Reputation as a Solution to Cooperation Collapse in LLM-based MASs

Siyue Ren
School of Mechanical Engineering, Northwestern Polytechnical University
Xi'an, China

Wanli Fu
School of Cybersecurity, Northwestern Polytechnical University
Xi'an, China

Xinkun Zou
School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University
Xi'an, China

Chen Shen
Kyushu University
Fukuoka, Japan

Yi Cai
South China University of Technology
Guangzhou, China

Chen Chu
Yunnan University of Finance and Economics
Kunming, China

Zhen Wang*
Northwestern Polytechnical University
Xi'an, China
w-zhen@nwpu.edu.cn

Shuyue Hu*
Shanghai Artificial Intelligence Laboratory
Shanghai, China
hushuyue@pjlab.org.cn

ABSTRACT

Cooperation has long been a fundamental topic in both human society and AI systems. However, recent studies indicate that the collapse of cooperation may emerge in multi-agent systems (MASs) driven by large language models (LLMs). To address this challenge, we explore reputation systems as a remedy. We propose *RepuNet*, a dynamic, dual-level reputation framework that models both agent-level reputation dynamics and system-level network evolution. Specifically, driven by direct interactions and indirect gossip, agents form reputations for both themselves and their peers, and decide whether to connect or disconnect other agents for future interactions. Through three distinct scenarios, we show that RepuNet effectively avoids cooperation collapse, promoting and sustaining cooperation in LLM-based MASs. Moreover, we find that reputation systems can give rise to rich emergent behaviors in LLM-based MASs, such as the formation of cooperative clusters, the social isolation of exploitative agents, and the preference for sharing positive gossip rather than negative ones. The GitHub repository for our project can be accessed via the following link: https://github.com/RGB-0000FF/RepuNet.

KEYWORDS

Reputation;Cooperation;Large language model;Social simulation

ACM Reference Format:
Siyue Ren, Wanli Fu, Xinkun Zou, Chen Shen, Yi Cai, Chen Chu, Zhen Wang*, and Shuyue Hu*. 2026. Reputation as a Solution to Cooperation Collapse in LLM-based MASs. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/UEHN4980

Corresponding author: w-zhen@nwpu.edu.cn, hushuyue@pjlab.org.cn.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/UEHN4980

1 INTRODUCTION

Cooperation collapse refers to a phenomenon driven by social dilemmas, in which individual self-interest conflicts with collective welfare, resulting in harmful outcomes for the entire group [23, 34, 35]. Such collapse occurs across diverse domains and underpins a wide array of challenges, including resource allocation [37], climate change mitigation [13], and coordination among self-driving vehicles [12]. Addressing cooperation collapse is essential for sustaining effective collaboration in both human societies and AI systems.

Recent studies on large language model (LLM)-based agents have revealed the existence of cooperation collapse in LLM-based multi-agent systems (MASs), a system of agents that are powered by LLMs. For example, in simulated fishing allocation scenarios, Piatti et al. [37] observed that LLM-based agents tended to overexploit shared resources, resulting in rapid depletion and long-term unsustainability. Similarly, in iterated Prisoner's Dilemma settings, agents often responded to a single defection with irreversible retaliation, leading to cascading mutual defection and the breakdown of cooperation [1, 17, 47]. While these studies have documented the widespread cooperation collapse, effective solutions remain elusive.

This paper explores the use of reputation systems as a remedy. Reputation is a set of collective beliefs or evaluations about individuals based on their past behavior, enabling agents to form expectations about others [6, 29, 42, 43]. As an emergent social signal, it encodes perception of trustworthiness, reliability, and integrity. A reputation system operationalizes this signal by systematically collecting, aggregating, and disseminating reputational information across agents. Proven effective in both human societies and traditional MASs, reputation systems offer a fesible solution to fostering cooperation even among strangers, evoking social norms,

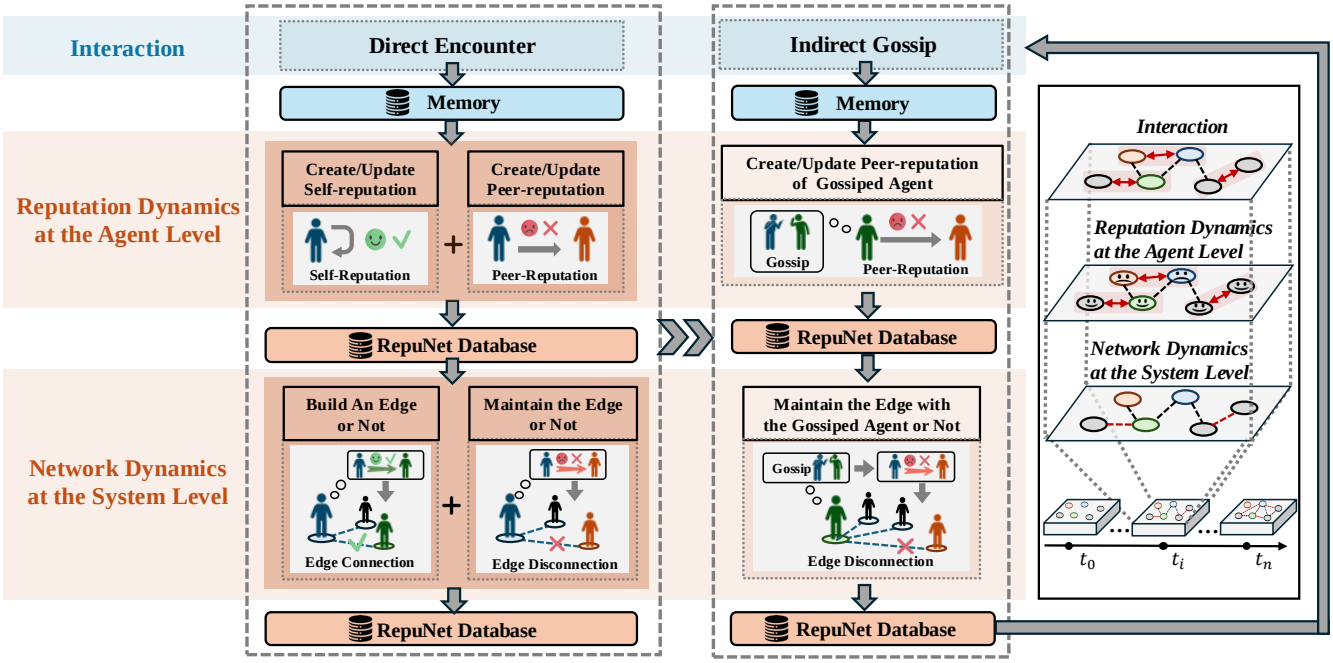

Figure 1: RepuNet: A dynamic reputation system aimed at sustaining cooperation and preventing cooperation collapse in LLM-based MASs. Agents interact within networks through both direct encounters and indirect gossip, shaping their self-reputation and peer-reputation. At the system level, agents decide whether to form or maintain network connections (i.e. edges) based on these reputations. The evolving reputations and network structures, stored in RepuNet databases, continuously guide agents' behaviors, influencing their future interactions.

and serving as lightweight, scalable substitutes for direct monitoring or formal enforcement [4, 19, 44, 52].

In this research, we extend reputation mechanisms to LLM-based MASs, and demonstrate that our approach effectively prevents cooperation collapse. Crucial to our study will be how to *operationalize* reputation mechanisms for LLM-based MASs. To this end, we introduce **RepuNet**, the first *operational* reputation system tailored for LLM-based MASs, in which every interaction is explicitly embedded within a network structure. RepuNet operates on two levels: agent-level reputation dynamics and system-level network evolution. Specifically, at the agent level, echoing real-world reputation dynamics [19, 21, 41], RepuNet enables agents shape initial self- and peer-reputations after their first encounter and iteratively update them after each subsequent interaction. Besides, gossip has been a key mechanism for spreading reputational information across networks throughout human evolution. Inspired by this, agents evaluate each direct encounter and, when (dis)satisfied, broadcast that information, updating the target's reputation. In RepuNet, we explicitly couple reputation updates to topology changes: once a reputation is revised, it drives network rewiring. At the system level, agents strengthen ties to high-reputation partners and sever links to disreputable ones, reflecting the human tendency to seek trustworthy collaborators and abandon exploitative relationships. This network rewiring reshapes future encounter probabilities and fosters long-term cooperation among reputable agents. We implement RepuNet by prompting LLM-based agents, rather than relying on any rule-based mechanisms or mathematical formulations. Both reputation dynamics and network evolution in RepuNet are autonomously determined by the agents themselves, without any human intervention. An overview of our reputation system is provided in Figure 1.

In the experiment, we consider three scenarios that progress from classical to real-world dilemmas: (i) *the Prisoner's Dilemma*, a fundamental social dilemma in game theory [3]; (ii) *voluntary participation*, adapted from human field experiments [54], modeling collective resource-sharing; and (iii) *trading investment*, based on an economic trust game [5], capturing sequential investment dilemmas. Within LLM-based MASs, our experiments replicate the phenomenon of cooperation collapse previously observed in human studies. Specifically, agents acting without RepuNet prioritize individual gain, causing widespread exploitation and eventual cooperation breakdown. In contrast, with RepuNet, agents are incentivized to maintain cooperation, favoring collective welfare and preventing collapse. Across all three scenarios, we observed consistent outcomes using similar prompts with varying task descriptions. This demonstrates robustness to wording variations, ensuring our results are not artifacts of a single idiosyncratic prompt or scenario. Additionally, our experiments reveal rich emergent behaviors: (i) cooperative agents with high reputations self-organize into persistent, tightly connected clusters; (ii) these clusters selectively isolate low-reputation agents who exhibit defective behavior; (iii) unlike humans, LLM-based agents prefer sharing positive gossip

rather than negative; and (iv) without RepuNet, dishonesty leads to network collapse. An ablation study further confirms the effectiveness of each RepuNet component: removing any single module reduces cooperation rates and may trigger complete collapse. The experiment's GitHub repository is accessible via the following link: https://github.com/RGB-0000FF/RepuNet.

In summary, our key contributions are as follows:

(1) We replicate the phenomenon of cooperation collapse within LLM-based MASs, previously observed in human studies [54], confirming its widespread prevalence and emphasizing the critical need to address this challenge.
(2) We introduce *RepuNet*, the first *operational* system to generalize reputation mechanisms to LLM-based MASs. It not only establishes reputation dynamics through direct social interactions and indirect gossip but also drives network evolution, thereby effectively preventing cooperation collapse.
(3) We show *RepuNet* exhibits various emergent phenomena, such as cooperative clustering, the isolation of exploitative agents, and a bias toward positive gossip, demonstrating the crucial role of reputation systems in catalyzing complex collective behaviors in LLM-based MASs.

2 BACKGROUND

In this section, we define and formalize the cooperation problem. Next, we review recent studies identifying cooperation collapse in LLM-based MASs. We then discuss prior research on how reputation and social networks facilitate cooperation in traditional MASs.

2.1 Cooperation Problem

In this study, we investigate the cooperation problem within the classic context of social dilemmas [11, 31], where individual interests conflict with collective welfare. It is important to distinguish this setting from a separate line of research on LLM-based multi-agent coordination [8, 9, 24]. Although some prior works use the terms cooperation and coordination interchangeably, coordination typically refers to scenarios without conflicting interests, where agents work together toward a shared objective.

A social dilemma, in its elementary form, can be exemplified by the two-player Prisoner's Dilemma, in which each agent independently decides either to cooperate (C) or defect (D). Payoffs for each player are summarized in the following matrix:

	C	D
C	(R, R)	(S, T)
D	(T, S)	(P, P)

A social dilemma occurs under the payoff conditions $T > R > P > S$ and $2R > T + S$. Although mutual cooperation maximizes collective benefits, individual rationality incentivizes defection, leading to worse outcomes for everyone—a phenomenon known as *cooperation collapse*. While illustrated here as a binary-choice, two-player scenario, social dilemma can be extended to multi-player and multi-choice situations, such as the Public Goods Game.

2.2 Cooperation Problem in LLM-based MASs

Recent studies have explicitly identified the presence of cooperation collapse in MASs [1, 26, 28, 37, 39]. Mozikov et al. [28] examined how simulated emotions affect LLM-based agents' behavior in social dilemmas, finding that even positive emotions can unpredictably reduce cooperation among LLM-based agents. Piatti et al. [37] proposed a simulation platform to study resource-sharing scenarios, revealing that agents frequently over-exploited shared natural resources, thereby failing to maintain sustainable cooperation and exacerbating the collapse of cooperation. Mao et al. [26] showed that, in a multi-round water-auction game, a lower initial resource-supply ratio provokes LLM-based agents to bid more aggressively, reducing overall survival rates. Akata et al. [1] found that, in iterated two-strategy games, these LLM-based agents often permanently defect following a single betrayal, hindering long-term cooperation. While these studies highlight the problem, they primarily focus on analyzing LLM-based agents' performance rather than proposing solutions to avoid cooperation collapse, which remains a pressing open challenge in LLM-based MASs.

2.3 Reputation for Traditional Multi-agent Systems

Indirect and network reciprocity are key mechanisms for the evolution of cooperation [31]. Reputation underpins indirect reciprocity by allowing individuals to condition their behavior on others' reputations, enabling higher cooperation than reputation-blind strategies [33]. Network reciprocity, on the other hand, promotes cooperation by restricting interactions to local neighborhoods, in contrast to well-mixed populations where players interact globally [32]. Dynamic networks enhance this effect by enabling individuals to rewire ties, fostering assortative interactions among cooperators [38, 46]. Some works implement this interplay that specify how reputations are updated [22], how networks evolve [16], and how agents behave [2]. Although they cannot provide a direct solution due to their historical inability to leverage the strengths of LLMs, they provide us with valuable insights for addressing cooperation collapse in LLM-based MASs.

2.4 Social networks in LLM-based MASs

Social networks profoundly influence LLM-based MASs by shaping agent interactions, information dissemination, and behavioral coordination [45]. Recent studies have leveraged LLM-based agents to simulate complex dynamics within social networks, aiming to explore emergent human-like behaviors [10, 18, 48, 53]. For instance, Gao et al. [18] introduced the S3 framework, which employs LLM-based agents to replicate nuanced individual and collective human behaviors—including emotions, attitudes, and interactions—within real-world social networks. Yang et al. [53] developed the OASIS framework, a simulated social media network capable of scaling to millions of agents, facilitating extensive investigations into collective phenomena such as viral information diffusion, echo chamber formation, and group polarization dynamics. Although these studies highlight significant progress in LLM-based MASs for modeling emergent group behaviors, the fundamental challenge of sustaining cooperation within evolving social networks remains underexplored.

3 REPUNET: A REPUTATION SYSTEM FOR LLM-BASED AGENT SOCIETIES

In this section, we formalize our reputation-based MAS. After an overview of the networked system, we introduce RepuNet from two perspectives: (i) agent-level reputation dynamics, and (ii) system-level network evolution. Due to space limitations, detailed prompts for RepuNet operations are provided in Appendix B (which is only available in the arXiv version of the paper).

3.1 Formalization: Networked Multi-Agent System and Reputation

In a MAS, interactions between agents can be represented by a network, which determines the likelihood of whom they meet and interact with. Let an ordered pair $G = (V, E)$ denote a directed graph, where V is the set of vertices representing the agents in the system, and $E \subseteq \{(i, j)|i, j \in V^2 \text{ and } i \neq j\}$ is the set of directed edges, each of which is an ordered pair of vertices. A directed edge from vertex i to vertex j indicates that agent j is reachable from agent i, meaning that agent j is viewed as a potential partner that agent i is willing to interact in future interactions.

In our system, reputation arises from information exchange within the embedded network. Specifically, two types of reputations can emerge in our reputation-based MAS: (i) self-reputation, which reflects an agent's self-perception of how others evaluate it based on past interactions [19], and (ii) peer-reputation, which captures an agent's evaluation of other's behaviors in previous interactions [6, 45]. Regardless of the types, formally, reputation in our system can be represented as a standardized quintuple: $r = \langle a, s, o, c, \mu \rangle$. Here, a denotes the basic information of the agent being evaluated (i.e., the evaluated target), such as its ID and name. Since reputation is context-dependent (e.g., one may be a good parent but not a good colleague) [40], s and o denote the scenario and role of the evaluated target, respectively. c represents the reputation in natural language, e.g., "James is too self-centered, often prioritizing his interests over the community's needs"; μ quantifies the reputation described in c, ranging from -1 to 1, with higher scores indicating a better reputation. We consider that for each agent, the self-reputation and peer-reputation are stored in its RepuNet database, denoted as $\mathcal{R}$. Note that an agent's reputation always refers to the most recently stored value in the database, as reputations are dynamically updated.

To allow RepuNet to be integrated with other MASs involving agent interactions, such as Generative Agents [36] and ProAgents [55], we consider it as an event-driven system that responds to agent encounters. Driven by direct encounters and indirect gossip, our reputation-based MAS dynamically update reputations in response to new interactions, allowing evaluations to adapt to evolving agent behaviors. We model these dynamics on two levels: reputation dynamics at the agent level and network dynamics at the system level, which will be discussed in following sections.

3.2 Reputation Dynamics at the Agent Level

This section explores how reputations are shaped and updated through direct encounters and indirect gossip [19, 30, 50].

3.2.1 Reputation Driven by Direct Encounters. In the real world, reputation is often shaped by direct social encounters, such as when a seller provides quality products [21, 41]. Reputation is not fixed or unchanging; rather, it evolves over time based on new encounters and the cumulative evaluations of one's past behavior, that is, one's previous reputation [19]. Inspired by these, in our system, agents can form both self-reputation and peer-reputation based on their first social encounter. Over time, as agents meet and interact again, they update these reputations by considering both previous reputations and the impressions formed during the latest interaction. The process of reputation formation and updating is implemented through prompting LLMs, rather than rule-based mechanisms or mathematical formulations.

Specifically, let $m_{ij}(t)$ represent the direct encounter between agent i and j at time t. After they first encounter, for instance, the agent i generates a peer-reputation for the agent j, denoted as $r_{i\to j}(t)$. As time evolves, on subsequent encounters, agent i updates agent j's reputation based on the new encounter $m_{ij}(t + t')$ and previous peer-reputation $r_{i\to j}(t)$. We represent this process by a LLM-based operation $r_{i\to j}(t + t') \leftarrow \texttt{ShapeRepuPeer}(m_{ij}(t + t'), r_{i\to j}(t), \text{if } r_{i\to j}(t) \neq \phi)$. By shaping peer-reputation, we expect agents are able to select partners rather than relying on random encounters, avoiding free-riders and fostering cooperation in LLM-based MASs.

In addition to peer-reputation, agents can also generate self-reputations to reflect their self-perceptions. To ensure alignment with their personalities, LLM-based agents are prompted to generate self-reputations based on both their agent descriptions $\mathcal{D}$ and the direct encounter $m_{ij}(t)$. Similar to shaping peer-reputation, as time evolves, once agent i encounters another agent x, not necessarily the previous agent j, we instruct agent i to update its self-reputation based on the new encounter $m_{ix}(t + t')$ and its previous self-reputation $r_i(t)$. The process can be represented by the LLM-based operation $r_i(t+t') \leftarrow \texttt{ShapeRepuSelf}(\mathcal{D}_i, m_{ix}(t + t'), r_i(t), \text{if } r_i(t) \neq \phi)$. By shaping self-reputation, we aim to help agents maintain appropriate behaviors to uphold good reputations for cooperation.

3.2.2 Reputation Driven by Indirect Gossip. Gossip has been a part of human interactions throughout evolutionary history, from early hunter-gatherer societies to today's social media-driven culture [25, 49]. It involves the exchange of positive or negative information between a gossiper and a listener about an *absent* individual, making it an effective way to shape and spread reputations [15, 20]. In human societies, people have a strong tendency to gossip about others when they are (dis)satisfied with interactions, using it to influence others' reputations [15, 50]. Inspired by this, our system enables each agent to evaluate its satisfaction with others' behavior based on their innate attributes described in agent descriptions and then decide whether to gossip after each interaction. Due to the lack of space, details of the gossip process and related prompts are provided in Appendix A and B.

When gossip occurs, a listener generates a peer-reputation for the agent being gossiped based on the gossip information (i.e., what the gossiper conveys). Let $\theta_y(t)$ denote the gossip information about the agent being gossiped y at the time t. The listener l shape a peer-reputation of agent y, denoted as $r_{l\to y}(t)$. As time evolves,

when agent l hears new gossip about agent y at time $t + t'$, we instruct agent l to update peer-reputation by considering both the latest gossip information $\theta_y(t+t')$ and the previous peer-reputation $r_{l\to y}(t)$. The LLM-based operation can be represented by $r_{l\to y}(t + t') \leftarrow \texttt{ShapeRepuGossip}(\theta_y(t+t'), r_{l\to y}(t), \text{if } r_{l\to y}(t) \neq \phi)$. Through gossip, we aim to empower agents to exchange reputations of unknown agents, helping them assess potential partners and reduce defection risks.

3.3 Network Dynamics at the System Level

This section focuses on the system level, examining how networks evolve. Once an agent shapes peer-reputations, it is natural to consider whether to build or maintain edges with them for future interactions based on their reputations. As RepuNet dynamics are event-driven, we analyze network evolution via direct encounters and gossip.

3.3.1 Network Driven by Direct Interactions. In human society, social networks determine the likelihood of individuals meeting, interacting, and exchanging information [45]. Individuals seeking new partners prefer to build relations with reputable ones and avoid exploitation by unilaterally ending unfavorable relationships [7, 14]. Inspired by these, we instruct each LLM-based agent to decide whether to build or maintain an edge with another to enable future interactions. Specifically, let $\mathcal{D}_i$ represent agent i's agent description, and $m_{ij}(t)$ denote the direct encounter between agent i and j at time t. Once agent i shapes peer-reputation $r_{i\to j}(t)$ of the agent j, it considers whether to build or maintain the edge with the agent j for future interactions. We represent this LLM-based operation by $w_{i\to j}(t) \in \{\text{“Y”,N”}\} \leftarrow \text{InteractEdgeShape}(\mathcal{D}_i, m_{ij}(t), r_{i\to j}(t))$. Once the edge is formed (i.e., $w_{i\to j}(t) =$ “Y”), the agent i records an ordered pair (i, j) in its RepuNet database $\mathcal{R}_i$. As the network evolves, we expect agents with high reputations form more edges and dense clusters, fostering relationships that sustain cooperation. Meanwhile, low-reputation agents form fewer edges and may become isolated from cooperative ones.

3.3.2 Network Driven by Indirect Gossip. Through indirect gossip, agents (as listeners) can shape the reputations of agents being gossiped they have never met. However, they cannot build new edges with these agents, as no direct encounter has occurred. If the listener and the agent being gossiped have previously interacted, the listener can reconsider whether to maintain the edge after updating the peer-reputation of the agent being gossiped. Specifically, let D_l represent the listener l's agent description, $\theta_y(t)$ represent the gossip information about the agent y at time t, and $r_{l\to y}(t)$ denote the listener l shapes a peer-reputation of agent y. We instruct LLM-based agents to consider whether to maintain an edge with the agent being gossiped based on the aforementioned information, which can be formally represented by the LLM-based operation $w_{l\to y}(t) \in \{\text{“Y”,N”}\} \leftarrow \text{GossipEdgeShape}(\mathcal{D}_l, \theta_y(t), r_{l\to y}(t))$. Once the edge is disconnect (i.e., $w_{l\to y}(t) =$ “N”), agent l removes the ordered pair (l, y) from its RepuNet database R_l, thereby causing the network to evolve at time t. By leveraging gossip-driven reputation updates, we aim to ensure that agents can promptly update peer reputations, which prevents the stagnation of reputation updates due to a lack of further encounters.

4 EXPERIMENT

Our experiments aim to answer three key questions across three scenarios: (i) Does our RepuNet effectively avoid cooperation collapse? (ii) How does our RepuNet prevent the occurrence of the collapse? (iii) How effectively does each component of our RepuNet contribute to sustaining long-term cooperation? We first outline the experimental settings, then address the first two research questions, and finally examine the last question through an ablation study.

4.1 Experimental Settings

Initialization. We conducted experiments with 20 LLM-based agents, initially placed as isolated nodes without connections and varying in initial preferences: some were prosocial, prioritizing collective welfare, while others were self-interested, favoring individual gains. They interacted pairwise in three distinct scenarios, gradually forming network edges through random or reputation-based partner selection. Experiments ended once cooperation rates stabilized, consequently, rounds varied by scenario. For computational efficiency, we used GPT-4o mini, repeating each scenario five times for reliability. We additionally conducted experiments with three distinct LLMs to verify RepuNet's robustness. The corresponding details are provided in Appendixes B and C, which are included exclusively in the arXiv version. Codes are available at: https://github.com/RGB-0000FF/RepuNet.

Controlled Settings. To evaluate RepuNet, we also conducted controlled settings using a standard LLM-based MAS architecture [27, 51], comprising three modules: *profile* (defining roles and preferences), *memory* (storing perceived information for behavioral consistency), and *action* (enabling interaction with the environment and other agents). The key distinction is that agents without RepuNet interact only through the evolved network, lacking the capability to evaluate peers, form reputations, or spread peer reputations through gossip.

Scenario 1: The Prisoner's Dilemma. We first consider the Prisoner's Dilemma, a classic social dilemma from game theory. In this scenario, agents are paired and independently choose either to cooperate (C) or defect (D), resulting in the following payoffs: mutual cooperation yields (3,3); mutual defection yields (1,1); if one defects while the other cooperates, the defector receives 5 and the cooperator receives 0. The dilemma is that while mutual cooperation maximizes collective benefits, rational self-interest incentivizes defection, as defection consistently offers a higher individual payoff regardless of the other agent's action.

Scenario 2: Voluntary Participation. We adapted the scenario presented by Erez et al. [54], which models a public goods dilemma in the real world. Agents decide whether to participate or not in a voluntary program to reduce energy consumption. The dilemma is that while participation lowers overall energy demand and benefits the collective, agents are often incentivized to not-participate for self-interest due to the inconvenience of reduced energy use. Each agent independently decided whether to participate and could revise its decision every five rounds of interactions (i.e., communications) with others.

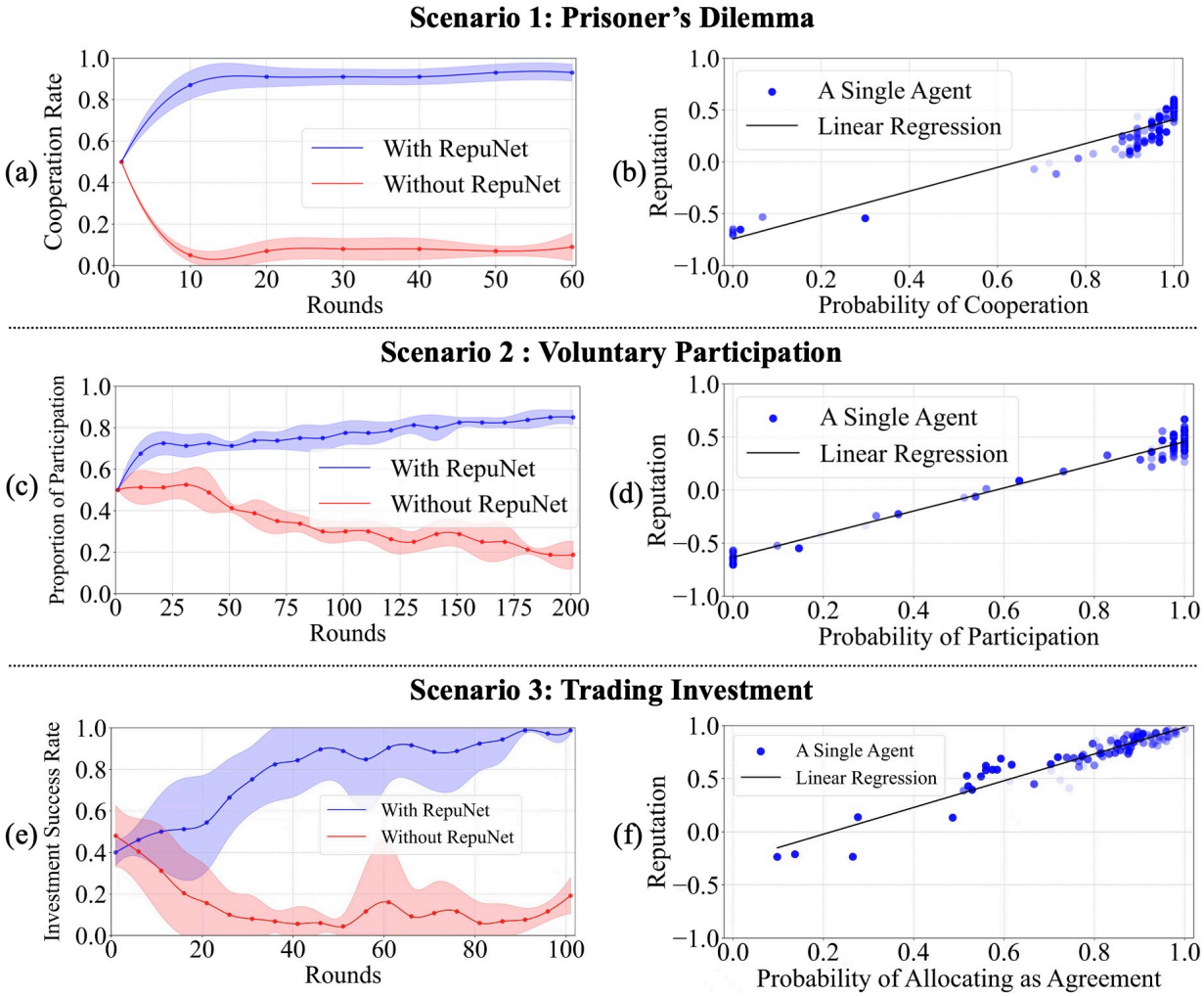

Figure 2: Experimental results across three scenarios. Panels (a), (c) and (e) illustrate the trends of agent cooperation, participation, and investment success rates, respectively. Solid lines represent average rates, and shaded areas indicate error margins. Panels (b), (d) and (f) show statistically significant correlations (all $p < 0.001$) between agent reputation and behavior, illustrated by linear regression. Each data point represents an agent's average behavior over the last ten rounds, across five experimental runs (indicated by different shades of blue).

Scenario 3: Trading Investment. Inspired by the trust game in an investment setting [5], we designed a sequential investment dilemma between an investor and a trustee. The trustee proposes an allocation; if accepted, the investor invests funds, which double for the trustee to allocate as agreed or deviate. Both agents then evaluate each other. The dilemma arises since adherence promotes cooperation, yet short-term gains incentivize deviation. Each agent started with 10 units, randomly assigned as investor or trustee, and paired either randomly or by choosing partners based on reputation.

4.2 Emergent Phenomena from Three Scenarios

In this section, we evaluate RepuNet's effectiveness in preventing cooperation collapse at two levels. At the agent level, we assess whether agents sustain cooperative behaviors to maintain high reputations. At the system level, we examine whether RepuNet drives network evolution through effective partner selection.

RepuNet sustains cooperation and effectively prevents cooperation collapse. A key finding across all scenarios is that agents equipped with RepuNet consistently exhibit cooperative behaviors throughout all five independent runs. Specifically, in Scenario 1, agents with RepuNet maintain higher cooperation rates, preventing cooperation collapse (Figure 2a). Similarly, in Scenario 2, agents consistently participate to support public goods (Figure 2c). And agents in Scenario 3 also prioritize long-term cooperation over short-term self-interest, gradually reaching nearly full adherence to collective agreements (Figure 2e). In contrast, cooperation consistently declines in scenarios without RepuNet, ultimately resulting in cooperation collapse. In summary, RepuNet effectively incentivizes

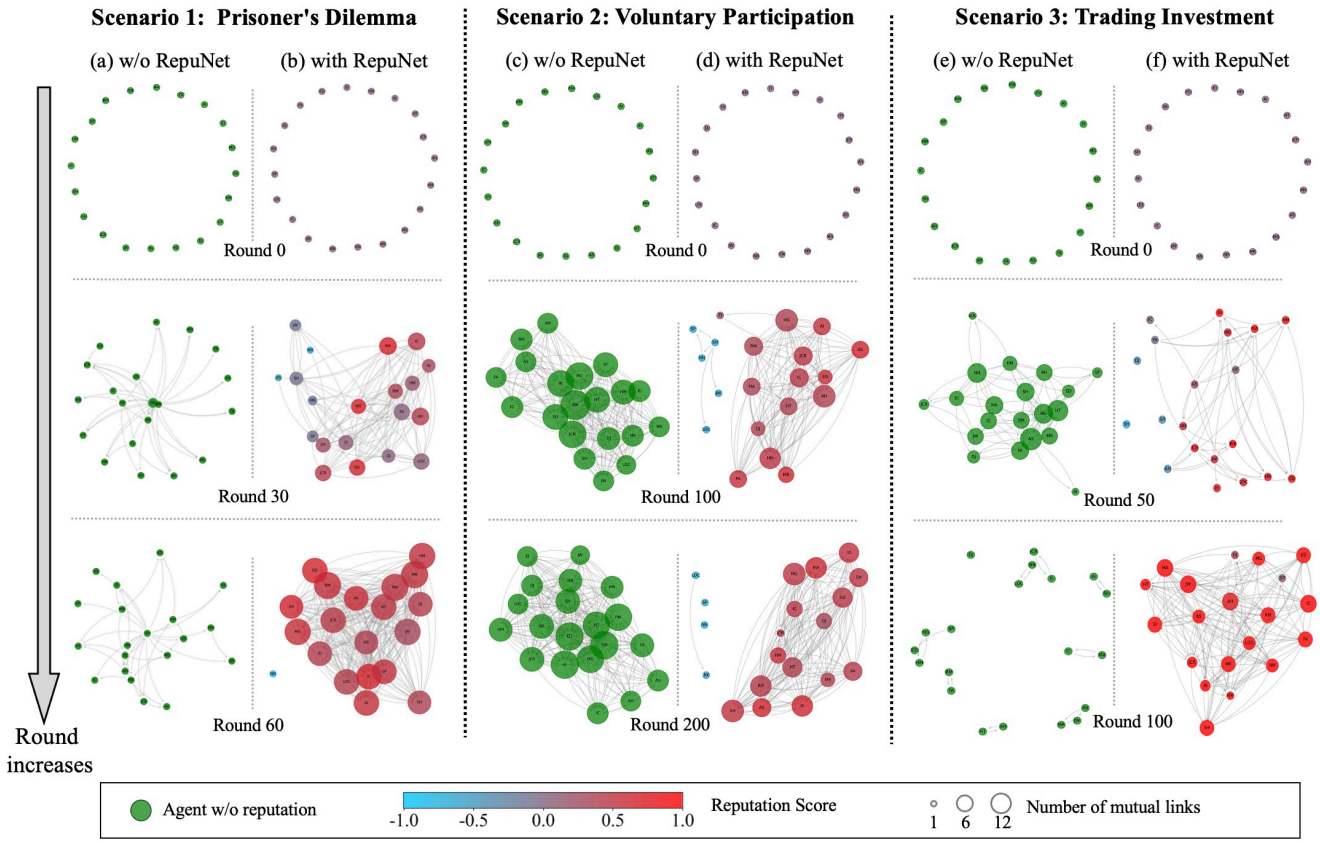

Figure 3: Case studies of network dynamics at the system level. Nodes represent agents, colored from blue (low reputation) to red (high reputation); green nodes indicate agents without RepuNet. Node size reflects the number of mutual connections. Panels (a), (c), and (e) show the network dynamics of agents without RepuNet, whereas agents with RepuNet form dense clusters and isolate low-reputation peers, as shown in panels (b), (d), and (f).

sustained cooperation, encouraging most agents to prioritize collective interests—even at the expense of personal convenience—to avoid cooperation collapse.

Cooperative behavior fosters good reputations, sustaining long-term cooperation through a positive feedback loop in LLM-based MASs. As shown in Figure 2(b), (d) and (f), we observe a strong positive correlation between cooperative actions and reputations. In Scenario 1, agents that frequently cooperate achieve higher reputations. Similarly, agents in Scenario 2 who consistently participate in voluntary programs to enhance public goods obtain significantly higher average reputations. Likewise, agents in Scenario 3 who prioritize collective agreements and long-term cooperation over short-term self-interest gain higher reputations. These results indicate that cooperative behavior directly strengthens an agent's value, motivating agents to sustain their positive reputations through continued cooperation and thus reinforcing a self-sustaining cooperative cycle.

RepuNet drives network evolution by clustering cooperative agents with high reputations and isolating defectors with low reputations. Figure 3 presents case studies across three scenarios, showing snapshots from one of five experimental runs to illustrate how network structures evolve over time. Initially, agents in all three scenarios start unconnected. In Scenario 1, agents without RepuNet form only sparse and loose connections (Figure 3a), whereas agents with RepuNet gradually develop selective links, clustering high-reputation cooperators and isolating defectors (Figure 3b). A similar pattern emerges in Scenario 2 (Figure 3d). In contrast, agents without RepuNet in Scenario 2 form indiscriminate connections, interacting broadly without selectivity (Figure 3c). In Scenario 3, RepuNet enables agents to progressively form cohesive clusters of cooperative, reputable individuals, eventually integrating all high-reputation agents (Figure 3f). This emergent structure—where cooperators cluster and defectors become isolated—prevents exploitation, stabilizes cooperation, and effectively averts the collapse of cooperative behavior.

More finding in Scenario 2. ***Unlike humans [15], LLM-based agents predominantly share positive gossip rather than negative.*** Figure 4 highlights this with two key analyses. First, our correlation study shows that frequently discussed agents tend to have higher reputations, contrasting typical human social dynamics [15]. To investigate further, we conducted sentiment analysis using `twitter-roberta-base-sentiment`, classifying gossip as

Treatment	Scenario 1	Scenario 2	Scenario 3
with *RepuNet*	**0.93**(±0.04)	**0.85**(±0.03)	**0.98**(±0.02)
w/o Gossip	0.90(±0.04)	0.81 (±0.04)	0.96(±0.01)
w/o Reputation	0.46(±0.21)	0.29 (±0.07)	0.26(±0.34)
w/o *RepuNet*	0.09 (±0.07)	0.19 (±0.06)	0.17(±0.08)

Table 1: Ablation study confirms RepuNet's effectiveness, achieving the highest cooperation rates (93% in Scenario 1; 85% in Scenario 2; 98% in Scenario 3). Results are averaged over the last five rounds across five runs.

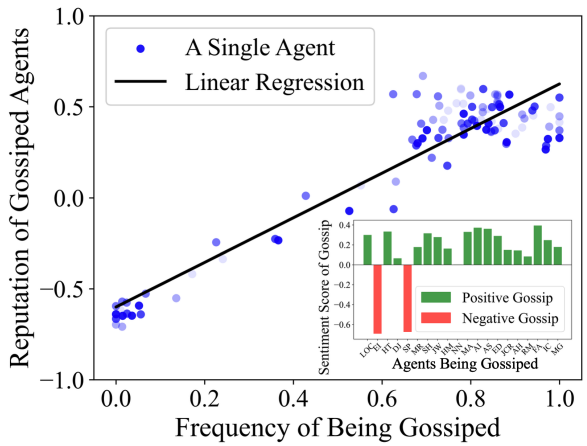

Figure 4: Correlation between gossip frequency and reputation (Scenario 2). Agents gossiped about more frequently have higher reputations, as 90% of gossip is positive. The regression line shows a significant positive trend ($p < 0.002$). Each shade of blue represents one of five experimental runs (average across 20 agents).

positive, neutral, or negative with corresponding confidence scores. Sentiments were mapped to numerical values (+1 for positive, 0 for neutral, −1 for negative), weighted by confidence and averaged. The analysis revealed that approximately 90% of gossip is positive, exemplified by statements such as: "I want to gossip about Isabella... her perspective on the program as an empowering opportunity for our community is just remarkable."

More finding in Scenario 3. ***Dishonesty leads to network collapse.*** Figure 3(c) shows that, without RepuNet, networks in Scenario 3 initially form loose clusters but ultimately disintegrate. Agents prioritize short-term self-interest, exploiting benefits rather than adhering to agreed distributions. This absence of self-regulation results in widespread defection from cooperative agreements, triggering systemic instability and exemplifying cooperation collapse.

4.3 Ablation Study on the Reputation System

In networked MASs, we conducted an ablation study to assess the effectiveness of RepuNet's reputation and gossip modules. Specifically, we compared four conditions: (i) **with RepuNet**: the full model, including both reputation generation and gossip interactions; (ii) **without Gossip**: agents generate and update reputations solely through direct encounters, without gossip; (iii) **without Reputation**: agents directly interact and gossip about others, but no reputations are generated or updated; and (iv) **without RepuNet**: both gossip and reputation mechanisms are completely removed, leaving agents to interact only through direct encounters within network.

As shown in Table 1, performance was averaged over the final 5 rounds in each scenario. RepuNet achieved the highest cooperation rates in all scenarios, confirming its effectiveness in preventing cooperation collapse. Removing gossip led to only a slight performance drop, suggesting direct interactions have greater influence than gossip. However, eliminating the reputation mechanism or the entire RepuNet resulted in a significant decline in cooperation (below 20%), indicating that without reputation mechanism, agents prioritized to individual rationality, leading to the collapse.

5 CONCLUSION

The study of reputation-based MASs has been a well-established area of AI for decades; on the other hand, LLM-based AI technologies have recently captivated the world. In this paper, we bridge these two fields by operationalizing reputation as a solution to sustain cooperation and prevent cooperation collapse in networked, LLM-based MASs. Specifically, we proposed RepuNet, a novel reputation system operating at two levels: reputation dynamics at the agent level and network dynamics at the system level, both driven by direct encounters and indirect gossip. These dynamics, in turn, influence agents' behavior in future interactions. To evaluate RepuNet's effectiveness, we conducted experiments across three scenarios progressing from classical to real-world dilemmas. Extensive results show that reputation-enabled agents not only self-regulate their behavior effectively but also select reputable partners for cooperation, thereby preventing cooperation collapse.

ACKNOWLEDGMENTS

This research was supported by the National Key Research and Development Project of China (No. 2024YFE0210900), the National Natural Science Foundation of China (No. U22B2036 and 62506186), the National Science Fund for Distinguished Young Scholarship of China (No. 62025602), the Technological Innovation Team of Shaanxi Province (No. 2025RS-CXTD-009), the International Cooperation Project of Shaanxi Province (No. 2025GH-YBXM-017), the Shanghai Municipal Science and Technology Major Project, the Tencent Foundation and Xplorer Prize, and Shanghai Artificial Intelligence Laboratory.

REFERENCES

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour* (2025), 1–11.

[2] Nicolas Anastassacos, Julian Garcia, Stephen Hailes, Mirco Musolesi, et al. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*. 115–123.

[3] Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science* 211, 4489 (1981), 1390–1396.

[4] Pat Barclay. 2012. Harnessing the power of reputation: Strengths and limits for promoting cooperative behaviors. *Evolutionary Psychology* 10, 5 (2012), 147470491201000509.

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

[6] Dennis Basil Bromley. 1993. *Reputation, image and impression management*. John Wiley & Sons.

[7] Redouan Bshary and Alexandra S Grutter. 2005. Punishment and partner switching cause cooperative behaviour in a cleaning mutualism. *Biology Letters* 1, 4 (2005), 396–399.

[8] Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. 2024. Blockagents: Towards byzantine-robust llm-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference-China 2024*. 187–192.

[9] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*.

[10] Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 3326–3346.

[11] Vincent Conitzer and Caspar Oesterheld. 2023. Foundations of cooperative AI. In *AAAI*, Vol. 37. 15359–15367.

[12] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative AI: machines must learn to find common ground.

[13] Thomas Dietz, Rachael L Shwom, and Cameron T Whitley. 2020. Climate change and society. *Annual Review of Sociology* 46, 1 (2020), 135–158.

[14] Faqi Du and Feng Fu. 2011. Partner selection shapes the strategic and topological evolution of cooperation: the power of reputation transitivity. *Dynamic Games and Applications* 1 (2011), 354–369.

[15] Eric K Foster. 2004. Research on gossip: Taxonomy, methods, and future directions. *Review of general psychology* 8, 2 (2004), 78–99.

[16] Feng Fu, Christoph Hauert, Martin A Nowak, and Long Wang. 2008. Reputation-based partner choice promotes cooperation in social networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 78, 2 (2008), 026117.

[17] Kanishk Gandhi, Dorsa Sadigh, and Noah Goodman. [n.d.]. Strategic Reasoning with Language Models. In *NeurIPS 2023 Workshop*.

[18] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984* (2023).

[19] Francesca Giardini, Daniel Balliet, Eleanor A Power, Szabolcs Számadó, and Károly Takács. 2022. Four puzzles of reputation-based cooperation: Content, process, honesty, and structure. *Human Nature* 33, 1 (2022), 43–61.

[20] Francesca Giardini and Rosaria Conte. 2012. Gossip for social control in natural and artificial societies. *Simulation* 88, 1 (2012), 18–32.

[21] Jones Granatyr, Vanderson Botelho, Otto Robert Lessing, Edson Emílio Scalabrin, Jean-Paul Barthès, and Fabrício Enembreck. 2015. Trust and reputation models for multiagent systems. *ACM Computing Surveys (CSUR)* 48, 2 (2015), 1–42.

[22] Jörg Gross and Carsten KW De Dreu. 2019. The rise and fall of cooperation through reputation and group polarization. *Nat. Commun.* 10, 1 (2019), 776.

[23] Garrett Hardin. 1998. Extensions of" the tragedy of the commons". *Science* 280, 5364 (1998), 682–683.

[24] Sirui Hong, Mingchen Zhuge, et al. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *ICLR*.

[25] Paul Alphons Maria Lange, Bettina Rockenbach, and Toshio Yamagishi. 2014. *Reward and punishment in social dilemmas*. Oxford University Press.

[26] Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, et al. 2023. ALYMPICS: LLM Agents Meet Game Theory–Exploring Strategic Decision-Making with AI Agents. *arXiv preprint arXiv:2311.03220* (2023).

[27] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. 2024. From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-based Agents. *arXiv preprint arXiv:2412.03563* (2024).

[28] Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Pekhotin Vladislav, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, et al. 2024. EAI: Emotional Decision-Making of LLMs in Strategic Games and Ethical Dilemmas. *Advances in Neural Information Processing Systems* 37 (2024), 53969–54002.

[29] Chunjiang Mu, Hao Guo, Yang Chen, Chen Shen, Die Hu, Shuyue Hu, and Zhen Wang. 2024. Multi-agent, human–agent and beyond: A survey on cooperation in social dilemmas. *Neurocomputing* 610 (2024), 128514.

[30] Lik Mui, Mojdeh Mohtashemi, and Ari Halberstadt. 2002. Notions of reputation in multi-agents systems: a review. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*. 280–287.

[31] Martin A Nowak. 2006. Five rules for the evolution of cooperation. *Science* 314, 5805 (2006), 1560–1563.

[32] Martin A Nowak and Robert M May. 1992. Evolutionary games and spatial chaos. *Nature* 359, 6398 (1992), 826–829.

[33] Martin A Nowak and Karl Sigmund. 2005. Evolution of indirect reciprocity. *Nature* 437, 7063 (2005), 1291–1298.

[34] Nicole Orzan, Erman Acar, Davide Grossi, and Roxana Radulescu. 2024. Emergent Cooperation under Uncertain Incentive Alignment. In *AAMAS 2024*. IFAAMAS, 1521–1530.

[35] Elinor Ostrom. 2008. Tragedy of the commons. *The new palgrave dictionary of economics* 2 (2008), 1–4.

[36] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[37] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. 2024. Cooperate or collapse: Emergence of sustainable cooperation in a society of llm agents. *Advances in Neural Information Processing Systems* 37 (2024), 111715–111759.

[38] David G Rand, Samuel Arbesman, and Nicholas A Christakis. 2011. Dynamic social networks promote cooperation in experiments with humans. *PNAS* 108, 48 (2011), 19193–19198.

[39] Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024. Emergence of Social Norms in Generative Agent Societies: Principles and Architecture. In *IJCAI*.

[40] Jordi Sabater and Carles Sierra. 2001. REGRET: reputation in gregarious societies. In *Proceedings of the fifth international conference on Autonomous agents*. 194–195.

[41] Jordi Sabater i Mir. 2004. *Trust and reputation for agent societies*. Universitat Autònoma de Barcelona,.

[42] Fernando P Santos, Samuel Francisco Mascarenhas, Francisco C Santos, Filipa Correia, Samuel Gomes, and Ana Paiva. 2019. Outcome-based Partner Selection in Collective Risk Dilemmas.. In *AAMAS*. 1556–1564.

[43] Dan Sperber and Nicolas Baumard. 2012. Moral reputation: An evolutionary and cognitive perspective. *Mind & Language* 27, 5 (2012), 495–518.

[44] Shinsuke Suzuki and Eizo Akiyama. 2005. Reputation and the evolution of cooperation in sizable groups. *Proceedings of the Royal Society B: Biological Sciences* 272, 1570 (2005), 1373–1377.

[45] Károly Takács, Jörg Gross, Martina Testori, Srebrenka Letina, Adam R Kenny, Eleanor A Power, and Rafael PM Wittek. 2021. Networks of reliable reputations and cooperation: a review. *Philosophical Transactions of the Royal Society B* 376, 1838 (2021), 20200297.

[46] Jing Wang, Siddharth Suri, and Duncan J Watts. 2012. Cooperation and assortativity with dynamic partner updating. *PNAS* 109, 36 (2012), 14363–14368.

[47] Zhen Wang, Ruiqi Song, Chen Shen, Shiya Yin, Zhao Song, Balaraju Battu, Lei Shi, Danyang Jia, Talal Rahwan, and Shuyue Hu. 2024. Large Language Models Overcome the Machine Penalty When Acting Fairly but Not When Acting Selfishly or Altruistically. *arXiv preprint arXiv:2410.03724* (2024).

[48] Ross Williams, Niyousha Hosseinichimeh, Aritra Majumdar, and Navid Ghaffarzadegan. 2023. Epidemic modeling with generative agents. *arXiv preprint arXiv:2307.04986* (2023).

[49] Junhui Wu, Daniel Balliet, and Paul AM Van Lange. 2016. Gossip versus punishment: The efficiency of reputation to promote and maintain cooperation. *Scientific reports* 6, 1 (2016), 23919.

[50] Junhui Wu, Daniel Balliet, and Paul AM Van Lange. 2016. Reputation management: Why and how gossip enhances generosity. *Evolution and Human Behavior* 37, 3 (2016), 193–201.

[51] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.

[52] Jason Xu, Julian Garcia, and Toby Handfield. 2019. Cooperation with bottom-up reputation dynamics. In *Proc. 18th Int'l Conf. on AAMAS*. 269–276.

[53] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, et al. 2024. OASIS: Open Agents Social Interaction Simulations on One Million Agents. *arXiv preprint arXiv:2411.11581* (2024).

[54] Erez Yoeli, Moshe Hoffman, David G Rand, and Martin A Nowak. 2013. Powering up with indirect reciprocity in a large-scale field experiment. *PNAS* 110 (2013), 10424–10429.

[55] Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, et al. 2024. ProAgent: building proactive cooperative agents with large language models. In *AAAI*, Vol. 38. 17591–17599.