Game Theory with Simulation in the Presence of Unpredictable Randomisation

Vojtěch Kovařík
Czech Technical University
Prague, Czech Republic
vojta.kovarik@gmail.com

Nathaniel Sauerberg
University of Texas
Austin, United States
njsauerberg@gmail.com

Lewis Hammond
University of Oxford
Oxford, United Kingdom
lewis.hammond@cs.ox.ac.uk

Vincent Conitzer
Carnegie-Mellon University
Pittsburgh, United States
conitzer@cs.cmu.edu

ABSTRACT

AI agents will be predictable in certain ways that traditional agents are not. Where and how can we leverage this predictability in order to improve social welfare? We study this question in a game-theoretic setting where one agent can pay a fixed cost to simulate the other in order to learn its mixed strategy. As a negative result, we prove that, in contrast to prior work on pure-strategy simulation, enabling mixed-strategy simulation may no longer lead to improved outcomes for both players in all so-called "generalised trust games". In fact, mixed-strategy simulation does not help in any game where the simulatee's action can depend on that of the simulator. We also show that, in general, deciding whether simulation introduces Pareto-improving Nash equilibria in a given game is NP-hard. As positive results, we establish that mixed-strategy simulation can improve social welfare if the simulator has the option to scale their level of trust, if the players face challenges with both trust and coordination, or if maintaining some level of privacy is essential for enabling cooperation.

KEYWORDS

AI agents; Simulation; Stackelberg games; Cooperative AI

ACM Reference Format:
Vojtěch Kovařík, Nathaniel Sauerberg, Lewis Hammond, and Vincent Conitzer. 2025. Game Theory with Simulation in the Presence of Unpredictable Randomisation. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

With the current pace of progress in AI, we are likely to increasingly see important interactions take place not only between humans, but also with and between AI agents [6, 7]. To ensure that the societal impact of these interactions is positive, it is important to understand the ways in which AI agents differ from humans [5]. This in turn can help us design interventions that promote socially desirable outcomes [19].

One important distinction between human and AI agents is that the behaviour of AI agents is determined by their source code, and can therefore – in certain cases – be reliably predicted [22]. This could be achieved, for example, by inspecting the AI's source code and reasoning about it, or by creating a copy of the AI and running it in a simulated environment. As these examples suggest, we will assume that predicting the AI's actions requires non-trivial effort, and is therefore associated with some *cost* [10, 15]. (Readers familiar with Stackelberg games [24] can think of this setting as one where the follower has to choose to pay some cost before they are allowed to see the leader's mixed commitment. For a more detailed discussion of related work, see Section 6.) For concreteness, this paper will discuss this general topic in terms of *simulating* the AI agent, though our results also apply to other forms of prediction.

As we will show, the ability to simulate agents before interacting with them can (provably) lead to increased trust and cooperation. More than a topic of merely theoretical interest, however, the availability of black-box access to the latest AI models and high-fidelity simulators could lead to simulation being a key tool in the safe and beneficial deployment of advanced AI agents [1, 20]. Importantly, in domains ranging from financial markets to public infrastructure, these agents will face *strategic incentives*, making it critical to understand the implications of simulation in *game-theoretic* settings.

An idealised variant of this setting was studied by Kovařík et al. [15], who assumed that the simulator is able to predict the AI's action perfectly. However, this assumption might often be unrealistic, not least because the AI might have access to a source of randomness that cannot be predicted by the simulator [14, 26]. As the following extended example shows, this difference has far-reaching consequences, which we explore in the remainder of the paper.

1.1 Illustrative Example

Alice, Bob, and his robots. Consider a setting in which Alice (player one) and Bob (player two) are due to interact in some particular situation, corresponding formally to some arbitrary two-player game $\mathcal{G}$. Instead of interacting directly, however, Bob will deploy a robot[1] that will act on his behalf. Moreover, Alice will have the option to analyse the robot, at some cost $c_{\text{sim}} > 0$. We will refer to this analysis as *simulation*.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[1] It is in no way important to the paper's results that the "robot" is physically embodied; we just use it here as evocative language.

In general, the robot can use randomness to determine which action to take. Correspondingly, we will distinguish between two types of simulation, depending on whether Alice is able to predict the robot's source of randomness or not. To capture this distinction formally, we can assume that the robot corresponds to some probability distribution σ_2 over pure strategies in $\mathcal{G}$. In *mixed-strategy simulation*, Alice learns the robot's mixed strategy σ_2. In *pure-strategy simulation*, Alice learns which pure strategy s_2 will be sampled by the robot when playing $\mathcal{G}$.

In this work, we assume that if Alice decides to simulate the robot, she will then best-respond to the revealed strategy, breaking ties in Bob's favour.

The simulation meta-game. When Alice has access to mixed-strategy simulation, Alice and Bob need to reason not only about the "base-game" $\mathcal{G}$, but also about the "simulation meta-game" $\mathcal{G}_{\text{sim}}$. In $\mathcal{G}_{\text{sim}}$, Alice must decide whether to simulate, and which strategy to use if she does not. Correspondingly, Bob must decide which robot – i.e., which mixed strategy – to select as his representative. When Bob randomises his choice, he is thus *mixing over mixed strategies*. And because Alice can simulate the robot but not Bob, Bob's overall strategy is not necessarily reducible to a single mixed strategy.[2]

Focusing on low simulation costs and strict Pareto improvements. Throughout the paper, we will assume that the simulation cost is not under the control of either of the players. However, for the purpose of interpreting the results, note that Bob in particular might be able to influence c_{sim} to some degree. For example, when Bob has a fully adversarial relationship with Alice, he might not share any details about the robot and even intentionally obfuscate its design to make the analysis harder. In contrast, when Bob wants Alice to trust him, he might make the robot easier to understand or even subsidise the simulation cost.[3] Because of this, in this paper we will primarily investigate the case where c_{sim} is *low but positive*. We will also focus on settings where simulation holds the promise of improving the outcome for *both* players. The primary candidates for such settings are games that revolve around trust, coordination, or both.

The trust game TG. The central example of a trust game is TG (Figure 1): Bob – or rather, his robot – approaches Alice with an investment opportunity. If she lends him \$100k, he will make \$40k in profit. Alice can Walk Out (WO) on Bob, terminating the game with payoffs $u_A = u_B = 0$. If Alice instead Trusts (T) Bob, he can either Defect (D) and steal Alice's money ($u_A = -100$, $u_B = 100$) or Cooperate (C), splitting the profits 50:50 with Alice ($u_A = u_B = 20$). Unfortunately, without simulation, Defect is a dominant strategy for Bob, so the only Nash equilibrium (NE) of TG is for Alice to Walk Out.

Pure-strategy simulation in TG. In the simulation variant TG_{sim} of TG, Bob attempts to earn Alice's trust by sharing the robot's specification with her. She then has the option to Simulate (sim) the robot for $c_{\text{sim}} = \$2\text{k}$ in order to learn which strategy it will employ. One might hope that this would reliably allow cooperation between Alice and Bob, yielding payoffs $u_A = 20 - 2 = 18$, $u_B = 20$. Unfortunately, such an outcome would not be stable, because Alice would be tempted to increase her profits by Trusting Bob blindly (and thus saving the simulation cost). This, in turn, creates a temptation for Bob to submit a robot that will Defect on Alice.

	Cooperate	Defect
Trust	20, 20	−100, 100
Walk Out	0, 0	0, 0

	Cooperate	Defect
Full Trust	20, 20	−100, 100
Partial Trust	10, 10	−25, 25
Walk Out	0, 0	0, 0

Figure 1: Trust game TG and its partial-trust extension PTG.

Pure-strategy simulation in generalised trust games. Kovařík et al. [15] show that in the pure-strategy simulation game $\text{TG}_{\text{p-sim}}$, it is an equilibrium for Alice to mix between Trust and Simulate, and for Bob to mix between (robots that) Cooperate and Defect. While this simulation equilibrium is not optimal in terms of social welfare (compared to what could be obtained in a world without strategic constraints), it constitutes a strict improvement over the original equilibrium for both Alice and Bob. The authors then show that a similar result holds in any *generalised* trust game (which they define as a game where giving Bob the ability to make credible pure-strategy commitments is guaranteed to strictly improve the utility for both players compared to any NE of the original game).

Mixed-strategy simulation is useless in TG. Unfortunately, this result no longer holds in the mixed-strategy simulation variant $\text{TG}_{\text{m-sim}}$. To see why, imagine that Bob is about to submit a robot which cooperates with Alice 100% of the time. He will then be tempted to replace this robot by one that only cooperates 99% of the time. If Alice simulates the robot, surely she will not Walk Out on him just because of the 1% defection chance; after all, her expected utility for Trusting is still positive. In fact, this reasoning shows that Bob can safely set the cooperation rate to $\frac{100+2}{100+20} = 85\%$, the lowest possible value where Alice still recoups all of her simulation costs. However, Bob can go even further. He can reason that once Alice has already simulated, she will treat the simulation cost as a sunk cost. Taken all together, this means that no matter what Alice does, he might as well replace all of his "100% cooperate robots" by "83.5% cooperate robots" (since a cooperation rate of $\frac{100}{100+20}$ is the lowest he can go while still giving Trust a positive expected value). Unfortunately, at this point, Alice no longer makes enough profit to recoup c_{sim}, so her only sensible options are to Trust Bob blindly or Walk Out. This effectively brings Alice and Bob back to TG *without simulation*, and the only Nash equilibrium of that game is for Alice to Walk Out.

Mixed-strategy simulation can help if Alice has good alternatives. In light of this negative result, one might wonder whether there are

[2] If Bob was certain that Alice has access to pure-strategy simulation, he might decide to use robots that play deterministically, since doing otherwise confers no benefit. This would make the additional level of randomisation unnecessary.

[3] However, in practice, it seems unlikely that Bob could achieve $c_{\text{sim}} = 0$ or even $c_{\text{sim}} < 0$. This is because even if Bob makes simulation as simple as possible and pays Alice to simulate, she will always be tempted to put lower than maximum effort into her analysis (to save effort or keep some of the subsidy for herself).

any games where mixed-strategy simulation is useful. Fortunately, it turns out that there are. To see this, consider an extension PTG (Figure 1) of the earlier trust game scenario TG where Alice has the additional option to trust Bob only *partially* (PT), *with the robot being unable to differentiate between* Partial *and* Full Trust. For example, she could secretly register her business in jurisdictions with higher taxes but more secure banking infrastructure, which would decrease the overall profits ($u_A(\text{PT}, \text{C}) = u_B(\text{PT}, \text{C}) = 10$) but reduce the robot's ability to steal from her ($u_A(\text{PT}, \text{D}) = -25$, $u_B(\text{PT}, \text{D}) = 25$). Bob could now repeat the same reasoning as before, concluding that any "100% cooperate robot" can be safely replaced by a "99% cooperate robot". However, he will now have to stop at $\frac{100-25}{20-10} \doteq 88.3\%$. This is because below this value, Alice's best response will switch from Full to Partial Trust, decreasing Bob's utility. (This shows the importance of $u_2(\text{FT}, \text{C})$ being higher than $u_2(\text{PT}, \text{C})$.) We can verify that the mixed-strategy simulation version of PTG has an NE where Bob mostly submits a "88.3% cooperate robot" but sometimes replaces it by a "100% defect robot", while Alice mixes between Simulating (after which, depending on the result, she plays either Full Trust or Walk Out) and blindly playing Partial Trust. Fortunately, the frequency of Bob using the "100% defect robot" is proportional to c_{sim}, so for any sufficiently low c_{sim} (e.g., for $c_{\text{sim}} = 2$), both players end up making a profit.

1.2 Outline and Contributions

Section 2 describes the standard notation for normal-form games and covers some classic game-theoretic concepts.

In Section 3, we formally define mixed-strategy simulation games $\mathcal{G}_{\text{m-sim}}$, contrast them with pure-strategy simulation games $\mathcal{G}_{\text{p-sim}}$, and establish their basic properties. In particular, we show that while $\mathcal{G}_{\text{m-sim}}$ is technically an infinite game, it can be reduced to a normal-form game whose size is at most exponential in the size of the base-game (Prop. 3.4).

Section 4 explores the computational aspects of mixed-strategy simulation. First, it establishes an upper bound on the complexity of finding an NE of $\mathcal{G}_{\text{m-sim}}$ (Prop. 4.1). Second, while determining the exact complexity of finding an NE of $\mathcal{G}_{\text{m-sim}}$ is left as an open problem, we observe that any NE of the original game still exists as an equilibrium of the simulation game, so finding *all* NE of a simulation game is NP-hard. Third, from the design point of view, a crucial question is whether enabling mixed-strategy simulation in a particular game introduces beneficial Nash equilibria – e.g., ones that result in a Pareto-improvement, an increase in social welfare, or an improvement in the utility of a particular player (relative to the equilibria of the original game $\mathcal{G}$). We show that answering any variant of this question is, in general, NP-hard (Thm. 4.2). This implies that one should not expect to be able to find a simple description of simulation's effects in *general* games. Consequently, we find it more promising to focus on identifying specific *classes* of games where simulation has easily describable effects.

In Section 5, we investigate the effects of simulation on the players' welfare. First, we extend the negative result for TG from Section 1.1, by showing that mixed-strategy simulation cannot help in any game where the robot observes Alice's base-game strategy before acting (Thm. 5.1). We then extend the positive result for PTG from Section 1.1, by describing a general class of games where Alice can vary her level of trust and mixed-strategy simulation allows her to profitably use the second-highest level of trust (Thm. 5.4). We also prove that mixed-strategy simulation is beneficial in a class of games involving elements of both trust and coordination (Thm. 5.9). Finally, we show that there are situations – those where the simulated agent finds it important to maintain their privacy – where mixed-strategy simulation is more socially beneficial than pure-strategy simulation (Thm. 5.11).

We conclude by reviewing the most closely related work (Sec. 6) and summarising the paper's findings (Sec. 7). The full proofs and all proof sketches can be found in the arXiv version of this text.

2 BACKGROUND

For a finite set X, $\Delta(X)$ denotes the **set of all probability distributions** over X. For a probability distribution ρ, $\text{supp}(\rho)$ denotes the **support** of ρ. We use P1 and P2 as shorthands for "player one" and "player two". When there is risk of confusion about which game a given object belongs to, we add superscript notation (e.g., $u^{\mathcal{G}}$ for utility in $\mathcal{G}$).

A two-player **normal-form game** (NFG) $\mathcal{G}$ is a triplet (S_1, S_2, u) where: $S := S_1 \times S_2 \neq \emptyset$ is a set of **pure strategy profiles** (finite, unless specified otherwise) and $u = (u_1, u_2) : S \to \mathbb{R}^2$ is the **utility function**. We will typically denote the elements of S_i (pure strategies) as s_i. A **mixed strategy** σ_i is a probability distribution over pure strategies. $\Sigma_i := \Delta(S_i)$ denotes the set of all mixed strategies. Since any pure strategy s_i can be identified with the mixed strategy $\sigma_i^{s_i}$ that selects s_i with probability 1, we sometimes view pure strategies as a subset of mixed strategies. A **subgame** of $\mathcal{G}$ is any game of the form $\mathcal{G}' = (S_1', S_2', u^{\mathcal{G}})$, where $S_i' \subseteq S_i^{\mathcal{G}}$.

With a light abuse of notation, we will overload the symbol u to also denote the *expected* utilities corresponding to mixed strategies. A strategy σ_1 is said to be a **best response** to a strategy σ_2 if $\sigma_1 \in \arg\max_{\sigma_1' \in \Sigma_1} u_1(\sigma_1', \sigma_2)$. We use $\text{br}(\sigma_2)$ to denote the (non-empty) set of all *pure* best responses to σ_2. Since the utility of the best-responding player is determined by the other player's strategy, we sometimes denote it as $u_1(\text{br}(\sigma_2), \sigma_2)$. (The analogous definitions apply when the roles of P1 and P2 are reversed.) A **Nash equilibrium** is a strategy profile $\sigma = (\sigma_1, \sigma_2)$ under which each player's strategy is a best response to the strategy of the other player. $\text{NE}(\mathcal{G})$ denotes the set of all Nash equilibria in $\mathcal{G}$.

A strategy s_1 is said to be an (opponent-)**favourable best response** to σ_2 if $s_1 \in \arg\max_{t_1 \in \text{br}(\sigma_2)} u_2(t_1, \sigma_2)$. We use $\text{f-br}(\sigma_2)$ to denote the (non-empty) set of all (pure) favourable best responses to σ_2. When one player uses a favourable best response, the utilities of *both* players are determined by the other player's strategy; this allows us to denote these utilities as $u_1(\text{f-br}(\sigma_2), \sigma_2)$ and $u_2(\text{f-br}(\sigma_2), \sigma_2)$.

A **Stackelberg game** [24] is a setting where one player, the leader, commits to a mixed strategy to which the other player, the follower, best-responds. *In this paper, we assume that P2 is the Stackelberg leader and P1 is the follower,* which better fits the assumption that P1 is the simulator. Formally, a Stackelberg game $\widehat{\mathcal{G}}$ corresponding to a base-game $\mathcal{G}$ works as follows. First, the leader selects a mixed strategy $\sigma_2 \in \Sigma_2^{\mathcal{G}}$. Afterwards, the follower selects a favourable best response $s_1 \in \text{f-br}^{\mathcal{G}}(\sigma_1)$, (i.e., *breaking ties in the leader's favour*). The players then receive payoffs

$u^{\widehat{\mathcal{G}}}(s_1, \sigma_2) := u^{\mathcal{G}}(s_1, \sigma_2)$. By **Stackelberg equilibrium** (SE) of $\mathcal{G}$, we mean any NE of the Stackelberg game $\widehat{\mathcal{G}}$.

We also also consider "pure Stackelberg games" where the leader is limited to committing to a *pure* strategy $s_2 \in S_2^{\mathcal{G}}$. However, to avoid the ambiguity of "pure Stackelberg equilibrium", we will refer to these games as **pure-commitment games** and to their NE as **pure-commitment equilibria**.

A strategy profile σ is said to be a **strict Pareto improvement** over ρ if it satisfies $u_1(\sigma) > u_1(\rho)$ and $u_2(\sigma) > u_2(\rho)$. A two-player game $\mathcal{G}$ is said to be a **generalised trust game** [15] if any pure-commitment equilibrium of $\mathcal{G}$ (with P2 as the leader) is a strict Pareto improvement over any Nash equilibrium of $\mathcal{G}$. (For a prototypical example of such a $\mathcal{G}$, see Figure 1.)

3 PURE- VS. MIXED-STRATEGY SIMULATION

In this section, we formally define mixed-strategy simulation, contrast it with pure-strategy simulation, and survey the basic properties of the corresponding games.

3.1 Definitions of Simulation Games

In Section 1.1, we informally described simulation games through a scenario in which Bob (P2) selects a robot that acts on his behalf and Alice (P1) has an option to pay a fixed cost to analyse the robot prior to interacting with it. If Alice takes advantage of this option, she learns the (pure or mixed) strategy that the robot is going to employ and best-responds to it, breaking ties in Bob's favour. Otherwise, the game proceeds as usual. We now give a formal counterpart to this description, the first part of which is a reformulation of that of Kovařík et al. [15].

Definition 3.1 (Pure- and mixed-strategy simulation). *The **mixed-** and **pure-strategy simulation games** $\mathcal{G}^{c_{\text{sim}}}_{\text{m-sim}}$ and $\mathcal{G}^{c_{\text{sim}}}_{\text{p-sim}}$ (or simply $\mathcal{G}_{\text{m-sim}}$ and $\mathcal{G}_{\text{p-sim}}$) corresponding to a two-player NFG $\mathcal{G}$ and **simulation cost** $c_{\text{sim}} > 0$ are defined as the (infinite) NFGs given by:*

$$S_1^{\mathcal{G}_{\text{m-sim}}} := S_1^{\mathcal{G}} \cup \{\text{m-sim}\}, \quad S_2^{\mathcal{G}_{\text{m-sim}}} := \Sigma_2^{\mathcal{G}},$$
$$S_1^{\mathcal{G}_{\text{p-sim}}} := S_1^{\mathcal{G}} \cup \{\text{p-sim}\}, \quad S_2^{\mathcal{G}_{\text{p-sim}}} := \Sigma_2^{\mathcal{G}},$$

where m-sim *and* p-sim *are new strategies of P1, called **mixed-** and **pure-strategy simulation** in $\mathcal{G}$, defined by*

$$u_1(\text{m-sim}, \sigma_2) := u_1^{\mathcal{G}}(\text{br}^{\mathcal{G}}(\sigma_2), \sigma_2) - c_{\text{sim}}$$
$$u_2(\text{m-sim}, \sigma_2) := u_2^{\mathcal{G}}(\text{f-br}^{\mathcal{G}}(\sigma_2), \sigma_2),$$

resp.

$$u_1(\text{p-sim}, \sigma_2) := \mathbb{E}_{s_2 \sim \sigma_2} u_1(\text{br}^{\mathcal{G}}(s_2), s_2) - c_{\text{sim}}$$
$$u_2(\text{p-sim}, \sigma_2) := \mathbb{E}_{s_2 \sim \sigma_2} u_2(\text{f-br}^{\mathcal{G}}(s_2), s_2).$$

*A **simulation equilibrium** is an* NE *in which P1 simulates with non-zero probability.*

[3]The phrases pure-and mixed-strategy simulation might suggest that the object that the simulation is being applied to is a pure, resp. mixed strategy. In fact, the two types of simulation are applicable to the same objects, but they differ in the information they reveal. In other words, pure and mixed-strategy simulation differ in the type of *output* they produce, not in the type of input they accept.

To illustrate the distinction between p-sim and m-sim, imagine that Alice and Bob play a game of rock-paper-scissors. If Bob employs a single robot, `Uniform-bot`, which uses an internal random number generator to play each of the three actions with probability $\frac{1}{3}$, applying mixed-strategy simulation m-sim will only tell Alice that the robot's strategy is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (which is entirely unhelpful). In contrast, applying pure-strategy simulation p-sim to `Uniform-bot` will predict the robot's exact action, allowing Alice to win the game every single time. If Bob instead randomises between `Rock-bot`, `Paper-bot`, and `Scissors-bot`, each of which can only use a single action, both p-sim and m-sim will reveal the robot's exact action.

When the distinction between pure- and mixed-strategy simulation does not matter, we will use the colloquial term simulation.

This example also shows that in a pure-strategy simulation game $\mathcal{G}_{\text{p-sim}}$, Bob cannot gain anything by randomising over multiple mixed strategies (since all utilities only depend on the *overall* distribution over $S_2^{\mathcal{G}}$). For the purposes of formal analysis of pure-strategy simulation games, this allows us to assume that Bob's space of pure strategies in $\mathcal{G}_{\text{p-sim}}$ is limited to $S_2^{\mathcal{G}_{\text{p-sim}}} = S_2^{\mathcal{G}}$. [4]

3.2 Randomising over Mixed Strategies

Throughout the paper, and in particular in some of the proofs, it will be crucial to be able to treat "mixtures over mixed strategies" differently from a standard mixed strategies (since the two respond differently to mixed-strategy simulation, as we saw earlier). To address this issue, we will refer to probability distributions over $\Sigma_2^{\mathcal{G}}$ as **meta-strategies** and denote them by symbols such as μ_2 (or m_2 when the meta-strategy is pure, i.e., when it puts all probability mass on a single $\sigma_2 \in \Sigma_2^{\mathcal{G}}$). We will use the hat symbol to indicate that a given mixed strategy is being used as a (pure) meta-strategy. For example, Bob's above-mentioned mixed meta-strategy of uniformly randomising between a `Rock-bot`, Paper-bot, and Scissors-bot could be formally written as $\mu_1 := \frac{1}{3}\widehat{\text{R}} + \frac{1}{3}\widehat{\text{P}} + \frac{1}{3}\widehat{\text{S}}$, while the pure meta-strategy of always using the `Uniform-bot` would correspond to $m_2 := \widehat{\frac{1}{3}\text{R} + \frac{1}{3}\text{P} + \frac{1}{3}\text{S}}$.

For a mixed meta-strategy $\mu_2 \in S_2^{\mathcal{G}_{\text{m-sim}}}$ and pure base-game strategy $s_2 \in S_2^{\mathcal{G}}$, we will use $\breve{\mu}_2(s_2)$ to denote the total probability that μ_2 puts on s_2. For example, if μ_2 represents Bob using `Uniform-bot` with probability 30% and `Rock-bot` $\widehat{\text{R}}$ with the remaining 70% probability, we have $\breve{\mu}_2(\text{R}) = 0.3 \cdot \frac{1}{3} + 0.7 \cdot 1 = 0.8$.

3.3 Basic Properties of Simulation Games

While this paper focuses on the implications of mixed-strategy simulation m-sim, pure-strategy simulation p-sim will be relevant for two reasons. First, it serves as an important baseline for comparison. Second, it can be a useful source of intuitions for the properties of mixed-strategy simulation — this is because m-sim in $\mathcal{G}$ can be also be understood as p-sim in the infinite game $(S_1^{\mathcal{G}}, \Sigma_2^{\mathcal{G}}, u)$.[5]

[4]Note that the same simplification – replacing $S_2^{\mathcal{G}_{\text{m-sim}}} \coloneqq \Sigma_2^{\mathcal{G}}$ by $S_2^{\mathcal{G}}$ – cannot be valid in $\mathcal{G}_{\text{m-sim}}$. This follows from the Trust Game example in Section 1.1 (which illustrates that $\mathcal{G}_{\text{p-sim}}$ and $\mathcal{G}_{\text{m-sim}}$ can have different properties).

[5]In light of the equivalence between $\mathcal{G}_{\text{m-sim}}$ and $(S_1^{\mathcal{G}}, \Sigma_2^{\mathcal{G}}, u)_{\text{p-sim}}$, we might hope to answer questions about mixed-strategy simulation by applying existing pure-strategy simulation theory to $(S_1^{\mathcal{G}}, \Sigma_2^{\mathcal{G}}, u)$. However, this strategy turns out to be inapplicable because the prior work requires the base game $\mathcal{G}$ to be finite.

The following lemma shows that – unlike in pure-strategy simulation games – any NE of $\mathcal{G}$ is also an NE of the simulation game $\mathcal{G}_{\text{m-sim}}$. This is because if P2 puts all probability mass on a single mixed strategy σ_2, costly simulation results in lower utility for P1 than directly best-responding to σ_2.

Lemma 3.2. *Identifying $\sigma \in \Sigma^{\mathcal{G}}$ with $(\sigma_1, \widehat{\sigma}_2) \in \Sigma^{\mathcal{G}_{\text{m-sim}}}$, we have $\text{NE}(\mathcal{G}) \subseteq \text{NE}(\mathcal{G}_{\text{m-sim}})$ for any $\mathcal{G}$.*

Our primary interest in simulation is to use it as a tool for improving the outcomes for the players. What, however, ought to be our metric of success, particularly in light of Lemma 3.2? When discussing the impacts of simulation on social welfare, we will primarily be concerned with whether enabling simulation **introduces Pareto-improving Nash equilibria**. Formally, this means that *there is some $c_0 > 0$ s.t. for every $c_{\text{sim}} < c_0$, the game $\mathcal{G}^{c_{\text{sim}}}_{\text{m-sim}}$ has a Nash equilibrium μ^* in which $u(\mu^*)$ is* strictly *higher, for both players, than the utility achievable in any NE of $\mathcal{G}$* (and similarly for p-sim).

Remark 3.3. Note that the fact that enabling simulation introduces Pareto-improving NE does not necessarily preclude simulation from also introducing new NE whose utility is *lower* than some, or even all, NE of the original game. While it is worth analysing when this occurs, such equilibrium selection problems are largely beyond the scope of the present work.

Strictly speaking, $\mathcal{G}_{\text{m-sim}}$ is defined as a game where P2 has infinitely many pure strategies, which could greatly complicate its analysis. Fortunately, the following result shows that it is always enough to consider a finite number of strategies for P2.

Proposition 3.4 (Reduction to a finite strategy space)**.** *For any finite $\mathcal{G}$, there exists a finite subgame $\mathcal{G}'_{\text{m-sim}}$ of $\mathcal{G}_{\text{m-sim}}$ s.t.:*

(i) $\text{NE}(\mathcal{G}'_{\text{m-sim}}) \subseteq \text{NE}(\mathcal{G}_{\text{m-sim}})$,

(ii) $\forall \mu \in \text{NE}(\mathcal{G}_{\text{m-sim}})\ \exists \mu' \in \text{NE}(\mathcal{G}'_{\text{m-sim}}) : u(\mu') = u(\mu)$.

Proof sketch. $\Sigma_2^{\mathcal{G}}$ can be expressed as a union of (non-closed) polytopes $\text{f-br}^{-1}(s_1) := \left\{\sigma_2 \in \Sigma_2^{\mathcal{G}} \mid \text{f-br}(\sigma_2) \ni s_1\right\}$, $s_1 \in S_1^{\mathcal{G}}$. We then have $u_2^{\mathcal{G}_{\text{m-sim}}}(\text{m-sim}, \sigma_2) = u_2^{\mathcal{G}}(s_1, \sigma_2)$ whenever σ_2 lies in $\text{f-br}^{-1}(s_1)$. P2 can then recover all relevant strategies by mixing over the vertices of the closure $\overline{\text{f-br}^{-1}(s_1)}$. □

4 COMPUTATIONAL RESULTS

In this section, we investigate the difficulty of analysing mixed-strategy simulation games. From Proposition 3.4, it follows that even though $\mathcal{G}_{\text{m-sim}}$ is defined as an infinite game, it can be solved in finite time. By "solving" a game, we mean any of: (a) finding one NE; (b) finding an NE that maximises social welfare or the utility of one of the players; or (c) finding all NE payoff profiles and some NE corresponding to each.

Proposition 4.1 (Upper bound on solving $\mathcal{G}_{\text{m-sim}}$)**.** *For any $\mathcal{G}$, solving $\mathcal{G}_{\text{m-sim}}$ is at most as difficult as solving a game $\widehat{\mathcal{G}}$ with $|S^{\widehat{\mathcal{G}}}| = O(|S_1^{\mathcal{G}}|^2 \cdot 2^{|S_2^{\mathcal{G}}|})$.*

Proof sketch. The non-trivial part is the size of $S_2^{\widehat{\mathcal{G}}}$. Proposition 3.4 shows that a suitable $S_2^{\widehat{\mathcal{G}}}$ can be obtained by splitting the $(|S_2^{\mathcal{G}}| - 1)$-dimensional simplex $\Sigma_2^{\mathcal{G}}$ into $|S_1^{\mathcal{G}}|$ convex polytopes and only considering the vertices of these polytopes. Estimating the number of these vertices yields the result. □

The fact that $\text{NE}(\mathcal{G}) \subseteq \text{NE}(\mathcal{G}_{\text{m-sim}})$ trivially implies that finding *all* NE of $\mathcal{G}_{\text{m-sim}}$ is at least as difficult as finding all NE of $\mathcal{G}$. The difficulty of finding simulation equilibria of $\mathcal{G}_{\text{m-sim}}$ depends on c_{sim}. When c_{sim} is prohibitively high, solving $\mathcal{G}_{\text{m-sim}}$ is equivalent to solving $\mathcal{G}$ (since Alice never simulates). For general c_{sim}, we leave determining the exact complexity of finding simulation equilibria of $\mathcal{G}_{\text{m-sim}}$ as an open problem.

From the perspective of a designer, arguably the most important question is whether enabling simulation is likely to lead to more socially beneficial outcomes in a given game. The following result shows that this is, in general, hard to determine.

Theorem 4.2 (Determining whether simulation helps is hard)**.** *Denote by $P_{\text{a}}, \ldots, P_{\text{e}}$ the problems of determining whether enabling* m-sim *introduces an* NE *which is strictly better than all* NE *of $\mathcal{G}$ in terms of (a) both players' utilities, (b) P1's utility, (c) P2's utility, (d) any strictly monotonic social welfare function (such as $u_1 + u_2$ or $u_1 \cdot u_2$), or (e) the egalitarian social welfare function* $\min\{u_1, u_2\}$.

For general games $\mathcal{G}$, each of the problems $P_{\text{a}}, \ldots, P_{\text{e}}$ is NP*-hard.*

For the purpose of this paper, Theorem 4.2 suggests that we should not expect to be able to find a concise description of the effects of enabling m-sim in general games. We will, therefore, instead focus on identifying particular classes of games where simulation has predictable effects.

5 EFFECTS OF SIMULATION ON PLAYERS' WELFARE

In this section, we describe specific classes of games where enabling mixed-strategy simulation does, and does not, lead to socially beneficial outcomes.

5.1 Drawbacks of an Overly Informed Co-Player

In Section 1.1, we saw that in the simple case of a 2×2 trust game[6], enabling p-sim introduces Pareto-improving NE but enabling m-sim does not. The following theorem is a generalisation of this negative result.

Theorem 5.1 (Simulating a perfectly informed player)**.** *Let $\mathcal{G}_0$ be a finite two-player game. Denote by $\mathcal{G}$ the game where:*

(i) First, P1 selects $s_1 \in S_1^{\mathcal{G}_0}$ and P2 observes P1's choice.

(ii) Next, P2 selects a pure strategy $s_2 \in S_2^{\mathcal{G}_0}$. We assume that P2 must select a Pareto-optimal response (but they are not required to best-respond).[7]

(iii) The players receive utilities $u_i^{\mathcal{G}_0}(s_1, s_2)$.

Then enabling m-sim *does not introduce Pareto-improving NE in $\mathcal{G}$.*

[6] Strictly speaking, Section 1.1 describes the trust game as a simultaneous-move game. However, note that this game is strategically equivalent to the game where P1 acts first (deciding between Trust and Walk Out), after which P2 observes P1's action and chooses whether to Cooperate or Defect. In other words, this 2×2 trust game is a normal-form representation of a game which satisfies the assumptions of Theorem 5.1.

[7] Formally, we require that if P2 selects s_2 in response to s_1, then any $s_2' \in S_2^{\mathcal{G}_0}$ must have either $u_1(s_1, s_2') \leq u_1(s_1, s_2)$ or $u_2(s_1, s_2') \leq u_2(s_1, s_2)$.

	Cooperate	Defect
T_1	$20, 20$	$-100, 100$
T_2	$10, 10$	$-20, 20$
T_3	$5, 3$	$-1, 6$
WO	$0, 0$	$0, 0$

T'_1	$20, 20$	$-100, 100$
T'_2	$10+1, 10-1$	$-30, 30$
$T'_{1.5}$	$\frac{10+20}{2}, \frac{10+20}{2}+1$	$\frac{(-20)+(-100)}{2}, \frac{20+100}{2}$

$T_{1.9}$	$11, 11$	$-99, 99$

Figure 2: *Top:* An illustration of a generalised partial-trust game $\mathcal{G}$ from Definition 5.2. *Middle:* Examples of strategies that would invalidate the technical conditions in the definition if we added them to $\mathcal{G}$. T'_1 fails (4a), since it does not have a unique value $u_1(T, C)$. T'_2 fails the requirement (4b), that any increase in $u_1(T, C)$ – here caused by going from T_2 to T'_2 – must also increase $u_2(T, C)$ (and $u_2(T, D)$, and decrease $u_1(T, D)$). $T'_{1.5}$ fails (5), since $T'_{1.5}$ yields the same u_1 as the convex combination $\frac{1}{2} \cdot T_1 + \frac{1}{2} \cdot T_2$ without also having the same u_2. *Bottom:* $T_{1.9}$ is an example of a strategy that would not invalidate any of the technical conditions from the definition, but adding it to $\mathcal{G}$ would break the assumption of "sufficiently high $u_2(FT, C)$" that is required for Theorem 5.4.

Proof sketch. When P2 can select s_2 as a function of P1's choice of s_1, they can increase the relative attractiveness of any fixed $s_1^* \in S_1$ by being maximally aggressive against any $s_1 \neq s_1^*$. Crucially, P2 can do this without lowering P1's utility of s_1^*. Moreover, P2 can then bring P1's utility for s_1^* all the way to their maxmin value – and any NE of $(\mathcal{G}_0)_{\text{m-sim}}$ will require P2 to do so. However, once P1 only gets their maxmin value, they have no reason to simulate, destroying any potential for simulation-based cooperation. □

5.2 Partial Trust

The following definition captures settings where Alice can vary the degree to which she trusts Bob, with more trust enabling better outcomes for both, but also making Alice more vulnerable to exploitation (for illustration, see Figure 2). The purpose of this section is to show that settings where such modulation of trust is possible can benefit from mixed-strategy simulation.

Definition 5.2 (Generalised Partial-Trust Game). *By a **generalised partial-trust game** (PTG), we mean any $\mathcal{G} = (S_1, S_2, u)$ that satisfies the conditions*

*(1) **P2 has two strategies:** P2 only has only two pure strategies, which we label* Cooperate *(*C*) and* Defect *(*D*);*

*(2) **P1 has a dedicated strategy for opting out of the game:** P1 has a strategy, which we label* Walk Out *(*WO*), for which $u(WO, C) = u(WO, D) = (0, 0)$;*

*(3) **Trust enables profits but is exploitable:** Any P1's strategy $T \neq WO$ ("trust") satisfies*

$$u_1(T, C) > u_1(WO, \cdot) = 0 > u_1(T, D)$$
$$u_2(T, D) > u_2(T, C) > u_2(WO, \cdot) = 0;$$

and the technical assumptions

*(4) **There is a straightforward hierarchy of trust:** (a) For any two strategies $T \neq T'$, we have $u_1(T, C) \neq u_1(T', C)$. (b) When $u_1(T, C) > u_1(T', C)$, we also have $u_2(T, C) > u_2(T', C)$, $u_1(T, D) < u_1(T', D)$, $u_2(T, D) > u_2(T', D)$;*

*(5) **P1 cannot use convex combinations for tie-breaking:** For any T, if a convex combination $\sigma_1 = \lambda s_1 + (1 - \lambda)t_1$ satisfies $u_1(T, \sigma_2) = u_1(\sigma_1, \sigma_2)$ for all σ_2, it must also satisfy $u_2(T, \sigma_2) = u_2(\sigma_1, \sigma_2)$ for all σ_2.*

To give an intuition for the conditions used in Definition 5.2, note that (3) ensures that non-zero payoffs can only be achieved when P1 Trusts P2, but P2 is always tempted to Defect, which makes P1 strictly worse off than if they Walk Out. The technical conditions (4a) and (5) ensure that once P1 decides on the tradeoff between potential gains from cooperation and exploitability, they have no room left for varying P2's payoffs. The technical condition (4b) ensures that higher cooperative gains for P1 go hand in hand with higher cooperative gains for P2 (but also increase P1's exploitability and P2's gains from defection).

The concept of a game with a gradation of trust can be extended in many ways, such as not having the default outcome be zero, not requiring that a higher degree of trust means that $u_2(T, D)$ is higher, giving P2 a hierarchy of cooperative and defective strategies, etc. However, to simplify the exposition, this paper will only consider the basic setup described in Definition 5.2. The following lemma summarises the basic properties of generalised partial-trust games.

Lemma 5.3. *Let $\mathcal{G}$ be a generalised partial-trust game. Then:*

(i) For any $\sigma \in \text{NE}(\mathcal{G})$, $\sigma_1(WO) = 1$;

(ii) The unique pure-commitment equilibrium of $\mathcal{G}$ is (FT, C), where $\{FT\} = \arg\max \left\{u_1(T, C) \mid T \in S_1^{\mathcal{G}}\right\}$. In particular, $\mathcal{G}$ is a generalised trust game.

If $u_2(FT, C)$ is sufficiently high relative to other payoffs, then:

(iii) The unique SE of $\mathcal{G}$ has the form $(FT, i^*[FT])$, where

$$i^*[FT] = \delta^*_{FT} \cdot D + (1 - \delta^*_{FT}) \cdot C,$$
$$\delta^*_{FT} = \max\{\delta \in [0, 1] \mid FT \in \text{br}(\delta D + (1 - \delta)C)\},$$

is the "optimal commitment that still incentivises FT";

(iv) The SE of $\mathcal{G}$ is a strict Pareto-improvement over NE of $\mathcal{G}$ if and only if there is $T \in S_1^{\mathcal{G}}$ s.t. $\frac{u_1(T,C)}{-u_1(T,D)} > \frac{u_1(FT,C)}{-u_1(FT,D)}$.

Proof sketch. The difficult part of Lemma 5.3 is (iv), which relies on the fact that P1 can trivially scale their level of trust by interpolating between any two actions, including FT and Walk Out. This lets P1 disregard any T that has a worse risk-benefit ratio than FT, making (iv) equivalent to the claim that: "giving P2 the ability to make mixed commitments results in a Pareto-improvement if and only if disregarding these redundant actions leaves P1 with more options than just FT and WO." This follows from (iii). □

In light of Lemma 5.3, a generalised PTG is said to be **non-trivial** when it satisfies the condition (iv). We can now prove a generalisation of the positive result from Section 1.1.

Theorem 5.4 (Simulation helps with partial trust)**.** *Let $\mathcal{G}$ be a non-trivial generalised partial-trust game. If $u_2(\mathrm{FT}, \mathrm{C})$ is sufficiently high relative to other payoffs in $\mathcal{G}$, enabling* m-sim *introduces Pareto-improving NE in $\mathcal{G}$.*

Proof sketch for Theorem 5.4. The key insight is that when P2 plays the Stackelberg equilibrium σ_2^{SE} of $\mathcal{G}$, P1 will be indifferent between FT and some other strategy PT, and the non-triviality condition ensures that PT ≠ WO. The proof then consists of showing that there is an equilibrium where P2 mixes between σ_2^{SE} of $\mathcal{G}$ and defecting with probability 100% and P1 mixes between PT and m-sim. The assumption on $u_2(\mathrm{FT}, \mathrm{C})$ ensures that P2 cannot improve their utility by switching to some intermediate level of defection. □

5.3 Trust and Coordination

We now investigate simulation in coordination games.

Definition 5.5 (Generalised coordination game). *By a **generalised coordination game**, we will mean a finite two-player $\mathcal{G}$ game where:*

- $S_i = \{a_i^1, \ldots, a_i^n\}$, *for some $n \geq 2$;*
- $u_1(a_1^k, a_2^l) = u_2(a_1^k, a_2^l) = 0$ *for $k \neq l$; and*
- $u_1(a_1^k, a_2^k),\ u_2(a_1^k, a_2^k) > 0$ *for any k.*

As a standard property of coordination games, we get that:

Lemma 5.6. *For any generalised coordination game,* $\mathrm{NE}(\mathcal{G}) = \{\sigma^K \mid K \subseteq \{1, \ldots, n\}\}$ *for some σ^K which satisfy: (i)* $\mathrm{supp}\,(\sigma_i^K) = \{a_i^k \mid k \in K\}$. *(ii) NE that mix over fewer actions yield higher payoffs. (That is, $\sigma^{K'}$ is a strict Pareto improvement over σ^K whenever $K' \subsetneq K$.)*

Recall that by Lemma 3.2, any NE of the original game also exists as an NE of the simulation game – in particular, enabling m-sim cannot prevent the existence of the (undesirable) NE where Bob only uses a single *mixed* "robot". However, enabling m-sim does have the potential to prevent miscoordination when Bob randomises over multiple robots that use incompatible strategies (e.g., when $\mu_2 = \frac{1}{2}\widehat{a}_2^1 + \frac{1}{2}\widehat{a}_2^2$). In addition to this fact, the following result shows that mixed-strategy simulation also introduces simulation equilibria that are better than miscoordination, but not as good as successful coordination at the players' favourite outcome.

Proposition 5.7 (Simulation in coordination games)**.** *Let $\mathcal{G}$ be a generalised coordination game and denote by $\sigma^{\{1,\ldots,n\}}$ its fully mixed NE. Then, for sufficiently low $\mathrm{c_{sim}}$, we have:*

(i) $\mathcal{G}_{\text{m-sim}}$ has some simulation equilibrium μ^*;

(ii) Any simulation equilibrium $\mu^* \in \mathrm{NE}(\mathcal{G}_{\text{m-sim}})$ satisfies

$$u_1(\sigma^{\{1,\ldots,n\}}) < u_1(\mu^*) < \max_k\, u_1(a_1^k, a_2^k)$$
$$u_2(\sigma^{\{1,\ldots,n\}}) < u_2(\mu^*) \leq \max_k\, u_2(a_1^k, a_2^k);$$

(iii) Unless $\mathcal{G}$ has multiple optimal pure commitments for P2, any such μ^* satisfies $u_2(\mu^*) < \max_k u_2(a_1^k, a_2^k)$.

	a_2^1	a_2^2	OO
a_1^1	20, 20 −99, 40 9, −99 9, −99	0, 0	0, 1
a_1^2	0, 0	20, 20 −99, 40 10, −99 10, −99	0, 1
OO	1, 0	1, 0	1, 1

Figure 3: Trust-and-coordination game, where coordinating on a joint action (a_1^k, a_2^k) leads the players to a trust subgame.

Proposition 5.7 shows that mixed-strategy simulation is able to prevent the worst equilibria, but does not introduce Pareto-improving NE in the stronger sense of allowing for an outcome that wouldn't be achievable through other means (i.e., by successful selection of a pure equilibrium). The following example and theorem show that the usefulness of mixed-strategy simulation increases when the players need to deal not only with coordination but also with issues of trust.

Definition 5.8 (Trust-and-coordination game). *By a **trust-and-coordination game**, we mean a game $\mathcal{G}$ which works as follows (for examples, see Figure 3 and the appendix).*

- *In the first stage, the players simultaneously select an action from the set $\{a_i^1, \ldots, a_i^n, \mathrm{OO}\}$.*
- *If the players select (a_1^k, a_1^l) for $k \neq l$, they receive "**b**ad" miscoordination payoffs $(\mathrm{B}_1, \mathrm{B}_2)$.*
- ***O**pting **O**ut of the game via* OO *yields $(\mathrm{B}_1, \mathrm{B}_2)$, with an additional reward ϵ for the player(s) who used* OO.
- *If the players coordinate on some (a_1^k, a_2^k), they enter the second stage of the game, where they play a subgame $\mathcal{G}_k$.*
- *Each $\mathcal{G}_k$ is a 2×2 trust game with actions* {Trust, Walk Out}*, resp.* {Cooperate, Defect}.
- *We denote the payoffs in $\mathcal{G}_k$ as*

$$u^{\mathcal{G}_k}(\mathrm{T}, \mathrm{C}) := (\mathrm{G}_1^k, \mathrm{G}_2^k), \quad u^{\mathcal{G}_k}(\mathrm{T}, \mathrm{D}) := (\mathrm{H}_1^k, \mathrm{A}_2^k),$$
$$u^{\mathcal{G}_k}(\mathrm{WO}, \mathrm{C}) = u(\mathrm{WO}, \mathrm{D}) := (\mathrm{N}_1^k, \mathrm{H}_2^k).$$

*(The naming is meant to be suggestive of **a**wesome, **g**ood, **n**eutral, and **h**orrible .) We assume that:*

$$\mathrm{H}_1^k < \mathrm{B}_1 < \mathrm{B}_1 + \epsilon < \mathrm{N}_1^k < \mathrm{G}_1^k$$
$$\mathrm{H}_2^k < \mathrm{B}_2 < \mathrm{B}_2 + \epsilon < \mathrm{G}_2^k < \mathrm{A}_2^k.$$

The only NE of $\mathcal{G}$ is for both players to Opt Out, yielding utilities $\mathrm{B}_i + \epsilon$. In contrast, the SE of $\mathcal{G}$ (with P2 as the leader) would consist of coordinating on one of the subgames $\mathcal{G}_k$ and then playing its SE, yielding utilities $u_1 = \mathrm{N}_1^k$, $u_2 > \mathrm{G}_2^k$. We will use $v_i^{\mathrm{SE}}(\mathcal{G}_k)$ to denote the expected utility corresponding to the SE (with P2 as the leader) of the subgame $\mathcal{G}_k$.

Theorem 5.9 (Simulation helps in trust-and-coordination games)**.** *Let $\mathcal{G}$ be a trust-and-coordination game. If*

(a) $\arg\max v_2^{\mathrm{SE}}(\mathcal{G}_k) \cap \arg\max v_1^{\mathrm{SE}}(\mathcal{G}_k) = \varnothing$*; and*

(b) The NE payoffs H_2^k of P2 in the subgames $\mathcal{G}_k$ are sufficiently low relative to the other payoffs in $\mathcal{G}$;

then enabling m-sim *introduces Pareto-improving NE.*

Proof sketch. The proof consists of showing that $\mathcal{G}_{\text{m-sim}}$ admits an NE where P2 mostly plays the Stackelberg equilibrium of "P1's favourite subgame" $\mathcal{G}_{k_1}$, but sometimes deviates to playing the SE "P2's favourite subgame" $\mathcal{G}_{k_2}$, while P1 mixes between their part of the SE of $\mathcal{G}_{k_1}$ and simulating. Because the only NE of $\mathcal{G}$ is the undesirable outcome (Opt Out, Opt Out), this simulation equilibrium will constitute a Pareto improvement over any NE of $\mathcal{G}$. □

Corollary 5.10. *Enabling* m-sim *introduces Pareto-improving NE in the trust-and-coordination game from Figure 3.*

5.4 Mixed-Strategy Simulation and Privacy

Kovařík et al. [15] show that pure-strategy simulation can sometimes be harmful to *both* players. An example that illustrates this dynamic is a scenario where Bob, after successfully cooperating with Alice, has to put all his profits into a password-protected account. While Alice could always attempt to guess Bob's password, doing so would typically be futile. However, if she had access to pure-strategy simulation, she would be able to predict Bob's password and steal his profits, so Bob would chose to not cooperate with Alice in the first place. In contrast, if Alice only had access to mixed-strategy simulation, Bob could protect his profits by using a randomly-generated password, thus preserving the possibility of cooperation with Alice.

In the appendix, we give a general construction which adds this "password-guessing" dynamic into any base-game, allowing us to derive the following result.

Theorem 5.11. *There are games where enabling* m-sim *introduces Pareto-improving NE, but pure-strategy simulation does not.*

6 RELATED WORK

Kovařík et al. [15] study a setting which is the closest to ours, but focus exclusively on the much stronger assumption of using pure-strategy simulation. Várdy [25] study the same setting (i.e., what we refer to as pure-strategy simulation), but approach it from a more traditional economics angle, focusing on the simulated agent's value of commitment rather than on Pareto-improvements. Simulation in the context of AI agents is also studied by Chen et al. [4], who distinguish between the "screening" and "disciplining" effects on the AI's behaviour. Unlike the present paper, this work assumes that the simulated agent is drawn from a fixed population (rather than being strategic).

Alternative formulations of games with simulation incorporate unpredictable randomisation in different ways. In *games with espionage* [21], for example, P1 can pay to gain a *probabilistic* signal based on P2's realised pure strategy. In *oracle games* [27], P1 can attempt to learn P2's pure strategy, but the success chance depends on the payment made by P1.

Harris et al. [12], study the problem of Bayesian persuasion under the assumption that the *leader* can simulate the follower a number of times in order to learn what their response will be.

Other related work exists in *incomplete* information games, where a player pays to learn what others *know* (rather than *do*). In mechanism design with *(partially) verifiable information*, an agent's strategy might be restricted by their private information [9], or this information might be revealed by the designer paying a cost [2, 23]. In some models of costly *information acquisition*, a player can pay to learn the others' observations of the hidden state [8, 13, 17].

Finally, some other works consider the possibility of *mutual* simulation, though they typically assume that simulation is not costly. Kovarik et al. [16] study a setting where the players observe the result of simulation jointly, but are uncertain as to whether they themselves might be in a simulation. Oesterheld [18] shows that in games played between AI agents with mutual access to each other's code [22], simulation can lead to cooperation. Another related approach is game theory with translucent players [11], which 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. Capraro and Halpern [3] show how translucency can lead to increased cooperation.

7 CONCLUSION

Strategic interactions involving AI agents are likely to become increasingly frequent and important. They may also be fundamentally different from more familiar interactions, as AI agents may – in theory – be more transparent than humans or human institutions. For example, it may be possible to simulate an AI agent, likely at a small cost.

In this paper, we studied the implications of costly simulation in the presence of unpredictable randomisation, showing that it is neither strictly weaker nor stronger than pure-strategy simulation (in terms of improving social welfare). While determining whether enabling mixed-strategy simulation is beneficial turns out to be NP-hard in general, we identified several classes of games where the effects of simulation can be predicted. Concretely, we showed that mixed-strategy simulation can lead to increased cooperation in games where the simulator needs to decide whether to trust the other player, when either the simulator has a more nuanced set of options than just full trust and no trust, or the players face challenges with both trust and coordination. However, we also saw that mixed-strategy simulation fails to foster cooperation if P2 observes P1's action before moving.

While mixed-strategy simulation is arguably a more realistic model than pure-strategy simulation, it is still limited in several important ways. Future work should explore generalisations such as dynamic games, games with incomplete information, and other forms of simulation (such as by generating multiple, stochastic samples). Other important directions include identifying further classes of game in which different forms of simulation can help, and in matching those to potential real-world domains of application. In so doing, we will be better equipped to reap the benefits and avoid the risks once AI agents become widespread.

ACKNOWLEDGMENTS

We are grateful to C. Oesterheld, E. Berker, J. Liu, I. Geffner, and D. Thomas for insightful discussions and to Cooperative AI Foundation, Polaris Ventures, and Long-term Future Fund for financial support. Vojtech Kovarik was partially funded by Czech Science Foundation grant no. GA22-26655S. Lewis Hammond acknowledges the support of an EPSRC Doctoral Training Partnership studentship (reference: 2218880).

REFERENCES

[1] Mikita Balesni, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy Scheurer, Charlotte Stix, Rusheb Shah, Nicholas Goldowsky-Dill, Dan Braun, Bilal Chughtai, Owain Evans, Daniel Kokotajlo, and Lucius Bushnaq. 2024. Towards evaluations-based safety cases for AI scheming. arXiv:2411.03336 [cs.CR] https://arxiv.org/abs/2411.03336

[2] Ian Ball and Deniz Kattwinkel. 2019. Probabilistic verification in mechanism design. In *Proceedings of the 2019 ACM Conference on Economics and Computation*. ACM New York, NY, USA, Phoenix, 389–390.

[3] Valerio Capraro and Joseph Y Halpern. 2019. Translucent players: Explaining cooperative behavior in social dilemmas. *Rationality and Society* 31, 4 (nov 2019), 371–408. https://doi.org/10.1177/1043463119885102

[4] Eric Chen, Alexis Ghersengorin, and Sami Petersen. 2024. Imperfect Recall as a Screening Device: Applications to Deceptive Alignment. (2024). unpublished.

[5] Vincent Conitzer and Caspar Oesterheld. 2023. Foundations of cooperative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. The AAAI Press, Washington, DC, USA, Washington DC, 15359–15367.

[6] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: Machines Must Learn to Find Common Ground. *Nature* 593, 7857 (May 2021), 33–36. https://doi.org/10.1038/d41586-021-01170-0

[7] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative ai. arXiv preprint arXiv:2012.08630.

[8] Tommaso Denti. 2023. Unrestricted information acquisition. *Theoretical Economics* 18, 3 (2023), 1101–1140. https://doi.org/10.3982/te4541

[9] Jerry R. Green and Jean-Jacques Laffont. 1986. Partially Verifiable Information and Mechanism Design. *The Review of Economic Studies* 53, 3 (July 1986), 447. https://doi.org/10.2307/2297639

[10] Joseph Y Halpern and Rafael Pass. 2008. Game theory with costly computation. arXiv preprint arXiv:0809.0024.

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

[12] Keegan Harris, Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. 2023. Algorithmic Persuasion Through Simulation. https://doi.org/10.48550/ARXIV.2311.18138 arXiv:2311.18138 [cs.GT] arXiv:2311.18138.

[13] Christian Hellwig and Laura Veldkamp. 2009. Knowing What Others Know: Coordination Motives in Information Acquisition. *Review of Economic Studies* 76, 1 (Jan. 2009), 223–251. https://doi.org/10.1111/j.1467-937x.2008.00515.x

[14] Marcin M Jacak, Piotr Jóźwiak, Jakub Niemczuk, and Janusz E Jacak. 2021. Quantum generators of random numbers. *Scientific Reports* 11, 1 (2021), 16108.

[15] Vojtěch Kovařík, Caspar Oesterheld, and Vincent Conitzer. 2023. Game theory with simulation of other players. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence, Macau, 2800–2807.

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

[17] D. P. Myatt and C. Wallace. 2011. Endogenous Information Acquisition in Coordination Games. *The Review of Economic Studies* 79, 1 (Oct. 2011), 340–374. https://doi.org/10.1093/restud/rdr018

[18] Caspar Oesterheld. 2019. Robust Program Equilibrium. *Theory and Decision* 86, 1 (2 2019), 143–159.

[19] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Andrea Vallone, Alexandre Passos, and David G. Robinson. 2023. *Practices for Governing Agentic AI Systems*. Technical Report. OpenAI.

[20] Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. arXiv preprint arXiv:2305.15324.

[21] Eilon Solan and Leeat Yariv. 2004. Games with espionage. *Games and Economic Behavior* 47, 1 (April 2004), 172–199. https://doi.org/10.1016/s0899-8256(03)00177-5

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

[23] Robert M. Townsend. 1979. Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory* 21, 2 (October 1979), 265–293. https://ideas.repec.org/a/eee/jetheo/v21y1979i2p265-293.html

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

[25] Felix Várdy. 2004. The value of commitment in Stackelberg games with observation costs. *Games and Economic Behavior* 49, 2 (Nov. 2004), 374–400. https://doi.org/10.1016/j.geb.2003.07.003

[26] Anbang Wang, Pu Li, Jianguo Zhang, Jianzhong Zhang, Lei Li, and Yuncai Wang. 2013. 4.5 Gbps high-speed real-time physical random bit generator. *Optics express* 21, 17 (2013), 20452–20462.

[27] Matthew J. Young and Andrew Belmonte. 2020. Simultaneous games with purchase of randomly supplied perfect information: Oracle Games. https://doi.org/10.48550/ARXIV.2002.08309 arXiv:2002.08309 [cs.GT] arXiv:2002.08309.