PNAS RESEARCH ARTICLE | SOCIAL SCIENCES APPLIED MATHEMATICS

On the optimal integration of intelligent agents into network systems to steer cooperation

Feng Fu[a,b,c], Xingru Chen[d], and Nicholas A. Christakis[e,f,1]

Affiliations are included on p. 6.

Contributed by Nicholas A. Christakis; received December 30, 2025; accepted February 5, 2026; reviewed by Alex McAvoy, Matjaz Perc, and David G. Rand

Sociotechnical networks, in which humans and technologies act as interacting entities (also known as "hybrid systems"), increasingly face perturbations by automated agents. What this implies for immersive steering of collective behavior, and how this shapes the stability and resilience of cooperation, remain unclear. Here, we extend evolutionary graph theory by incorporating a distinct type of node representing embedded intelligent agents, namely, algorithmic nodes that autonomously implement prescribed behavioral responses during interactions. These agents are randomly placed within a social network and exert local influence in their neighborhood in social dilemma games. Individual behavior changes are driven by evolutionary dynamics. We derive closed-form analytical results characterizing evolutionary stability and long-run cooperation levels, and show that there exists an optimal, intermediate prevalence of intelligent agents that best promotes cooperation. Our work offers insights into the optimal alignment of human populations with respect to the social good using intelligent agents.

hybrid human–AI systems | cooperative AI | game theory | immersive control

Significance

The advent of large language models, and, more generally, algorithms and AI, poses sociotechnical challenges. Such hybrid human–AI systems underscore the critical need to study collective behavior dynamics and steering control for the greater good, with wide applications in smart governance, crowd management, and intelligent control. Using evolutionary graph theory, we study embedded intelligent agents that exert neutral or differential influences on their local neighborhoods of normal individuals. We show that an intermediate, optimal presence of such agents best promotes cooperation in social networks. Our results offer a theoretical framework for understanding how embedded AI agents can steer cooperation in hybrid human–AI networks.

Multi-agent, "hybrid" systems of humans and AI-endowed agents and optimal distributed intelligent control have been central foci in the interdisciplinary study of decision-making, collective action, and control in network systems, accompanied by a wealth of theoretical progress and applications across fields ranging from statistical physics to computational social science (1–30). Among these efforts, focus has been on various networked dynamical systems under realistic conditions, including involving stochasticity and uncertainty, with applications in diverse areas including self-driving cars (10, 13, 31, 32), robotics (16), democratic consensus and misinformation (33–36), and adversary defense technologies (23, 37). The role of game theory as a framework for approaching these problems has also grown, particularly due to the strategic interactions and learning dynamics among individuals that can take place in such systems (17, 18, 21, 37–39).

In multiagent learning systems (38), integrating game theory is helpful to thoroughly understand distributed intelligent control and optimization (21), addressing problems including but not limited to consensus (29), specialization (27), formation control (17, 23), and cooperative control (22). Recently, there has been increasing interest in merging game theory with AI (6, 39–41), particularly in foundational questions of reinforcement learning (5, 24, 26, 42). The emergence of large language models (LLMs) and, more broadly, AI algorithms, has underscored the importance of understanding their societal and technological impacts in hybrid Human–AI systems to ensure the greater good (4, 28, 32, 43–46). This consideration leads to an open question regarding the optimal control of collective behavior dynamics in hybrid network systems, in which populations are immersed in interactions with intelligent bots (47, 48), such as misinformation fact-checking bots (49) or virtual conversational agents like ChatGPT (28). Here, we focus on studying collective behavior dynamics modeled by evolutionary game theory and explore an optimal steering control problem in population systems with intelligent bots—namely, determining how many bots to deploy and how strongly they should influence their neighbors to optimally steer the system toward desired prosocial outcomes with the highest possible level of cooperation.

The increasing availability of data in this era of big data, coupled with breakthroughs in AI technologies such as deep learning, reinforcement learning, and LLMs, has facilitated widespread adoption in domains ranging from commercial to educational (10, 14, 16, 24, 26, 27, 50). A particularly important emerging direction is to leverage AI technologies to solve collective action problems, including promoting cooperation (7, 8, 51–56) in social networks through exogenous control, for example, by deploying automated agents (bots) that interact with human subjects (4, 50, 57). However, effecting

Author contributions: F.F., X.C., and N.A.C. designed research; F.F., X.C., and N.A.C. performed research; F.F. and X.C. contributed new analytic tools; F.F., X.C., and N.A.C. analyzed data; and F.F., X.C., and N.A.C. wrote the paper.

Reviewers: A.M., The University of North Carolina at Chapel Hill; M.P., Univerza v Mariboru; and D.G.R., Cornell University Cornell Ann S. Bowers College of Computing and Information Science.

Competing interest statement: Co-author N.A.C. and reviewer D.G.R. were coauthors on a 2026 opinion (36).

[1]To whom correspondence may be addressed. Email: nicholas.christakis@yale.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2537939123/-/DCSupplemental.

Published March 16, 2026.

bottom–up behavior and opinion change remains a challenging objective (19, 58, 59). Traditional control theory, applied to networked systems, often assumes little or no agency of individual nodes in problems of distributed control or considers control with a central perspective (3, 60). This top–down assumption is often misaligned with bottom–up opinion and behavior change, where individuals retain autonomy and respond strategically to interventions and to one another, rather than complying blindly.

Because the dynamics of hybrid systems are inherently stochastic, in contrast to deterministic dynamical systems, we must differentiate between stability and population state characterizations (61). To measure stability, we borrow the concept of the fixation probability, widely used in mathematical population genetics (62). The fixation probability gives the likelihood that a population system (resident type, say B) can be overtaken by a new type (mutant, say A). The comparison of fixation probabilities and their relative ratios reveals the long-term behavior of the stochastic system under a low mutation limit (63, 64). This limit assumes that perturbations to the strategy composition of normal individuals (i.e., endogenous changes in strategy type, not the exogenous deployment of bots), such as mutations or explorations as in the reinforcement learning literature (10, 16, 24), are sufficiently infrequent (65).

We also introduce an immersive control concept that essentially introduces heterogeneity of individuals into the systems at hand. Rather than directly controlling the response behavior or opinions of normal individuals (those who act in their own interests) (59), our modeling framework assumes that influence on normal individuals, such as differential behavioral responses that act as reward or punishment feedback, is exerted through a distinct type of individuals: intelligent agents (or bots) with specific characteristics dependent on the context or problem under consideration. We refer to this in-network deployment of bots that continuously interact with and influence normal individuals in a hybrid system as "immersive control": control implemented through agents embedded within, rather than external to, the social network.

Here, "intelligent" is used strictly in an operational and domain-specific sense. We use it to denote algorithmic nodes that autonomously execute a specified behavioral rule or policy, which may condition on local neighborhood information, with the explicit design objective of promoting cooperation (i.e., "social intelligence" in the narrow sense of steering outcomes toward prosocial equilibria). The term is not intended to imply human-level cognition or general, open-ended intelligence; it simply distinguishes these policy-driven agents from normal individuals whose behavior evolves via payoff-based social imitation.

Our interest lies in how to facilitate and promote the greater good in society at large. We use a classic noncooperative game, the Prisoner's Dilemma, to characterize individual behavior strategies and their interactions, which are further structured by a social network. Interestingly, our results show that immersive control via embedded intelligent agents leads to a clear "Goldilocks" effect for cooperation. Too few bots, and they cannot protect cooperators from being exploited; too many, and they start to interfere with the natural formation and spread of cooperation, ultimately making things worse. In between, however, there is an optimal, intermediate level of bot presence that most effectively helps cooperation take hold and persist in a hybrid network (of certain types) evolving dynamically.

Model

We are primarily interested in steering, and optimizing, cooperation through intelligent agents in hybrid human–AI systems, as has been recently empirically studied (4, 50, 57). It is natural to use social dilemma games such as the Prisoner's Dilemma to model interactions between individuals (66) (as a model for other collective action challenges), while also synergistically combining game theory and AI (39).

As illustrated in Fig. 1, we focus specifically on steering collective behavior dynamics with the objective of enhancing and optimizing social network cooperation. To this end, we study stochastic evolutionary game dynamics in structured populations using a symmetric 2×2 game between normal agents, namely, strategies C (cooperate) and D (defect), interspersed within a network with intelligent agents or bots (Fig. 1). The resulting hybrid network system consists of heterogeneous types of individuals, determined by the prevalence of intelligent agents and their specific, preprogrammed influence on normal individuals in their neighborhoods.

The normal individuals are, at any time step, either cooperators (C) or defectors (D). The intelligent agents (E) exert unilateral influence on their neighbors through behavior-specific responses, quantified by α_C and α_D, respectively. The pairwise interactions in each round can be represented as a symmetric normal-form game:

$$\begin{array}{c|cc|c} & \mathrm{C} & \mathrm{D} & \mathrm{E} \\ \hline \mathrm{C} & R & S & \alpha_C \\ \mathrm{D} & T & P & \alpha_D \end{array}. \qquad [1]$$

The social network structure, together with the underlying payoff structure, has a nontrivial effect on cooperation. As shown below and detailed in *SI Appendix*, we consider social dilemmas on a variety of networks beyond the degree-regular graphs that are the focus of our closed-form analysis. Among social dilemmas in which individual and collective interests are at odds, the Prisoner's Dilemma satisfies $T > R > P > S$, whereas the snowdrift

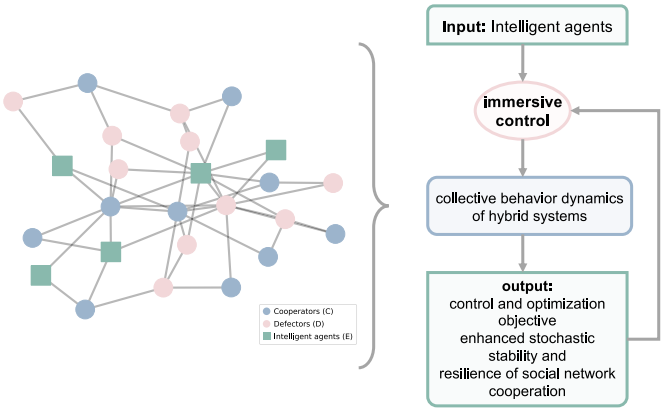

Fig. 1. Steering cooperation in hybrid network systems via intelligent agents. We study stochastic evolutionary games on social networks in which normal individuals (circles) play a symmetric 2×2 game with cooperation (C, blue) and defection (D, red), while intelligent agents (or bots, E, green, squares) are randomly embedded as a distinct additional node type. The extent of heterogeneity in node behavioral types (C, D, or E) across the network is controlled by the prevalence of bots, which exert unilateral influence on their immediate neighbors. By varying this prevalence and the strength of influence that subsequently impacts the payoffs of C and D, we obtain an immersive control of the hybrid network, allowing us to investigate how embedded intelligent agents can steer collective behavior dynamics and enhance cooperation.

game satisfies $T > R > S > P$. The Prisoner's Dilemma is a paradigm for studying prosocial cooperation: In the absence of any mechanism, defection is individually favored even though cooperation is socially optimal. The snowdrift game admits an interior evolutionarily stable strategy (ESS) with a stable mixture of cooperation and defection, but still yields a level of cooperation below the social optimum (67).

Normal individuals update their strategies through social imitation based on payoffs (*Materials and Methods*). In our base model, we restrict imitation to occur only among normal individuals; intelligent agents affect the interactions and payoffs of normal individuals, but not their social imitation. We relax this assumption in model extensions detailed in *SI Appendix* by allowing normal individuals to also imitate the behavioral strategies of intelligent agents, thereby introducing the dual impacts of intelligent agents on cooperation through an interaction/payoff channel (shaping the payoffs that drive selection gradient) and an imitation channel (serving as additional role models for social learning and imitation). To characterize and quantify the impact of the prevalence p_e of intelligent agents on the system's cooperation, we focus on the long-run abundance λ_C of cooperators among normal individuals, and we derive closed-form analytical results for λ_C.

Results

The quantity ρ_C (and analogously ρ_D) describes the probability that a single cooperator (defector) reaches fixation within the subpopulation of normal individuals, starting from an initial state in which all other normal individuals are defectors (cooperators). Thus, ρ_C measures how likely cooperation is to spread and take over a network of "normal" (i.e., non-AI) and otherwise defecting individuals. We derive closed-form analytical results that quantify the impact of intelligent agents through the fixation probability ρ_C and the long-run cooperation level λ_C based on the fixation probabilities using an extended pair approximation method (68) (see *SI Appendix* for details). *SI Appendix*, Fig. S1 shows excellent agreement between our analytical predictions and simulation results for different values of p_e on lattice networks (*SI Appendix*).

In the low-mutation limit, the long-run cooperation level λ_C can be expressed in terms of the ratio ρ_C/ρ_D as

$$\lambda_C = \frac{\rho_C}{\rho_C + \rho_D}. \quad [2]$$

As shown in *SI Appendix*, we obtain a closed-form expression for ρ_C (and ρ_D) up to first order in the strength of selection β. From these expressions, we derive λ_C up to first order in the effective selection strength βN_e, where N_e denotes the total size of the network (including both normal individuals and intelligent agents):

$$\lambda_C = \frac{1}{2} + \frac{\beta N_e \mathcal{C}(k, p_e)}{8(k-1)} \mathcal{F}(k, R, S, T, P, p_e, \alpha_C, \alpha_D) + \mathcal{O}(\beta^2). \quad [3]$$

Here, $\mathcal{C}(k, p_e)$ is a shorthand for the conditional expectation $\mathbb{E}[1 - 1/m \mid m > 0]$, with m denoting the total number of C and D neighbors of a focal normal individual (excluding intelligent agents). As intelligent agents occupy independently with probability p_e each neighboring site, we obtain

$$\mathcal{C}(k, p_e) = \frac{\sum_{m=1}^{k} \left(\frac{m-1}{m}\right) \binom{k}{m} (1-p_e)^m p_e^{k-m}}{1 - p_e^k}. \quad [4]$$

The factor $\mathcal{F}(k, R, S, T, P, p_e, \alpha_C, \alpha_D)$ captures the net effect of selection and depends quadratically on the fraction p_e of intelligent agents:

$$\mathcal{F}(\cdot) = [k(1-p_e) + p_e] \Big\{ (k+1)R + (k-1)S - (k-1)T - (k+1)P + p_e(k-1)\big[2(\alpha_C - \alpha_D) + T + P - R - S\big] \Big\}. \quad [5]$$

In the absence of intelligent agents ($p_e \to 0$), we have $\mathcal{C}(k, p_e) \to (k-1)/k$ and recover the classic result on strategy selection for games on regular networks (62):

$$\lambda_C = \frac{1}{2} + \frac{\beta N_e}{8} \big[(k+1)R + (k-1)S - (k-1)T - (k+1)P\big]. \quad [6]$$

From this expression, cooperation is disfavored, $\lambda_C < 1/2$, when

$$(k+1)R + (k-1)S - (k-1)T - (k+1)P < 0,$$

meaning that network reciprocity alone is insufficient to support cooperation.

To "rescue" cooperation in such cases, we can derive the minimal presence of intelligent agents (acting in one of two representative ways, as noted below) that is required to ensure that cooperation is favored in the long run, $\lambda_C > 1/2$:

$$p_e > p_e^* = \frac{(k-1)T + (k+1)P - (k+1)R - (k-1)S}{(k-1)\big[2(\alpha_C - \alpha_D) + T + P - R - S\big]}. \quad [7]$$

To specify how bots exert influence, we consider two representative types: neutral and prosocial. Neutral bots exert no differential (and zero) influence, so $\alpha_C = \alpha_D = 0$. For prosocial bots, we consider fair-mined Tit-for-Tat (TFT)-like bots that reward cooperation with cooperation and punish defection with defection, leading to $\alpha_C = R$ and $\alpha_D = P$. In general, prosocial bots satisfy $\Delta\alpha = \alpha_C - \alpha_D > 0$, where $\Delta\alpha$ quantifies the differential strength of the bots' prosocial influence.

Because the selection term $\mathcal{F}(\cdot)$ is quadratic in p_e and the accompanying factor $\mathcal{C}(k, p_e)$ decreases as more intelligent agents are added, their interaction yields an optimal intermediate value of p_e^h. In other words, cooperation is maximized at an intermediate prevalence of intelligent agents: not too few, but not too many (Fig. 2).

Intuitively, once the fraction of intelligent agents p_e exceeds the critical threshold p_e^*, cooperation among normal individuals rises above 1/2 (i.e., $\lambda_C > 1/2$); however, in the limit $p_e \to 1$, $\mathcal{C}(k, p_e) \to 0$, so $\lambda_C \to 1/2$ again (*SI Appendix*). Therefore, λ_C must peak at an intermediate prevalence, implying

$$p_e^* < p_e^h < p_{\mathcal{F}} < 1, \quad [8]$$

where $p_{\mathcal{F}}$ is the maximizer (vertex) of the parabola $\mathcal{F}(\cdot)$:

$$p_{\mathcal{F}} = \frac{2k(\alpha_C - \alpha_D) + (2k-1)(T-S) + (2k+1)(P-R)}{2(k-1)\big[2(\alpha_C - \alpha_D) + T - R + P - S\big]}.$$

Take neutral agents as an example (Fig. 2*A*). Even though they exert zero influence, at low p_e their presence reduces direct contact between C and D, as well as the average neighborhood size for

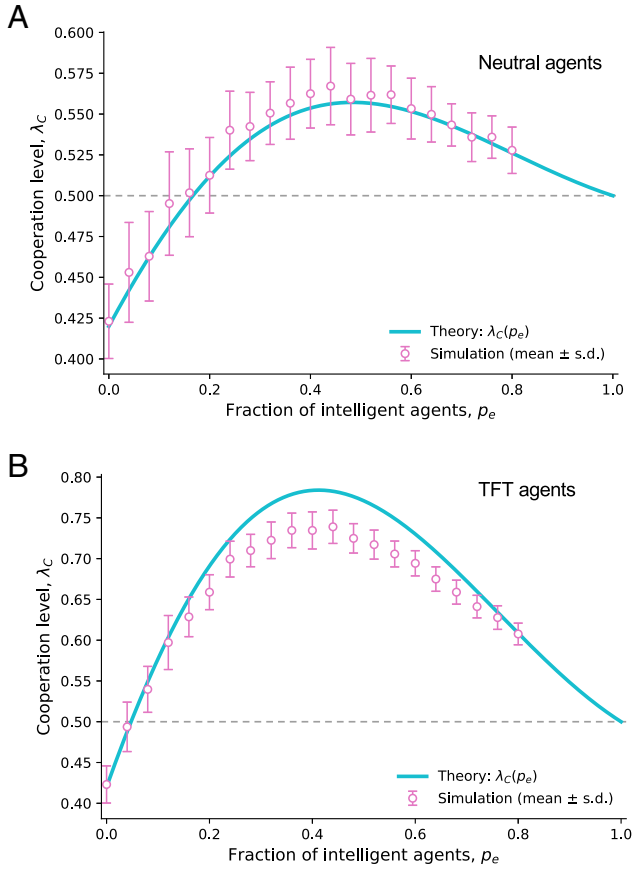

Fig. 2. Optimal prevalence of intelligent agents for promoting cooperation. There exists an intermediate optimal fraction p_e^h of randomly placed bots (*A*: neutral, *B*: TFT-like) that best promotes cooperation. Simulation results agree well with theoretical predictions, with discrepancies arising for large values of p_e. The horizontal dashed line marks $\lambda_C = 1/2$. Parameters: random regular graph of size 800 with degree $k = 4$, strength of selection $\beta = 0.002$, and mutation rate $\mu = 0.0001$; donation game with normalized payoffs $R = 1$, $S = -u$, $T = 1 + u$, and $P = 0$, where $u = c/(b - c) = 0.4$ represents the ratio of cost to the net benefit of mutual cooperation. For each realized instance of randomly placed bots with a given fraction p_e, the cooperation level is averaged over 4×10^9 time steps and then further averaged over 40 independent realizations; the error bars indicate the SD. Panel (*A*): $\alpha_C = \alpha_D = 0$ (neutral bots); Panel (*B*): $\alpha_C = R$, $\alpha_D = P$ (TFT-like bots).

normal individuals, which helps cooperation prevail. Once p_e becomes too large, however, widely interspersed bots can impede the clustering of cooperators and hinder the spread of cooperation. The presence of TFT-like bots yields a similar optimality: their differential effects acting as reward or punishment feedback on their neighborhoods lead to a more pronounced impact on promoting cooperation (Fig. 2*B*).

Our results hold for general social dilemmas with an arbitrary payoff structure. The critical threshold p_e^* ensuring $\lambda_C > 1/2$ takes a particularly simple form for donation games with $R = b - c$, $S = -c$, $T = b$, and $P = 0$. When the network effects alone are insufficient to support cooperation [for example, $b/c < k$ (55)], the minimal presence of TFT-like bots ($\alpha_C = b - c$ and $\alpha_D = 0$) is given by

$$p_e^* = \max\left\{0, \frac{(c/b)\,k - 1}{k - 1}\right\},$$

which scales linearly with the cost-to-benefit ratio c/b and approaches the extreme value 1 as $c/b \to 1$. In the case of neutral bots, $\alpha_C = 0$ and $\alpha_D = 0$, we obtain

$$p_e^* = \max\left\{0, \frac{(c/b)\,k - 1}{(c/b)(k - 1)}\right\},$$

which requires an even larger threshold (Fig. 3*A*). The optimal value of p_e^h that maximizes the cooperation level exhibits a similar trend (Fig. 3*B*). For TFT bots, the optimum lies slightly above c/b and increases with c/b in an accelerating (convex) manner, whereas for neutral bots, the optimal p_e^h increases with c/b in a decelerating (concave) fashion.

Fig. 2 illustrates that the cooperation level increases with the fraction of bots up to an optimal intermediate value, $p_e \in [0.4, 0.5]$, and then declines as p_e increases further beyond this optimum on random regular graphs (see *SI Appendix*, Fig. S2 for spatial lattices in *SI Appendix*). The optimum presence of TFT-like agents is smaller than that of neutral agents (cf. Fig. 2 *A* and *B*). Our analytical predictions agree well with simulation results. Discrepancies arise especially for very high values of $p_e > 0.5$, where bots outnumber normal individuals in the system. This deviation is expected: In this regime, the effective

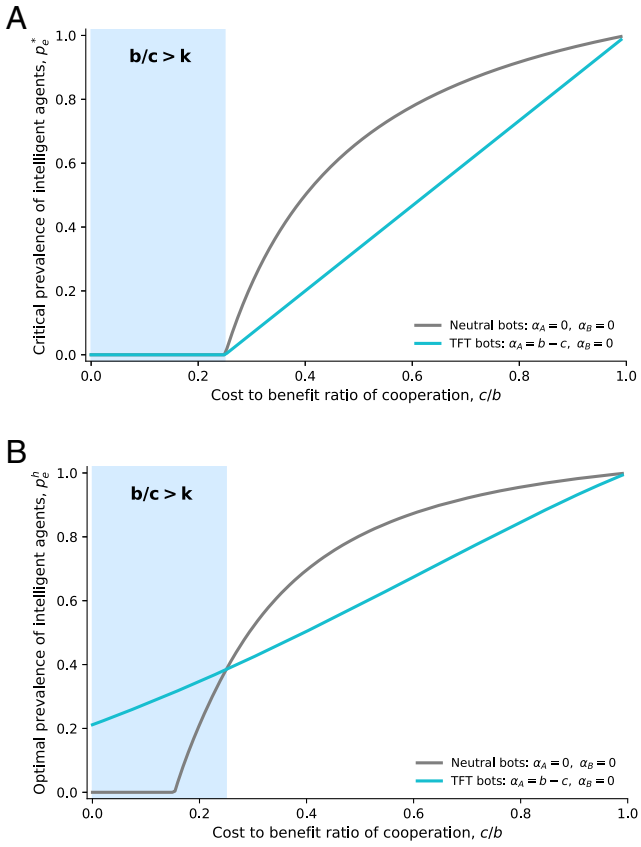

Fig. 3. Comparing neutral bots and TFT-like bots. Panels (*A* and *B*) show minimum and optimal thresholds of randomly placed bots for promoting cooperation as a function of the ratio of cost-to-benefit c/b. Panel (*A*) shows the critical thresholds of randomly placed bots needed to promote cooperation, namely to achieve a long-term cooperation level $\lambda_C > 1/2$. Panel (*B*) shows the optimal thresholds of randomly placed bots that best promote cooperation, achieving the highest possible cooperation level. Even neutral bots that exert zero net effects on their neighborhood can support cooperation, and there exists an intermediate optimal fraction of intelligent agents that best promotes cooperation. TFT-like bots have similar impacts but require both a smaller critical threshold of presence and a smaller optimal fraction. Parameters: lattice degree $k = 4$; donation game with payoffs $R = b - c$, $S = -c$, $T = b$, and $P = 0$.

population size $N_e(1 - p_e)$ becomes so small that the large-population approximation underlying our analytical method breaks down. The existence of an optimum p_e is robust across different strengths of the social dilemma, such as varying c/b values, and across different game types, including the snowdrift game (*SI Appendix*, Fig. S3).

These results thus reveal a "Goldilocks" effect in the prevalence of embedded intelligent agents. When bots are too few—below the minimal threshold required—they cannot provide sufficient shielding for cooperators against defectors. Conversely, when the system is saturated with bots, they can inadvertently harm cooperation by disrupting the dynamic clustering of cooperators and by blocking pathways through which defectors could be converted into cooperators. An intermediate prevalence of bots therefore yields the highest cooperation levels: enough to protect and support cooperative clusters, but not so many as to hinder beneficial strategy updates from defection to cooperation. Notably, neutral intelligent agents act effectively as a uniform dilution of the social ties binding normal individuals, thereby promoting cooperation, albeit to a lesser extent than prosocial, TFT-like agents (Fig. 3). Taken together, this Goldilocks effect exists not only when bots exert prosocial influence (TFT-like agents) but also when they are purely neutral (neutral bots).

Discussion

Our results demonstrate that steering and optimizing collective behavior dynamics can be achieved by integrating intelligent bots as a form of distributed or crowd-sourced control into an underlying social network system. Interestingly, our work shows the existence of an optimal prevalence of intelligent bots for enhancing the system's long-term cooperation level.

This Goldilocks effect has direct implications for the design of human–AI social interventions. It shows that simply increasing the number or strength of automated agents does not guarantee better collective outcomes; beyond a certain point, additional intervention becomes counterproductive. The same intelligent agents that can stabilize cooperative norms can, when deployed in excess or without regard to network effects, inadvertently erode the social dynamics that sustain cooperation. Our theoretical framework can serve as a guiding principle for future human–AI, group level interactions, informing how intervention bots are integrated into social platforms to support, rather than suppress, organic pathways to cooperative behavior.

It is worth noting that the optimal prevalence of bots (p_e^b) for promoting cooperation depends not only on the type of social dilemmas (such as the Prisoner's Dilemma versus snowdrift games) and their difficulty (characterized, for example, by the cost-to-benefit ratio in donation games), but also on the differential impact of bots $\Delta\alpha$. This dependence suggests an intricate potential tradeoff between steering cost and optimization objectives for social interventions, which is not fully understood and warrants further research.

In light of this, we integrate LLM-powered agents into our hybrid simulations under more realistic conditions, including heterogeneous network topologies and sophisticated policies for intelligent agents. The policies of LLM agents can be strategy-dependent on the behavioral choices of their normal neighbors: being generous (like generous TFT, slow to punish but fast to reciprocate), harsh (similar to TFT but with a limited punishment window), or insistent (always cooperate). The placement of these LLM players can be either random or targeted (such as hub-centered deployment). Consistent with our analytical results presented above, there exists a critical threshold of LLM-based agents, along with an optimal level, required to promote cooperation, depending on the policing strength of the LLM agents (see *SI Appendix*, Fig. S6 and Movies S1–S3 for animations of such real-time hybrid simulations).

We also demonstrate the feasibility of leveraging real-time LLM-powered agents to shape the cooperative behavior of normal (non-LLM) individuals in small yet more realistic social networks, thereby paving the way for future human–AI social network cooperation experiments informed by our theoretical results. Although actual human strategy updating may differ from the social imitation process considered here (69, 70), our present framework can be extended to optimize the steering role played by intelligent agents accordingly. Recent empirical studies have demonstrated the effectiveness of belief and attitude change driven by conversational bots (33), suggesting that behavioral nudging through strategy-dependent responses may be synergistically combined with human–AI dialogues to improve cooperation in social network interventions.

Previous studies (primarily using simulations) have shown how a variety of network perturbations, by introducing empty sites (71), zealots (72), and loners (73), can impact the underlying cooperation dynamics. Our present work advances this line of research by providing analytical insights into the optimal prevalence of intelligent agents, with a focus on social network interventions. Our results not only shed light on when and why intelligent agents can optimally steer cooperation, but also inform testable principles for choosing how many bots to deploy and how strongly they should influence their neighbors in order to achieve a desired cooperative outcome.

Thus, our work represents a step toward a theoretical foundation for recent empirical work on hybrid human–AI systems. Pertinent examples include using noisy bots to improve collective action problems such as distributed network coloring (4), deep learning algorithm-based bots for making personal recommendations, such as creating or breaking a tie in network games (50, 57), and semiautomatic steering and braking systems in human coordination experiments with respect to driving (32).

To further the study of optimal control of hybrid network systems, we envisage several promising extensions for future work. First, to steer learning dynamics in repeated interactions effectively, we can take advantage of the class of unbending strategies (74), known to be able to steer the behavioral responses of their payoff-maximizing opponents, even those commonly believed to be formidable extortionate zero-determinant players, also referred to as extortioners, who can control payoff relationships in a specific way (75). Second, to account for a multitude of interactions from different domains—alongside possible extensions to higher-order interactions (7, 8, 76)—we can use the concept of multiplex games, where each layer is characterized by a different type of game, leading to game dynamics on multilayer networks (77–79). Third, the perturbation approach based on the limit of weak selection assumes that the contribution of payoffs from game interactions is infinitesimal, which may not hold in more realistic situations. Therefore, studying nonweak selection will likely be fruitful (80). Last, our current model considers intelligent bots as randomly placed across the network systems. However, not all network positions have equal impact. An exciting extension of the present work would be to strategically place bots in certain network positions (57) (see *SI Appendix*, Fig. S5 for a hub-based placement of LLM-powered agents in degree-heterogeneous networks in *SI Appendix*). Determining how to choose these strategic network positions

for the bots is a highly nontrivial yet important question to address (81).

In sum, we have applied the framework of evolutionary game theory to incorporate immersive control via intelligent bots seamlessly embedded in the network system, aiming to control and optimize the system's stability and the resilience of cooperation. The insights that emerge can be used to enhance cooperation and fairness for the greater good (74, 82), particularly in hybrid human–AI systems. This approach can be applied to a variety of real-world scenarios, including reputation (83) and behavioral norms (84). And this is highly relevant in the era of large language models (85, 86), such as conversational bots (28, 43), and pervasive social media algorithms that influence decision-making (87), often as distorted by misinformation and fake news (49, 88). Considering their impact, addressing AI system alignment problems is crucial to ensure that hybrid systems serve humanity effectively.

Materials and Methods

Extended Pair Approximation Method. We study stochastic game dynamics between normal individuals using either of the two strategies (C versus D) on a network of total size N_e embedded with intelligent agents (E). Without loss of generality, we consider randomly placed intelligent agents across the network, and their prevalence is given by p_e. The unilateral influence exerted by intelligent agents on their neighbors depends on the type of normal individuals, which can be characterized by α_C and α_D, respectively. As such, intelligent agents remain unchanged as a means of immersive control, shaping the evolutionary dynamics exclusively among normal individuals. In the current work, we assume that the interaction graph and the replacement (imitation) graph among normal individuals are identical. To obtain closed-form results, we focus on k-regular networks, such as random regular graphs (Fig. 2) and spatial lattices (*SI Appendix*, Figs. S1–S3). Our results are readily extended to more general scenarios (63), including degree-heterogeneous networks (see *SI Appendix*, Fig. S6 for details). For convenience, the frequency of type X is denoted by p_X, and that of XY pairs is referred to as p_{XY}, where X, Y = {C, D}. Also, we use the conditional probability $q_{X|Y}$ to represent the local density, namely, the conditional probability of finding an X neighbor given a Y individual (68).

The payoff of an individual depends not only on their own strategy but also on the strategies of their neighbors, including intelligent agents. Moreover, the corresponding fitness f of the individual is an exponential function of the accumulated payoff π, that is, $f = \exp(\beta\pi)$, where $\beta > 0$ is the intensity of selection. For a focal C individual, the expected payoff of a neighboring C or D individual can be calculated as

$$\begin{cases} \pi_C^C = R + (k_C^C - 1)(q_{C|C}R + q_{D|C}S + p_e\alpha_C), \\ \pi_C^D = T + (k_C^D - 1)(q_{C|D}T + q_{D|D}P + p_e\alpha_D), \end{cases} \quad [9]$$

where k_C^C and k_C^D are the network degrees of the neighboring C and D individuals. We note that the terms involving $p_e\alpha_C$ and $p_e\alpha_D$ summarize the control impact exerted by the presence of intelligent agents in the network. As an example, we have $k_C^C = k_C^D = k$ in a regular network such as lattice populations with degree k. And the fitnesses of these two types of neighbors are $f_C^C = \exp[\beta\pi_C^C]$ and $f_C^D = \exp[\beta\pi_C^D]$. Analogously, the fitnesses of neighboring C and D individuals of a focal D individual, denoted f_D^C and f_D^D, can be written similarly.

Regarding strategy updating for normal individuals, we focus on the death-birth update rule, which resembles a payoff-based social imitation process (62). That said, we restrict our evolutionary dynamics to the subpopulation of normal individuals. At each step, a focal individual, C or D, is randomly selected and imitates the strategy of a particular neighbor, chosen with probability proportional to their fitness. To be more specific, we assume that a focal C individual at the moment has n_C^C many C neighbors and n_C^D many D neighbors. Based on the updating rule, the probability that the number of cooperators increases or decreases by one is given by

$$\begin{cases} T_C^+ = \mathbb{E}\left[p_D n_D^C f_D^C/(n_D^C f_D^C + n_D^D f_D^D) \;\middle|\; \text{C and D individuals}\right], \\ T_C^- = \mathbb{E}\left[p_C n_C^D f_C^D/(n_C^C f_C^C + n_C^D f_C^D) \;\middle|\; \text{C and D individuals}\right]. \end{cases} \quad [10]$$

Denote by $N = N_e(1 - p_e)$ the subpopulation size of normal individuals. The fixation probability ρ_C is given by ref. 89

$$\rho_C = \frac{1}{1 + \sum_{i=1}^{N-1}\prod_{j=1}^{i}\frac{T_C^-(j)}{T_C^+(j)}}. \quad [11]$$

The ratio ρ_C/ρ_D is given by

$$\frac{\rho_C}{\rho_D} = \prod_{i=1}^{N-1}\frac{T_C^+(i)}{T_C^-(i)}. \quad [12]$$

Abundance of Cooperators Based on Fixation Probabilities. In the low-mutation limit ($\mu \to 0$, denoting the probability that a focal individual randomly chooses C or D with 1/2 probability instead of imitating the chosen neighbor's strategy), the long-run abundance λ_C of cooperators is

$$\lambda_C = \frac{\rho_C}{\rho_C + \rho_D}. \quad [13]$$

As detailed in *SI Appendix*, under the weak selection limit ($\beta \ll 1$), we derive closed-form results for ρ_C and λ_C that quantify the impact of intelligent agents. Our analytical predictions agree well with the agent-based simulation results (see Fig. 2 and *SI Appendix*, Fig. S1–S3, and *SI Appendix* for additional implementation details).

Extensions. In *SI Appendix*, we also consider a variety of extensions, including LLM-powered intelligent agents with different policy structures and their targeted placement in degree-heterogeneous networks.

Data, Materials, and Software Availability. Code has been deposited in GitHub (https://github.com/fufeng/multiLLM) (90). All other data are included in the manuscript and/or supporting information.

ACKNOWLEDGMENTS. This work is supported by the NOMIS Foundation and the Pershing Square Foundation. F.F. is grateful for the generous support of a Dartmouth Senior Faculty Grant and Scholarly Innovation and Advancement award. X.C. acknowledges the support from the National Natural Science Foundation of China (grant no. 12526528) and the Beijing Natural Science Foundation (grant no. 1244045).

Author affiliations: [a]Department of Mathematics, Dartmouth College, Hanover, NH 03755; [b]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, NH 03756; [c]Department of Applied Mathematics, School of Engineering & Applied Science, Yale University, New Haven, CT 06520; [d]School of Artificial Intelligence, Beihang University, Beijing 100191, China; [e]Human Nature Lab, Yale University, New Haven, CT 06520; and [f]Department of Sociology, Yale University, New Haven, CT 06520

1. M. Granovetter, Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
2. D. J. Watts, S. H. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
3. Y. Y. Liu, J. J. Slotine, A. L. Barabási, Controllability of complex networks. *Nature* **473**, 167–173 (2011).
4. H. Shirado, N. A. Christakis, Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370–374 (2017).
5. W. Barfuss *et al.*, Collective cooperative intelligence. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319948121 (2025).
6. U. M. Sehwag, A. McAvoy, J. B. Plotkin, Collective artificial intelligence and evolutionary dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2505860122 (2025).

7. U. Alvarez-Rodriguez *et al*., Evolutionary dynamics of higher-order interactions in social networks. *Nat. Hum. Behav.* **5**, 586–595 (2021).
8. F. Battiston *et al*., Higher-order interactions shape collective human behaviour. *Nat. Hum. Behav.* **9**, 1–17 (2025).
9. T. Zhou, Y. Zhang, On the stability and robust stability of networked dynamic systems. *IEEE Trans. Autom. Control.* **61**, 1595–1600 (2015).
10. K. Yuan *et al*., Evolutionary decision-making and planning for autonomous driving based on safe and rational exploration and exploitation. *Engineering* **33**, 108–120 (2024).
11. Y. Li, J. Chen, E. Tuncel, W. Su, MIMO control over additive white noise channels: Stabilization and tracking by LTI controllers. *IEEE Trans. Autom. Control.* **61**, 1281–1296 (2015).
12. X. Zhao, F. Deng, Stabilization of systems by delayed noisy states. *IEEE Trans. Autom. Control.* **69**, 3764–3779 (2023).
13. Z. Ye, D. Zhang, C. Deng, H. Yan, G. Feng, Finite-time resilient sliding mode control of nonlinear UMV systems subject to dos attacks. *Automatica* **156**, 111170 (2023).
14. X. Zhang, R. Gao, C. Zhu, C. Liu, S. Mei, Ultra-short-term prediction of regional photovoltaic power based on dynamic graph convolutional neural network. *Electr. Power Syst. Res.* **226**, 109965 (2024).
15. Z. M. Wang, K. Z. Liu, S. X. Wen, X. M. Sun, Data-driven switched model predictive control without terminal ingredients. *IEEE Trans. Autom. Sci. Eng.* **21**, 4247–4260 (2023).
16. T. Zhang *et al*., Residual reinforcement learning for motion control of a bionic exploration robot-robodact. *IEEE Trans. Instrum. Meas.* **72**, 1–13 (2023).
17. Y. Li, X. Hu, A differential game approach to intrinsic formation control. *Automatica* **136**, 110077 (2022).
18. N. Liu, L. Guo, Stochastic adaptive linear quadratic differential games. *IEEE Trans. Autom. Control.* **69**, 1066–1073 (2023).
19. R. Gao, G. H. Yang, Resilient cluster consensus for discrete-time high-order multi-agent systems against malicious adversaries. *Automatica* **159**, 111382 (2024).
20. Y. P. Tian, X. Yu, Robust learning control for a class of nonlinear systems with periodic and aperiodic uncertainties. *Automatica* **39**, 1957–1966 (2003).
21. J. Liu, W. Wang, J. Xu, H. Zhang, Stackelberg strategy of two-player stochastic difference game with time delay. *Int. J. Control. Autom. Syst.* **21**, 2904–2915 (2023).
22. W. Gao, Z. P. Jiang, Data-driven cooperative output regulation of multi-agent systems under distributed denial of service attacks. *Sci. China Inf. Sci.* **66**, 190201 (2023).
23. P. Zhou, B. M. Chen, Distributed optimal solutions for multiagent pursuit-evasion games for capture and formation control. *IEEE Trans. Ind. Electron.* **71**, 5224–5234 (2023).
24. Y. Yang, W. Cao, L. Guo, C. Gan, M. Wu, "Reinforcement learning with reward shaping and hybrid exploration in sparse reward scenes" in *2023 IEEE 6th International Conference on Industrial Cyber-Physical Systems (ICPS)* (IEEE, 2023), pp. 1–6.
25. Y. Yu, X. Li, L. Li, L. Xie, Distributed online optimization for heterogeneous linear multi-agent systems with coupled constraints. *Automatica* **159**, 111407 (2024).
26. S. A. A. Rizvi, Z. Lin, A note on state parameterizations in output feedback reinforcement learning control of linear systems. *IEEE Trans. Autom. Control.* **68**, 6200–6207 (2022).
27. C. Li *et al*., Celebrating diversity with subtask specialization in shared multiagent reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* **36**, 2051–2065 (2023).
28. H. Chen *et al*., Feedback is all you need: From ChatGPT to autonomous driving. *Sci. China Inf. Sci.* **66**, 1–3 (2023).
29. C. Bernardo *et al*., Achieving consensus in multilateral international negotiations: The case study of the 2015 Paris agreement on climate change. *Sci. Adv.* **7**, eabg8068 (2021).
30. X. Guan, C. Song, Y. C. Ho, Q. Zhao, Constrained ordinal optimization-a feasibility model based approach. *Discret. Event Dyn. Syst.* **16**, 279–299 (2006).
31. J. F. Bonnefon, A. Shariff, I. Rahwan, The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
32. H. Shirado, S. Kasahara, N. A. Christakis, Emergence and collapse of reciprocity in semiautomatic driving coordination experiments with humans. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2307804120 (2023).
33. H. Lin *et al*., Persuading voters using human-artificial intelligence dialogues. *Nature* **648**, 394–401 (2025).
34. K. Hackenburg *et al*., The levers of political persuasion with conversational artificial intelligence. *Science* **390**, eaea3884 (2025).
35. N. Köbis *et al*., Delegation to artificial intelligence can increase dishonest behaviour. *Nature* **646**, 126–134 (2025).
36. D. T. Schroeder *et al*., How malicious ai swarms can threaten democracy. *Science* **391**, 354–357 (2026).
37. Z. Cheng, G. Chen, Y. Hong, Zero-determinant strategy in stochastic stackelberg asymmetric security game. *Sci. Rep.* **13**, 11308 (2023).
38. K. Tuyls, S. Parsons, What evolutionary game theory tells us about multiagent learning. *Artif. Intell.* **171**, 406–416 (2007).
39. A. Traulsen, N. E. Glynatsi, The future of theoretical evolutionary game theory. *Philos. Trans. R. Soc. B* **378**, 20210508 (2023).
40. L. Wang, F. Fu, X. Chen, Mathematics of multi-agent learning systems at the interface of game theory and artificial intelligence. *Sci. China Inf. Sci.* **67**, 166201 (2024).
41. T. A. Han, A. Antonioni, A. S. Teixeira, Focus on game theory and AI in complex systems. *J. Phys. Complex.* **6**, 030201 (2025).
42. Y. Xu *et al*., Reinforcement learning and collective cooperation on higher-order networks. *Knowl. Based Syst.* **301**, 112326 (2024).
43. M. L. Traeger, S. Strohkorb Sebo, M. Jung, B. Scassellati, N. A. Christakis, Vulnerable robots positively shape human conversational dynamics in a human-robot team. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 6370–6375 (2020).
44. A. Ueshima, M. I. Jones, N. A. Christakis, Simple autonomous agents can enhance creative semantic discovery by human groups. *Nat. Commun.* **15**, 5212 (2024).
45. F. P. Santos, Prosocial dynamics in multiagent systems. *AI Mag.* **45**, 131–138 (2024).
46. D. Pedreschi *et al*., Human-AI coevolution. *Artif. Intell.* **339**, 104244 (2025).
47. C. Shen, Z. He, L. Shi, Z. Wang, J. Tanimoto, Prosocial punishment bots breed social punishment in human players. *J. R. Soc. Interface* **21**, 20240019 (2024).
48. A. Harrell, M. L. Traeger, Reputation-based reciprocity in human-bot and human-human networks. *PNAS Nexus* **4**, pgaf150 (2025).
49. M. Mosleh, C. Martel, D. Eckles, D. G. Rand, Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022761118 (2021).
50. K. R. McKee *et al*., Scaffolding cooperation in human groups with deep reinforcement learning. *Nat. Hum. Behav.* **7**, 1787–1796 (2023).
51. J. H. Fowler, N. A. Christakis, Cooperative behavior cascades in human social networks. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 5334–5338 (2010).
52. D. G. Rand, S. Arbesman, N. A. Christakis, Dynamic social networks promote cooperation in experiments with humans. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19193–19198 (2011).
53. D. G. Rand, M. A. Nowak, Human cooperation. *Trends Cogn. Sci.* **17**, 413–425 (2013).
54. H. Shirado, F. Fu, J. H. Fowler, N. A. Christakis, Quality versus quantity of social ties in experimental cooperative networks. *Nat. Commun.* **4**, 2814 (2013).
55. D. G. Rand, M. A. Nowak, J. H. Fowler, N. A. Christakis, Static network structure can stabilize human cooperation. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 17093–17098 (2014).
56. A. Nishi, H. Shirado, D. G. Rand, N. A. Christakis, Inequality and visibility of wealth in experimental social networks. *Nature* **526**, 426–429 (2015).
57. H. Shirado, N. A. Christakis, Network engineering using autonomous agents increases cooperation in human groups. *iScience* **23**, 101438 (2020).
58. X. Wang, F. Fu, Eco-evolutionary dynamics with environmental feedback: Cooperation in a changing world. *Europhys. Lett.* **132**, 10001 (2020).
59. B. Wu, J. Du, L. Wang, "Bridging the gap between opinion dynamics and evolutionary game theory: Some equivalence results" in *2020 39th Chinese Control Conference (CCC)* (IEEE, 2020), pp. 6707–6714.
60. G. Yan *et al*., Spectrum of controlling and observing complex networks. *Nat. Phys.* **11**, 779–786 (2015).
61. C. E. Tarnita, H. Ohtsuki, T. Antal, F. Fu, M. A. Nowak, Strategy selection in structured populations. *J. Theor. Biol.* **259**, 570–581 (2009).
62. H. Ohtsuki, C. Hauert, E. Lieberman, M. A. Nowak, A simple rule for the evolution of cooperation on graphs and social networks. *Nature* **441**, 502–505 (2006).
63. B. Allen *et al*., Evolutionary dynamics on any population structure. *Nature* **544**, 227–230 (2017).
64. A. McAvoy, B. Allen, Fixation probabilities in evolutionary dynamics under weak selection. *J. Math. Biol.* **82**, 14 (2021).
65. D. Fudenberg, L. A. Imhof, Imitation processes with small mutations. *J. Econ. Theory* **131**, 251–262 (2006).
66. M. Harper *et al*., Reinforcement learning produces dominant strategies for the iterated prisoner's dilemma. *PLoS One* **12**, e0188046 (2017).
67. M. Doebeli, C. Hauert, Models of cooperation based on the prisoner's dilemma and the snowdrift game. *Ecol. Lett.* **8**, 748–766 (2005).
68. F. Fu, L. Wang, M. A. Nowak, C. Hauert, Evolutionary dynamics on graphs: Efficient method for weak selection. *Phys. Rev. E* **79**, 046707 (2009).
69. A. Traulsen, D. Semmann, R. D. Sommerfeld, H. J. Krambeck, M. Milinski, Human strategy updating in evolutionary games. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 2962–2966 (2010).
70. C. Gracia-Lázaro *et al*., Heterogeneous networks do not promote cooperation when humans play a prisoner's dilemma. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12922–12926 (2012).
71. M. A. Nowak, S. Bonhoeffer, R. M. May, More spatial games. *Int. J. Bifurc. Chaos* **4**, 33–56 (1994).
72. N. Masuda, Evolution of cooperation driven by zealots. *Sci. Rep.* **2**, 646 (2012).
73. G. Szabó, J. Vukov, Cooperation for volunteering and partially random partnerships. *Phys. Rev. E* **69**, 036107 (2004).
74. X. Chen, F. Fu, Outlearning extortioners: Unbending strategies can foster reciprocal fairness and cooperation. *PNAS Nexus* **2**, pgad176 (2023).
75. A. McAvoy *et al*., Unilateral incentive alignment in two-agent stochastic games. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2319927121 (2025).
76. D. Wang, P. Yi, G. Yan, F. Fu, Evolutionary dynamics of pairwise and group cooperation in heterogeneous social networks. *IEEE Trans. Netw. Sci. Eng.* **13**, 5074–5091 (2025).
77. T. Khoo, F. Fu, S. Pauls, Spillover modes in multiplex games: Double-edged effects on cooperation and their coevolution. *Sci. Repo.* **8**, 6922 (2018).
78. Q. Su, A. McAvoy, Y. Mori, J. B. Plotkin, Evolution of prosocial behaviours in multilayer populations. *Nat. Hum. Behav.* **6**, 338–348 (2022).
79. N. Nakis, S. Lehmann, N. A. Christakis, M. Mørup, Modeling roles and trade-offs in multiplex networks. arXiv [Preprint] (2025). http://arxiv.org/abs/2508.05488 (Accessed 3 February 2026).
80. B. Wu, J. García, C. Hauert, A. Traulsen, Extrapolating weak selection in evolutionary games. *PLoS Comput. Biol.* **9**, e1003381 (2013).
81. M. I. Jones, S. D. Pauls, F. Fu, Containing misinformation: Modeling spatial games of fake news. *PNAS Nexus* **3**, pgae090 (2024).
82. A. McAvoy, J. Kates-Harbeck, K. Chatterjee, C. Hilbe, Evolutionary instability of selfish learning in repeated games. *PNAS Nexus* **1**, pgac141 (2022).
83. Y. Murase, C. Hilbe, Computational evolution of social norms in well-mixed and group-structured populations. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2406885121 (2024).
84. L. L. Cavalli-Sforza, M. W. Feldman, *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton University Press, Princeton, NJ, 1981).
85. M. Perc, Counterfeit judgments in large language models. *Proc. Natl. Acad. Sci. U.S.A.* **122**, e2528527122 (2025).
86. A. S. Pires, L. Samson, S. Ghebreab, F. P. Santos, How large language models judge and influence human cooperation. arXiv [Preprint] (2025). http://arxiv.org/abs/2507.00088 (Accessed 3 February 2026).
87. V. V. Vasconcelos *et al*., Segregation and clustering of preferences erode socially beneficial coordination. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102153118 (2021).
88. A. J. Stewart *et al*., Information gerrymandering and undemocratic decisions. *Nature* **573**, 117–121 (2019).
89. M. A. Nowak, A. Sasaki, C. Taylor, D. Fudenberg, Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
90. F. Fu, X. Chen, N. Christakis, Agent-based simulation code for hybrid systems. GitHub. https://github.com/fufeng/multiLLM. Deposited 4 March 2026.