https://doi.org/10.1038/s44260-025-00065-9

Promoting cooperation in the public goods game using artificial intelligent agents

Check for updates

Arend Hintze[1,2] ✉ & Christoph Adami[2,3,4,5]

The *tragedy of the commons* describes a dilemma where individually rational actions degrade shared resources. Strategies to overcome this issue remain limited. We examine how artificial intelligence (AI) agents can foster cooperation in Public Goods games, moving beyond traditional regulations. We test three scenarios: (1) **Mandatory Cooperation Policy**, requiring AI agents always to cooperate; (2) **Player-Controlled Cooperation Policy**, allowing players to control AI cooperation rates; and (3) **Agents Mimic Players**, where AI agents imitate human behaviors. Using computational evolutionary modeling, we show that only AI agents mimicking player behavior effectively reduce the synergy threshold needed for cooperation, resolving the dilemma. These findings suggest that designing AI agents to replicate human behaviors can enhance cooperation and improve collective welfare.

The *tragedy of the commons*[1] refers to a fundamental social dilemma where individual rational actions lead to collectively undesired outcomes. In systems involving shared resources or public goods – which are formally defined as goods that are both non-excludable (individuals cannot be prevented from accessing them) and non-rivalrous (one person's consumption does not reduce availability for others) – individuals who cooperate by contributing are often disadvantaged compared to defectors who withhold their contributions[2,3]. The critical feature of public goods is their "even distribution": once created through collective contributions, the amplified resource pool is divided equally among all participants regardless of their individual contribution levels[4,5]. This distribution mechanism creates the fundamental tension where defectors achieve higher personal gains than their cooperative counterparts because they receive the same share of benefits without incurring the cost of contribution. This imbalance between contribution and reward undermines cooperative behavior and threatens the sustainability of shared resources, posing significant challenges across disciplines such as economics, environmental science, and public policy [6,7].

There are multiple theoretical frameworks for analyzing strategic interactions. Classical game theory[8–10] provides powerful analytical tools for identifying equilibria, including in settings with incomplete information[11], repeated interactions[12], and dynamic sequential play[9]. Evolutionary game theory offers a complementary perspective[13,14], particularly suited for modeling how strategies emerge and adapt over time through population-level selection dynamics rather than individual optimization. Agent-based implementations of EGT are especially valuable when analyzing systems with finite populations, non-zero mutation rates, stochastic decisions, and heterogeneous agents[15].

For our investigation, we are interested in understanding how cooperation emerges and stabilizes when AI agents with different behavioral policies are gradually integrated into human populations. This is fundamentally a question about adaptive processes and convergence: starting from uncertain initial conditions where players have unknown strategies, how does the system evolve toward stable behavioral patterns? This process-oriented perspective, where strategies are not pre-computed but emerge through iterated interactions, selection, and mutation, makes evolutionary game theory the appropriate framework for our analysis.

To study the emergence of cooperation in groups, EGT most commonly turns to the "Public Goods" game. At its core, this game presents participants in a group with a choice: to cooperate (C) by contributing a cost (or ante, here set to 1 without loss of generality) to the (common) public good, or to defect (D) by withholding that contribution. The total contributions to the common are amplified by a multiplication factor r, symbolizing the *synergy* between cooperating parties. This enhanced public good is then *evenly* distributed among all participants, regardless of their contributions (see Fig. 1). In such a game, defectors always obtain higher rewards than cooperators in the same group, even though they could reap an even higher return if everyone cooperated. Once the synergy factor r becomes high enough ($r \geq k+1$, where $k+1$ is the group size), a defecting player fares worse than if they had cooperated instead. Therefore, to promote cooperation we ask: "How can we influence the game so that cooperation becomes the rational choice for synergies where $r < k+1$?".

Although numerous mechanisms have been proposed to promote cooperation in two-player games – including the *green beard effect*[16], reciprocity[17], spatial dynamics[18], group selection[19], costly punishment[4,20–22], institutional incentives[23] and resource redistribution[24], only a few have been

[1]Department of MicroData Analytics, Dalarna University, Falun, Sweden. [2]BEACON Center for the Study of Evolution in Action, East Lansing, MI, USA. [3]Department of Microbiology, Genetics, and Immunology, Michigan State University, East Lansing, MI, USA. [4]Program in Ecology, Evolution, and Behavior, Michigan State University, East Lansing, MI, USA. [5]Department of Physics and Astronomy, Michigan State University, East Lansing, MI, USA. ✉e-mail: ahz@du.se

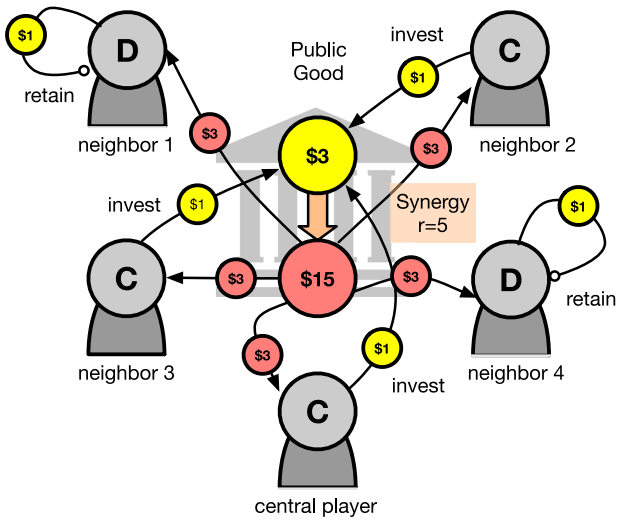

Fig. 1 | Illustration of the Public Goods game. A central player and its ($k = 4$) neighbors engaging in the Public Goods game. The three cooperators (labeled "C") contribute \$1 to the public goods pool (invest), while two defectors ("D") withhold their \$1 investment (retain). A synergy multiplier (here, $r = 5$) is applied to the sum of contributions, which raises the value of the public goods pool (here, \$15). This pool is redistributed equally to all players.

shown to work in groups where cooperation is harder to achieve and maintain[25–28]. While these established mechanisms effectively alter the conditions of the Public Goods game to promote cooperation, these approaches do not account for the transformative impact that AI integration could have on human societies.

Recent studies have begun to explore this intersection: AI and machine learning techniques have been used to study Public Goods games, for example, by using reinforcement learning to analyze how group size affects cooperation[29] and investigating how uncertainty and communication shape outcomes[30]. Delegating decisions to artificial agents has also been shown to foster prosocial behavior in collective dilemmas[31], while research on heterogeneous human populations indicates that diversity can support cooperation[32]. However, explicitly considering mixed populations of humans and AI agents within the framework of the Public Goods game remains underexplored. Recognizing this gap, we pursue a novel exploration of how AI agents can be strategically leveraged to enhance cooperative behavior.

Governments are fundamentally responsible for establishing institutions that address the tragedy of the commons [3]. By enacting legislation that incorporates mechanisms like costly punishment and resource redistribution, they create regulations, norms, and incentives that make cooperation more advantageous than defection[20,23,33]. For instance, legal penalties for free-riding discourage defection, while taxation and subsidies can redistribute resources to promote fairness[1]. By integrating these strategies, governments align individual interests with collective welfare, fostering sustainable cooperation [6].

In the rapidly evolving intersection of technology and society, we identified a critical opportunity presented by the integration of artificial intelligence (AI) agents into socio-economic systems [34]. Unlike humans, whose strategies are complex and often unpredictable, AI agents offer a malleable platform that can be precisely programmed or influenced[35], a realization that opens new avenues for promoting cooperation. By controlling AI agents' propensity to cooperate, we can enhance overall system cooperation. This presents an opportunity for governments to enact rules specifically governing AI behavior[36], mandating cooperative strategies and using AI as tools to promote collective interests. As a consequence, it could be argued that the role of government extends to overseeing AI behavior in social dilemmas[37]. By legislating AI conduct to favor cooperation, governments can regulate AI's impact on society and harness it to encourage cooperative behavior among all participants[38].

Importantly, in our framework, the fundamental equal distribution principle of public goods is maintained regardless of whether participants are human players or AI agents – both receive identical shares of the amplified contribution pool, preserving the core dilemma structure while allowing us to examine how different AI behavioral policies affect cooperative dynamics. This approach ensures that our results reflect genuine changes in cooperative incentives rather than artifacts of differential reward structures. This approach integrates technological advancements with policy interventions, aiming to create systems where cooperation is the rational and advantageous choice for all agents involved.

In the economics literature, goods are traditionally classified along two dimensions: excludability (whether individuals can be prevented from accessing the good) and rivalry (whether one person's consumption reduces availability for others)[3]. *Public goods* are both non-excludable and non-rivalrous (e.g., national defense, clean air), while *common-pool resources* are non-excludable but rivalrous (e.g., fisheries, grazing lands)[3]. Despite this conceptual distinction, both types of goods share a fundamental characteristic: they create a *cooperation dilemma* where individual rational actions can lead to collectively undesired outcomes[1,20].

Our focus in this study is on the cooperation dilemma itself—the tension between individual and collective interests—rather than on the specific properties of excludability or rivalry. The Public Goods game, despite its name, has been widely used to model cooperation problems in both public goods provision and common-pool resource management[3–5]. The game captures the essential strategic structure of these social dilemmas: groups of cooperators can achieve better collective outcomes than groups of defectors, yet individual defectors always outperform cooperators in the same group.

To illustrate how governments might interact with AI companies and their technologies in contexts that involve cooperation dilemmas, consider the fictional company "Edison" which manufactures self-driving cars. While these vehicles must comply with all traffic laws and regulations, there are behaviors not explicitly governed by legislation. For instance, when a driver attempts to merge onto a busy road without the right of way, self-driving cars could, within legal parameters, choose not to yield. However, Edison could program its cars to act more cooperatively, allowing others to merge smoothly without compromising safety. This added layer of courteous behavior enhances overall traffic flow and exemplifies how AI can be designed to promote cooperation beyond what regulations mandate. One could debate whether smooth traffic flow represents a pure public good (benefiting all without rivalry) or involves common-pool resource dynamics (where road space is rivalrous). Regardless of this classification, the core issue remains the same: individual vehicles (and their manufacturers) face a choice between cooperative behavior that benefits the collective and self-interested behavior that prioritizes individual advantage.

Furthermore, Edison might program its cars to be especially accommodating to other vehicles of the same brand, effectively using cooperative behavior as a form of brand promotion. Such actions are currently outside the scope of governmental influence but have significant implications for public cooperation and competition. Similar cooperation dilemmas exist in other AI applications[39]: social media algorithms that must balance individual user engagement against collective information quality[40], AI-controlled energy grids managing the trade-off between individual consumption and grid stability, or AI systems managing shared computational resources where one agent's resource use affects availability for others. In each case, regardless of whether the underlying resource is technically a public good or a common-pool resource, the fundamental cooperation dilemma persists: AI agents can be designed to act selfishly or cooperatively, with significant consequences for collective outcomes.

The advent of AI thus opens new opportunities and raises questions about regulations or policies that promote cooperative behavior in situations akin to the Public Goods game. Here, we investigate different policy scenarios that influence AI agent behavior and their effectiveness in promoting cooperation. We assume a population of humans, referred to as *players*, who engage in the Public Goods game, while a fraction of them is replaced by AI

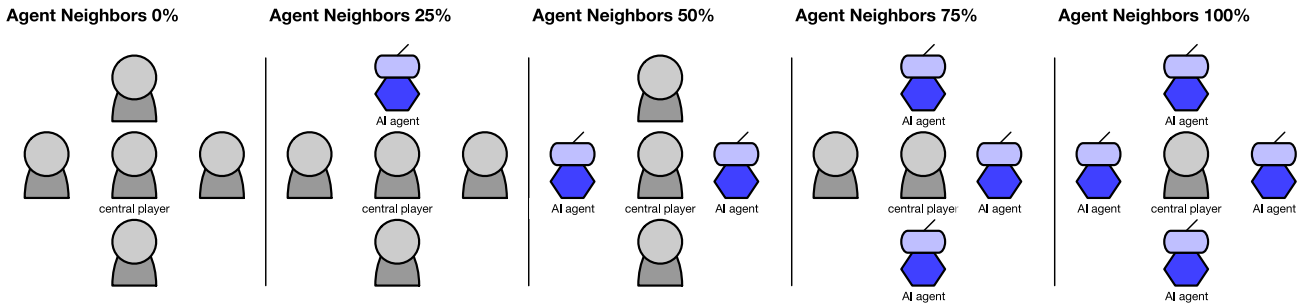

Fig. 2 | Illustration of AI Agent density in the neighborhood of the central player. Each panel illustrates the effect of players being replaced by AI agents, ranging from 0% to 100% replacement in increments of 25% (ρ_A = 0.0, 0.25, 0.5, 0.75, 1.0 respectively). Not that the central player is never replaced as ρ_A only pertains to the neighborhood.

agents. Since the extent to which AI will replace human decision-makers is uncertain, we experimentally control the density of these agents (ρ_A) (see Fig. 2 middle and right panel for an illustration). In practice, we use a computational evolutionary model to test our hypotheses (see Fig. 3 for details).

In principle, regulatory bodies could either mandate that agents behave in a certain way, or allow AI companies and users deploying such agents to determine their behavior. Beyond these two extremes, we propose a third mechanism that requires agents to emulate the behavior of human players in the game – a mimicking strategy. Human behavior is modeled here using evolutionary game theory mechanisms[15], simplifying player behavior to a probability of cooperating or defecting. Through natural selection, players adjust their strategies based on the payoffs they receive in different situations. We will show that neither of the first two mechanisms – forcing behavior or not regulating it – promotes cooperation. In contrast, our proposal for agents to mimic player behavior promotes cooperation among players. Further, the more players are replaced by AI agents, the more players are encouraged to cooperate, which makes this proposal sustainable for a future in which we will find an increasing number of AI agents.

Results

We test the impact that different policies regulating AI agents that participate in the Public Goods game have on player behavior using an agent-based evolutionary computational model. A population of players, each defined by a probability to cooperate (i.e., contribute) in a Public Goods game, evolves under selection and mutation over thousands of generations. In each generation, players are grouped into small sets (of size $k+1$) to play the Public Goods game multiple times, from each player's perspective. Payoffs accumulate based on individual contributions and a synergy factor r that amplifies the individual contributions. This amplified capital constitutes the public good, which is distributed equally to all individuals in a group (see Methods for more details on the game and its implementation).

Each player's likelihood of cooperating evolves under fitness-proportional selection and mutation, where payoffs that accumulate from multiple rounds of the Public Goods game determine the probability of transmitting a given strategy to the next generation. Crucially, when AI agents replace a fraction of players, they do not accumulate payoffs themselves or are replaced, and thus do not evolve their own strategy. Instead, they follow a policy set by the experimenter.

For each experimental condition, we conduct 100 independent evolutionary experiments. After 10,000 generations, the evolutionary line of descent (LOD)[15] is reconstructed, and population averages are extracted. Population averages comport with the values seen on the LOD, indicating that under the conditions tested, the population evolves to a unique steady state. Observe that multiple stable attractors are only possible in games with an interior unstable fixed point, as in the coordination game. The effective payoff matrix for the Public Goods game is that of a domination game (see 15, Eq. (26)), which only has a single fixed point. This agrees with our previous experiments studying other evolutionary games[15,41,42]. We now compare the analytical predictions to these experiments.

In mathematical terms, consider a single player in a group with k other players. Let N_C represent the number of cooperators and N_D the number of defectors among those k neighbors. Then, the payoffs for the cooperator and defector can be written as

$$P_C = r\frac{N_C+1}{k+1} - 1 \tag{1}$$

$$P_D = r\frac{N_C}{k+1} \tag{2}$$

For cooperation to be favorable, the payoff for a cooperator in a group of cooperators should exceed that of a defector among defectors, that is, $P_C(N_C = k) - P_D(N_C = 0) > 0$, which implies $r > 1$. A defector outscores a cooperator if $P_D - P_C > 0$, which implies $r < k+1$. However, independently of r it is favorable for all players to cooperate simultaneously ($N_C = k$), which indicates that a *dilemma* exists for all r between 1 and $k+1$ [21]. The parameter r becomes a *critical parameter* controlling the game's outcome, and the point at which $P_D - P_C = 0$ is a *critical* point (the relationship to the critical points of condensed matter physics is made more clear in ref. 21).

Mandatory Cooperation Policy for AI Agents

A governing body might attempt to promote cooperation by forcing AI-controlled agents to always cooperate, with the idea of ensuring that at a minimum the agents behave well and support the public good, potentially promoting cooperation among human players (see Fig. 4 left panel). However, the calculation that bounded the region in which a dilemma is present showed that the region does *not* depend on the density of cooperators, which makes it unlikely that influencing the density of cooperators will change the boundaries and, therefore, alleviate the dilemma. We experimentally confirm this prediction by creating an evolutionary scenario that implements the mandatory cooperation policy for AI agents.

Key Finding: We find the probability of cooperation after evolutionary optimization to be **unaffected by the infusion of cooperating AI agents** (see Fig. 5A and C), even though the frequency of cooperating agents rises (see Fig. 5B). **Mandating AI agents to always cooperate only increases the cooperation within the population in proportion to the density of AI agentsρ_A, but does not encourage players to cooperate**. This demonstrates that simply adding cooperative actors to the population is insufficient to resolve the underlying dilemma.

Player-controlled agent cooperation policy

In this scenario, the government does not regulate agent behavior in any way. Instead, the AI companies set their agent's policies, or whoever owns or deploys an AI agent does so (see Fig. 4 middle panel). To model this, each player still has a probability of cooperating (p_C) that can adapt over the course of evolution to the conditions of the current scenario. In addition, each player also has at their disposal an auxiliary probability (p_{AC}) that is used to determine the probability of an agent cooperating within its neighborhood. This auxiliary probability is also optimized by

Computational Model of the Public Goods Game

Players and Population Structure: We consider a population of N individuals (here $N = 1,024$), where each player i is characterized by a cooperation probability $p_{C,i} \in [0,1]$ that determines their likelihood of contributing to the public good in any given interaction. A fraction $\rho_A \in [0,1]$ of the population consists of AI agents whose behavior follows externally determined policies, while the remaining fraction $(1-\rho_A)$ represents human players whose strategies evolve over time.

Action Space and Group Formation: In each round, players face a binary choice: **Cooperate** (C) by contributing a fixed cost of 1 unit, or **Defect** (D) by contributing nothing. Players are randomly assigned to groups of size $k+1$ (here $k = 4$, yielding groups of 5 players), ensuring well-mixed interactions that avoid spatial clustering effects that might independently promote cooperation.

Payoff Structure: For a focal player in a group where N_C neighbors cooperate and $N_D = k - N_C$ neighbors , the payoffs are calculated by Equations (1) and (2), where $r > 1$ is the synergy factor that amplifies collective contributions. The term $(N_C + 1)$ in the cooperator equation accounts for the focal player's own contribution.

Evolutionary Dynamics: Human players' cooperation probabilities evolve through: **Selection**: Each generation, players reproduce with a probability proportional to their accumulated payoffs (Roulette-Wheel selection); **Mutation**: With probability $\mu = 0.01$, a player's cooperation probability is replaced by a value drawn uniformly from $[0,1]$; **Payoff accumulation**: Each player participates in $(k+1)$ games as the focal player to ensure balanced exposure across different group compositions.

AI Agent Integration: AI agents do not evolve or accumulate payoffs for reproductive purposes. Instead, they follow one of three policy frameworks.

Parameter Justification: Group size $(k = 4)$ is chosen because it can be easily implemented on a grid topology (where each player interacts with its four neighbors) and has been successfully used in our previous evolutionary game theory studies [15, 21]. Increasing group size only shifts the critical synergy factor, but otherwise does not affect the results.

Fig. 3 | Computational model framework for the Public Goods game with AI agents.

evolution. As a consequence, depending on the AI agent density, the density of cooperators around the central player is determined by the likelihood of other (non-AI agent) players and its probability reserved for agent cooperation (p_{AC}).

The player's payoff, however, is now not only dependent on their own cooperation probability p_C but also on their auxiliary probability for agent cooperation (p_{AC}), so evolution is optimizing *both* probabilities. If it is, for example, beneficial for a player to cooperate while also surrounding itself with cooperating agents, both probabilities should be maximized ($p_C = 1.0$ and $p_{AC} = 1.0$).

Note that the central player only pays its own ante when cooperating. In contrast, an AI agent would pay for its own cost to cooperate even though the

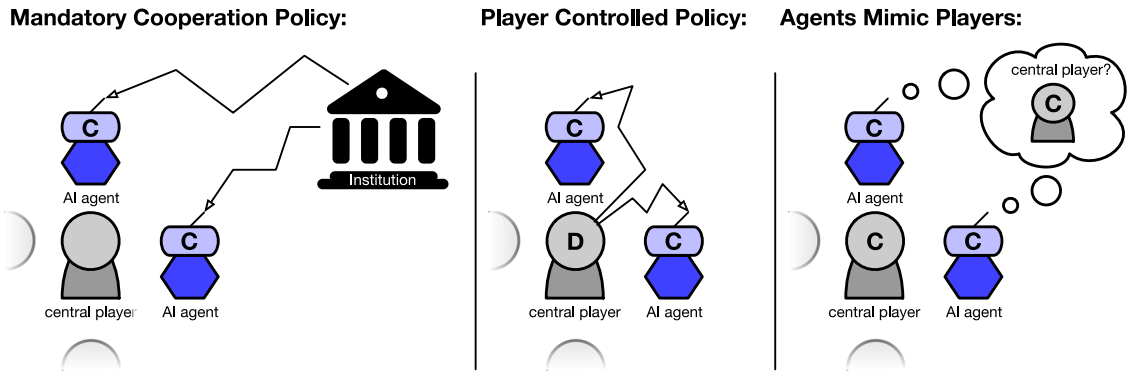

Fig. 4 | Illustration of three different policies for AI agent behavior. The left panel shows the "Mandatory Cooperation Policy for Agents"; here, a hypothetical institution that mandates that all AI agents are to always cooperate. The middle panel shows the "Player Controlled Agent Policy"; here, the central player can determine the likelihood of peripheral AI agents cooperating. However, the decision of the central player to cooperate is independent of the AI agents' policy. The right panel shows the policy where "Agents Mimic Players"; here, the peripheral AI agents use the central players' likelihood to cooperate to determine their own choice.

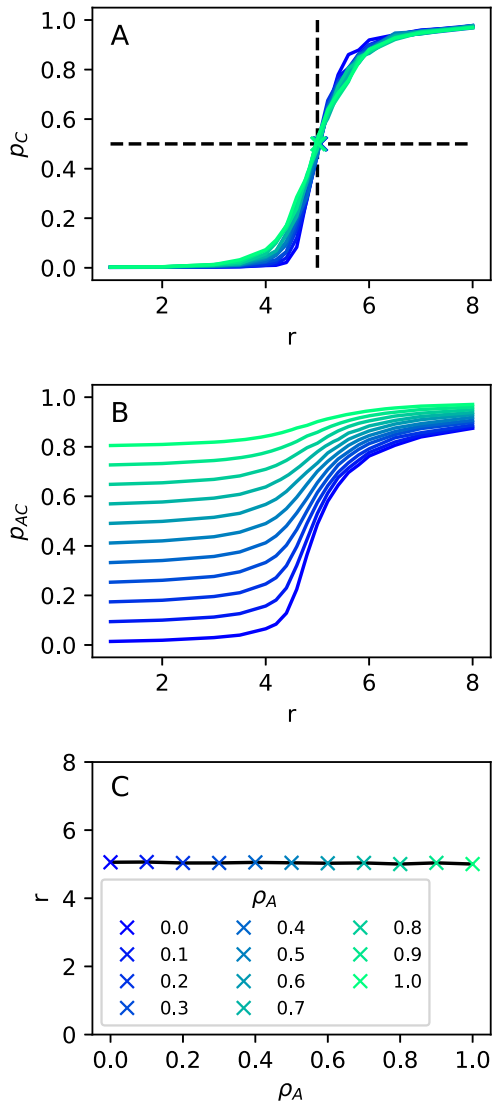

Fig. 5 | Evolution of cooperation when cooperation of agents is mandated, as a function of synergy factors r. A Average (human) player probability of cooperation p_C after evolution (100 independent replicates) as a function of r, for different AI agent densities ρ_A indicated by the color code in (C). **B** Overall cooperation frequency within the population (players and AI agents). **C** Critical point at which cooperation evolves (y-axis), given the different probabilities of encountering AI agents (ρ_A) (x-axis) that are mandated to cooperate.

central player controls its probability of cooperating. As an illustration, if an AI-driven car is programmed by a driver to give way to them coming from an on-ramp, the AI agent having to wait does not cost the driver anything.

Key Finding: While this scenario can increase the likelihood of cooperators, we again do not expect this to reduce the synergy threshold for the player to engage in cooperation following the previous logic. We indeed find **the synergy factor to be unaffected** (see Fig. 6A and C). While the controlled agents become cooperators the moment they appear in the neighborhood ($\rho_A > 0.0$) as this maximizes the payoff of the player controlling them ($P_D(N_C) < P_D(N_C + 1)$) (see Fig. 6 B), **this scenario does not affect the probability of the central player to cooperate, and thus the dilemma is not lifted**. This reveals that allowing unrestricted control over AI agents leads to exploitation rather than genuine cooperation.

Agents mimic players

Having analyzed the shortcomings of both mandating cooperation and allowing players to control AI agents, we recognize the need for a fundamentally different policy. These initial strategies fall short of removing the dilemma because they do not alter the critical point at which cooperation becomes advantageous for players.

From previous experience, we hypothesized that incentivizing players to cooperate requires increasing the payoff associated with cooperative actions. This realization led us to propose an innovative policy where AI agents mimic the behavior of the central player. In this paradigm, if the central player chooses to cooperate, the neighboring agents (with density ρ_A) will also cooperate (see Fig. 4, right panel). This approach establishes a reciprocal relationship between agents and players, effectively reshaping the incentives and promoting cooperative behavior.

Although the Public Goods game is not an iterative game where agents can estimate the likelihood of players cooperating based on previous actions (that is, where players can use information to shape their future actions[41]), we can model such a mechanism by forcing the AI agents' cooperation probability to be identical to that of the central player. Thus, depending on the agent density ρ_A, players will encounter groups where agents mirror player behavior. If the central player is a cooperator, the likelihood of having cooperating neighbors increases with the AI agent density. This effectively enhances the expected payoff for cooperators, as they are more likely to be in groups with other cooperators.

If ρ_A is the likelihood that a peripheral player mimics the strategy of the focal player, then if the focal player is a cooperator, the number of defectors is reduced from N_D to

$$N'_D = (1 - \rho_A)N_D. \quad (3)$$

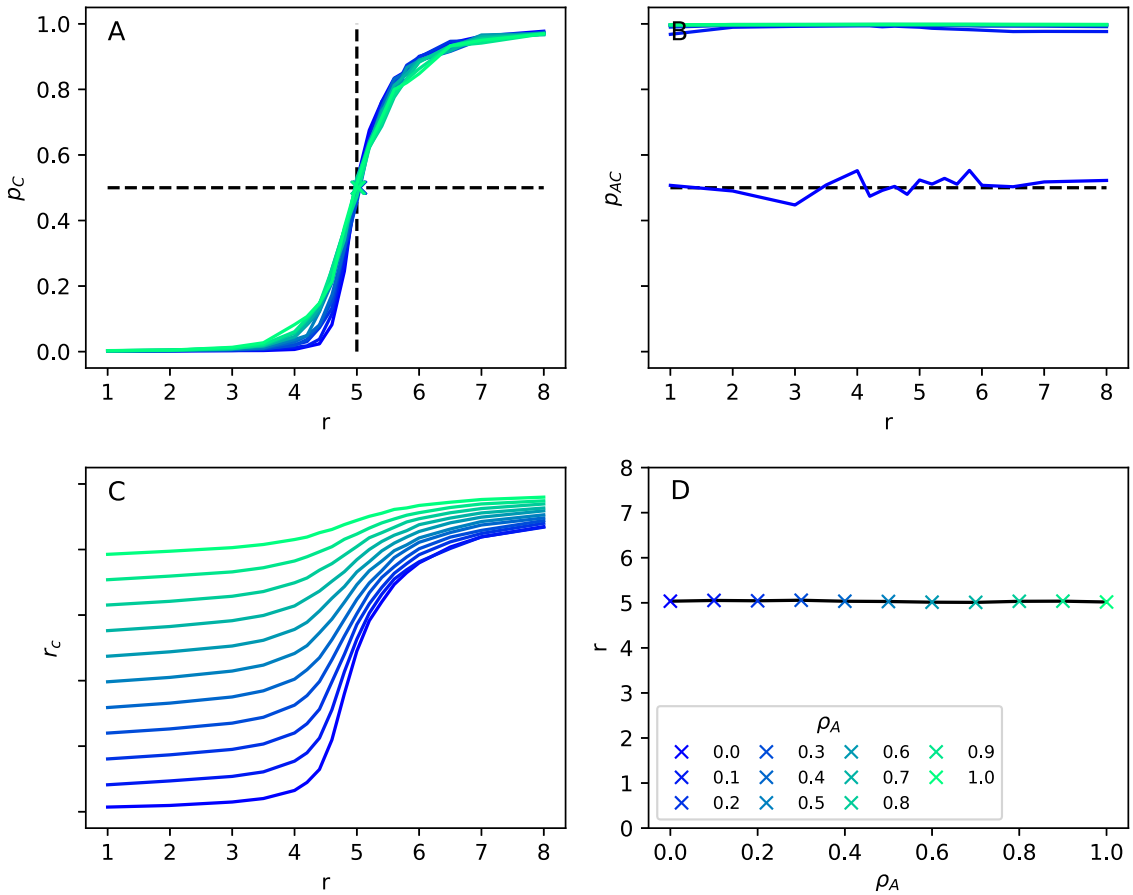

Fig. 6 | Evolution of cooperation when agent behavior can evolve under the control of players for the same experimental setup (except the choice of actions) as for Fig. 5. A Central player's probability to cooperate p_C at the end of evolutionary adaptation as a function of synergy parameter r and AI agent density ρ_A (color code in panel D). **B** Evolved likelihood for an AI agent to cooperate p_{AC} at the end of evolution as a function of synergy r and AI agent density ρ_A. The horizontal dashed line depicts the expected $p_{AC} = 0.5$ probability for evolutionary drift. **C** Population frequency of cooperation (all players and AI agents) as a function of r and ρ_A. **D** Critical synergy value r (y-axis) at which the probability of cooperation for players exceeds 0.5 for all experimental conditions of ρ_A (x-axis).

On the other hand, if the focal player is a defector, the number of cooperators is reduced by mimicry to

$$N_C' = (1 - \rho_A)N_C. \tag{4}$$

We can now calculate the cooperation condition $P_C - P_D \geq 0$ using these substitutions. We find

$$\frac{r(k - N_D' + 1)}{k+1} - 1 \geq \frac{N_C'}{k+1}\ , \tag{5}$$

which simplifies to

$$r \geq \frac{k+1}{\rho_A \cdot k + 1}\ . \tag{6}$$

Key Finding: Surprisingly, this relation is **independent of the number of cooperators and defectors in each neighborhood and only depends on the mimicry probability** (here given by the AI agent density ρ_A) and the group size. It reduces to the dilemma $r \geq k + 1$ in the limit $\rho_A = 0$, and **completely eliminates the dilemma when** $\rho_A = 1$. Figure 7 shows the numerical experiments implementing this scenario and **confirms the prediction** Eq. (6). Simply put, **agents that mimic player behavior reduce the synergy threshold** r **so that cooperation can evolve at much-reduced synergies, promoting collaboration within the group and thus changing player behavior**. This represents the only policy tested that successfully resolves the cooperation dilemma.

Discussion

We explored the impact of different policy scenarios on promoting cooperation in the Public Goods game, specifically in the context of integrating AI agents into human social systems. Our findings highlight the nuanced role that AI agents can play in influencing human cooperative behavior, depending on how their actions are regulated or programmed.

We first examined the **Mandatory Cooperation Policy for AI Agents**, where AI agents are forced institutionally to always cooperate. Contrary to intuitive expectations, this policy did not enhance cooperation among human players. While the overall frequency of cooperative actions increased due to the agents' enforced cooperation, the critical threshold for human players to find cooperation advantageous remained unchanged. This outcome aligns with theoretical predictions that the critical point in the Public Goods game is independent of the frequency of cooperators in a player's immediate environment[1,21]. The inability of this policy to shift the cooperative dynamics (and thus lift the dilemma) suggests that simply injecting cooperative agents into the system is insufficient to influence human behavior in the desired manner.

Next, we investigated the **Player-Controlled Agent Cooperation Policy**, where human players determine the cooperation probability of AI agents in their vicinity. Despite the potential for players to benefit from surrounding themselves with cooperative agents, the evolutionarily optimized strategies did not lead to increased cooperation among players themselves. Players evolved to maximize their own payoffs by programming agents to cooperate while choosing to defect themselves, exploiting the cooperative agents without reciprocating. This mirrors real-world scenarios where individuals might leverage AI systems for personal gain without

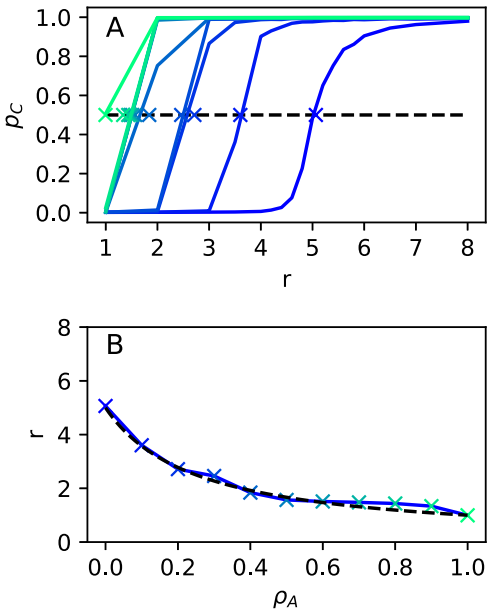

Fig. 7 | Evolution of cooperation when agent behavior mimics player behavior, for different agent densities ρ_A, for the same experimental setup (except the choice of actions) as used for Fig. 5. A players' probability of cooperating at the end of evolutionary adaptation p_C as a function of r and AI agent density ρ_A (color code in Figs. 5 and 6). Critical points denoted by `x'. **B** Critical synergy $r$ at which the probability of cooperation for players exceeds 0.5, obtained from crossover points in panel A marked as `x'. The dashed line represents the predicted synergy factor r for cooperation to evolve given by Eq. (6).

contributing to the collective good, highlighting a potential pitfall in allowing unrestricted control over AI behavior.

In contrast, the **Agents Mimic Players** policy yielded a significant shift in cooperative dynamics and was successful in removing the dilemma. By programming AI agents to mirror the cooperation probability of the central player, we introduced a mechanism akin to conditional cooperation or reciprocity[17]. This policy effectively lowered the synergy threshold (r) required for cooperation to be advantageous for players (see Eq. (6)). As AI agents copy the players' strategies, cooperative players find themselves in increasingly cooperative groups, enhancing their payoffs and promoting the evolution of cooperation. This outcome underscores the power of reciprocity and conditional strategies (or more precisely, the power of information) in fostering cooperative behavior[6,20].

Note that this mimicking strategy could also have AI agents copying the defective behavior of players. However, the mimicking policy of AI agents shifts the critical synergy threshold so that it becomes more lucrative for the players to cooperate. Thus, the "selfish" thing to do in this scenario is to cooperate.

The success of the mimicry policy suggests that AI agents can serve as catalysts for cooperation when their behavior is contingent upon that of human players. By aligning agents' actions with players' choices, we create a feedback loop that rewards cooperative behavior and penalizes defection. This mechanism leverages principles from evolutionary game theory, where strategies that promote mutual cooperation can become stable under certain conditions[15,17]. While the mechanism that creates mimicry is not specified, we can imagine that AI agents use information gathered from observing player actions to mimic the players, thus establishing again that it is information that promotes cooperation[43,44].

Our findings offer vital insights for policymakers and institutions aiming to foster cooperative behavior in an era increasingly dominated by AI technologies. By introducing and validating a novel policy where AI agents mimic human behavior, we provide a transformative framework that could redefine strategies for enhancing cooperation in complex social systems. Mandating cooperation from AI agents or allowing users to control agent behavior without oversight may not lead to the desired increase in human cooperation. Instead, policies that encourage AI systems to respond adaptively to human behavior –specifically by mimicking or reciprocating actions – could foster more cooperative interactions.

The approach outlined here aligns with concepts of responsive regulation and adaptive governance, where rules and systems are designed to be flexible and responsive to the behaviors of individuals within the system[3,36]. By integrating AI agents that promote reciprocity, governments and institutions can harness technology to enhance social welfare without imposing rigid mandates that may be ineffective or counterproductive.

The sharp behavioral transitions observed in our models reflect the non-linear dynamics characteristic of complex adaptive systems, where modest changes in interaction rules can produce qualitatively different collective outcomes[42,43]. This underscores the broader complexity science insight that emergent system-level properties often cannot be predicted from individual agent behaviors alone [44].

The use of evolutionary game theory (EGT) as the framework for modeling human behavior has important limitations. EGT assumes that strategies evolve over time based on their relative success, akin to natural selection in biological systems. However, human decision-making is influenced by a myriad of factors beyond simple payoff maximization, including emotions, social norms, and cognitive biases[45]. Consequently, the extent to which EGT can accurately capture the complexities of human cooperation is debatable. While EGT has been successfully applied in economics and biology to model strategic interactions [13], its assumptions may oversimplify the nuanced processes underlying human behavior.

Moreover, the agents in our simulations are controlled via parameters that directly influence their likelihood to cooperate or defect. In scenarios like "Mandatory Cooperation" or "Player-Controlled Agents," we impose external rules or allow players to evolve control over agents, respectively. However, in real-world settings, the mechanisms by which human behavior would determine agent actions are not straightforward. Autonomous agents, particularly those powered by advanced AIs such as large language models, operate on the basis of complex algorithms and vast datasets[46]. Directly coupling player behavior to agent actions, as done in our models, may not accurately reflect how such systems function in practice. However, governmental rulings on AI policies can pave the path to integrate AI agents such that they promote cooperation in the ways exemplified here.

To address the discrepancy between models and reality, future research could explore more sophisticated models where agent behavior emerges from interactions with players rather than being directly controlled. For instance, incorporating reinforcement learning mechanisms could allow agents to adapt their strategies based on observed player actions over time[47]. This approach would better mirror the iterative and dynamic nature of human-agent interactions, capturing feedback loops and mutual adaptation. It also presents an opportunity to determine agent behavior. As mentioned before, reinforcement learning based on the heuristic to *do upon others as done upon yourself* can implement the mimicking strategy for AI agents, thus promoting cooperation among humans.

Another methodological limitation lies in the assumption of well-mixed populations and the absence of spatial structure in our simulations. In reality, interactions often occur within networks where the structure can significantly impact the evolution of cooperation[48,49]. Incorporating network topology into the model could provide a more accurate representation of social dynamics, potentially altering the outcomes observed in our study. However, we remark that spatial proximity in itself tends to promote cooperation, so we do not expect structured populations to negate our findings.

While our study focuses on three specific AI agent policies: mandated cooperation, player-controlled cooperation, and mimicry, many other known mechanisms for promoting cooperation remain unexplored within this framework. Mechanisms such as direct and indirect reciprocity, kin selection, the green beard effect, and structured interactions (e.g., spatial or network-based populations) have all been shown to support cooperation in both theoretical and empirical settings. However, integrating these into our

model would significantly expand the design space. Additionally, adapting concepts like kinship or green beard effects to artificial agents raises non-trivial questions – what constitutes "kin" among AI agents: shared ownership, shared training data, or brand affiliation? What would serve as a "green beard" signal in robots?

These extensions are beyond the scope of this work but represent promising directions for future exploration and could contribute to understanding mixed human-AI societies. Their omission might limit the generalizability of our findings, but it also highlights the importance of continued research on how diverse social and structural conditions interact with AI behavior.

We note that the "Agents Mimic Players" scenario presupposes that agents can perfectly observe and replicate player behavior without error. In practical applications, noise and imperfect information are inevitable[50]. Introducing stochastic elements or uncertainties into the agents' observation and imitation processes could yield more realistic results, shedding light on how robust the cooperative dynamics are to such imperfections. In addition, our model simplifies the cost and benefit structure of cooperation and defection to a single parameter, the synergy factor r. While this abstraction is useful for theoretical exploration, real-world public goods scenarios involve more complex and often non-linear cost-benefit relationships[3]. Extending the model to include variable costs, diminishing returns, or threshold effects could enhance its applicability to real situations.

Finally, ethical considerations regarding the deployment of autonomous agents that can influence human behavior should not be overlooked. The implementation of agents designed to promote cooperation raises questions about manipulation, consent, and the potential for unintended consequences[51]. However, promoting cooperation is in the interest of all and thus might be easily accepted. Further, ethical AI guidelines generally seek to prevent malicious or deceptive behavior and usually explicitly try to make AI behave "well", which mimicry encourages as long as the player they are emulating behaves well.

This study examined the potential of artificially intelligent agents to promote cooperation in Public Goods games by testing three different scenarios: *Mandatory Cooperation Policy for AI Agents*, *Player-Controlled Agent Cooperation Policy*, and *Agents Mimic Players*. Our computational simulations revealed that only when agents mimic player behavior does the critical synergy threshold r for the emergence of cooperation decrease, thus lifting the dilemma. This finding suggests that agents designed to reciprocate player actions can foster cooperative behavior among humans.

While our model simplifies complex real-world interactions, it provides a theoretical foundation for how AI agents might be leveraged to enhance cooperation in societal dilemmas. The assumption that agents can perfectly mimic player behavior is an idealization; in practice, implementing such strategies may involve challenges related to imperfect information and stochasticity. Future research should explore more sophisticated models that incorporate learning mechanisms, uncertainties, and network structures to better reflect the intricacies of human-agent interactions.

Ethical considerations are paramount when deploying AI agents that can influence human behavior. Ensuring that such agents promote cooperation without infringing on individual autonomy or causing unintended consequences is crucial. As AI continues to integrate into various aspects of daily life, understanding and guiding its impact on social dynamics becomes increasingly important.

In conclusion, our study illuminates a pivotal role for AI agents as catalysts for cooperation in social dilemmas. By uncovering and validating a novel mechanism wherein agents mimic player behavior, we offer a new paradigm that could profoundly impact how cooperative behavior is fostered in increasingly complex societal interactions. By designing agents that reciprocate human actions, we can create environments that encourage cooperative behavior, ultimately contributing to more harmonious and efficient societies.

Methods

Computational evolutionary model

We model the Public Goods game using evolutionary game theory. Each player's behavior is characterized by an individual probability to cooperate (p_C) in a group of k players (see Fig. 1, here we use $k = 4$). The player population is initialized with random strategies ($p_C = 0.5$). The score that each player accumulates by playing multiple ($k + 1$) games with others in a group in each generation determines their likelihood of propagating to the next generation via Roulette Wheel selection[52]. At each generation, a player's cooperation probability may mutate (with rate $\mu = 0.01$). In that case, a new p_C is drawn from a uniform distribution [0.0, 1.0], which will determine how likely the player will cooperate in the next generation. We use a population of 1024 players arranged on a 32×32 grid, which allows us to control the number of interacting neighbors more easily. However, players are replaced independently of their location at every generation, ensuring a well-mixed population.

In each round, some players could be replaced by AI agents and would thus miss out on possibly receiving a payoff. Therefore, each player engages in five games as the "central" player together with k other players or agents. In these games, only the score for the central player accumulates to avoid imbalances due to missed games among the neighbors. The peripheral players or agents receive no payoff until it is their turn to be the central player.[53,54].

Experiments are run for 10,000 generations, and the line of descent (LOD), as well as the number of actions (C or D) taken by the entire population in each generation are recorded. The LOD connects a random player in the last generation through its direct ancestors back to a player in the initial population [15]. Every mutation that fixed in the population (as it led to a selective sweep) will be on this line, which tells the tale of this particular evolutionary history. Since the LOD also contains the most recent common ancestor (MRCA, which is relatively close to the end of the line) we always ignore the last 25% of the LOD in our analyses so as only to include the history beyond the MRCA of the population.

In one scenario where players are allowed to control AI agents, players possess a second probability that also undergoes evolutionary adaptation. The only difference is that this second probability controls AI agents rather than the player's own behavior.

Data availability

The code to create all data and the analysis scripts are available at: https://osf.io/npsmr/.

Code availability

The code will be made available on OpenScienceFramework.

Received: 13 December 2024; Accepted: 24 November 2025;
Published online: 22 January 2026

References

1. Hardin, G. The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *Science* **162**, 1243–1248 (1968).
2. Samuelson, P. A. The pure theory of public expenditure. *Rev. Econ. Stat.* 387–389 (1954).
3. Ostrom, E. *Governing the commons: The evolution of institutions for collective action* (Cambridge University Press, Cambridge, MA, 1990).
4. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
5. Ledyard, J. O. et al. *Public goods: A survey of experimental research* (Division of the Humanities and Social Sciences, California Inst. of Technology, 1994).
6. Ostrom, E. Collective action and the evolution of social norms. *J. Econ.c Perspect.* **14**, 137–158 (2000).

7. Dietz, T., Ostrom, E. & Stern, P. C. The struggle to govern the commons. *Science* **302**, 1907–1912 (2003).
8. Von Neumann, J. & Morgenstern, O. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior* (Princeton university press, 2007).
9. Fudenberg, D. & Tirole, J. *Game theory* (MIT press, 1991).
10. Osborne, M. J. & Rubinstein, A. *A course in game theory* (MIT Press, 1994).
11. Harsanyi, J. C. Games with incomplete information played by "Bayesian" players, I–III Part I. The basic model. *Manag. Sci.* **14**, 159–182 (1967).
12. Fudenberg, D. & Maskin, E. The folk theorem in repeated games with discounting or with incomplete information. *Econometrica* **54**, 533–554 (1986).
13. Maynard Smith, J. *Evolution and the Theory of Games* (Cambridge University Press, Cambridge, UK, 1982).
14. Weibull, J. W. *Evolutionary game theory* (MIT Press, Cambridge, MA, 1995).
15. Adami, C., Schossau, J. & Hintze, A. Evolutionary game theory using agent-based methods. *Phys. Life Rev.* **19**, 1–26 (2016).
16. Hamilton, W. D. The evolution of social behavior. *J. theor. Biol.* **7**, 1–52 (1964).
17. Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).
18. Helbing, D., Szolnoki, A., Perc, M. & Szabó, G. Punish, but not too hard: how costly punishment spreads in the spatial public goods game. *N. J. Phys.* **12**, 083005 (2010).
19. Killingback, T., Bieri, J. & Flatt, T. Evolution in group-structured populations can resolve the tragedy of the commons. *Proc. R. Soc. B: Biol. Sci.* **273**, 1477–1481 (2006).
20. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
21. Hintze, A. & Adami, C. Punishment in public goods games leads to meta-stable phase transitions and hysteresis. *Phys. Biol.* **12**, 046005 (2015).
22. Raihani, N. J. & Bshary, R. Punishment: one tool, many uses. *Evolut. Hum. Sci.* **1**, e12 (2019).
23. Dong, Y., Zhang, B. & Tao, Y. The dynamics of human behavior in the public goods game with institutional incentives. *Sci. Rep.* **6**, 28809 (2016).
24. Hintze, A. et al. Inclusive groups can avoid the tragedy of the commons. *Sci. Rep.* **10**, 22392 (2020).
25. Bach, L. A., Helvik, T. & Christiansen, F. B. The evolution of n-player cooperation-threshold games and ess bifurcations. *J. Theor. Biol.* **238**, 426–434 (2006).
26. Gokhale, C. S. & Traulsen, A. Evolutionary games in the multiverse. *Proc. Natl Acad. Sci.* **107**, 5500–5504 (2010).
27. Hilbe, C., Wu, B., Traulsen, A. & Nowak, M. A. Cooperation and control in multiplayer social dilemmas. *Proc. Natl Acad. Sci.* **111**, 16425–16430 (2014).
28. Peña, J., Wu, B., Arranz, J. & Traulsen, A. Evolutionary games of multiplayer cooperation on graphs. *PLoS Comput. Biol.* **12**, e1005059 (2016).
29. Tamura, K. & Morita, S. Analysing public goods games using reinforcement learning: effect of increasing group size on cooperation. *R. Soc. Open Sci.* **11**, 241195 (2024).
30. Orzan, N., Acar, E., Grossi, D. & Rădulescu, R. Learning in public goods games: the effects of uncertainty and communication on cooperation. *Neural Comput. Appl.* 1–34 (2025).
31. Fernández Domingos, E. et al. Delegation to artificial agents fosters prosocial behaviors in the collective risk dilemma. *Sci. Rep.* **12**, 8492 (2022).
32. Wang, X. et al. Cooperation and coordination in heterogeneous populations. *Philos. Trans. R. Soc. B* **378**, 20210504 (2023).
33. Bowles, S. Policies designed for self-interested citizens may undermine "the moral sentiments": Evidence from economic experiments. *Science* **320**, 1605–1609 (2008).
34. Rahwan, I. et al. Machine behaviour. *Nature* **568**, 477–486 (2019).
35. Cath, C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos. Trans. R. Soc. A* **376**, 20180080 (2018).
36. Floridi, L. & Cowls, J. A unified framework of five principles for ai in society. In Carta, S. (ed.) *Machine Learning and the City: Applications in Architecture and Urban Design*, 535–545 (John Wiley & Sons, Hoboken, NJ, 2022).
37. Taddeo, M. & Floridi, L. How AI can be a force for good. *Science* **361**, 751–752 (2018).
38. Hughes, E. et al. Inequity aversion improves cooperation in intertemporal social dilemmas. In Bengio, S. et al. (eds.) *Advances in Neural Information Processing Systems 31 (NeuIPS 2018)* (Curran Associates, Inc., Red Hook, NY, 2018).
39. Hintze, A. & Dunn, P. T. Whose interests will AI serve? Autonomous agents in infrastructure use. *J. Mega Infrastruct. Sustain. Dev.* **2**, 21–36 (2022).
40. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
41. Adami, C. *The Evolution of Biological Information* (Princeton University Press, Princeton, N.J., 2024).
42. Holland, J. H. Hidden order. *Bus. Week-Domest. Ed.* **21**, 93–136 (1995).
43. Bar-Yam, Y. *Making things work: Solving complex problems in a complex world* (NECSI Knowledge Press, Cambridge, MA, 2016).
44. Epstein, J. M. *Generative social science: Studies in agent-based computational modeling* (Princeton University Press, Princeton, NJ, 2006).
45. Camerer, C. F. *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, Princeton, NJ, 2011).
46. Rafailov, R. et al. Direct preference optimization: Your language model is secretly a reward model. In Oh, A. *et al*. (eds.) *Thirty-seventh Conference on Neural Information Processing Systems*, 53728–53741 (Curran Associates, Inc., Red Hook, NY, 2023).
47. Sutton, R. S. *Reinforcement learning: An introduction* (MIT Press, Cambridge, MA, 2018).
48. Santos, F. C., Pacheco, J. M. & Lenaerts, T. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proc. Natl Acad. Sci.* **103**, 3490–3494 (2006).
49. Szabó, G. & Fáth, G. Evolutionary games on graphs. *Phys. Rep.* **446**, 97–216 (2007).
50. Fudenberg, D. & Levine, D. K. Learning and equilibrium. *Annu. Rev. Econ.* **1**, 385–420 (2009).
51. Bryson, J. J. Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics Inf. Technol.* **20**, 15–26 (2018).
52. Goldberg, D. E.*Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley Publishing Company, Reading, MA, 1989).
53. Iliopoulos, D., Hintze, A. & Adami, C. Critical dynamics in the evolution of stochastic strategies for the iterated prisoner's dilemma. *PLoS Comput. Biol.* **6**, e1000948 (2010).
54. Adami, C. & Hintze, A. Evolutionary instability of zero-determinant strategies demonstrates that winning is not everything. *Nat. Commun.* **4**, 2193 (2013).

Acknowledgements

This work was supported in part through computational resources and services provided by the Institute for Cyber-Enabled Research at Michigan State University.

Author contributions

A.H. conceived the research idea and experimental framework, designed the experiments and developed the methodology, performed the experiments, and analyzed the data. C.A. was responsible for formal analysis. A.H. and C.A. contributed to writing the manuscript by drafting and revising it. Both

authors supervised and coordinated the research work. All authors reviewed the manuscript.

Funding
Open access funding provided by Dalarna University.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Arend Hintze.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025