resp. r^{-1}_d*

$$\frac{u_2(\pi^B_1, d') - u_2(\pi^B)}{u_2(\pi^B) - u_2(\text{br}, d')} \text{ resp. } \frac{u_2(\text{br}, d') - u_2(\pi^B)}{u_2(\pi^B) - u_2(\pi^B_1, d')}$$

among all $d' \in \mathcal{A}_2$ that satisfy $(D^{>}_2)$, resp. $(D^{<}_2)$.

In a generic game, these conditions even imply that both the baseline and deviation policies are pure.

Theorem 1 (Equilibria with binary supports)**.** *Let $\mathcal{G}$ be a game with generic payoffs and $\subset \in (0, e_1)$. Then all NE of $\mathcal{G}^{\subset}_{\text{sim}}$ are either pure or have supports of size two.*

5 Computational Aspects

We now investigate the difficulty of solving simulation games. Since many of the results hold for multiple solution concepts, we formulate them using the phrase "solving a game", with the understanding that this refers to either finding all Nash equilibria, or a single NE, or a single NE with a specific property (e.g., one with the highest social welfare). For a specific game $\mathcal{G}$, we will also use $-\infty < 0 < e_1 < \cdots < e_k < \infty$ to denote the breakpoints of $\mathcal{G}_{\text{sim}}$ (given by Proposition 11).

As an upper bound on the complexity of solving simulation games, their definition immediately yields that:

Proposition 16 (Simulation games are no harder than general games)**.** *Solving $\mathcal{G}^{\subset}_{\text{sim}}$ is at most as difficult as solving a normal-form game where* P1 *has one more action than in $\mathcal{G}$.*

For extreme values of $\subset$, Prop. 7 implies the following:

Proposition 17 (Solving $\mathcal{G}_{\text{sim}}$ for extreme $\subset$)**.** *(i) For $\subset < 0$, the time complexity of solving $\mathcal{G}^{\subset}_{\text{sim}}$ is $O(|\mathcal{A}|)$.*
(ii) For $\subset > e_k$, the time-complexity of solving $\mathcal{G}^{\subset}_{\text{sim}}$ is the same as the time-complexity of solving $\mathcal{G}$.

In contrast with Proposition 17 (ii), finding the equilibria at low simulation costs is straightforward if we restrict our attention to generic-payoff NFGs:

Theorem 2 (Cheap-simulation equilibria in generic games)**.** *Let $\mathcal{G}$ be a NFG with generic payoffs and $\subset \in (0, e_1)$. Then the time complexity of finding all equilibria of $\mathcal{G}^{\subset}_{\text{sim}}$ is $O(|\mathcal{A}|)$.*

Finally, it is also generally difficult to determine whether simulation is beneficial or not:

Theorem 3. *For a general NFG $\mathcal{G}$ and $\subset \in \mathbb{R}$, it is co-NP-hard to determine whether there is $\pi \in \text{NE}(\mathcal{G}^{\subset}_{\text{sim}}) \setminus \text{NE}(\mathcal{G})$ s.t. $\forall \rho \in \text{NE}(\mathcal{G}) : u_1(\pi) \geq u_1(\rho)$ & $u_2(\pi) \geq u_2(\rho)$.*

6 Effects on Players' Welfare

As we saw in Theorem 3, there is no simple method for determining whether introducing simulation into a general game will be socially beneficial. However, this does not rule out the possibility of identifying particular sub-classes of games where simulation is useful or harmful. We now first confirm the hypothesis that simulation is beneficial in settings where the only obstacle to cooperation is the missing trust in the simulated player. We then give specific examples to illustrate that in general games, simulation can also benefit either player at the cost of the other, or even be harmful to both.

6.1 Simulation in Generalized Trust Games

We now show that when the *only* obstacle to cooperation is the lack of trust in the possibly-simulated player, simulation enables equilibria that improve the outcome for both players.

Definition 18 (Generalized trust games)**.** *A game $\mathcal{G}$ is said to be a **generalized trust game** if any pure-commitment equilibrium (where* P2 *is the leader) is a strict Pareto improvement over any $\pi \in \text{NE}(\mathcal{G})$.*

Theorem 4 (Simulation in trust games helps)**.** *Let $\mathcal{G}$ be a generalized trust game that admits no best-response utility tiebreaking by* P1*. Then for all sufficiently low c, $\mathcal{G}^c_{\text{sim}}$ admits a Nash equilibrium with $\pi_1(S) > 0$ that is a strict Pareto improvement over any NE of $\mathcal{G}$.*

Proof sketch. We construct a NE where P2 mixes between their optimal commitment b (from the pure-commitment equilibrium corresponding to $\mathcal{G}$) and some deviation d while P1 mixes between their best-response to b and simulating. We show that (a, b) forms the baseline policy of this simulation equilibrium, which implies that as $c \to 0_+$, this NE eventually strictly Pareto-improves any NE of $\mathcal{G}$. (And the fact that (a, b) cannot be a NE of $\mathcal{G}$ ensures the existence of a suitable d.) □

6.2 Simulation in General Games

We now investigate the relationship between simulation cost and the players' payoffs in *general* games. We start by listing the two general trends that we are aware of.

The first of the general results is that for the extreme values of c, the situation is always predictable: For $c < 0$, P1 always simulates (Prop. 7) and making simulation cheaper will increase their utility without otherwise affecting the outcome. Similarly, when c is already so high that P1 never simulates, any further increase of c makes no additional difference.

Second, if P2 could choose the value of c, they would generally be indifferent between all the values within a specific interval (e_i, e_{i+1}). Indeed, this follows from Proposition 11, which implies that P2's utility remains constant between any two breakpoints of $\mathcal{G}_{\text{sim}}$.

The Examples 19-23 illustrate that the players might both agree and disagree about their preferred value of c, and this value might be both low and high.

Example 19 (Both players prefer cheap simulation)**.** In the Alice and Bobble game from Figure 2, each player's favoured NE exists for $c = 0$.

Example 20 (Only simulator prefers cheap simulation)**.** Consider the "unfair guess-the-number game" where each player picks an integer between 1 and N. If the numbers match, P2 pays 1 to P1. Otherwise, P1 pays 1 to P2. In this game, P2 clearly prefers simulation to be prohibitively costly while P1 prefers as low c as possible.

In fact, Example 20 extends to all zero-sum games:

Proposition 21. *If a zero-sum $\mathcal{G}$ has NE utilities $(v, -v)$, then $\forall c\, \forall \pi \in \text{NE}(\mathcal{G}^c_{\text{sim}})$: $u_1(\pi) \geq v$, $u_2(\pi) \leq -v$.*

Example 22 (Only simulator prefers expensive simulation)**.** In the commitment game (Figure 3), introducing free simulation creates a second NE in which P1 is strictly worse off and stops the original NE from being trembling-hand perfect. If simulation were subsidized, the original simulator-preferred NE would disappear completely. (In fact, with $c > 0$ that is not prohibitively costly, the situation is similar to the $c = 0$ case.) In summary, this shows that simulation can hurt the simulator, even when using it is free (or even subsidized) and voluntary.

Example 23 (Both players prefer expensive simulation)**.** Consider a Joint Project game where P1 proposes that P2 collaborates with them on a startup. If P2 accepts, their business will be successful, yielding utilities $u_1 = u_2 = 100$. P2 then picks a secure password ($pw \in \{1, \dots, 26\}^4$) and puts their profit in a savings account protected by that password. Finally, P1 can either do nothing or try to guess P2's password ($g \in \{1, \dots, 26\}^4$) and steal their money. Successfully guessing the password would result in utilities $u_1 = 200$, $u_2 = -10$, where the -10 comes from opportunity costs. However, if P1 guesses wrong, they will be caught and sent to jail, yielding utilities $u_1 = -999$, $u_2 = 123$ [Smith *et al.*, 2009].

Without simulation, the NE of this game is for the players to collaborate and for P1 to not attempt to guess the password. However, with cheap enough simulation, P1 would simulate P2's choice of password and steal their money — and P2, expecting this, would not agree to the collaboration in the first place. As a result, both players would prefer simulation to be prohibitively expensive.

Example 24 (The preferences depend on equilibrium selection)**.** Consider various mixed-motive games such as the Threat Game (e.g., [Clifton, 2020, Sec. 3-4]), Battle of the Sexes, or Chicken (e.g., [Shoham and Leyton-Brown, 2008]). Generally, these games have one pure NE that favours P1, a second pure NE that favours P2, and a mixed NE that is strictly worse than either of the pure equilibria for both P1 and P2. By introducing subsidized simulation into such a game, we eliminate both the simulator-favoured pure NE and the dispreferred mixed NE. This can be bad, neutral, or even good news for the simulator, depending on which of the NE would have been selected in the original game. Somewhat relatedly, introducing subsidized simulation destroys the suboptimal equilibria in Stag Hunt and Coordination Game (e.g., [Shoham and Leyton-Brown, 2008]).

Beyond the examples above, players might even prefer neither $c = 0$ nor $c = \infty$ but rather something inbetween:

Example 25 (The preferred c is non-extreme)**.** Informally, the underlying idea behind the example is that the game should

have the potential for a positive-sum interaction, but also be unfair towards P1 if they never simulate and unfair towards P2 if P1 always simulates. If we then give each player the option to opt out, the only way either of the players can profit is if simulation is neither free nor prohibitively expensive. For a detailed proof, see Example 27 in Appendix A.

7 Related Work

In terms of the formal framework, our work is closest to the literature on games with commitment [Conitzer and Sandholm, 2006; von Stengel and Zamir, 2010]. This is typically modelled as a Stackelberg game [von Stackelberg, 1934], where one player commits to a strategy while the other player only selects their strategy after seeing the commitment. In particular, [Letchford *et al.*, 2014] investigates how much the committing player can gain from committing. Commitment in a Stackelberg game is always observed. (An exception is [Korzhyk *et al.*, 2011], which assumes a fixed chance of commitment observation.) In contrast, the simulation considered in this paper would correspond to a setting where (1) one player pays for having *the other player* make a (pure) commitment and (2) the latter player does not know whether their commitment is observed, as the probability of it being observed is a parameter controlled by the observer. Ultimately, these differences imply that the Stackelberg game results are highly relevant as inspiration, but they are unlikely to have immediate technical implications for our setting (except for when $c < 0$).

In terms of motivation, the setting that is the closest to our paper is open-source game theory and program equilibria [McAfee, 1984; Howard, 1988; Rubinstein, 1998], [Tennenholtz, 2004, Sec. 10.4]. In program games, two (or more) players each choose a program that will play on their behalf in the game, and these programs can read each other. To highlight the connection to the present paper, note that one approach to attaining cooperative play in this formalism is to have the programs simulate each other [Oesterheld, 2019]. The setting of the program equilibrium literature differs from ours in two important ways. First, the program equilibrium literature assumes that both players have access to the other player's strategy. (Much of the literature addresses the difficulties of mutual simulation or analysis, e.g., see [Barasz *et al.*, 2014; Critch, 2019; Critch *et al.*, 2022; Oesterheld, 2022] in addition to the above.) Second, with the exception of time discounting [Fortnow, 2009], the program equilibrium formalism assumes that access to the other player's code is without cost.

Another approach to simulation is game theory with translucent players [Halpern and Pass, 2018]. This framework assumes that the players tentatively settle on some strategy from which they can deviate, but doing so has some chance of being visible to the other player. In our terminology, this corresponds to a setting where each player always performs free but unreliable simulation of the other player.

8 Discussion

Summary In this paper, we considered how the traditional game-theoretic setting changes when one player obtains the ability to run an accurate but costly simulation of the other. We established some basic properties of the resulting simulation games. We saw that (between breakpoint values of which there can be only finitely many), their equilibria change piecewise constantly/linearly (for P1/P2) with the simulation cost. Additionally, the value of information of simulating is often equal to the simulation cost. These properties had strong implications for the equilibria of games with cheap simulation and allowed us to prove several deeper results. Our initial hope was that simulation could counter a lack of trust — and this turned out to be true. However, we also saw that the effects of simulation can be ambiguous, or even harmful to both players. This suggests that before introducing simulation to a new setting (or changing its cost), one should determine whether doing so is likely to be beneficial or not. Fortunately, our analysis revealed that for the very general class of normal-form games with generic payoffs, this can be done cheaply.

Future Work The future work directions we find particularly promising are the following: First, the results on generic-payoff NFGs cover the normal-form representations of some, but not all, extensive-form games. Extending these results to EFGs thus constitutes a natural next step. Second, we saw that the cost of simulation that results in the socially-optimal outcome varies between games. It might therefore be beneficial to learn how to tailor the simulation cost to the specific game, and to what value. Third, we assumed that simulation predicts not only the simulated agent's policy, but also the result of any of their randomization — i.e., their precise action. Whether this assumption makes sense depends on the precise setting, but in any case, by considering mixtures over *behavioral* strategies [Halpern and Pass, 2021], it might be possible to go beyond this assumption while recovering most of our results. Finally, our work assumes that simulation is perfectly reliable, captures all parts of the other agent, and is only available to one agent but not the other. Ultimately, it will be necessary to go beyond these assumptions. We hope that progress in this direction can be made by developing a framework that encompasses both our work and some of the formalisms discussed in Section 7 (and in particular the work on program equilibria).

Limitations The simulation approach to cooperation has various limitations. Apart from the obstacles implied by the future work above, there is the issue of making sure that the agent we are simulating is the same as the agent we end up interacting with — for example, the other party might try to feed us fake source code, or change it after sharing it. Moreover, the simulated party needs to be willing to its policy, which might be in tension with retaining privacy, trade secrets, etc. Finally, the simulation approach relies on the simulated agents being unable to differentiate between simulation and reality. This might be difficult to achieve as the relevant situations become more complicated and AI agents grow in capability.

Acknowledgments

We are grateful to Emanuel Tewolde for pointing out the connection between Proposition 11 and parametrized linear programming, Zuzana Kovarikova for help with Lemma 14, Lewis Hammond for discussions and feedback on an earlier version of the text, and Emin Berker for feedback on the camera-ready

version. We would also like to thank an anonymous IJCAI reviewer for their suggestions regarding the definition of simulation games and inspiring questions regarding the difficulty of determining the usefulness of simulation. We thank the Cooperative AI Foundation, Polaris Ventures (formerly the Center for Emerging Risk Research) and Jaan Tallinn's donor-advised fund at Founders Pledge for financial support.

References

[Adler and Monteiro, 1992] Ilan Adler and Renato DC Monteiro. A geometric view of parametric linear programming. *Algorithmica*, 8(1):161–176, 1992.

[Barasz *et al.*, 2014] Mihaly Barasz, Paul Christiano, Benja Fallenstein, Marcello Herreshoff, Patrick LaVictoire, and Eliezer Yudkowsky. Robust cooperation in the prisoner's dilemma: Program equilibrium via provability logic. *arXiv preprint arXiv:1401.5577*, 2014.

[Berg *et al.*, 1995] Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[Clifton, 2020] Jesse Clifton. Cooperation, conflict, and transformative artificial intelligence: A research agenda. *Effective Altruism Foundation, March*, 4, 2020.

[Conitzer and Sandholm, 2006] Vincent Conitzer and Tuomas Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 82–90, 2006.

[Conitzer and Sandholm, 2008] Vincent Conitzer and Tuomas Sandholm. New complexity results about nash equilibria. *Games and Economic Behavior*, 63(2):621–641, 2008.

[Critch *et al.*, 2022] Andrew Critch, Michael Dennis, and Stuart Russell. Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory. *arXiv preprint arXiv:2208.07006*, 2022.

[Critch, 2019] Andrew Critch. A parametric, resource-bounded generalization of Löb's theorem, and a robust cooperation criterion for open-source game theory. *Journal of Symbolic Logic*, 84(4):1368–1381, 12 2019.

[Fortnow, 2009] Lance Fortnow. Program equilibria and discounted computation time. In *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pages 128–133, 2009.

[Halpern and Pass, 2018] Joseph Y Halpern and Rafael Pass. Game theory with translucent players. *International Journal of Game Theory*, 47(3):949–976, 2018.

[Halpern and Pass, 2021] Joseph Y Halpern and Rafael Pass. Sequential equilibrium in games of imperfect recall. *ACM Transactions on Economics and Computation*, 9(4):1–26, 2021.

[Howard, 1988] J. V. Howard. Cooperation in the prisoner's dilemma. *Theory and Decision*, 24:203–213, 5 1988.

[Korzhyk *et al.*, 2011] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. Solving stackelberg games with uncertain observability. In *AAMAS*, pages 1013–1020, 2011.

[Letchford *et al.*, 2014] Joshua Letchford, Dmytro Korzhyk, and Vincent Conitzer. On the value of commitment. *Autonomous Agents and Multi-Agent Systems*, 28(6):986–1016, 2014.

[McAfee, 1984] R. Preston McAfee. Effective computability in economic decisions. https://vita.mc4f.ee/PDF/EffectiveComputability.pdf, 1984. Accessed: 2022-12-14.

[Oesterheld, 2019] Caspar Oesterheld. Robust program equilibrium. *Theory and Decision*, 86(1):143–159, 2 2019.

[Oesterheld, 2022] Caspar Oesterheld. A note on the compatibility of different robust program equilibria of the prisoner's dilemma. *arXiv preprint arXiv:2211.05057*, 2022.

[Rubinstein, 1998] Ariel Rubinstein. *Modeling Bounded Rationality*. Zeuthen Lecture Book Series. The MIT Press, 1998.

[Rudin, 1987] Walter Rudin. *Real and Complex Analysis*. Mathematics. McGraw.. HiII, 1987.

[Shoham and Leyton-Brown, 2008] Yoav Shoham and Kevin Leyton-Brown. *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

[Smith *et al.*, 2009] Richard H Smith, Caitlin AJ Powell, David JY Combs, and David Ryan Schurtz. Exploring the when and why of Schadenfreude. *Social and Personality Psychology Compass*, 3(4):530–546, 2009.

[Tennenholtz, 2004] Moshe Tennenholtz. Program equilibrium. *Games and Economic Behavior*, 49(2):363–373, 11 2004.

[von Stackelberg, 1934] Heinrich von Stackelberg. *Marktform und Gleichgewicht*. Springer, 1934.

[von Stengel and Zamir, 2010] Bernhard von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010.