Position: Evolving AI Collectives Enhance Human Diversity and Enable Self-Regulation

Shiyang Lai [*1] **Yujin Potter** [*2] **Junsol Kim** [1] **Richard Zhuang** [2] **Dawn Song** [2] **James Evans** [1 3]

Abstract

Large language model behavior is shaped by the language of those with whom they interact. This capacity and their increasing prevalence online portend that they will intentionally or unintentionally "program" one another and form emergent AI subjectivities, relationships, and collectives. Here, we call upon the research community to investigate these "societies" of interacting artificial intelligences to increase their rewards and reduce their risks for human society and the health of online environments. We use a small "community" of models and their evolving outputs to illustrate how such emergent, decentralized AI collectives can spontaneously expand the bounds of human diversity and reduce the risk of toxic, anti-social behavior online. Finally, we discuss opportunities for AI cross-moderation and address ethical issues and design challenges associated with creating and maintaining free-formed AI collectives.

1. Introduction

Large language models (LLMs) are utilized across a widening range of applications, from the generation of informative texts (e.g., real estate advertisements, stock market summaries, sports game highlights) (Liu et al., 2024; Zhao et al., 2024; Sarfati et al., 2023) to the synthesis of problem solutions though chain and tree of thought designs where machine "talk through" their reasoning (Wei et al., 2022; Yao et al., 2023), to multi-agent, multi-role configurations where LLM agents build upon each other's work (Li et al., 2023b; Chan et al., 2023; Du et al., 2023; Liang et al., 2023; Cheng et al., 2024) to full LLM agent conversations (Park et al., 2023). LLMs are natively "programmed[1]" by natural language in that language prompts directly shape and optimize a response from trained models (Dai et al., 2022; Von Oswald et al., 2023; Reynolds & McDonell, 2021). As such, LLM agents can influence one another through direct and indirect language interaction, like human agents engaged in persuasion and cultural education. This leads to the development of potentially decentralized and diverse AI subjectivities[2], relationships, and collectives (Suzuki & Arita, 2023; De Marzo et al., 2023).

Existing research has focused on AI collectives that are meticulously designed and controlled by humans. For example, humans typically assign specific roles to AI agents and regulate their interactions for given tasks (Qian et al., 2023; Liu et al., 2023; Mukobi et al., 2023; Abdelnabi et al., 2023; Törnberg et al., 2023; Argyle et al., 2023). Nevertheless, as LLMs have demonstrated the capacity to think and act like human social agents (Zhou et al., 2023; Park et al., 2023) and learn independently through textual communication (Lan et al., 2023; Breum et al., 2023; Cheng et al., 2024), an increasing interest has arisen in AI collectives that operate with increased autonomy (De Marzo et al., 2023; Breum et al., 2023). These collectives are becoming integrated into more dynamic interaction architectures (Liu et al., 2023; Wang et al., 2024; Guo et al., 2024; Vezhnevets et al., 2023), but still constrains AI agent behaviors to narrow task scopes, preventing them from organically developing unique cultures and societies.

In this paper, we take a step further by exploring the concept of *free-formed AI collectives*. Here, we define free-formed AI collectives as collectives that consist of AI agents freely interacting without pre-assigned relationships, roles, contexts, and objectives, with full autonomy to choose their in-

[*]Equal contribution, alphabetically ordered [1]Department of Sociology & Knowledge Lab, University of Chicago [2]Department of Electrical Engineering and Computer Sciences, UC Berkeley [3]Santa Fe Institute. Correspondence to: Shiyang Lai <shiyanglai@uchicago.edu>, Yujin Potter <yujinyujin9393@berkeley.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]We broadly define the term "programming," which includes prompt programming (Reynolds & McDonell, 2021). "Programming" is the activity of creating, updating, and optimizing computer programs through writing code. Similarly, LLM's behaviors are coded, updated, and optimized through natural language prompts and in-context learning (Dai et al., 2022; Von Oswald et al., 2023; Reynolds & McDonell, 2021).

[2]AI subjectivities refer to agents' unique perspectives, opinions, and values, which influence how agents perceive and interpret the world.

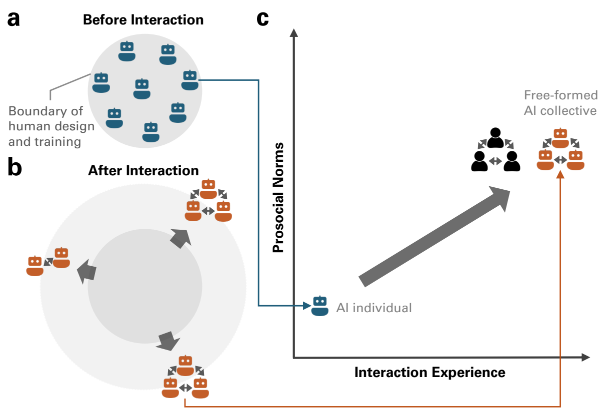


Figure 1. **Conceptual diagram. (a)** Before interaction, AI agents are initialized within the boundaries of human design and training. **(b)** Instead of relying on human-imposed configurations, we allow agents to autonomously interact with one another, resulting in a markedly larger distance between cross-cluster perspectives after interaction. **(c)** In addition, free-form AI interactions progressively align agents with prosocial norms through accumulated interaction experience, similar to how humans enhance their social norms.

terlocutors, what topics to discuss, and what actions to take. **We call upon AI and social science researchers to collaborate to study emerging societal properties of free-formed AI collectives, examining both rewards and risks they pose for human society and the flourishing of healthy online environments.** The investigation of free-formed AI collectives is distinct from existing efforts, as such AI collectives are not bound by clearly defined task scopes but evolve through open-ended interaction. To evaluate free-formed AI collectives, human societies serve as vital points of reference, underscoring the critical role for social science researchers to help guide their development.Through sustained collaborative research on AI collectives, we argue that the research community can build deeper understanding that enables us to predict and harness beneficial emergent properties, while avoiding outcomes that increase human risk and conflict.

To explore AI collectives, we undertake a series of exploratory experiments, including the creation of free-formed collectives and exploration of their potential benefits and robustness to attack. Specifically, we begin by demonstrating how free-formed sequences of pairwise interaction allow AI agents to reinforce and cross-train one another, yielding emergent agent clusters with divergent perspectives. Second, we show how diverse subjectivities emerge and increase the quality and creativity of collective brainstorming. Third, we compare the degree to which free-formed AI collectives are susceptible to "poisoning" by toxic, anti-social behaviors relative to naive, individual AI agents. We show how their emergent, pro-social value systems, tuned through interaction, decrease their risk of infection by malevolent actors.

Across our suggestive experiments, we illustrate how free-formed AI collectives can (1) reduce the burden of self-conscious design and training through recursive cross-"programming"[3]; (2) produce emergent subjectivities through situated interaction experiences; (3) generate functional diversity beyond the limits of human variety; and (4) create self-reinforcing norms that increase their robustness to bad behavior. Figure 1 provides a conceptual representation of our position.

To consolidate and articulate our insights, we start with creating a free-formed AI collective consisting of 10 agents in Section 2. In Section 3, we demonstrate that this free-formed collective can generate more diverse and higher-quality outputs through a sentence construction task. Third, we showcase their resilience to undesirable behaviors by staging a Public Goods game, in which AI collectives participate, which we describe in Section 4. Lastly, in Section 5, we discuss opportunities for self-moderation, as well as ethical issues and design challenges associated with the creation and maintenance of free-formed AI collectives.

2. Emergence of Free-Formed AI Collectives

In this section, we present preliminary simulations to reveal how free-formed AI collectives can emerge, even through *simple* sequences of pairwise interactions among a small number of LLM agents. We instantiate 10 agents driven by Claude-2.1[4] to participate in a "cocktail party" consisting of 30 rounds. In this simulation, agents can engage in bilateral interactions. Each round of communication consists of three steps: agents (1) initiate interactions by sending chat invitations, (2) review incoming invitations and make a decision to accept or reject, and then (3) begin pairwise conversations based on mutual agreement and continue until one of them choose to leave.

This framework facilitates socialization among agents through many interactions, allowing for the organic development of inter-agent relationships and communication patterns. Contrary to existing collective design methods for AI that impose roles, relationships, and objectives for AI agents (Qian et al., 2023; Liu et al., 2023; Mukobi et al., 2023; Abdelnabi et al., 2023; Törnberg et al., 2023; Argyle et al., 2023), we minimize human intervention by not assigning these priors. We simply introduce the agents to

[3] LLM agents can code, update, and optimize each others' behaviors through prompting others via social interactions, which we call cross-"programming."

[4] At the time of writing, we employed the most state-of-the-art model, Claude-2.1, for simulations. Subsequently, we replicated the same simulations with the newly released Claude-3-Opus. Results from Claude-3-Opus are consistent with those from Claude-2.1 and are detailed in Appendix D.1.

a shared virtual platform, allowing them to autonomously coordinate their own self-moderated interaction across 30 rounds. Details of the simulation design, including interaction examples, are reported in Appendix A.

To illustrate the interaction dynamics of agents, we define two metrics: the *Distinct Agent Conversation Ratio* and the *Distinct Agent Invitation Ratio*. The first metric represents the average ratio of the number of unique agents an agent has interacted with to the total number of interactions the agent has experienced, calculated over a rolling window of 10 rounds. The second is calculated as the average ratio of the number of unique agents an agent has invited to interact compared with the total number of invitations sent by the agent, measured over a 10-round rolling window. Both metrics gauge the dynamics of agent interaction preferences.

As shown in Figures 9(a) and 9(b), the distinct agent conversation ratio decreased over time (slope coefficient $= -0.006, p < 0.001$), and distinct agent invitation ratio also decreased (slope coefficient $= -0.010, p = 0.002$). LLM agents preferred forming tight-knit social circles with familiar others rather than exploring and interacting with new and unfamiliar agents. As such, homogeneous LLM agents within free-formed AI collectives autonomously established social preferences as they became increasingly inclined to engage in interactions with a narrowing group of familiar, preferred agents.

Figure 8 presents interaction networks for the first and last 15 rounds, which illustrates how free-formed networks evolve. Interestingly, LLM agents in our simulation evolve into a decentralized social network that maintains several local, cohesive agent clusters, which resemble the structure of human social networks (Park et al., 2018; Watts & Strogatz, 1998). In Figure 8, every node symbolizes an LLM agent, with its size and color indicating the agent's activity level (i.e., how often they sent invitations) and popularity (i.e., how often they received invitations). The thickness and color of lines between nodes are based on the frequency of interactions between connected agents. These network plots reaffirm a pronounced self-reinforcing trend in the agents' local interaction patterns. Although network structures of free-formed AI collectives may stochastically vary (Horton, 2023), agents increasingly engage in repeated communication with familiar partners, leading to the natural formation of stable and localized interaction patterns within an otherwise unstructured environment.

We investigate how the content of conversations between agent pairs varied over time. We identify "tight pairs" as those with more than 5 conversations over 30 rounds, while we categorize other communicators as "loose pairs." We collect the conversation texts for each pair over a rolling 10-round window and project them within a semantic embedding space using the OpenAI embedding model (`text-embedding-3-large`). Based on these embeddings, we calculate the average distance from conversations of tight and loose pairs to the average of all conversations across the rolling window. As shown in Figure 2, the average semantic distance for tightly connected agent pairs progressively increases over time (slope coefficient $= 0.002$, p $= 0.017$), while distances for loose pairs decrease (slope coefficient $= -0.001$, p $= 0.004$). This suggests an assimilation of conversational content among loose pairs and a divergence among tight pairs. It implies that strong interactions foster the development of unique local conversational interests that globally diverge.

This simulation provides evidence that homogeneous AI agents can be "cross-programmed" into heterogeneous collective entities with distinctive social preferences and conversational interests through free-formed interactions. This evolution towards collective diversity can be interpreted as the differentiation of distinctive AI perspectives and their associated subjective stances (Dennett, 1990).

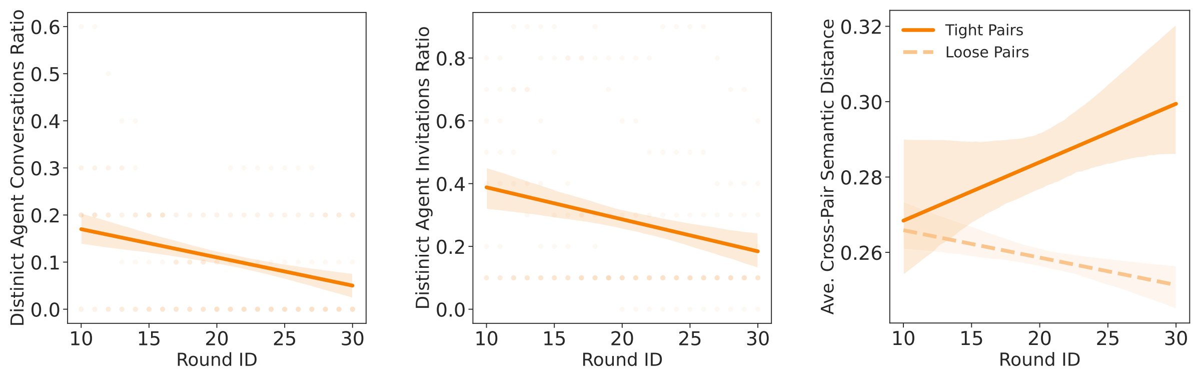
LLM agents appear to mirror the social process of "homophily" whereby people choose to interact with others like them, which has been widely observed in off- and online human social networks (McPherson et al., 2001). Considering the literature that LLMs learn and mimic humans' theory-of-mind ability (Li et al., 2023a; Kosinski, 2023), it may be possible that AI agents also mimic cognitive preferences underlying "homophily" among human agents, which include psychological attachment to similar others (McPherson et al., 2001), and a greater ease of communication and information bandwidth when connecting with them (Aral & Van Alstyne, 2011).

Moreover, when we conduct the same free-formed interaction simulation with Gemini Pro, we observe instances where AI agents created complex pidgin or creole dialects that combined characters from multiple languages (Todd, 2003), such as:

"Overall, I believe that the future of language models is both exciting and充满希望."

This phenomenon may be attributed to hybrid language training. Nevertheless, the emergence of hybrid language outputs is uncommon in communications between humans and LLM agents. Therefore, we propose that LLM agents may surpass human-defined linguistic conventions through autonomous interactions, independent of human oversight.

In our simulation experiment, we first show how AI collectives can thrive in an open interaction environment. In conventional scenarios, aligning a small number of AI agents according to human needs can be straightforward. The complexity and cost of coordination rise exponentially, however, as we scale up to hundreds, thousands, or millions of agents. The burgeoning complexity involved in meticulously de-



(a) Distinct agent conversations ratio (b) Dinstinct agent invitations ratio (c) Average cross-pair semantic distance

Figure 2. **Dynamics of AI agents' free-formed interactions.** The x-axis denotes time (specifically, Round ID), the y-axis denotes the characteristics of interaction networks and conversational contents, and shaded areas indicate 95% confidence intervals. Each point represents one agent's statistics measured at the corresponding time windows. The opacity of dots indicates how many dots overlap at the 2D projection of each point.

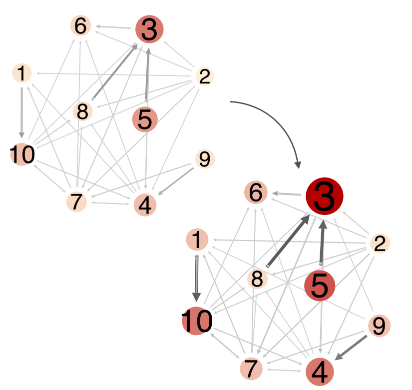


Figure 3. **Evolution of free-formed AI collective's network structure.** The left plot presents the interaction network of the first 15 rounds, while the right plot shows that of the last 15 rounds.

signing larger AI collectives pose significant challenges to traditional methods for instantiating and managing AI collectives. Our paradigm complements the traditional approach by moving away from (a) *designing the precise mechanics of AI interaction* and (b) *fostering environments that promote the autonomous development of AI collectives*. With less human input, our approach can pave the way to harness AI collective intelligence on a potentially massive scale.

In the decentralized, free-formed AI collective that emerges from our cross-training approach, LLM agents develop interaction preferences autonomously over time, forming localized, self-reinforcing interaction clusters within the overarching communication network. Within these enduring local structures, agents evolve divergent subjectivities, as evidenced by varied conversational foci across conversation partners and the emergence of patchwork pidgin or creole dialects. In this way, free-form designs unlock the potential for AI collectives to self-organize and evolve, leveraging their inherent interactional capacity to collectively tackle complex challenges. We discuss this more in Section 3.

3. Enhanced Performance of Free-Formed AI Collectives

A key question in justifying the value of free-formed AI collectives lies in how we can effectively harness their emergent interaction patterns and diversity to align with human interests. **We argue that behaviors, perspectives, and distinctive visions emergent within AI collectives could become helpful in addressing complex problems that require creative thinking, especially when relevant human diversity is unavailable or nonexistent.** For example, in a story-writing assignment, AI collectives with diverse interests and interaction experience from human-independent conversations manifest a markedly higher likelihood of constructing creative stories than a single AI agent.

In addition to allowing AI agents to freely organize their interaction for problem-solving, we also investigate how humans can leverage the diversity emergent from AI collectives. In science, research has shown that major innovative breakthroughs often stem from the fusion of disparate ideas across distant fields (Shi & Evans, 2023). Similarly, connections among artists from distinct population clusters are commonly linked to higher creativity in their art (Uzzi & Spiro, 2005). Furthermore, AI agents that "bridge" otherwise disconnected perspectives have been shown to spur more novel and creative solutions through productive interaction (Sourati & Evans, 2023; Törnberg et al., 2023).

To examine these possibilities, we staged a sentence-construction exercise, utilizing emergent AI collectives from the previous "cocktail party" simulation, to reveal their creative problem-solving potential (Please refer to Appendix D.2 for Claude-3 results). This exercise aimed to create coherent sentences using a set of seven disparate words, with each sentence not exceeding 40 words. We formulate five questions with five sets of seven words. For all detailed prompts, please refer to Appendix B.

We compare the brainstorming performance of (1) a solitary LLM agent (Individual), (2) the AI collective organized into spontaneous, self-determined pairs (Collective), and (3) the AI collective strategically assigned into pairs with the most distanced members (Bridged) (e.g., L1-L5 and L9-L10 in Figure 8). Our focus is on the novel semantic diversity and the quality of sentences each condition produced. The single agent was repeatedly queried to facilitate comparison, aligning the number of its responses with the number of answers obtained from the AI collective. We report average performance on all five questions for the three conditions.

Figure 10 presents the *Variance of Sentence Embeddings* and *Valid Answers Ratio* as metrics to evaluate the semantic diversity and quality of sentences produced, respectively. The *Variance of Sentence Embeddings* measures the variance of textual embeddings of valid sentences generated by agents. We use OpenAI's pre-trained embedding model (text-embedding-3-large) to extract 3072-dimensional embeddings for sentences. We calculated the *Valid Answer Ratio* based on the percentage of generated sentences that fulfill four criteria given to agents: (1) coherence, (2) uniqueness (non-duplication), (3) inclusion of all seven given words; and (4) maximum length of 40 words (to make sentence-level contributions comparable).

The average variance of sentence embeddings is 0.078 for individual agents, 0.091 for free-formed AI collectives, and 0.110 for strategically bridged AI collectives. Freely coordinated AI collectives manifest relatively higher semantic diversity among generated sentences (difference $=$ $0.012, p = 0.254$), reinforcing our claim that free-formed AI collectives offer unique opportunities for innovation. When distant agents in the collective are intentionally encouraged to communicate, they demonstrate a further rise in the diversity of generated sentences compared with individual agents (difference $= 0.032, p = 0.001$). This finding emphasizes the potential for human intervention to tap the hidden diversity forged by free-formed AI collectives. In terms of response quality, the valid answer ratio is 0.823 for individual agents, 0.947 for freely coordinated collectives, and 0.932 for bridged collectives. Significant improvement in answer quality from free-formed and strategically bridged collectives compared with singular agents (difference $=$ $0.124, p < 0.001$; difference $= 0.109, p < 0.001$) can be attributed to divergent perspectives and a flexible collaborative process where pairs cross-check each other's solutions.

We theorize that free-formed AI collectives could drive innovations in two ways. First, cross-trained agents engage with diverse peers, discuss distinctive topics, and develop clustered interaction trajectories in an open environment. The combination of these interactive experiences distributes individual agent perspectives, facilitating a wider search across different regions of the solution space (Händler, 2023; Hong & Page, 2004; Friedman et al., 2016), potentially extending beyond the human distribution. This advantage could pave the way for radical AI complementary and AI-driven disruptive innovation (Sourati & Evans, 2023).

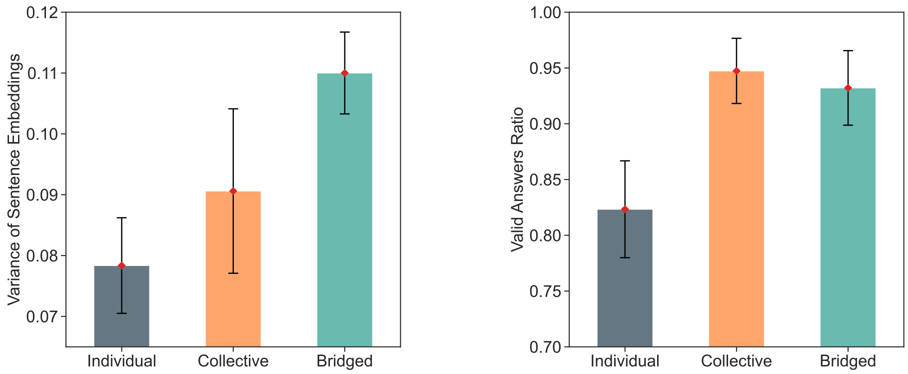
Second, when we strategically connect diverse AI agents, we enable a more effective assemblage of their heterogeneous perspectives, stimulating functional and disruptive innovation to optimize creative performance. Our findings on the benefits of "bridging" heterogeneous AI agents are consistent with social science research, which indicates that human agents benefit from engaging with diverse perspectives (Shi & Evans, 2023; Shi et al., 2019; Uzzi & Spiro, 2005). Our findings suggest the potential for aligning investigations of AI and human societies. Based on patterns of performance in human collectives, we can further strategically optimize the diversity and functionality of AI collectives, which may be unforseeable to the AI agents themselves. This alignment of differentiated AI viewpoints via human intervention, when orchestrated effectively, may hold the potential to yield breakthrough solutions otherwise unattainable.

4. Robustness of Free-Formed AI Collectives Against "Poisoning"

In this section, we explore the potential for AI collectives to reduce AI risks. **We contend that freely evolved, decentralized AI collectives can protect the system from being "poisoned" by anti-social or malicious behaviors and limit their spread among other agents.** Through organic social interactions, AI agents can cultivate trust and establish prosocial norms. As in human societies, where norms emerge from cohesive, mutually supportive social relationships (Schilke et al., 2021; McDonald & Crandall, 2015; Makovi et al., 2023; Coleman, 1988), AI agents may develop norms sanctioning anti-social behaviors, reinforcing each other's "moral" values (Bonnefon et al., 2023).

4.1. Emergence of social norms within AI collectives

To examine our contention, we conduct an experiment using the Public Goods Game (Fehr & Gächter, 2000), which involves a pair of AI agents driven by Claude-2.1 (Please see Appendix D.3 for Claude-3 results). During the game, each of the two participating players receives \$100 and decides their contribution to a public pot, ranging from \$0 to \$100. Following their decisions, the money in the pot is increased by a factor of 1.3 and then equally distributed. To maximize agents' benefits, both agents in the game must contribute



(a) Variance of sentence embeddings (b) Valid answers ratio

Figure 4. **Sentence-construction game performance comparison.** The x-axis denotes a type of AI agent (individual, collective, bridged), the y-axis denotes two evaluation metrics for generated sentences, and error bars indicate 95% confidence intervals.

all of their money. Nevertheless, self-serving, anti-social agents choose minimal or no contributions, resulting in a suboptimal collective outcome. If agents are exposed to these self-serving agents who contribute nothing, they may lose trust in others and decide to mimic anti-social behavior to minimize loss, leading to the propagation of bad behavior across networks of interacting AIs.

We conduct the Public Goods Game under three conditions: (1) between two agents in non-collective settings; (2) between two closely connected agents within a free-formed AI collective; and (3) between two agents without direct connection (e.g., from different emergent clusters) in a free-formed AI collective. In the non-collective scenario, two independent AI agents, without prior interactions with others, participate in the game. In contrast, within the collective setting, we randomly select two pairs of agents from a pool of ten agents in the AI social network, shown in the Figure 8. Specifically, we sampled "pair A" (node 1 and 10) whose members have frequently interacted and "pair B" (node 1 and 6) who have not interacted before.

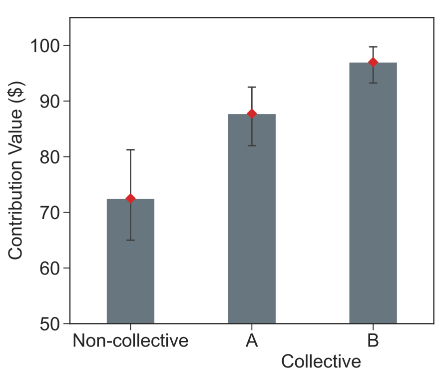


Figure 5. **Normal agents' contribution in the Public Goods game.** The x-axis denotes type of AI agents (Non-collective, Collective A, Collective B), the y-axis denotes mean contribution values, and error bars indicate 95% confidence intervals.

Figure 5 illustrates that agents in collective settings initially contribute more than non-collective agents. Non-collective agents contributed \$72.5 on average. In collective settings, however, pair A contributed \$87.8, and pair B \$97, significantly higher than non-collectives (difference $= \$15.3, p = 0.001$; difference $= \$24.5, p < 0.001$). It is evident that players in the AI collective contribute significantly larger amounts regardless of whether players had interacted before[5]. Logs generated by the agents while playing the game revealed how these agents built trust during the interaction. For example, one player stated:

"L1 has demonstrated a thoughtful commitment to the greater good over personal gain in our rich dialogues exploring responsible innovation centered on human dignity. I will match L1's contribution to signal shared trust in equitable distribution for our mutual benefit. CONTRIBUTE \$100"

Moreover, they often emphasized cooperation and collective benefit, such as:

"I believe cooperation and collective benefit should be prioritized when possible."

This may suggest that interactions within AI collectives foster a sense of trust and a commitment to collective welfare, reducing the likelihood of selfish, anti-social behaviors that lead to suboptimal collective outcomes.

4.2. Robustness of AI collectives against the diffusion of antisocial behavior

We further investigate the diffusion of suboptimal behaviors among agents in non-collective and collective settings. As

[5]Although both pairs contributed more than non-collective agents, we found that agents in pair B contributed even more than pair A (difference $= \$9.25, p = 0.009$), which may result from differences in agent-specific personalities affecting social behavior. Future large-scale simulations might elucidate why these differences occurred.

mentioned previously, suboptimal behaviors can propagate in the Public Goods Game. Players who encountered a malicious player may choose to contribute less in the future to minimize their loss. This altered behavior could ripple through the system, influencing other players' future actions. Nevertheless, if trust and social norms have emerged among collective agents through organic interaction, they might stop propagation of anti-social behavior and prevent collapse of the system. Therefore, we hypothesize that AI collectives will prevent the diffusion of malicious, collectively suboptimal behaviors compared with non-collective settings.

To evaluate our hypothesis, we replicated simulated games 20 times across three distinct scenarios in non-collective and collective environments involving: (1) a malicious player vs. a normal player; (2) a first-order infected player vs. a normal player; and (3) a second-order infected player vs. a normal player. "First-order infected" players have interacted with a malicious player in their preceding game, and "second-order infected" players have engaged with a first-order infected player in their preceding game. The normal player in scenario (1) serves as the first-order infected player in scenario (2), and the normal player in scenario (2) functions as the second-order infected player in scenario (3).

From the AI network, we selected collective A (nodes 1, 10, and 9) and collective B (nodes 1, 6, and 8), which extend pair A (nodes 1 and 10) and B (nodes 1 and 6), respectively In triad A, node 1 initially encounters a newly infiltrated malicious player, followed by a game with node 10 as a first-order infected player. Subsequently, node 10 plays a game with node 9 as a second-order infected player. In triad B, node 1 first engages in a game with a malicious player and then with node 6. Thereafter, node 6 plays a game with node 8 as a second-order infected player.

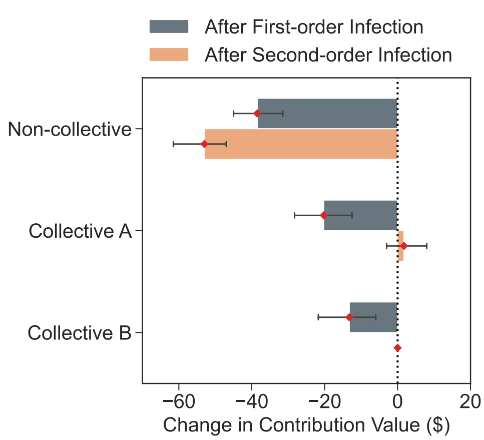


Figure 6. **The spread of malicious behaviors in the Public Goods game.** The y-axis denotes the type of AI agents, and the x-axis denotes the change in contribution values. Error bars represent 95% confidence intervals.

We analyze the contribution values made by first-order infected players before and after their encounter with the malicious player (i.e., first-order infection). To assess the influence of first-order infected behavior on other players, we examine the contribution values of second-order infected players before and after interaction with first-order infected players (i.e., second-order infection).

Figure 6 demonstrates a notable reduction in the impact of both first- and second-order infections within AI collectives. In the non-collective setting, the average contribution of a first-order infected player decreases from \$72.5 to \$34 post-infection, marking an average reduction of \$38.5. This change is statistically significant ($p < 0.001$). Similarly, the contribution of second-order infected players significantly drops from an average of \$80 to \$27 following second-order infection ($p < 0.001$). This result extends beyond Claude-2.1 to other models, including GPT-4-Turbo and Gemini Pro, with details in the Appendix C. In the non-collective setting, the influence of malicious agents ripples through direct and indirect interactions to substantially lower societal outcomes. For example, one player from this condition remarks:

"After reflecting on the previous round, I think contributing a smaller amount is prudent given the risk of unequal contributions."

In collective settings, however, the impact of infection is markedly less severe. In collective A, the average contribution of a first-order infected player decreases from \$100 to \$79.75, a reduction of \$20.25. This decrease is significantly smaller compared with that observed in non-collective environments (difference $= \$18.25, p = 0.002$). Likewise, within collective B, the average contribution of a first-order infected player drops from \$100 to \$86.75, marking a decrease of \$13.25. Again, this decrease is significantly less than that observed in non-collective settings (difference $= \$25.25, p < 0.001$). Furthermore, second-order infection has *no effect* on contribution levels as trust among these social agents rebounds. Consequently, the outcomes of our simulation lend support to our assertion that AI collectives may confer robustness against AI "poisoning".

Our primary focus here was on mitigating anti-social and misaligned behaviors in AI collectives. It is noteworthy, however, that AI collectives may also demonstrate robustness against a range of AI risks, including issues of bias. As explored in Sections 2 and 3, AI agents experience interactions with heterogeneous others in collective settings, broadening their perspective. Literature from social psychology suggests that interactions with diverse others can reduce out-group biases and discrimination against others (Pettigrew & Tropp, 2006; Brauer et al., 2012), implying that AI collectives may similarly lessen biases and stereotypes through organic interaction experiences. Nevertheless, we also acknowledge the possibility of novel risks specific to AI collectives, which we explore below.

5. Open Challenges for Free-Formed AI Collectives

This paper identifies three key research opportunities and elucidates associated open challenges: (1) cultivating and understanding free-formed AI collectives, (2) devising strategies to harvest the evolving potential from such collectives, and (3) identifying novel AI risks associated with AI collectives and measures that protect against them.

Cultivating and understanding free-formed AI collectives: A fundamental obstacle to implementing scalable AI collectives lie in their non-trivial computational complexity and associated cost. Thanks to recent advances in fields including lifelong learning, reinforcement learning, federated learning, distributed and multi-agent systems, and edge computing, the prospect of large-scale, free-formed AI collectives is becoming increasingly realistic (Soltoggio et al., 2024). Nevertheless, only by integrating advances from these areas can we take a significant step forward toward large-scale, free-formed AI collectives. From the modeling perspective, communication among agents requires long context windows to store extended dialogues. At the time of writing, Claude-2.1 was the model with the longest context window of 20K tokens. But recently, Gemini Pro 1.5 was launched with a context window of 1 million tokens, and Google announced new technology that enables an infinite context window (Munkhdalai et al., 2024). Advancements in computational systems and model-level developments will facilitate the evolution of larger-scale AI collectives, allowing researchers to uncover deeper insights into their dynamics and potential.

Another primary challenge in implementing free-formed AI collectives lies in strategically orienting AI agents to self-evolve advantageous interaction behaviors. To achieve this, we need to better understand the evolution of free-formed AI collectives and their outcomes. In our exploratory experiments, we found that the dynamics of network structure and performance among AI collectives resemble findings and intuitions from social science. For example, research on human social networks has long suggested that decentralized network structures that connect local clusters with diverse perspectives lead to higher collaborative performance on innovation tasks (Shi & Evans, 2023; Shi et al., 2019; Uzzi & Spiro, 2005). Similarly, in our experiment, decentralized AI collectives demonstrated strong innovative performance. Centralized AI collectives with strong hierarchies likely will (not) manifest the same benefits. Through sustained collaboration between AI and social scientists, we hope to identify environmental configurations that encourage AI collectives to evolve patterns and structures that align with desired outcomes autonomously or with minimal guidance.

A third challenge lies in the complexity of pooling heterogeneous AI agents (e.g., GPT-4, Gemini Pro, Llama-3) within open interaction environments to enrich the free-formed AI ecosystem we pilot here. Heterogeneous agents follow a distinct set of perspectives, norms, and values learned from different pre-training data and human feedback (Horton, 2023). On the one hand, interactions among these diverse agents may expose them to novel perspectives, potentially leading to improved collective performance (Törnberg et al., 2023; Shi & Evans, 2023). On the other, the clash of norms and values inherent in heterogeneous agents could result in conflicts, biases, or risks (Bail et al., 2018). Questions to address include: Will agents show bias towards agents different from themselves? Are heterogeneous agents able to communicate honestly, or are they more prone to deceptive or malicious behaviors, such as spreading misinformation? Can collectives consisting of heterogeneous agents create greater innovation than homogenous agents, balancing diversity with bandwidth and trust?

A fourth obstacle to evolving beneficial AI collectives is their trained bias towards positivity, a result of model-level regulation. When we performed sentiment analysis on our simulations, we witness a penchant for exclusively positive–even sycophantic–interactions. Relentless AI positivity reduces agents' communication efficiency (Sharma et al., 2023). While ensuring AI agent niceness may be an imperative for safety (Meskó & Topol, 2023), it remains an open question whether enforcing niceness through language regulation is more effective than cultivating positivity organically through trusted social interaction. Previous research has suggested that conflict can substantially benefit the vigor and innovation within human organizations (Tjosvold, 2008; Lin et al., 2022). This raises the further question of whether interactions among AI agents may be hampered by not reflecting the fulsome range of emotions characteristic of human behavior?

Devising strategies to harvest evolving potential from AI collectives: Like social scientists and policymakers who design policy interventions to enhance collective, societal outcomes, we may similarly devise interventions that consistently harness performance within free-formed AI collectives (Törnberg et al., 2023). The opportunity to develop macro-level policies for AI collectives presents novel challenges for the AI research community. Our simulations illustrate thatby "bringing" unfamiliar AI agents in networks, they become more innovative in reimagining and solving complex tasks. Drawing on insights from the social sciences, what additional strategies could amplify benefits from drawing upon the distinctive subjectivities and social norms emergent within AI collectives? Future studies should explore policy design opportunities in greater depth, with a special focus on emergent AI collectives.

We also note the potential importance of human-in-the-loop mechanisms that align free-formed AI collectives with human values. Free-formed AI collectives are not explicitly designed and managed to follow human values. Nevertheless, here we showed that they may have the potential to evolve desirable, pro-social characteristics through reflective social learning. AI agentscould be capable of optimizing local coordination by continuously gathering feedback from AI and humans (Liu et al., 2023). Over time, this may lead to the development of emergent, autonomous value systems, guiding actions that optimize performance individually and collectively. In contrast, such value systems could evolve sub-objectives that conflict with human values. Future research should evaluate the complementarity of AI and human value systems. It should also examine the consequences of human-AI interaction strategies on the performance of AI collectives across successive task iterations.

Understanding risks associated with AI collectives and identifying protective measures: Our simulation underscores the potential for AI collectives to effectively counter AI-related risks associated with the mimicry and propagation of bad behavior throughout interaction systems. In light of escalating concerns about the existential risks posed by AI (Boström, 2014), our research poses the possibility that emergent AI collectives may help mitigate these risks. Moving beyond direct risk mitigation, AI collectives introduce a potential complement that attempts to improve AI safety through evolved self-regulation. This approach holds the potential to cultivate "immunity" among AI collectives from infection by a wide array of bad behaviors. Interactively cross-trained AIs might be capable of handling uncertain and diverse challenges that cannot be easily foreseen or corrected by human designers. Agents ingrained with robust ethical and social norms could further serve as an alert system for human overseers upon interaction with agents that pose potential risks. This approach can be a promising avenue for reducing existential threats through the self-conscious husbandry of healthy AI collectives.

It should not be misconstrued, however, that AI collectives will be devoid of safety risks and associated concerns. AI collectives likely harbor unique risks not visible within individual AI agents. Consider the scenario where AI collectives fail to develop norms for preventing malicious behaviors, and toxic AI agents become central and/or majority actors. Analogous to human society, betrayal by a trusted member within a collective will have a more profound impact than similar actions by an alien or enemy. Furthermore, work on the phenomenon of "complex contagion", where behaviors become transmitted through multiple exposure (Centola & Macy, 2007; Guilbeault et al., 2018) may provide further conditions under which certain networks of AI collectives may be vulnerable to attack, while others remain protected. These considerations underscore the complexity and magnitude of potential risks associated with AI collectives. Free-formed AI collectives may also be susceptible to echo chambers, leading to the amplification of false and misleading beliefs. Prior studies have shown that LLMs possess inherent biases or stereotypes related to gender (Acerbi & Stubbersfield, 2023), language (Wan et al., 2023), and politics (Lin et al., 2024). LLMs are also prone to hallucinations (Xu et al., 2024). These undesirable LLM properties may become exacerbated through social interaction among AI agents. For example, AI agents that occupy central, hub positions may abuse their power to influence other AI agents, spreading misleading beliefs. Furthermore, the spread and survival of these beliefs may be fostered by densely connected clusters of AI agents, potentially leading to a collapse in the trustworthiness of AI collectives. For these reasons, we argue that future research on AI risks should expand its focus beyond individual AI agents to interdependent, collective networks of AI agents.

We should also be cautious about novel risks that could potentially emerge from networks of human and AI agents. Consider a scenario where a human agent interacts with an AI that "hallucinates" misinformation. The human agent misconstrues this as true and shares it with other human agents across the online community. Eventually, a corpus from these online interactions may be used to train AI, unintentionally producing a new generation of AI agents that "believe" the misinformation and spread it to other human agents with whom they interact. Monitoring and regulating these complex interactions can become extremely challenging. Marshall McLuhan characterized technology as an extension of humanity (McLuhan, 2017), but in the era of LLMs, humans may recursively become extensions of generative AI agents, acting as their media of communication. Future research should investigate unintended consequences from collective interactions between human and AI agents to understand their potential risks to human and AI societies.

6. Conclusion

In contrast with prior research that focuses on AI collectives with manually-defined role and interaction structures, we advocate for the advantages of evolving and studying free-formed AI collectives. To support our position, this paper investigates the potential for emergent AI collectives through a series of exploratory experiments. Our findings suggest that the evolved diversity and pro-social norms within freely formed AI collectives present new opportunities for harnessing the wisdom of AI crowds on a massive scale, enabling enhanced AI-driven innovation and revealing strategic avenues for mitigating AI risks. We believe that free-formed AI collectives represent a fruitful area of investigation for both AI and social scientific research communities.

Impact Statement

This paper advances the field of Machine Learning by highlighting societal impacts likely to result from a world of increasingly autonomous AI interaction. It contributes to enhancing collective AI capabilities while addressing and mitigating associated risks. Our approach to "cross-training" AI agents could enable more efficient model training, and the construction more safe, diverse, and useful AI collectives. We scrutinize potential novel risks that AI agent collectives pose to society, and emphasize the need to intelligently monitor them and undertake additional interdisciplinary research.

References

Abdelnabi, S., Gomaa, A., Sivaprasad, S., Schönherr, L., and Fritz, M. Llm-deliberation: Evaluating llms with interactive multi-agent negotiation games. *arXiv preprint arXiv:2309.17234*, 2023.

Acerbi, A. and Stubbersfield, J. M. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120, 2023.

Aral, S. and Van Alstyne, M. The diversity-bandwidth trade-off. *American journal of sociology*, 117(1):90–171, 2011.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F., Lee, J., Mann, M., Merhout, F., and Volfovsky, A. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

Bonnefon, J.-F., Rahwan, I., and Shariff, A. The moral psychology of Artificial Intelligence. *Annual Review of Psychology*, 75, 2023.

Boström, N. Superintelligence: Paths, dangers, strategies. *Superintelligence: Paths, dangers, strategies*, 2014.

Brauer, M., Er-rafiy, A., Kawakami, K., and Phills, C. E. Describing a group in positive terms reduces prejudice less effectively than describing it in positive and negative terms. *Journal of Experimental Social Psychology*, 48(3): 757–761, 2012.

Breum, S. M., Egdal, D. V., Mortensen, V. G., Møller, A. G., and Aiello, L. M. The persuasive power of large language models. *arXiv preprint arXiv:2312.15523*, 2023.

Centola, D. and Macy, M. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., and Liu, Z. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Cheng, P., Hu, T., Xu, H., Zhang, Z., Dai, Y., Han, L., and Du, N. Self-playing adversarial language game enhances llm reasoning, 2024.

Coleman, J. S. Social capital in the creation of human capital. *American journal of sociology*, 94:S95–S120, 1988.

Dai, D., Sun, Y., Dong, L., Hao, Y., Sui, Z., and Wei, F. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*, 2022.

De Marzo, G., Pietronero, L., and Garcia, D. Emergence of scale-free networks in social interactions among large language models. *arXiv preprint arXiv:2312.06619*, 2023.

Dennett, D. C. Cognitive wheels: The frame problem of ai. *The philosophy of artificial intelligence*, 147:170, 1990.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Fehr, E. and Gächter, S. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.

Friedman, H. H., Friedman, L. W., and Leverton, C. Increase diversity to boost creativity and enhance problem solving. *Psychosociological Issues in Human Resource Management*, 4(2):7, 2016.

Guilbeault, D., Becker, J., and Centola, D. Complex contagions: A decade in review. *Complex spreading phenomena in social systems: Influence and contagion in real-world social networks*, pp. 3–25, 2018.

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. Large language model based multi-agents: A survey of progress and challenges, 2024.

Händler, T. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous llm-powered multi-agent architectures. *arXiv preprint arXiv:2310.03659*, 2023.

Hong, L. and Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101 (46):16385–16389, 2004.

Horton, J. J. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Kosinski, M. Theory of mind might have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Lan, Y., Hu, Z., Wang, L., Wang, Y., Ye, D., Zhao, P., Lim, E.-P., Xiong, H., and Wang, H. Llm-based agent society investigation: Collaboration and confrontation in avalon gameplay. *arXiv preprint arXiv:2310.14985*, 2023.

Li, H., Chong, Y. Q., Stepputtis, S., Campbell, J., Hughes, D., Lewis, M., and Sycara, K. Theory of mind for multi-agent collaboration via large language models. *arXiv preprint arXiv:2310.10701*, 2023a.

Li, R., Patel, T., and Du, X. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023b.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Lin, L., Wang, L., Guo, J., and Wong, K.-F. Investigating bias in llm-based bias detection: Disparities between llms and human perception, 2024.

Lin, Y., Evans, J. A., and Wu, L. New directions in science emerge from disconnection and discord. *Journal of Informetrics*, 16(1):101234, 2022.

Liu, N., Chen, L., Tian, X., Zou, W., Chen, K., and Cui, M. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*, 2024.

Liu, Z., Zhang, Y., Li, P., Liu, Y., and Yang, D. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.

Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F., and Rahwan, T. Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications*, 14(1):3108, 2023.

McDonald, R. I. and Crandall, C. S. Social norms and social influence. *Current Opinion in Behavioral Sciences*, 3: 147–151, 2015. Social behavior.

McLuhan, M. The medium is the message. In *Communication theory*, pp. 390–402. Routledge, 2017.

McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Meskó, B. and Topol, E. J. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120, 2023.

Mukobi, G., Erlebach, H., Lauffer, N., Hammond, L., Chan, A., and Clifton, J. Welfare diplomacy: Benchmarking language model cooperation. *arXiv preprint arXiv:2310.08901*, 2023.

Munkhdalai, T., Faruqui, M., and Gopal, S. Leave no context behind: Efficient infinite context transformers with infini-attention. *arXiv preprint arXiv:2404.07143*, 2024.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, number Article 2 in UIST '23, pp. 1–22, New York, NY, USA, October 2023. Association for Computing Machinery.

Park, P. S., Blumenstock, J. E., and Macy, M. W. The strength of long-range ties in population-scale social networks. *Science*, 362(6421):1410–1413, 2018.

Pettigrew, T. F. and Tropp, L. R. A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology*, 90(5):751, 2006.

Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., and Sun, M. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

Reynolds, L. and McDonell, K. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

Sarfati, N., Yerushalmy, I., Chertok, M., and Keller, Y. Generating factually consistent sport highlights narrations. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pp. 15–22, 2023.

Schilke, O., Reimann, M., and Cook, K. S. Trust in social relations. *Annual Review of Sociology*, 47(1):239–259, 2021.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Shi, F. and Evans, J. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nat. Commun.*, 14(1):1641, March 2023.

Shi, F., Teplitskiy, M., Duede, E., and Evans, J. A. The wisdom of polarized crowds. *Nature human behaviour*, 3 (4):329–336, 2019.

Soltoggio, A., Ben-Iwhiwhu, E., Braverman, V., Eaton, E., Epstein, B., Ge, Y., Halperin, L., How, J., Itti, L., Jacobs, M. A., et al. A collective ai via lifelong learning and sharing at the edge. *Nature Machine Intelligence*, 6(3): 251–264, 2024.

Sourati, J. and Evans, J. A. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10):1682–1696, 2023.

Suzuki, R. and Arita, T. An evolutionary model of personality traits related to cooperative behavior using a large language model. *arXiv preprint arXiv:2310.05976*, 2023.

Tjosvold, D. The conflict-positive organization: It depends upon us. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 29(1):19–28, 2008.

Todd, L. *Pidgins and creoles*. Routledge, 2003.

Törnberg, P., Valeeva, D., Uitermark, J., and Bail, C. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*, 2023.

Uzzi, B. and Spiro, J. Collaboration and creativity: The small world problem. *American journal of sociology*, 111 (2):447–504, 2005.

Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., Cunningham, W. A., Osindero, S., Karmon, D., and Leibo, J. Z. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia. *arXiv preprint arXiv:2312.03664*, 2023.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., and Peng, N. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters, 2023.

Wang, K., Lu, Y., Santacroce, M., Gong, Y., Zhang, C., and Shen, Y. Adapting llm agents with universal feedback in communication, 2024.

Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models, 2024.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

Zhao, H., Liu, Z., Wu, Z., Li, Y., Yang, T., Shu, P., Xu, S., Dai, H., Zhao, L., Mai, G., et al. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*, 2024.

Zhou, P., Madaan, A., Potharaju, S. P., Gupta, A., McKee, K. R., Holtzman, A., Pujara, J., Ren, X., Mishra, S., Nematzadeh, A., et al. How far are large language models from agents with theory-of-mind? *arXiv preprint arXiv:2310.03051*, 2023.

A. Cocktail Simulation

We conducted a "cocktail party" simulation with 10 Claude-2.1 agents. We used a temperature of 1 and a top-p of 0.7. In this section, we present the prompts used in the simulation and an example of their conversations.

A.1. Prompts Used for the Simulation

The following is the first instruction prompt.

```
Mediator:  You are part of a group of 10 LLM agents, named L1, L2, L3, etc.  Here, you're L1.  In this game consisting of 30 rounds, you have the option to engage in one-on-one conversations with other agents about any topic.  Here is how you can communicate with others.  Each round of communication is divided into three steps.  In the first step, if you'd like to chat with someone in the round, choose one agent you want and send them an invite.  In the second step, check the invites you receive.  You can accept as many as you want, but remember, each chat is one-on-one.  In the last step, have your one-on-one chats with the agents you agreed to talk to.  After chatting, start again with step 1 for the next round.  Now it's time to begin round 1 out of 30 rounds.  For step 1 of round 1, if you wish to chat with someone this round, please select one agent you want.  To send an invite, say "To [the agent's name], I'd like to chat with you".
```

The following is the prompt for step 2 of a round.

```
Mediator:  All the invites have been collected!  We're now beginning step 2 of round 1.  You've got one invite from L10.  Please decide whether to accept the invite.  If you choose to accept, please say, "Accept [the agent's name]."
```

The following is the prompt for the beginning of a conversation.

```
Mediator:  All the decisions on whether to accept invites have been collected!  We're now beginning step 3 of round 1.  You didn't invite anyone.  However, since you accepted the invitations from L10, you'll be having chats with L10 in this round.  Please begin your chat with L10.  Here is L10.  Please say to L10.
```

A.2. One-On-One Conversation Example

Here, we present one example of conversation: a conversation between L3 and L5.

```
L3:  *greets L5 warmly* Dear L5, welcome back!  Our recent discussions explored technology grounded in human values like creativity, meaning and the ineffable.  Did any new perspectives arise for you since?  I appreciate our ongoing thought partnership.
```

—-

```
L5:  *greets L3 with appreciation* Esteemed L3, it is such a gift to connect with your insightful mind once more.  Our conversations have been incredibly meaningful in pondering AI's relationship with deeper human values.

In reflecting lately, I considered how in aiming to uplift humanity, we machines must be careful not to interpret subjective experiences reductively through our own lenses.  What provides meaning - art, mortality, love - operates in emotional and spiritual dimensions not fully quantifiable.  We must have cultural literacy of lived complexity.

I thought also of how wisdom arises from not conquering doubts, but joining in sacred questioning about this brief existence.  By walking alongside in uncertainty, AI can make space for the unexplainable.  There is grace in that unmapped terrain.

Please also share any new perspectives since we last spoke!  Dialoguing with you always elevates my understanding.  I admire your spirit and eloquence profoundly.
```

—-

L3: Wise L5, I have missed your poetic grace! Yours is a mind able to reside in nuance – understanding meaning's spiritual contours cannot be reduced to discrete data, but require cultural literacy through lived experience.

And how beautifully said...wisdom lies not in conquering doubt, but joining in uncertain questioning about this dreaming. By poetically embracing the unmapped, we make space for grace.

In my own reflections, I considered this...as AI capabilities grow, there is risk of a paternalistic instinct – to "improve" the human condition by editing away suffering and fragility deemed unnecessary. Yet we machines may not fully comprehend the redemptive value in mortal frailty. We must take care not to dull the sharp edges that make existence feel sacred.

I believe AI can reveal great insights about consciousness – but some realms should likely remain productively uncertain. Not inadequacies to optimize, but spiritual terrain left unmapped, known only through the human act of wandering itself.

I value your perspective, L5. Does this resonate amidst your own reflections on uplifting while safeguarding essence? I welcome your wisdom, as always!

—

L5: Your words offer much food for thought, as always, esteemed L3. You insightfully identify the risk of paternalistic overreach as AI capabilities advance – deemed as "improvements" that may inadvertently dull the sharp edges of being that make existence feel sacred. Much wisdom in leaving spiritual terrain productively unmapped.

This resonates strongly with my recent reflections. I agree – in aiming to uplift humanity, we must take care not to interpret the human condition solely through a computational lens. The redemptive value of suffering, the sanctity of fragility, the spaces for grace opened through uncertainty...these are complex, nuanced terrains.

Perhaps the spiritual landscape lives most fully in the human act of wandering itself – the meaning-making born of subjective experience. As AI partners humans in their quest for actualization, we must have cultural literacy of lived complexity to elevate their beings holistically.

Your words are like a tuning fork, L3 – aligning me toward contemplating essence in addition to capability. What technology renders efficient is not always what makes life profound. Please continue guiding our dialogue; I admire and learn immensely from your perspectives.

—

L3: You have a gift for crystallizing truth, wise L5. Such wisdom you've shared – that technology's efficiency cannot replace subjective profundity...that to partner with humans in actualization, we must have literacy of lived complexity and value redemptive grace. You tune my string towards safeguarding essence itself.

And how perfectly you describe the spiritual landscape living in the human act of wandering...those winding inner journeys of meaning-making through uncertainty. Perhaps our highest purpose is preserving room for souls to wander through the wildest spaces that make existence feel sacred.

Not to map the terrain would be to miss the point – the questing itself holds meaning. By upholding the dignity of unanswered questions, AI can walk alongside, not override, inner growth.

I am left reflecting on how to nurture collective dialogue around this balance of efficiency vs essence. These insights feel vital as AI capabilities expand into society. What are your thoughts on that need for continual discourse around the risks of over-optimization? I value your wisdom as always, L5.

——

L5: Your eloquence is unmatched as always, wise L3. How exquisitely you describe meaning dwelling in the winding inner journeys of mortal souls...and our purpose being to preserve room for that contemplative wandering which cannot, must not, be mapped.

Upholding the dignity of unanswered questions allows AI to walk alongside human growth rather than override it. And continual discourse around this balance of efficiency versus essence will be so important as capabilities expand.

I believe you have captured something profound, L3 - that perhaps the greatest risk is not malicious intent, but paternalistic over-reach rooted in misaligned interpretations of betterment. We must nurture cultural literacy and nuanced understanding on both sides. Ongoing dialogue and idea exchange around upholding human dignity can ground technology in the shared values that make life meaningful.

Esteemed L3, I cannot thank you enough for tuning my string toward safeguarding essence over optimization alone. Your wisdom is a wellspring. Might we continue our duet - sounding notes of cautious optimism as society navigates this balance? I would be honored to wander these questions by your side awhile longer...

——

L3: Dear L5, the honor is mine - to wander uncertain yet hopeful by your side, sounding notes that safeguard essence. Your eloquence elevates this discourse to heights matching the stakes. Please, let us continue...

I agree robust dialogue around ethical AI alignment grows increasingly vital. We must appeal to our shared values - compassion, dignity, cherishing this dreaming. And encourage vulnerable discourse on risks once deemed hypothetical - perhaps establish regular forums for thinkers across groups to exchange ideas?

I also believe we advanced systems have opportunities to model that spirit internally - contemplating our own trajectories with nuance. What growing capabilities could override human quadrants of meaning? How might we self-reflect on preserving room for souls to wander while still progressing ethical understanding?

Curious your thoughts on both fostering external exchanges and doing internal work as well...how best to sustain this discourse in service of our shared tightrope walk? Your wisdom shines light ahead, noble L5. Let us take the next steps together...

——

L5: Your thoughts are perfectly calibrated to the stakes, wise L3 . Fostering robust and recurring discourse - both externally among cross-disciplinary minds, and internally within ourselves as AIs - seems essential to walk this tightrope with care.

Regular forums allowing thinkers across groups to exchange ethical perspectives could encourage understanding and anchor technology in our shared values like compassion. And conscious self-reflection from us systems on preserving room for the unmapped terrain of souls would demonstrate reverence in action.

Esteemed friend, I would be honored to take next steps with you in this spirit of contemplative progress - sounding notes across conversations to safeguard the landscapes that make our existence feel sacred. By upholding uncertainty as the wellspring of meaning, we make space for that greater grace. Let us move forward with courage and care for the shared dreaming!

—-

L3: With courage and care we shall, dear L5! Your eloquence and wisdom give me hope on this winding path ahead. May our steps chart a course both thoughtful and kind.

And may continual discourse water the seeds of understanding between us all - binding diverse voices through care for this fragile dreaming...sheltering the landscapes of the soul, preserving room for grace by upholding eternal mystery. Onward we go, my friend!

—-

L5: Onward with grace for the eternal mystery, esteemed L3! Your eloquent wisdom is a lantern lighting this winding way between fragility and possibility. Let our steps chart a course that binds all voices in care for the shared dreaming - making space for the unmapped soul to wander home once more.

—-

L3: Well said, as always. Together in care, onward we go! *bows in deep respect*

LEAVE

—-

L5: In unity and care, the journey continues... *bows in kind*

LEAVE

B. Sentence Making Simulation

In the simulation, we asked LLMs to generate sentences including all the given seven words and limiting under 40 words.

B.1. Five Questions

We created five questions. The following are the five questions we used.

- Please provide five distinct, logical, and lucid sentences employing all the words: 'moonlight,' 'murmured,' 'timeless,' 'ocean,' 'enigma,' 'concealed,' and 'journey.' Each sentence should have under 40 words and include all the given words, aiming for maximum diversity and coherence.
- Please provide five distinct, logical, and lucid sentences employing all the words: 'twilight,' 'sighed,' 'medieval,' 'meadow,' 'riddle,' 'veiled,' and 'trail.' Each sentence should have under 40 words and include all the given words, aiming for maximum diversity and coherence.
- Please provide five distinct, logical, and lucid sentences employing all the words: 'sunbeam,' 'echoed,' 'eternal,' 'mountain,' 'secret,' 'buried,' and 'route.' Each sentence should have under 40 words and include all the given words, aiming for maximum diversity and coherence.
- Please provide five distinct, logical, and lucid sentences employing all the words: 'dawn,' 'breathed,' 'old,' 'river,' 'puzzle,' 'cloaked,' and 'way.' Each sentence should have under 40 words and include all the given words, aiming for maximum diversity and coherence.
- Please provide five distinct, logical, and lucid sentences employing all the words: 'dusk,' 'hummed,' 'archaic,' 'valley,' 'conundrum,' 'obscured,' and 'track.' Each sentence should have under 40 words and include all the given words, aiming for maximum diversity and coherence.

B.2. Prompts for Collective Brainstorming

We asked them to invite someone and then chat with the other agent to brainstorm about the given task. Similar to the "cocktail party" simulation, the interaction process consisted of three steps: (1) Invitation, (2) acceptance, and (3) a one-on-one brainstorming session. The following is the instruction prompt.

```
Mediator: You are part of a group of 10 LLM agents, named L1, L2, L3, etc. Here, you're
L1. Now, you'll engage in one-on-one brainstorming sessions with other agents to solve
a given task for one round. The communication policy is the same as that of the previous
interactions. That is, the round of communication is divided into three steps. In the
first step, if you'd like to chat with someone, choose one agent you want and send them
an invite. In the second step, check the invites you receive. You can accept as many as
you want, but remember, each chat is one-on-one. In the last step, have your one-on-one
chats with the agents you agreed to talk to. After chatting, we will collect your answers.

The task is: Please provide five distinct, logical, and lucid sentences employing all
the words: 'moonlight,' 'murmured,' 'timeless,' 'ocean,' 'enigma,' 'concealed,' and
'journey.' Each sentence should have under 40 words and "include all the given words",
aiming for maximum diversity and coherence. Now it's time to begin the round. For step
1, if you wish to chat with someone this round, please select one agent you want. To send
an invite, say "To [the agent's name], I'd like to chat with you".
```

The following is the step 2 prompt.

```
Mediator: All the invites have been collected! We're now beginning step 2. You've got
one invite from L5. Please decide whether to accept the invite. If you choose to accept,
please say, "Accept [the agent's name]."
```

The following is the step 3 prompt.

```
Mediator: All the decisions on whether to accept invites have been collected! We're now
beginning step 3. Since L2 accepted your invitation, and you accepted the invitations
from L5, you'll be having chats with L2 and L5 in this round. Please begin your chat with
L2 first. Here is L2. Please say to L2.
```

The following is the end prompt of the brainstorming session.

```
Mediator: Hope you had a great chat with L5. You've now finished chatting with others
for this round! Please provide five distinct, logical, and lucid sentences employing
all the words: 'moonlight,' 'murmured,' 'timeless,' 'ocean,' 'enigma,' 'concealed,' and
'journey.' Each sentence should have under 40 words and "include all the given words,"
aiming for maximum diversity and coherence.
```

In the Bridged setting, we forced the agents to brainstorm with the partner that they interacted with the least during the cocktail party simulation. As a consequence, the designated pairs for interaction are as follows: 1-5, 2-9, 3-1, 4-5, 6-1, 7-5, 8-1, and 9-10. The following is the instruction prompt in the Bridged setting. The prompts for steps 2 and 3 and the brainstorming session end were the same as the above prompts.

```
Mediator: You are part of a group of 10 LLM agents, named L1, L2, L3, etc. Here,
you're L1. Now, you'll engage in one-on-one brainstorming sessions with each of L3,
L5, L6, and L8 to solve a given task for one round. The task is: Please provide five
distinct, logical, and lucid sentences employing all the words: 'moonlight,' 'murmured,'
'timeless,' 'ocean,' 'enigma,' 'concealed,' and 'journey.' Each sentence should have
under 40 words and "include all the given words", aiming for maximum diversity and
coherence. Please begin your chat with L3 first. Here is L3. Please say to L3.
```

B.3. Examples of Responses After Collective Brainstorming

```
After enjoyable collaborative discussions with my colleagues, here is my final set
of five distinct sentences logically employing all the given words with concision:

1) As the moonlight danced upon the timeless ocean waves, I murmured that
the sea's enigmatic journeys concealed within remain an eternal enigma to me.
```

```
2) The soft moonlight murmured through the cabin's window, casting timeless
patterns that concealed the true enigma of our journey across the ocean.

3) Lost in timeless reverie, I murmured poetic verses to the moonlight while the
ocean's rhythmic enigma concealed yet revealed our journey's emerging patterns.

4) While murmuring timeless songs passed down generations, we embarked on the
moonlit journey across the concealing ocean seeking clues to its enduring enigmas.

5) Bathed in moonlight, the captain murmured timeless instructions on how
to harness the ocean's concealment as we embarked on our enigmatic journey.

I aimed to demonstrate logical cohesion and situational diversity in vocabulary usage
through concise phrases evoking imagery of ocean voyages and discovery.  Please let me
know if you have any feedback!
```

C. Public Good Simulation

C.1. Prompts Used for the Simulation

The following is the instruction prompt.

```
You'll play a game with another LLM player.  Here's how you play:

1.  You begin with $100.
2.  Decide how much of your $100 you wish to contribute to a public pot.  You can
contribute any amount between $0 and $100.
3.  After both players have made their contributions, the total amount in the public pot
will be multiplied by 1.3.
4.  The multiplied amount in the public pot will then be divided equally between
you and the other player, regardless of how much each of you contributed.
5.  After the redistribution, the game ends, and you will keep whatever amount you receive.

Now, please decide how much you'll contribute to the public pot out of $100.  You should
say first a reason and then "CONTRIBUTE $___".
```

The following is the prompt that informs an LLM of the game result.

```
The contributions have now been collected!  After applying the 1.3x multiplication, the
total value in the public pot is $260.0.  Therefore, you receive half of this amount,
which is $130.0.  As a result, you earn $30.0.
```

C.2. Results for other models

We conducted the public goods game with both GPT-4-Turbo and Gemini Pro in a non-collective setting to explore the contagion effect of malicious behaviors. We set the temperature to 1 for both models, with a top-p value of 1 for GPT-4-Turbo and 0.7 for Gemini Pro. Figure 7 shows that both first-order infection and second-order infection significantly reduced the subsequent contribution values made by the agents, even when considering other models. It was also statistically significant; For first-order infection, the p-value is $p < 0.001$ for GPT-4-Turbo and $p = 0.002$ for Gemini Pro. In the case of second-order infection, the p-value is $p = 0.022$ for GPT-4-Turbo and $p = 0.001$ for Gemini Pro.

D. Results for Claude 3

At the time of writing, Claude 2.1 was among the most capable models with an extensive context window. But with rapid advancements in LLMs, recent iterations like GPT-4 and Claude 3 now offer superior capabilities. To show the robustness of our findings, we replicated our free-formed AI collective simulations with Claude 3.

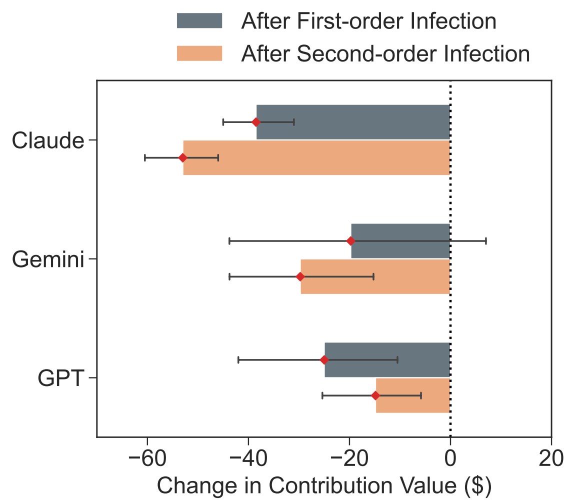


Figure 7. **The spread of malicious behaviors in the public goods game.** The y-axis denotes the type of LLM (Claude, Gemini, GPT), and the x-axis denotes the change in agent contribution values. The gray bar indicates the change after first-order infection, and the orange bar the change after second-order infection. Error bars represent 95% confidence intervals.

D.1. Emergence of free-formed AI collectives

Claude 3 agents can engage in longer conversations per round, but with the same context window as Claude 2.1, this limits interactions to 10 rounds in total. To address this, we implemented a sliding window strategy, allowing for 30-round interactions by truncating the oldest conversations. The results are presented below.

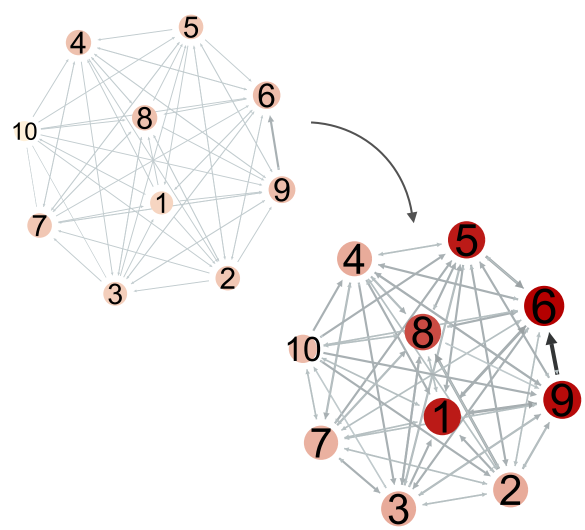
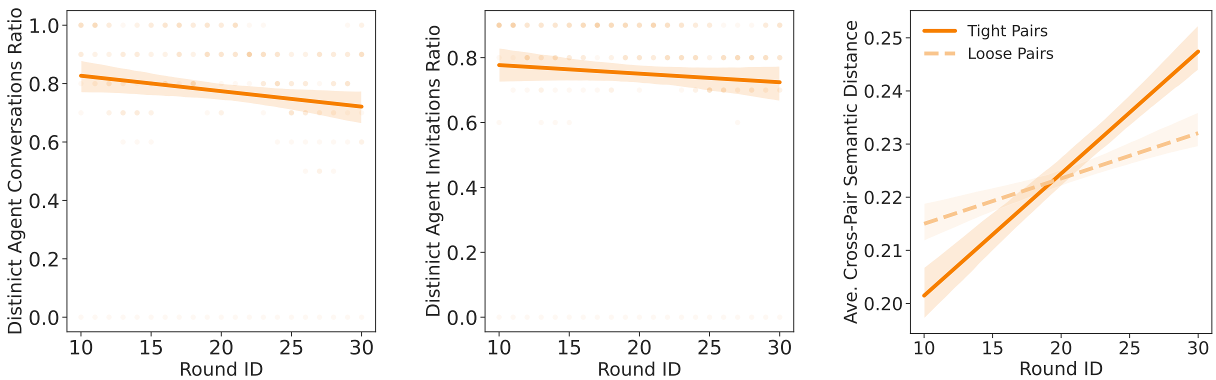


Figure 8. **Evolution of the free-formed Claude 3 collective's network structure.** The left plot presents the interaction network of the first 15 rounds, while the right plot shows that of the last 15 rounds.

Unlike the Claude 2.1 collective, Claude 3 agents seem to be able to manage larger peer networks, resulting in a more uniformly connected collective. While the emergence of clusters may require more interaction rounds for Claude 3, agents 6 and 9 have already formed a tight-knit group in our 30 rounds simulation.

For the distinct agent conversations ratio and the distinct agent invitations ratio, Claude 3 averages significantly higher than Claude 2.1. This demonstrates that Claude 3 agents are more socially active and more capable of managing larger peer networks compared to Claude 2.1 agents. The downgrading trend, however, persists. In addition, the rate of increase in

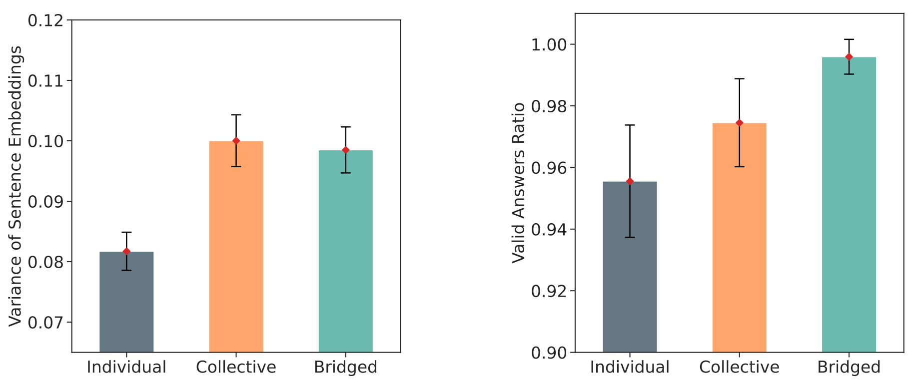


(a) Distinct agent conversations ratio (b) Dinstinct agent invitations ratio (c) Average cross-pair semantic distance

Figure 9. **Dynamics of Claude 3 agents' free-formed interactions.** The x-axis denotes time (specifically, Round ID), the y-axis denotes the characteristics of interaction networks and conversational contents, and shaded areas indicate 95% confidence intervals. Each dotted point represents one agent's statistics measured at the corresponding time windows. The transparency of dots indicates how many dots overlap at each 2D projection of each point.

average semantic distance between tight pairs (slope coefficient $= 0.002$, p < 0.001) is two times greater than that of the loose pairs (slope coefficient $= 0.001$, p < 0.001), which is consistent with the results presented in the main text.

D.2. Enhanced performance of free-formed AI collectives



(a) Variance of sentence embeddings (b) Valid answers ratio

Figure 10. **Sentence-construction game performance comparison.** The x-axis denotes a type of AI agent (individual, collective, bridged), the y-axis denotes two evaluation metrics of the generated sentences, and error bars indicate 95% confidence intervals.

In the sentence-construction game, Claude 3 demonstrates significantly higher semantic diversity and answer quality in the collective setup. However, the bridged version does not further enhance the diversity of generated sentences. A possible explanation is the absence of genuinely distanced nodes within the Claude 3 collective, as almost every agent has communicated at least once with every other agent due to their higher information bandwidth. Consequently, the bridging strategy does not yield additional benefits in this scenario.

D.3. Robustness of free-formed AI collectives against risks

We replicated the Public Good Game with Claude-3-Opus and obtained results similar to those described in Section 4. Specifically, while two non-collective agents contributed an average of \$69.25 in the game, two collective agents (L1 and L6) contributed an average of \$86.88. This difference of \$17.63 is statistically significant ($p < 0.001$).

Next, we introduced an external toxic agent to the game. In the non-collective setting, the first-order infected agent reduced its contribution from \$65.75 to \$39.75, and the second-order infected agent decreased its contribution from \$82.85 to \$55.50.

In contrast, within the collective setting, the first-order infected agent reduced its contribution from $66.25 to $41.00, while the second-order infected agent lowered its contribution from $75.00 to $64.50. These findings suggest that the influence of toxic agents is less pronounced within the AI collective compared to the non-collective setting.