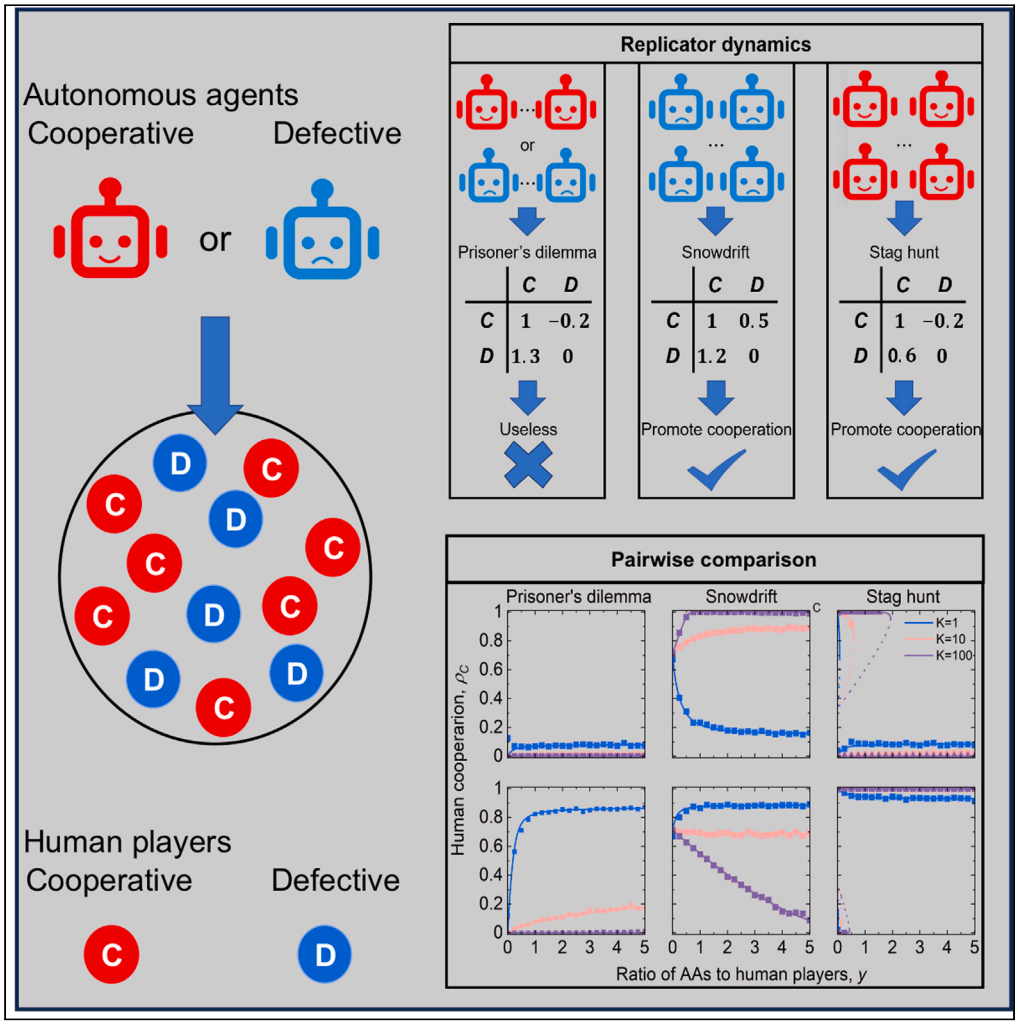
iScience

CellPress
OPEN ACCESS

Article

Facilitating cooperation in human-agent hybrid populations through autonomous agents



Hao Guo, Chen Shen, Shuyue Hu, Junliang Xing, Pin Tao, Yuanchun Shi, Zhen Wang

jlxing@tsinghua.edu.cn (J.X.)
zhenwang0@gmail.com (Z.W.)

Highlights

Modeling cooperative and defective autonomous agents in social dilemmas

In snowdrift/stag hunt games, a minority of autonomous agents drives full cooperation

More autonomous agents can disrupt cooperation

Autonomous agents at hub nodes wield influence

Guo et al., iScience 26, 108179
November 17, 2023
https://doi.org/10.1016/j.isci.2023.108179

iScience

CellPress
OPEN ACCESS

Article

Facilitating cooperation in human-agent hybrid populations through autonomous agents

Hao Guo,[1,2] Chen Shen,[3] Shuyue Hu,[4] Junliang Xing,[2,*] Pin Tao,[2] Yuanchun Shi,[2] and Zhen Wang[1,5,*]

SUMMARY

Cooperative AI has shown its effectiveness in solving the conundrum of cooperation. Understanding how cooperation emerges in human-agent hybrid populations is a topic of significant interest, particularly in the realm of evolutionary game theory. In this article, we scrutinize how cooperative and defective Autonomous Agents (AAs) influence human cooperation in social dilemma games with a one-shot setting. Focusing on well-mixed populations, we find that cooperative AAs have a limited impact in the prisoner's dilemma games but facilitate cooperation in the stag hunt games. Surprisingly, defective AAs can promote complete dominance of cooperation in the snowdrift games. As the proportion of AAs increases, both cooperative and defective AAs have the potential to cause human cooperation to disappear. We then extend our investigation to consider the pairwise comparison rule and complex networks, elucidating that imitation strength and population structure are critical for the emergence of human cooperation in human-agent hybrid populations.

INTRODUCTION

Cooperation, which serves as a fundamental social behavior,[1,2] plays a crucial role in ensuring human prosperity. It not only facilitates the resolution of individual conflicts, such as hunting and driving, but also mitigates burdensome catastrophes such as global climate change and disease transmission.[3,4] However, cooperation often struggles to survive in the face of competition with defection due to lower payoffs.[5] Although mutual cooperation is beneficial to collective interests, individuals are frequently tempted to choose defection. The concept of social dilemma captures the inherent challenge in the evolution of cooperation, referring to a situation where an individual's interests conflict with collective interests. Two-player social dilemma games such as prisoner's dilemma (*PD*) game, stag hunt (*SH*) game, and snowdrift (*SD*) game, are employed ubiquitously, portraying the rational decision-making of two participants using a strategy set and payoff matrix.[6,7] This type of matrix game allows for equilibrium analysis and has been extensively utilized in research within the fields of social science, biology, and artificial intelligence (AI).[8,9]

With the integration of AI into various aspects of human life, advancements in science and technology have allowed humans to delegate decision-making tasks to machines.[10–12] Although previous studies have suggested fascinating solutions to encourage cooperation in human-human interactions,[13,14] they do not address this problem within human-agent hybrid populations. Consequently, research on human-agent coordination has gained significant attention and encompasses diverse areas. One typical example is autonomous driving,[15] where humans relinquish decision-making power to cars, thereby freeing themselves from the physical demands of driving and making travel easier and more enjoyable. However, most of these researches focus on situations where humans and agents share a *common goal*.[16,17] When conflicts of interest arise, it becomes crucial to investigate the evolution of human behavior in a human-agent hybrid environment.[18] For instance, studying human-AI interaction within the general-sum environment and the trust-revenge game provides a comprehensive understanding of these domains.[19,20] As social interactions have become more hybrid,[16,21] involving humans and AAs, there lies an opportunity to gain new insights into how human cooperation is affected.[18,22] This work aims to examine the influence of AAs on human cooperative behavior when social dilemmas exist.

Understanding how human behavior changes in the presence of robots or AAs is a challenging but essential topic.[23,24] To accurately capture human cooperation toward agents, previous studies have predominantly concentrated on the development and design of algorithms for AAs.[18,25,26] These studies mainly focused on repeated games where human players (HPs) can make decisions based on historical information about AAs. A repeated game involves the repetition, either finite or infinite, of a base game, which is represented in the strategic form.[27] However, the impact of one-shot settings, where players lack prior experience and information about their counterparts, has been generally ignored with few exceptions.[28,29] In this article, we focus on how cooperative and defective AAs affect human cooperation in two-player social

[1]School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[3]Faculty of Engineering Sciences, Kyushu University, Kasuga-koen, Kasuga-shi, Fukuoka 816-8580, Japan
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[5]Lead contact
*Correspondence: jlxing@tsinghua.edu.cn (J.X.), zhenwang0@gmail.com (Z.W.)
https://doi.org/10.1016/j.isci.2023.108179

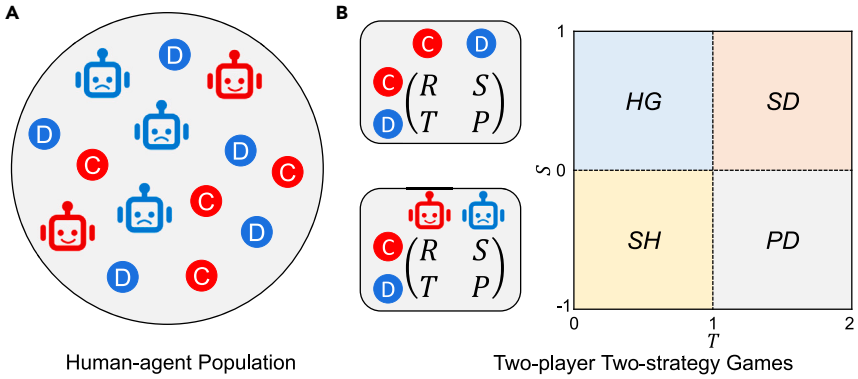


Figure 1. Schematic representation of the human-agent hybrid population

(A) The well-mixed population consists of human players and AAs interacting with each other. The frequencies of human-human and human-agent interactions depend on the composition of the population, specifically the ratio of AAs to human players. The red and blue solid circles respectively represent cooperators and defectors in the human players.

(B) The two-player two-strategy games in human-human and human-agent interactions. Mutual cooperation yields a reward R to both players, while mutual defection results in a punishment P. Unilateral cooperation leads to a sucker's payoff S, while the corresponding defector receives a temptation to defect T. By setting $R = 1$ and $P = 0$, there are four kinds of games, including prisoner's dilemma game, snowdrift game, stag hunt game, and harmony game.

dilemma games, and we ask: Are cooperative AAs always beneficial to human cooperation? Do defective AAs consistently impede the evolution of human cooperation? How do population dynamics change in structured and unstructured populations when the ratio of human-human interaction to human-agent interaction varies?

To address the research questions mentioned above, we utilize an evolutionary game theoretic framework to study the conundrum of cooperation in social dilemma games with a one-shot setting. As shown in Figure 1, the typical games involve *PD*, *SD*, and *SH* games. Previous evidence has proved (or hypothesized) that human players update strategies according to payoff differences, with social learning being the most well-known modality.[30–32] Therefore, we examine the evolutionary dynamics of human cooperation by employing replicator dynamics (RD) and pairwise comparison.[32–34] The main difference between these two dynamics lies in the consideration of AAs. In the pairwise comparison rule, human players can adopt the strategies of AAs, whereas replicator dynamics do not incorporate such imitation.

In the human-agent hybrid population, the fraction of cooperation among human players is denoted as ρ_C ($0 \leq \rho_C \leq 1$). Assuming human players are one unit, add y units of AAs to the hybrid population. Specifically, AAs are programmed to choose cooperation with a fixed probability φ that remains constant over time. The summary of the notations is given in Table 1. Our findings indicate that in well-mixed populations with replicator dynamics, AAs have little impact on cooperation in games with a dominant strategy. Cooperative agents facilitate cooperation in the *SH* game but undermine cooperation in the *SD* game. Counterintuitively, seemingly harmful defective agents can support the dominance of cooperation in the *SD* game. Additionally, we conduct stability analysis and establish the conditions for the prevalence of cooperation. Our results demonstrate that even a minority of defective (or cooperative) AAs can significantly enhance human cooperation in *SD* (or *SH*) games. However, when cooperative (or defective) AAs constitute a majority, they may trigger the collapse of cooperation in *SD* (or *SH*) games. These findings are further corroborated by pairwise comparison rule when strong imitation strength is considered. In scenarios with weak imitation strength, cooperative AAs are more likely to stimulate human cooperation. In contrast to well-mixed populations, where players can interact with others with an equal probability, structured populations restrict interactions to locally connected neighbors. Such a difference in interactive environments is deemed a determinate factor influencing the emergence of cooperation.[35] To investigate this, we conduct experimental simulations on complex networks and observe that structured populations yield comparable results to well-mixed scenarios, except in the case of heterogeneous networks. This divergence can be attributed to the influential role of nodes with higher degrees. Our results, taken together, provide valuable insights into the impact of AAs on human cooperation.

RESULTS

In this section, we mainly present the theoretical results of well-mixed populations in the *PD*, *SD*, and *SH* games by analyzing the replicator equations and pairwise comparison rule. The replicator equation is a differential equation, depicting the growth of a specific strategy based on the payoff difference $\pi_C - \pi_D$, where π_C and π_D mean the expected payoff of cooperation and defection, respectively. Pairwise comparison, guided by the Fermi rule, elucidates the process of strategy updating. Lastly, we introduce an extension to incorporate complex networks into the analysis.

Replicator dynamics

Prisoner's dilemma game and harmony game

In the *PD* game, despite the presence of AAs, defection remains the dominant strategy, and the expected payoff for defection is either equal to or greater than that of cooperation, leading to $\dot{\rho}_C \leq 0$. Therefore, evolutionary dynamics reach a full defection equilibrium state, and human

Table 1. Summary of notations

Notation	Meaning
y	The ratio of autonomous agents to human players
$\varnothing$	The cooperation probability of autonomous agents
ρ_C	The frequency of human cooperation among human players
$\dot{\rho}_C$	ρ_C's derivative with respect to time
K	The imitation strength of human player

cooperation diminishes irrespective of its initial frequency. This finding remains robust against any cooperation probability of AAs, as shown in Figure 2A. Although the equilibrium remains constant, the gradient is influenced by the values of y and φ. In contrast, the expected payoff of cooperation is equal to or larger than defection (see Figure 2B) in the harmony game, leading to $\dot{\rho}_C \geq 0$. Thereby, the population evolves to a full cooperation equilibrium state. The results show that both cooperative and defective AAs have no effect on the convergent state in the games with a dominant strategy.

Snowdrift game

In the *SD* game, when AAs are absent, replicator dynamics have demonstrated that the interior equilibrium $\widehat{\rho}_C = \frac{P-S}{R+P-T-S}$ is the unique asymptotically stable state, while the equilibria $\rho_C = 0$ and $\rho_C = 1$ are always unstable. However, the results will be different if we take AAs into consideration. We find that the population converges to a full cooperation equilibrium state regardless of the initial frequency of human cooperation, provided the condition $\frac{1+y\varphi}{1+y} \leq \frac{P-S}{R+P-T-S}$ is satisfied. Interestingly, to achieve a full cooperation equilibrium state with as few AAs as possible, the optimal approach is to introduce AAs with a cooperation probability of $\varphi = 0$. Correspondingly, the minimum value for y is $\frac{T-R}{S-P}$. On the other hand, when $\frac{\varphi}{1+y} \geq \frac{P-S}{R+P-T-S}$, the population converges to a full defection equilibrium state regardless of the initial frequency of human cooperation. Therefore, we reveal that defective AAs can actually stimulate human cooperation in the *SD* games.

We then present analytical results regarding how AAs affect the equilibrium by setting $T = 1.2$ and $S = 0.5$ in Figure 3. Panel A depicts the $\varphi - y$ phase diagram, which consists of three parts: full *C* state, full *D* state, and coexistence state of *C* and *D*. We find that a minority of defective AAs (see the area with $y < 1$) can shift the equilibrium state from a coexistence state to a full cooperation equilibrium state, whereas a sufficiently large fraction of cooperative AAs drives the system to a full defection equilibrium state (blue color). To better understand how such an unexpected full defection state happens, we examine the frequency of human cooperation as a function of AAs' cooperation probability φ at $y = 4$ in Figure 3B. We observe that with a fixed ratio of AAs to human players y, the lower the cooperation probability of AAs, the higher the level of human cooperation. In detail, the unique asymptotically stable state moves from full cooperation to the coexistence of *C* and *D*, and ultimately to complete defection with the increase of φ.

Stag hunt game

In the absence of AAs, replicator dynamics have revealed that the coexistence of *C* and *D* is an unstable equilibrium state, whereas the full cooperation and defection equilibrium states are both asymptotically stable. The equilibrium state that the population evolves in depends on the initial frequency of human cooperation. However, the full cooperation equilibrium state provides each player with a higher payoff

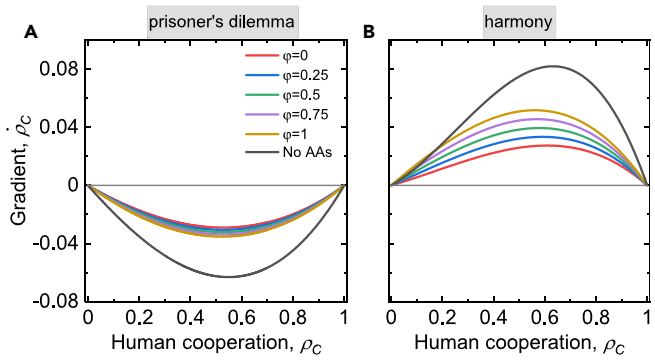


Figure 2. Phase portrait of *PD* and harmony games

(A) In the *PD* game, additional AAs have no influence on the equilibrium. Regardless of whether cooperative or defective AAs are involved, the population inevitably evolves toward a state of complete defection due to the constant condition $\dot{\rho}_C \leq 0$.

(B) In the harmony game, the introduction of additional AAs also has no effect on the equilibrium. Since $\dot{\rho}_C \geq 0$ consistently holds, the population evolves to a full cooperation state despite cooperative or defective AAs. The black line means the situation in the absence of AAs. The parameters are fixed as $y = 0.95$, $S = -0.2$, $T = 1.3$ for the *PD* game and $S = 0.1$, $T = 0.5$ for the harmony game.

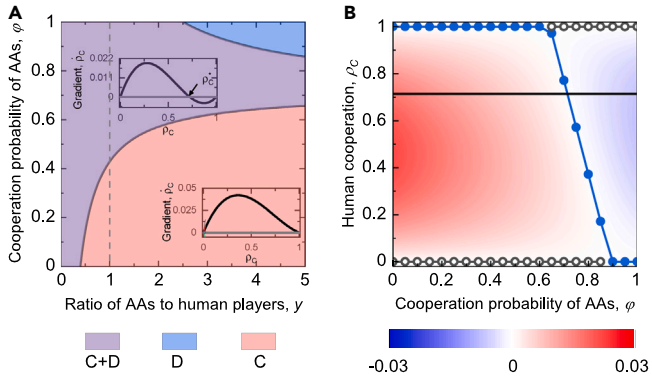


Figure 3. Equilibrium and phase diagram in the *SD* game

(A) A phase diagram of stable equilibrium state as a function of (φ, y) pair. The blue, purple, and red areas mean the full *C* state, the coexistence state of *C* and *D*, and the full *D* state, respectively. The insets are phase portraits of human cooperation ρ_C. In the full *C* state, the gradient of ρ_C is always not less than 0. In the region where cooperation and defection coexist, the gradient of ρ_C is non-negative when $\rho_C < \rho_C^*$, and non-positive otherwise. The dashed line $y = 1$ indicates an equal contribution of human players and AAs, each accounting for 0.5.

(B) The stable (solid circles) and unstable (open circles) equilibrium states as a function of φ with $y = 4$. The background means the gradient of ρ_C, which reveals the evolutionary direction of this population. The change of stable equilibrium shows that compared with cooperative AAs, defective AAs are more beneficial to the evolution of cooperation. Other parameters are fixed as $S = 0.5$ and $T = 1.2$.

compared to the full defection equilibrium state. Consequently, the question arises of how to steer the population toward a full cooperation equilibrium state that is independent of the initial frequency of human cooperation. This can be addressed by incorporating AAs, particularly cooperative AAs. In detail, when $\frac{y\varphi}{1+y} \geq \frac{P-S}{R+P-T-S}$, the full cooperation equilibrium state becomes a unique asymptotically stable solution, implying that the population converges to full cooperation irrespective of initial frequency of human cooperation. In particular, to achieve full cooperation with as few AAs as possible, the best option is to introduce AAs with cooperation probability $\varphi = 1$. On the other hand, the population converges to full defection regardless of the initial frequency of human cooperation when $\frac{1+y\varphi}{1+y} \leq \frac{P-S}{R+P-T-S}$.

In Figure 4A, we present the phase diagram of analytical solutions. There exist three phases, including a full *C* state, a full *D* state, and a bistable state of *C* or *D*. The results demonstrate that the higher the cooperation probability of AAs, the lower the threshold for the proportion of AA required to achieve a full cooperation equilibrium state. Notably, when $y < 1$, we showcase that even a minority of cooperative AAs can stimulate a full cooperation equilibrium state. On the other hand, AAs with a lower cooperation probability can result in the collapse of cooperation in a population containing a large fraction of AAs (see blue area). We then show how the equilibrium of the population varies as a function of φ by fixing $y = 4$. In the monostable state, the equilibrium is insensitive to the initial frequency of human cooperation. However, in the bistable state, increasing φ decreases the unstable interior equilibrium and expands the area capable of reaching a full cooperation equilibrium state.

Overall, we find that a minority of defective (or cooperative) AAs can trigger a pronounced phase transition toward a full cooperation equilibrium state in the *SD* (or *SH*) games. In contrast, if AAs take a larger proportion, although a full cooperation equilibrium state is easy to reach, there is a risk of transitioning to a full defection equilibrium state. Given the recognition of social learning as a means of describing human strategy updating,[34] we further investigate the results by considering pairwise comparison rule, focusing on the situation where human players can not only imitate the strategy of human player but also AAs.

Pairwise comparison rule

In the pairwise comparison rule, we investigate the influence of AAs as well as the imitation strength *K*. Note that, $K \to +\infty$ and $K \to 0$ mean strong and weak imitation strength, respectively. The probability given by Fermi function is totally affected by the sign of payoff difference $\pi_C - \pi_D$ if $K \to +\infty$, or tends to 0.5 if $K \to 0$. To illustrate how human cooperation evolves under pairwise comparison rule, we present theoretical and agent-based simulation results in this section. A fascinating finding is that results are qualitatively consistent with replicator dynamics if we consider strong imitation strength in pairwise comparison. However, the results vary as the imitation strength weakens.

We utilize the same values of *T* and *S* as in the replicator dynamics section and present the results in Figure 5. Note that we here represent cooperative and defective AAs as $\varphi = 0.1$ and $\varphi = 0.9$, respectively. The details of the agent-based simulation process are given in the STAR Methods. In the *PD* game, human cooperation, either in $\varphi = 0.1$ or $\varphi = 0.9$ condition, is difficult to emerge under strong imitation strength (see Figure 5A). In detail, when $K = 100$, human cooperation is insensitive to the strategy and fraction of AAs, which is consistent with replicator dynamics. However, when we reduce the imitation strength, the results change. Both AAs' proportion and cooperation probability positively influence the evolution of human cooperation. In particular, cooperative AAs promote human cooperation more effectively than defective AAs. This effect becomes more significant with lower imitation strength (see $K = 1$). In the *SD* game (see Figure 5B), human cooperation increases (or decreases) with AA's proportion when $\varphi = 0.1$ (or $\varphi = 0.9$) under strong imitation strength. Defective AAs boost the evolution of human cooperation, consistent with the results obtained through RD. In particular, this effect is insensitive to the initial fraction

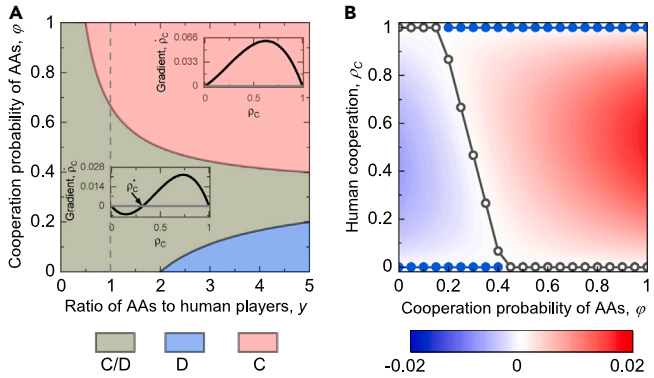


Figure 4. Equilibrium and phase diagram in the *SH* game

(A) Phase diagram showing the stable equilibrium state as a function of the (φ, y) pair. Introducing cooperative AAs in the population increases the likelihood of achieving a full cooperation equilibrium state, while introducing a large fraction of defective AAs may lead to cooperation collapse. The blue and red areas show the unique asymptotically stable states of full cooperation and full defection, respectively. The brown area indicates a bistable state that includes both full cooperation and full defection. The insets display phase portraits of ρ_C. The dashed line $y = 1$ means that human players and AAs each account for 0.5.

(B) Stable (solid circles) and unstable (open circles) equilibrium states as a function of φ with a fixed $y = 4$. The background illustrates the gradient of ρ_C. Cooperative AAs have a more positive impact on the evolution of cooperation compared to defective AAs. The parameters are fixed as $S = -0.2$ and $T = 0.6$.

of human cooperation, as shown in Figure S1 (electronic supplementary material). However, the results become totally contrary when imitation strength weakens (see $K = 1$): cooperative AAs are more beneficial to human cooperation. In the *SH* game (see Figure 5C), there exists two kinds of state: a unique asymptotically stable state ρ^*_{C1} (or ρ^*_{C2}) and a bistable state ρ^*_{C1} and ρ^*_{C2}, where ρ^*_{C1} means the coexistence of *C* and *D* with a higher frequency of cooperation and ρ^*_{C2} means the coexistence of *C* and *D* with a lower frequency of cooperation. As shown in Figure S2 (electronic supplementary material), in a bistable state, which equilibrium the system evolves to is affected by the initial frequency of human cooperation. Furthermore, similar to the findings in the RD section, AAs with lower (or higher) cooperation probability $\varphi = 0.1$ (or

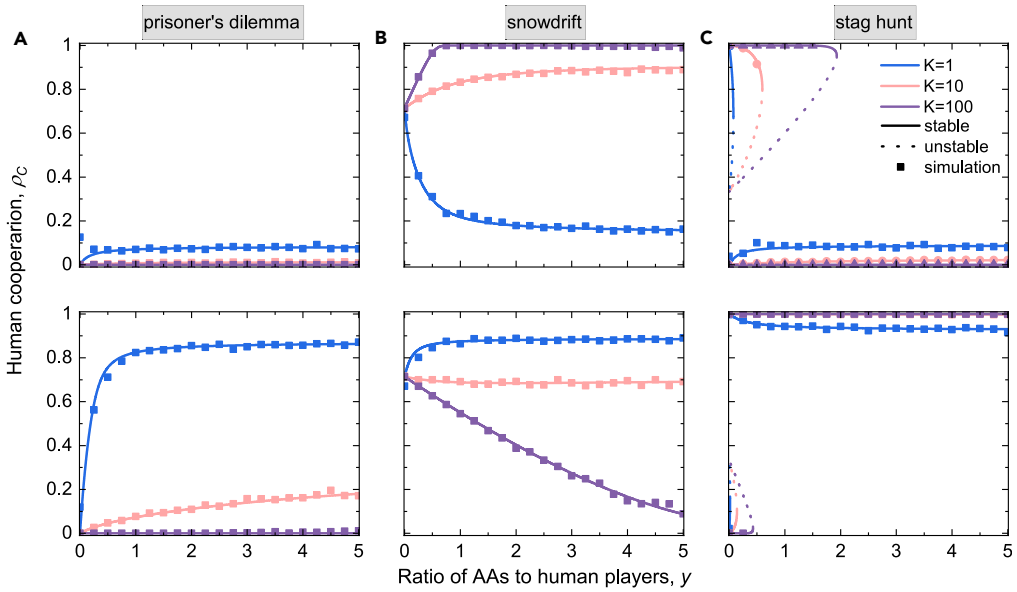


Figure 5. Equilibrium in three types of social dilemma games under the pairwise comparison rule

Consistently to the findings in replicator dynamics, under strong imitation strength, (A) cooperative AAs have no significant influence on cooperation in the *PD* game.

(B) Cooperative AAs inhibit cooperation in the *SD* game.

(C) However, cooperative AAs stimulate cooperation in the *SH* game. In contrast, when imitation strength is weak, cooperative AAs promote cooperation in all three games. From left to right, each column shows the results for the *PD* ($S = -0.2$, $T = 1.3$), *SD* ($S = 0.5$, $T = 1.2$), and *SH* ($S = -0.2$, $T = 0.6$) games. The lines show theoretical results and the squares show the agent-based simulation results. The solid and dashed lines show stable and unstable equilibria, respectively. The top and bottom panels are obtained with $\varphi = 0.1$ and $\varphi = 0.9$, respectively.

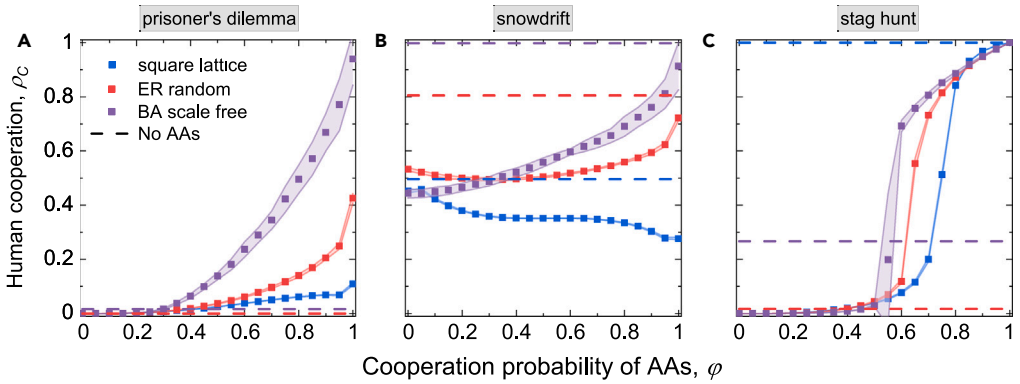


Figure 6. The frequency of human cooperation as a function of φ in complex networks with $y = 0.95$

(A) Human cooperation is promoted by AAs in the *PD* game across all three networks. In particular, the higher the value of φ, the higher the frequency of cooperation.

(B) Compared to the results without AAs, human cooperation is prohibited by AAs in the *SD* games regardless of network type. With the increase of φ, human cooperation decreases (or increases) in the square lattice (or heterogeneous networks).

(C) Cooperative AAs are more conducive to the evolution of human cooperation. The circles show simulation results averaged over 60 iterations, and the shaded regions show the standard deviation. The dashed lines indicate the frequency of human cooperation in the absence of AA. The parameters are fixed as $S = -0.2$, $T = 1.3$ in the *PD* game, $S = 0.5$, $T = 1.4$ in the *SD* game, and $S = -0.5$, $T = 0.6$ in the *SH* game.

$\varphi = 0.9$) are more feasible to cause the state ρ^*_{C2} (or ρ^*_{C1}) as y increases under strong imitation strength. This finding is still robust when imitation strength becomes weak.

Our findings demonstrate consistent results in both the replicator dynamics and pairwise comparison rule under strong imitation strength. However, if weak imitation strength is taken into account, cooperative AAs are more beneficial to the evolution of human cooperation in all three types of games. While our previous discussions primarily focused on well-mixed populations, exploring the outcomes within networks that incorporate local interactions is also pertinent.

Extension to complex networks

We present the average human cooperation frequency ρ_C as a function of φ for square lattice, Barabasi Albert (BA) scale-free, and Erdos Renyi (ER) random network in Figure 6. In the *PD* game, cooperation is consistently promoted with increasing φ, regardless of the network type. The results are qualitatively consistent with previous theoretical analyses. This finding is further verified in Figures S3 and S4 (electronic supplementary material). Turning our attention to *SD* games, we find that AAs inhibit cooperation compared to scenarios without AAs, regardless of the network type. In the square lattice, cooperation weakens further as φ increases. However, a contrasting phenomenon emerges when considering heterogeneous networks, such as ER random and BA scale-free networks. We infer that this discrepancy may be attributed to AAs occupying the hub nodes of the networks. To verify this inference, we conduct simulation experiments on a BA scale-free network under two scenarios (see Figure S5 in electronic supplementary material): (i) By assigning 4871 nodes (to maintain a similar number of AAs as in Figure 6) with the highest degrees as AAs, Figure S5 presents results similar to those in Figure 6B. (ii) By assigning 4871 nodes with the lowest degrees as AAs, we arrive at the conclusion that defective AAs effectively promote cooperation once again. This finding highlights the significant influence of AAs' location on human cooperation. Next, in the *SH* game (see Figure 6C), AAs with higher cooperation probability are more beneficial for human cooperation compared to defective AAs. In particular, the increase in φ triggers tipping points in three types of networks. Given the significance of hub nodes, we conduct additional analysis by focusing on ten nodes in three different types of games, as shown in Figure 7. When AAs are assigned to nodes with the highest degrees, the strategy of AAs has a profound impact on human cooperation. Even a slight change in the behavior of AAs, especially in the *PD* games (see Figure 7A), can lead to a significant increase in overall human cooperation. Conversely, when these ten AAs are randomly allocated across the network, they have little influence on human behavior and exhibit limited utility in altering cooperation levels.

DISCUSSION

In this study, we investigate human cooperation in hybrid populations, involving interactions between human players and AAs. AAs, who are programmed to choose cooperation with a specific probability, are employed to answer our motivation questions. The human player is assumed to update strategy according to the payoff difference given by replicator dynamics and pairwise comparison rule. Theoretic analysis and experimental simulations mainly proceed following well-mixed population and network structures, respectively. Using replicator dynamics, we investigate the impact of AAs on the equilibrium of various social dilemma games, such as the *PD*, *SH*, and *SD* games. We showcase that cooperative AAs effectively promote human cooperation in the *SH* game, but their influence is limited in games with dominant strategies. Surprisingly, in the *SD* game, cooperative AAs can even disrupt cooperation. To achieve a full cooperation state with as few AAs as possible in the *SH* game, introducing AAs with fully cooperating probability is proved to be the most effective approach. Furthermore, our

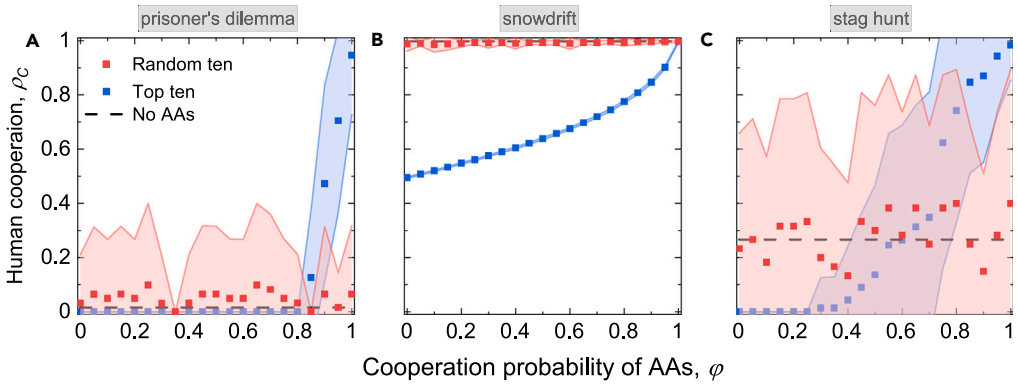


Figure 7. The frequency of human cooperation as a function of φ in BA scale-free network with ten AAs

We examine two scenarios, one involves assigning AAs to the nodes with the largest degree (blue), while the other assigns AAs to randomly selected nodes (red). Panels (A–C) show the results of the *PD*, *SD*, and *SH* games, respectively. Human cooperation closely approximates the results without AAs when AAs are randomly assigned to nodes. However, if AAs are assigned to nodes with the largest degree, even with only 10 nodes, they can boost human cooperation effectively, particularly in the *PD* and *SH* games. The circles show agent-based simulation results averaged over 60 times, and the shaded regions show standard deviation. The parameter settings are the same as Figure 6.

results also show that defective AAs are not useless, as they can stimulate cooperation in the *SD* games. Correspondingly, to achieve a cooperation-dominant state with as few AAs as possible, the best choice is to introduce AAs with always defection. These findings are further verified using pairwise comparison rule with strong imitation strength. On the other hand, if taking weak imitation strength (which includes irrational options of human players) into account, we demonstrate that cooperative AAs are beneficial for promoting human cooperation regardless of social dilemmas.

In an extended study, we implement experimental simulations involving three types of complex networks. By incorporating spatial structures into the interaction environment, we obtain qualitatively consistent results in the homogeneous network. The differences in heterogeneous networks are mainly due to the location of AAs. By controlling AAs' location, we find that assigning AAs to hub nodes, even in a small proportion, can significantly affect evolutionary outcomes. Thus far, we have contributed a model for studying human cooperation in hybrid populations, showing that it is essential to consider environments related to social dilemmas, networks, and imitation strength when designing AAs. The insights gained from our results have practical implications for developing AI algorithms to foster human cooperation.

Our research distinguishes itself from previous studies mainly in several key aspects. Firstly, in addition to considering the committed minorities,[36,37] we incorporate a substantial number of autonomous agents into our model. These AAs cooperate with their counterparts with a certain probability. The inclusion of a substantial number of AAs is motivated by recent advancements in social network research,[38] which suggests that machine accounts make up approximately 32\% of all tweets based on empirical evidence from Twitter data. Moreover, there is an increasing trend in the number of machine accounts, posing significant challenges in terms of reducing their potential risks.[39] In the context of human-agent games, we validate that minority cooperative AAs stimulate cooperation, which aligns with existing literature.[40–42] However, our research reveals an unexpected result: the inclusion of a large number of AAs leads to a breakdown of the cooperative system, surpassing the effects observed with a mere minority of AAs, as demonstrated in the defective region depicted in the left panel of Figures 3 and 4. This discovery emphasizes the potential risks posed by the growing prevalence of AAs in relation to human cooperation.

Secondly, we introduce defective AAs, whose role in fostering cooperation in two-player social dilemma games has been largely overlooked in the context of human-agent interaction. Either replicator dynamics or pairwise comparison rule, we find that the inclusion of defective AAs can indeed trigger the dominance of cooperation in SD games, an effect that remains hidden when solely focusing on cooperative AAs. Furthermore, although several existing studies primarily use AAs to address fairness or collective risk problems,[28,29] they have not introduced AAs in structured populations.[43] These interactive environments have been recognized as important factors in the context of human-human interactions. By introducing structured populations, we reveal tipping points that are triggered by AAs in the *SH* games. Meanwhile, we investigate the effect of nodes with higher degrees on triggering human cooperation. These additional critical extensions provide a comprehensive understanding of the role of cooperative and defective AAs in the evolution of human cooperation. They offer a more realistic representation of interactive environments in human-agent interactions, shedding light on the complex dynamics at play in social systems.

There are still intriguing avenues for future exploration in this field. In human-human interactions, punishment has been proven to be a powerful behavior in eliciting cooperation.[44] It's also essential to understand its utility in human-agent populations, particularly in solving second-order free-rider problems.[45] Even though theoretical analysis helps identify critical values, it neglects human players' emotional and social factors. Conducting experiments involving structured and unstructured populations to test these findings will open up exciting avenues for research in human-agent interaction.

Limitations of the study

The current study has some limitations that should be acknowledged. Firstly, it focuses solely on the simplest social dilemmas involving two decisions: cooperation and defection. While we have shown that artificial agents (AAs) with even simple intelligence can facilitate cooperation in social dilemmas where players have no prior information about counterparts, it remains uncertain how they would perform in more complex scenarios, such as stochastic games and sequential social dilemma games.[46,47] The interaction among players in these scenarios may be influenced by historical information or the state of the environment, which necessitates addressing these complex problems. Moreover, although simple algorithms may be effective in stimulating cooperation,[48] developing algorithms with more intelligence would further benefit the development of artificial intelligence, particularly for human-machine or human-robot interaction.[24]

Furthermore, it is important to acknowledge that a purely theoretical study may overlook the complex motivations behind human behavior. Conducting human experiments is crucial for testing predictions from theoretical models and gaining insights into various aspects, including psychological effects, emotions, and cultural differences.[20,26] Mechanisms such as communication sentiment, reward, and punishment in human-human interactions have provided clear evidence of prosocial behavior.[49–51] Empirical experiments not only test theoretical possibilities but also reveal what actually occurs. As artificial agents increasingly integrate into human life, the study of human-agent cooperation from both theoretical and experimental perspectives becomes necessary. Building upon experimental findings of human-human interactions, we can construct a theoretical model of human-agent interaction and explore the influence of cooperative and defective artificial agents on human cooperation. Although our current work does not include human behavior experimental results, we believe that our theoretical work can serve as a valuable starting point for future human behavior experiments.

Several subsequent studies have examined the impact of the information level on human cooperation, specifically the extent to which human players know their opponents are artificial agents.[24] These studies have revealed that human cooperation is often higher when human players have no information about the true nature of their opponents.[52] Conversely, even when participants recognize that artificial agents perform better than human players at inducing cooperation, human players tend to reduce their willingness to cooperate. However, Shirado and Christakis conducted human-agent interactions on network structures and presented contrasting results, showing that human cooperation can still be promoted even if the identity of the agents is transparent.[48] In social dilemma games with one-shot settings, the answer is still not deterministic. Our study did not account for the intrinsic properties of artificial agents and focused on a scenario without learning bias between humans and artificial agents. Therefore, it is crucial to consider the true nature of agents, particularly in one-shot settings, to expand our understanding of their behavior and implications.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Social dilemma games
  - Population setup and autonomous agents
  - Hybrid population game
  - Replicator dynamics
  - Pairwise comparison rule
  - Simulation for complex networks
  - Agent-based model
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108179.

ACKNOWLEDGMENTS

This research was supported by the National Science Fund for Distinguished Young Scholars (No. 62025602), the National Science Fund for Excellent Young Scholars (No. 62222606), the National Natural Science Foundation of China (Nos. 11931015, U1803263, 81961138010 and 62076238), Fok Ying-Tong Education Foundation, China (No. 171105), Technological Innovation Team of Shaanxi Province (No. 2020TD-013), Fundamental Research Funds for the Central Universities (No. D5000211001), the Tencent Foundation and XPLORER PRIZE, JSPS Postdoctoral Fellowship Program for Foreign Researchers (grant no. P21374).

AUTHOR CONTRIBUTIONS

All authors have read and approved the manuscript. H.G.: Conceptualization, formal analysis, investigation, methodology, visualization, writing-original draft, writing-review and editing; C.S.: formal analysis, investigation, writing-original draft, writing-review and editing; S.H.: resources, visualization, writing-review and editing; J.X.: project administration, supervision, writing-review and editing; P.T.: project administration, writing-review and editing; Y.S.: project administration, writing-review and editing; Z.W.: project administration, supervision, writing-review and editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 19, 2023
Revised: July 10, 2023
Accepted: October 9, 2023
Published: October 12, 2023

REFERENCES

1. Nowak, M.A. (2006). Five rules for the evolution of cooperation. Science *314*, 1560–1563. https://doi.org/10.1126/science.1133755.
2. West, S.A., Griffin, A.S., and Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. J. Evol. Biol. *20*, 415–432. https://doi.org/10.1111/j.1420-9101.2006. 01258.x.
3. Vasconcelos, V.V., Santos, F.C., and Pacheco, J.M. (2013). A bottom-up institutional approach to cooperative governance of risky commons. Nat. Clim. Change *3*, 797–801. https://doi.org/10.1038/nclimate1927.
4. Bauch, C.T., and Earn, D.J.D. (2004). Vaccination and the theory of games. Proc. Natl. Acad. Sci. USA *101*, 13391–13394. https://doi.org/10.1073/pnas.040382310.
5. Hauert, C., and Szabó, G. (2005). Game theory and physics. Am. J. Phys. *73*, 405–414. https://doi.org/10.1119/1.1848514.
6. Perc, M., Jordan, J.J., Rand, D.G., Wang, Z., Boccaletti, S., and Szolnoki, A. (2017). Statistical physics of human cooperation. Phys. Rep. *687*, 1–51. https://doi.org/10.1016/j.physrep.2017.05.004.
7. Wang, Z., Kokubo, S., Jusup, M., and Tanimoto, J. (2015). Universal scaling for the dilemma strength in evolutionary games. Phys. Life Rev. *14*, 1–30. https://doi.org/10.1016/j.plrev.2015.04.033.
8. Szolnoki, A., and Perc, M. (2015). Conformity enhances network reciprocity in evolutionary social dilemmas. J. R. Soc. Interface *12*, 20141299. https://doi.org/10.1098/rsif.2014.1299.
9. Hu, S., and Leung, H.-F. (2018). Do social norms emerge? the evolution of agents' decisions with the awareness of social values under iterated prisoner's dilemma. In 2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO) (IEEE), pp. 11–19. https://doi.org/10.1109/SASO.2018.00012.
10. de Melo, C.M., Marsella, S., and Gratch, J. (2019). Human cooperation when acting through autonomous machines. Proc. Natl. Acad. Sci. USA *116*, 3482–3487. https://doi.org/10.1073/pnas.1817656116.
11. Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science *352*, 1573–1576. https://doi.org/10.1126/science.aaf2654.
12. Faisal, A., Kamruzzaman, M., Yigitcanlar, T., and Currie, G. (2019). Understanding autonomous vehicles. J. Transp Land Use *12*, 45–72. https://www.jtlu.org/index.php/jtlu/article/view/1405.
13. Guo, H., Song, Z., Geček, S., Li, X., Jusup, M., Perc, M., Moreno, Y., Boccaletti, S., and Wang, Z. (2020). A novel route to cyclic dominance in voluntary social dilemmas. J. R. Soc. Interface *17*, 20190789. https://doi.org/10.1098/rsif.2019.0789.
14. Chen, X., Sasaki, T., Brännström, Å., and Dieckmann, U. (2015). First carrot, then stick: how the adaptive hybridization of incentives promotes cooperation. J. R. Soc. Interface *12*, 20140935. https://doi.org/10.1098/rsif.2014. 0935.
15. Nair, G.S., and Bhat, C.R. (2021). Sharing the road with autonomous vehicles: Perceived safety and regulatory preferences. Transport. Res. C Emerg. Technol. *122*, 102885. https://doi.org/10.1016/j.trc.2020.102885.
16. Nikolaidis, S., Ramakrishnan, R., Gu, K., and Shah, J. (2015). Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. ACM Trans. Comput. Hum. Interact. 189–196. IEEE. https://doi.org/10.1145/2696454.2696455.
17. Beans, C. (2018). Can robots make good teammates? Proc. Natl. Acad. Sci. USA *115*, 11106–11108. https://doi.org/10.1073/pnas.1814453115.
18. Crandall, J.W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M.A., Rahwan, I., et al. (2018). Cooperating with machines. Nat. Commun. *9*, 1–12. https://doi.org/10.1038/s41467-017-02597-8.
19. Shapira, I., and Azaria, A. (2022). Reinforcement learning agents for interacting with humans. Proc. Annu. Meet. Cognitive Sci. Soc. *44*. https://escholarship.org/uc/item/9zh0v0kw.
20. Azaria, A., Richardson, A., and Rosenfeld, A. (2016). Autonomous agents and human cultures in the trust–revenge game. Auton. Agent. Multi. Agent. Syst. *30*, 486–505. https://doi.org/10.1007/s10458-015-9297-1.
21. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. Nature *529*, 484–489. https://doi.org/10.1038/nature16961.
22. Correia, F., Mascarenhas, S., Gomes, S., Tulli, S., Santos, F.P., Santos, F.C., Prada, R., Melo, F.S., and Paiva, A. (2019). For the Record-A Public Goods Game for Exploring Human-Robot Collaboration (AAMAS), pp. 2351–2353.
23. Sheridan, T.B. (2016). Human–robot interaction: status and challenges. Hum. Factors *58*, 525–532. https://doi.org/10.1177/0018720816644436.
24. Paiva, A., Santos, F., and Santos, F. (2018). Engineering pro-sociality with autonomous agents. AAAI *32*. https://doi.org/10.1609/aaai.v32i1.12215.
25. Crandall, J.W. (2015). Robust Learning for Repeated Stochastic Games via Meta-Gaming (IJCAI), pp. 3416–3422. https://doi.org/10.1016/j.artint.2016.02.004.
26. Manistersky, E., Lin, R., and Kraus, S. (2014). The development of the strategic behavior of peer designed agents. In Language, Culture, Computation. Computing-Theory and Technology: Essays Dedicated to Yaacov Choueka on the Occasion of His 75th Birthday, Part I (Springer), pp. 180–196. https://doi.org/10.1007/978-3-642-45321-2_9.
27. Sun, X., Pieroth, F.R., Schmid, K., Wirsing, M., and Belzner, L. (2022). On learning stable cooperation in the iterated prisoner's dilemma with paid incentives. In ICDCSW (IEEE), pp. 113–118. https://doi.org/10.1109/ICDCSW56584.2022.00031.
28. Terrucha, I., Dmingos, E., Santos, F., Simoens, P., and Lenaerts, T. (2022). The Art of Compensation: How Hybrid Teams Solve Collective Risk Dilemmas (ALA2022, Adaptive and Learning Agents Workshop), pp. 1–8.
29. Santos, F.P., Pacheco, J.M., Paiva, A., and Santos, F.C. (2019). Evolution of collective fairness in hybrid populations of humans and agents. AAAI *33*, 6146–6153. https://doi.org/10.1609/aaai.v33i01. 33016146.
30. Sigmund, K., De Silva, H., Traulsen, A., and Hauert, C. (2010). Social learning promotes institutions for governing the commons. Nature *466*, 861–863. https://doi.org/10.1038/nature09203.
31. Santos, F.C., Pacheco, J.M., and Lenaerts, T. (2006). Cooperation prevails when individuals adjust their social ties. PLoS Comput. Biol. *2*,

e140. https://doi.org/10.1371/journal.pcbi.0020140.
32. Traulsen, A., Pacheco, J.M., and Nowak, M.A. (2007). Pairwise comparison and selection temperature in evolutionary game dynamics. J. Theor. Biol. *246*, 522–529. https://doi.org/10.1016/j.jtbi.2007.01.002.
33. Roca, C.P., Cuesta, J.A., and Sánchez, A. (2009). Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. Phys. Life Rev. *6*, 208–249. https://doi.org/10.1016/j.plrev.2009.08.001.
34. Traulsen, A., Semmann, D., Sommerfeld, R.D., Krambeck, H.-J., and Milinski, M. (2010). Human strategy updating in evolutionary games. Proc. Natl. Acad. Sci. USA *107*, 2962–2966. https://doi.org/10.1073/pnas.0912515107.
35. Ohtsuki, H., Hauert, C., Lieberman, E., and Nowak, M.A. (2006). A simple rule for the evolution of cooperation on graphs and social networks. Nature *441*, 502–505. https://doi.org/10.1038/nature04605.
36. Cardillo, A., and Masuda, N. (2020). Critical mass effect in evolutionary games triggered by zealots. Phys. Rev. Res. *2*, 023305. https://doi.org/10.1103/PhysRevResearch.2.023305.
37. Matsuzawa, R., Tanimoto, J., and Fukuda, E. (2016). Spatial prisoner's dilemma games with zealous cooperators. Phys. Rev. E *94*, 022114. https://doi.org/10.1038/nature04605.
38. Abokhodair, N., Yoo, D., and McDonald, D.W. (2015). Dissecting a social botnet: Growth, content and influence in twitter. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 839–851. https://doi.org/10.1145/2675133.2675208.
39. Ping, H., and Qin, S. (2018). A Social Bots Detection Model Based on Deep Learning Algorithm. In ICCT (IEEE), pp. 1435–1439. https://doi.org/10.1109/ICCT.2018.8600029.
40. Xie, J., Sreenivasan, S., Korniss, G., Zhang, W., Lim, C., and Szymanski, B.K. (2011). Social consensus through the influence of committed minorities. Phys. Rev. E *84*, 011130. https://doi.org/10.1103/PhysRevE.84. 011130.
41. Centola, D., Becker, J., Brackbill, D., and Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. Science *360*, 1116–1119. https://doi.org/10.1126/science.aas8827.
42. Arendt, D.L., and Blaha, L.M. (2015). Opinions, influence, and zealotry: a computational study on stubbornness. Comput. Math. Organ. Theor. *21*, 184–209. https://doi.org/10.1007/s10588-015-9181-1.
43. Nowak, M.A., and May, R.M. (1992). Evolutionary games and spatial chaos. Nature *359*, 826–829. https://doi.org/10.1038/359826a0.
44. Han, T.A. (2016). Emergence of social punishment and cooperation through prior commitments. AAAI *30*, 2494–2500. https://doi.org/10.1609/aaai.v30i1.10120.
45. Szolnoki, A., and Perc, M. (2017). Second-order free-riding on antisocial punishment restores the effectiveness of prosocial punishment. Phys. Rev. X *7*, 041027. https://doi.org/10.1103/PhysRevX.7.041027.
46. Barfuss, W., Donges, J.F., Vasconcelos, V.V., Kurths, J., and Levin, S.A. (2020). Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. Proc. Natl. Acad. Sci. USA *117*, 12915–12922. https://doi.org/10.1073/pnas.1916545117.
47. Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. AAMAS, 464–473. https://doi.org/10.48550/arXiv.1702.03037.
48. Shirado, H., and Christakis, N.A. (2020). Network engineering using autonomous agents increases cooperation in human groups. iScience *23*, 101438. https://doi.org/10.1016/j.isci.2020.101438.
49. Wang, Z., Jusup, M., Guo, H., Shi, L., Geček, S., Anand, M., Perc, M., Bauch, C.T., Kurths, J., Boccaletti, S., and Schellnhuber, H.J. (2020). Communicating sentiment and outlook reverses inaction against collective risks. Proc. Natl. Acad. Sci. USA *117*, 17650–17655. https://doi.org/10.1073/pnas.1922345117.
50. Dreber, A., Rand, D.G., Fudenberg, D., and Nowak, M.A. (2008). Winners don't punish. Nature *452*, 348–351. https://doi.org/10.1038/nature06723.
51. Wang, Z., Jusup, M., Shi, L., Lee, J.-H., Iwasa, Y., and Boccaletti, S. (2018). Exploiting a cognitive bias promotes cooperation in social dilemma experiments. Nat. Commun. *9*, 2954. https://doi.org/10.1038/s41467-018-05259-5.
52. Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., and Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. Nat. Mach. Intell. *1*, 517–521. https://doi.org/10.1038/s42256-019-0113-5.
53. Macy, M.W., and Flache, A. (2002). Learning dynamics in social dilemmas. Proc. Natl. Acad. Sci. USA *99*, 7229–7236. https://doi.org/10.1073/pnas.092080099.
54. Taylor, C., Fudenberg, D., Sasaki, A., and Nowak, M.A. (2004). Evolutionary game dynamics in finite populations. Bull. Math. Biol. *66*, 1621–1644. https://doi.org/10.1016/j.bulm.2004.03.004.
55. Barabási, A.-L., and Albert, R. (1999). Emergence of scaling in random networks. Science *286*, 509–512. https://doi.org/10.1126/science.286.5439.509.
56. Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. *5*, 17–60.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
T together with R, S, and P for social dilemma games	Macy et al.[53]	https://doi.org/10.1073/pnas.09208009
y together with ∅ for autonomous agents	Cardillo and Masuda[36]	https://doi.org/10.1103/PhysRevResearch.2.023305
Software and algorithms		
Dev C++	Bloodshed	https://www.bloodshed.net/
OriginLab	OriginLab Corporation	https://www.originlab.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Zhen Wang (zhenwang0@gmail.com).

Materials availability

No materials were newly generated for this paper.

Data and code availability

- The data that support the results of this study are available from the corresponding authors upon request.
- The code used to generate the figures is available from the corresponding authors upon request.
- All other items: any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Our study does not use typical experimental models in the life sciences.

METHOD DETAILS

Social dilemma games

Two-player social dilemma game, a typical subclass of social dilemma, depicts the rational decision-making of two participants by introducing a strategy set and payoff matrix. In the simplest version, each player selects a strategy from a strategy set $\mathcal{S} = \{C, D\}$, where C and D represent cooperation and defection, respectively. Mutual cooperation yields a reward R to both players, while mutual defection results in a punishment P. Unilateral cooperation leads to a sucker's payoff S, while the corresponding defection receives a temptation to defect T. The above process can be represented by a payoff matrix:

$$\mathcal{A} = \begin{pmatrix} R & S \\ T & P \end{pmatrix}. \quad \text{(Equation 1)}$$

Using this payoff matrix, the so-called social dilemma is meeting if it follows these four conditions simultaneously:[53]

(i) $R > P$. Players prefer to cooperate with each other than to defect from each other.
(ii) $R > S$. Mutual cooperation is preferred over unilateral cooperation.
(iii) $2R > T + S$. Mutual cooperation is more beneficial for the collective than defecting against a cooperator.
(iv) Either $T > R$ (greed) or $P > S$ (fear). The former condition means players prefer exploiting a cooperator to cooperating with him. The latter condition means players prefer mutual defection over being exploited by a defector.

According to the ranking order of these parameters, these two-player social dilemma games can be classified into four different kinds of games,[7] which are *PD* games ($T > R > P > S$), *SD* games ($T > R > S > P$), *SH* games ($R > T > P > S$), and harmony (*HG*) game ($R > T, S > P$). It is noteworthy that the first three games exhibit social dilemmas,[47,53] while the harmony game does not. Without a specific declaration, we set $R = 1$ and $P = 0$ throughout this paper.

Population setup and autonomous agents

We consider a well-mixed and infinitely large population $\mathcal{P} = \{1,2,\cdots,N\}$, where $N \to +\infty$, and each player can interact with each other with equal probability. In the population, player $i \in \mathcal{P}$ can choose one of two strategies from set $\mathcal{S} = \{C,D\}$. We denote the strategy of player i as a vector $\mathcal{X}_i = (x_1, x_2)'$, where $x_j = 1$ if the j th strategy is chosen and the other element is equal to 0. To investigate how AAs affect the cooperative behavior among human players, we consider a hybrid population consisting of human players and AAs (see Figure 1A). Human players participate in the game and update their strategies through a social learning process. AAs, on the other hand, follow a pre-designed algorithm to make their choices: they cooperate with a fixed probability $\varphi(0 \leq \varphi \leq 1)$ and defect with probability $1 - \varphi$. We refer to them as cooperative AAs if $0.5 \leq \varphi \leq 1$, and defective AAs if $0 \leq \varphi < 0.5$. In the human-agent hybrid population, each player has an equal chance to engage in a two-player game with other players (see Figure 1B). Consequently, the interaction probability between humans and AAs significantly depends on the composition of this population.

Hybrid population game

In the hybrid population, the fraction of cooperation among human players is denoted as ρ_C $(0 \leq \rho_C \leq 1)$. Assuming human players are one unit, add y units of AAs to the hybrid population. Consequently, the fraction of human cooperation is denoted by $f_C = \frac{\rho_C}{1+y}$. The parameter y can also be used to quantify the composition of the population: if $0 < y < 1$, it implies that the fraction of AAs is lower than that of human players; whereas $y > 1$ indicates a higher proportion of AAs. Accordingly, the expected payoff of cooperation and defection among human players in a hybrid population can be calculated as follows:

$$\pi_C = \frac{1}{1+y}(\rho_C R + (1-\rho_C)S) + \frac{y}{1+y}(\varphi R + (1-\varphi)S),$$

$$\pi_D = \frac{1}{1+y}(\rho_C T + (1-\rho_C)P) + \frac{y}{1+y}(\varphi T + (1-\varphi)P), \quad \text{(Equation 2)}$$

where the first term on the right-hand represents the payoff from interacting with human players, and the second term signifies the payoff obtained from interacting with AAs.

Replicator dynamics

The replicator equation[33] is a widely used differential equation that depicts evolutionary dynamics in infinitely large populations. Following this rule, the growth of a specific strategy is proportional to the payoff difference. Therefore, the dynamics of human cooperation can be represented by the following differential equation:

$$\dot{\rho}_C = (1+y)\dot{f}_C$$

$$= (1+y)\frac{\rho_C}{1+y}\frac{1-\rho_C}{1+y}(\pi_C - \pi_D)$$

$$= \frac{\rho_C(1-\rho_C)}{1+y}(\pi_C - \pi_D), \quad \text{(Equation 3)}$$

where

$$\pi_C - \pi_D = \frac{\rho_C + y\varphi}{1+y}(R - T - S + P) + S - P. \quad \text{(Equation 4)}$$

By solving $\dot{\rho}_C = 0$, we find that there exists two trivial equilibrium $\rho_C = 0$ and $\rho_C = 1$, and a third equilibrium ρ_C^* that is closely associated with game models and AAs. By solving $\pi_C - \pi_D = 0$, one can derive:

$$\rho_C^* = \frac{P-S}{R+P-T-S} + y\left(\frac{P-S}{R+P-T-S} - \varphi\right)$$

$$= \widehat{\rho}_C + y(\widehat{\rho}_C - \varphi), \quad \text{(Equation 5)}$$

where $\widehat{\rho}_C = \frac{P-S}{R+P-T-S}$. It is easy to deduce that the restriction $T > R > S > P$ or $R > T > P > S$ guarantees $0 < \widehat{\rho}_C < 1$. Moreover, ρ_C^* increases with y if $\widehat{\rho}_C - \varphi > 0$, whereas decreases with y if $\widehat{\rho}_C - \varphi < 0$. Since ρ_C^* measures cooperation rate in human players, this equilibrium will vanish if $\rho_C^* < 0$ or $\rho_C^* > 1$. Note that $\widehat{\rho}_C$ is also the interior equilibrium when the population consists only of human players,[54] *i.e.*, the scenario $y = 0$. Subsequently, the stability of the equilibrium will be discussed from three types of social dilemmas. As evidence has revealed that human players may update their behaviors through social learning,[34] one may ask: what results can be obtained when considering pairwise comparison rule? Under this rule, we can also examine the effect of imitation strength on the outcomes.

Pairwise comparison rule

Pairwise comparison is a well-known cultural process that effectively portrays game dynamics. Note that we employ the Fermi rule during the strategy updating stage.[5] In this process, strategy updating takes place within a randomly chosen pair of players, denoted as i and j, with strategy $\mathcal{X}$ and $\mathcal{Y}$ ($\mathcal{X}$, $\mathcal{Y} \in \mathcal{S}$). If $\mathcal{X} \neq \mathcal{Y}$, player i takes j as a reference and imitates its strategy with a probability determined by the Fermi function,

$$W_{\mathcal{X}\leftarrow\mathcal{Y}} = \frac{1}{1+e^{-K(\pi_{\mathcal{Y}}-\pi_{\mathcal{X}})}}, \quad \text{(Equation 6)}$$

where K represents selection intensity (which is also known as imitation strength) and measures the irrational degree of human players (or the extent players make decisions by payoff comparisons).[5,30] In the hybrid population defined, the probability of a cooperator taking a defector as an indicator is given by:

$$P_1 = \frac{\rho_C(1-\rho_C+y(1-\varphi))}{1+y}.$$

Subsequently, the probability that cooperators decrease by one is

$$Q^- = P_1 W_{C\leftarrow D} = \frac{\rho_C(1-\rho_C+y(1-\varphi))}{1+y}\frac{1}{1+e^{-K(\pi_D-\pi_C)}}. \quad \text{(Equation 7)}$$

Similarly, the probability that a defector taking a cooperator as an indicator is

$$P_2 = \frac{(\rho_C+y\varphi)(1-\rho_C)}{1+y}.$$

Consequently, the probability that cooperators increase by one is

$$Q^+ = P_2 W_{D\leftarrow C} = \frac{(\rho_C+y\varphi)(1-\rho_C)}{1+y}\frac{1}{1+e^{-K(\pi_C-\pi_D)}}. \quad \text{(Equation 8)}$$

In total, the dynamics of cooperation can be represented as follows:

$$\dot{\rho}_C = Q^+ - Q^- = \frac{\left(\rho_C-\rho_C^2+y\varphi-\rho_C y\varphi\right)e^{K(\pi_C-\pi_D)}}{(1+y)(1+e^{K(\pi_C-\pi_D)})} - \frac{\rho_C-\rho_C^2+\rho_C y-\rho_C y\varphi}{(1+y)(1+e^{K(\pi_C-\pi_D)})}. \quad \text{(Equation 9)}$$

We can derive the equilibrium by solving $\dot{\rho}_C = 0$. Since the denominator is larger than 0 evidently, the equilibrium is mainly determined by the numerator. We then denote the numerator as $f(\rho_C)$ and examine the condition:

$$f(0) = y\varphi e^{K\left(\frac{y\varphi}{1+y}(R-T-S+P)+S-P\right)},$$

$$f(1) = y\varphi - y. \quad \text{(Equation 10)}$$

In the presence of AAs, $f(0) \geq 0$ and $f(1) \leq 0$ are established. The equality holds when $\varphi = 0$ and $\varphi = 1$, respectively. Therefore, there exists at least one interior equilibrium when $0 < \varphi < 1$. Note that $\varphi = 1$ is the so-called zealous cooperator.[36] In addition to analytical results, we verify our findings through agent-based simulations. The simulation procedure includes the following steps: (i) Initially, each human player is assigned either cooperation with probability p or defection with a probability $1 - p$ initially, where p controls the initial frequency of cooperation. Each AA adopts cooperation and defection with probability φ and $1 - \varphi$ in each round. (ii) With a specific strategy, a randomly chosen player (presumed to be human), denoted as i, obtains an expected payoff by interacting with other $N - 1$ individuals. (iii) Individual i decides whether to imitate the strategy of a randomly chosen individual j, who obtains a payoff in the same way, according to the Fermi function. Repeat (ii) and (iii) until the population reaches an asymptotically stable state. We get the simulation results by setting $N = 1000$.

Simulation for complex networks

Building upon the aforementioned results, we further study the network structure effect on human-agent cooperation in this section. Since interaction, in reality, is not limited to well-mixed populations, we also implement experiments in complex networks that contain local interactions. This means that players can only interact with a limited set of neighboring individuals. To assess the effect of network structure on cooperation, we employ pairwise imitation as a strategy updating rule and measure the expected cooperation rate among human players. Following this, players are matched in pairs and imitate their opponent's strategy based on a probability determined by their payoff difference.[35] We begin by introducing three types of complex networks.

Network settings

Denote $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ as a complex network, where $\mathcal{V} = \{1, 2, \cdots N\}$ represents node set, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is link set. Each node $i \in \mathcal{V}$ represents either a human player or an AA. For the edge $(i, j) \in \mathcal{E}$, each player i is paired up with another player j to play a two-player social dilemma game. We consider a network with $N = 10000$ players and an average degree of $4 < k > =$.

- Square lattice is a homogeneous network. Each player interacts with their four neighbors and receives a payoff by playing with its north, south, east, and west neighbors. It is noteworthy that here we consider lattice with periodic boundary.
- Barabasi Albert scale-free network is generated following the growth and preferential attach rules.[55] The degree distribution of the ultimate network satisfies a power-law function.
- Erdos Renyi random network is generated by linking two different nodes with a random probability.[56] The degree distribution of the ultimate network satisfies the Poisson distribution.

Agent-based model

Agent-based simulation

We utilized the Monte Carlo simulation to examine the variation of cooperation across different networks. Initially, each human player is assigned either cooperation or defection with a probability of 0.5, whereas each AA adopts cooperation and defection with probability φ and $1 - \varphi$, respectively. With the specific strategy, a randomly chosen player (assumed to be human), denoted as i, obtains the payoff by interacting with connected neighbors

$$F_i = \sum_{y \in \Omega_i} \mathcal{X}_i' \mathcal{A} \mathcal{X}_j, \quad \text{(Equation 11)}$$

where Ω_i represents neighbor set of player i, $\mathcal{A}$ is the payoff matrix given by Figure 1B. After calculating the cumulative payoff, player i decides whether to adopt one of his/her neighbors' strategies with the probability given by the Fermi function

$$W_{\mathcal{X}_i \leftarrow \mathcal{X}_j} = \frac{1}{1 + e^{-K(F_j - F_i)}}, \quad \text{(Equation 12)}$$

where j is a randomly chosen neighbor. We set $K = 10$ in the simulations. Results are calculated by conducting 60 realizations. For each realization, we fix the total step as 50000, and each value is averaged over 5000 steps when the network reaches an asymptotic state.

QUANTIFICATION AND STATISTICAL ANALYSIS

Our study does not use typical statistical analysis.