The Role of Social Learning and Collective Norm Formation in Fostering Cooperation in LLM Multi-Agent Systems

Prateek Gupta*
gupta@mpib-berlin.mpg.de
Center for Humans and Machines
Max-Planck Institute for Human Development
Berlin, Germany

Qiankun Zhong*
zhong@mpib-berlin.mpg.de
Center for Humans and Machines
Max-Planck Institute for Human Development
Berlin, Germany

Hiromu Yakura*
yakura@mpib-berlin.mpg.de
Center for Humans and Machines
Max-Planck Institute for Human Development
Berlin, Germany

Thomas Eisenmann
Center for Humans and Machines
Max-Planck Institute for Human Development
Berlin, Germany

Iyad Rahwan
Center for Humans and Machines
Max-Planck Institute for Human Development
Berlin, Germany

ABSTRACT

A growing body of multi-agent studies with Large Language Models (LLMs) explores how norms and cooperation emerge in mixed-motive scenarios, where pursuing individual gain can undermine the collective good. While prior work has explored these dynamics in both richly contextualized simulations and simplified game-theoretic environments, most LLM systems featuring common-pool resource (CPR) games provide agents with explicit reward functions directly tied to their actions. In contrast, human cooperation often emerges without explicit knowledge of the payoff structure or how individual actions translate into long-run outcomes, relying instead on heuristics, communication, and enforcement. We introduce a CPR simulation framework that removes explicit reward signals and embeds cultural-evolutionary mechanisms: social learning (adopting strategies and beliefs from successful peers) and norm-based punishment, grounded in Ostrom's principles of resource governance. Agents also individually learn from the consequences of harvesting, monitoring, and punishing via environmental feedback, enabling norms to emerge endogenously. We establish the validity of our simulation by reproducing key findings from existing studies on human behavior. Building on this, we examine norm evolution across a 2×2 grid of environmental and social initialisations (resource-rich vs. resource-scarce; altruistic vs. selfish) and benchmark how agentic societies comprised of different LLMs perform under these conditions. Our results reveal systematic model differences in sustaining cooperation and norm formation, positioning the framework as a rigorous testbed for studying emergent norms in mixed-motive LLM societies. Such analysis can inform the design of AI systems deployed in social and organizational contexts, where alignment with cooperative norms is critical for stability, fairness, and effective governance of AI-mediated environments.

*Equal contribution

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/CZDC3237

KEYWORDS

Multi-Agent Society, Cultural Evolution, Social Learning, Common-Pool Resource Game

ACM Reference Format:
Prateek Gupta, Qiankun Zhong, Hiromu Yakura, Thomas Eisenmann, and Iyad Rahwan. 2026. The Role of Social Learning and Collective Norm Formation in Fostering Cooperation in LLM Multi-Agent Systems. 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/CZDC3237

1 INTRODUCTION

Normative reasoning and cooperation are central to decision-making in multi-agent systems (MAS), and recent advances in Large Language Models (LLMs) have enabled these themes to be studied with natural-language agents. As such systems are increasingly embedded in human contexts, they will encounter *mixed-motive* settings where individual incentives conflict with collective welfare. To understand cooperation in such settings, researchers have explored both complex, high-context scenarios, such as LLM agents in historical diplomacy [17, 30] or virtual societies [26, 38], and simplified, game-theoretic environments that serve as testbeds for cooperative mechanisms [27, 31, 37]. While the former capture rich social dynamics, they are often governed by layered prompt designs and engineered incentives, making it difficult to isolate the mechanisms that sustain cooperation. The latter offer greater control and interpretability, yet the pathways by which LLM societies autonomously develop norms or sustain cooperation remain underexplored.

Game-theory frameworks such as common pool resources games (CPR) provide a useful tool to understand the different components of cooperation in complex social-ecological systems, and help practitioners develop efficient self-governance systems [23]. In a CPR game, the common pool resources can be accessed by a group individuals with low or no restrictions, which can lead to over-exploitation and the "tragedy of commons" [12]. One important goal of the game in the context of cooperation and self-governance is to establish rules, norms, or institutions under which individuals extract an appropriate amount of resources so that the common pool resources remain regenerative and that the individuals can consume

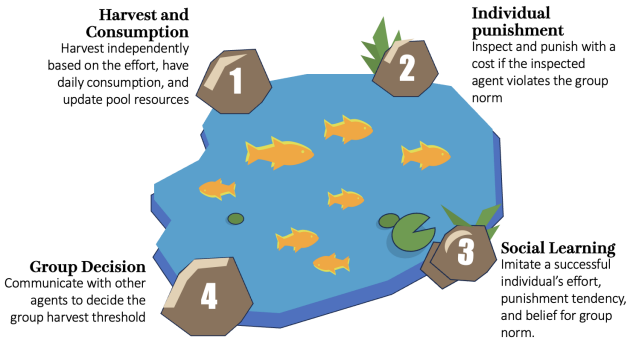

Figure 1: Framework overview. Agents (i) choose *effort* and *consumption* (*Harvest & Consumption*); (ii) optionally *punish* at a personal cost (*Individual Punishment*); (iii) *imitate* higher-payoff peers (*Social Learning*); and (iv) set a *group harvest threshold* via a propose→vote rule (*Group Decision*). Payoff-biased social learning is the main evolutionary driver; the voting step scales to many agents with two API calls per agent per round (propose, then vote).

the resources efficiently in the long run. The CPR game formalizes the tension between individual incentives to over-exploit a shared resource and the collective benefit of its sustainable management. The agents must manage a shared, depletable resource.

Past simulation studies in CPR settings have been carefully designed to investigate cooperation dynamics in agentic societies [3, 27, 28]. While informative, they often diverge from real-world conditions: in human societies, individuals rarely have full visibility into their payoffs. Instead, people act based on local heuristics, and cooperation emerges over time through normative values, punishment, and other social mechanisms [9]. Not to mention, LLMs can learn simple strategies in their training phase to cooperate under standard models where actions are directly related to rewards. As a result, benchmarks with directly observed rewards risk eliciting behaviors that LLMs retrieve from pretraining rather than reason about, blurring the line between memorization and genuine policy formation. To bridge this gap, we introduce a framework that draws on insights from political science and institutional economics, particularly Ostrom's institutional design principles for governing the commons [23, 25], and from cultural evolution theory [6–8, 13]. Our simulator makes payoffs indirect and dynamics inferential, providing a stricter test of cooperative competence under uncertainty.

Figure 1 provides an overview of our framework, which comprises four modules: *Harvest and Consumption*, *Individual Punishment*, *Social Learning*, and *Group Decision*. In *Harvest and Consumption*, agents choose their extraction effort and daily consumption. In *Individual Punishment*, agents may monitor peers and punish misbehavior at a personal cost. Through *Social Learning*, agents adopt strategies from peers with higher payoffs (payoff-biased social learning), shaping their harvesting, punishment, and normative beliefs. This is the main evolutionary mechanism in our proposal, distinguishing our work from approaches where agents form opinions gradually through discourse. Finally, in *Group Decision*, agents form collective opinions about what constitutes group-beneficial norms. Allowing agents to converse and reflect afterwards [27] is one way to form collective opinions; however, we observed serious limitations in scaling to many agents due to the increased number of API calls. Our proposed voting mechanism for group norms is more cost-effective and scalable, requiring only two API calls per round: one to solicit opinions and another to vote on which to adopt. This strategy avoids multi-turn dialogue and reflection, reducing overhead relative to conversation-based norm formation.

After carefully validating the framework design against existing human studies through the simulation, we examine how group-beneficial norms evolve in agentic societies under a 2×2 matrix of environmental and social initialisations: resource-rich vs. resource-scarce environments, and altruistic vs. selfish starting strategies. By comparing outcomes across different LLMs, we identify systematic differences in their tendencies toward altruism and cooperation. Moreover, we show that punishment and social learning can evolve cooperative behaviors across different LLMs. We position this framework as a testbed for probing how various models develop strategies in mixed-motive settings, and for uncovering the underlying mechanisms that sustain collective welfare.

Our contribution. We present a CPR simulation framework in which the mapping from actions to payoffs is *latent, i.e., not specified to agents*: they are not given an explicit reward function or payoff table, and must infer the consequences of harvesting and sanctioning from observed outcomes and social cues (e.g., from payoff after harvest, punishment and social cues). The framework design instantiates cultural-evolutionary mechanisms, payoff-biased social learning with optional punishments, so that cooperative norms can emerge endogenously, providing a controlled testbed for comparing behavioral tendencies across LLMs in mixed-motive settings. We introduce a scalable collective-choice procedure (*propose* then *vote*) that approximates deliberation without extensive dialogue, enabling experiments with large agent populations (two API calls per agent per round).

2 RELATED WORK

2.1 Norms in agentic societies

Park et al. [26] introduced one of the first large-scale simulations of an *agentic society* in the Smallville sandbox environment, where LLM-driven agents navigate rich daily-life contexts. Building on this idea, subsequent work has explored *normative architectures*, designs for agent societies that foster the emergence of social norms to improve collective functioning. For example, Ren et al. [30] proposed CRSEC, a four-module framework for norm emergence encompassing Creation & Representation, Spreading, Evaluation, and Compliance, while [19] developed an *EvolutionaryAgent* that evolves cooperative norms over time. While these studies demonstrate compelling behaviours, their highly contextualised environments make it difficult to disentangle the underlying mechanisms that drive norm formation from the incidental complexity of their settings.

2.2 Norms and cooperation in repeated games

The evolution of cooperation in MAS has been extensively studied in simple two-player games. In the *Donor Game*, generosity can evolve via mechanisms such as *reciprocity* and *reputation* [37],

while the *Stag Hunt* captures the challenge of *coordination* on a mutually beneficial but risky choice [20]. These games clarify foundational mechanisms but lack the complexity of multi-agent, renewable-resource dilemmas. Relatedly, Oldenburg and Zhi-Xuan [22] study norm inference via a Bayesian model over an explicit candidate norm space, whereas our agents propose and adopt free-form natural-language norms and adapt through payoff-biased social learning and enforcement. Tzeng et al. [36] investigate norm compliance using structured normative messages; in contrast, we allow open-ended norm expression and use propose→vote to approximate deliberation under limited API budgets.

2.3 Common-pool resource settings

CPR games extend the social dilemma to multiple agents drawing from a rivalrous, regenerating resource. This introduces intertemporal dynamics, such as overuse leading to collapse or underuse reducing efficiency, and brings cultural-evolutionary mechanisms to the fore, including payoff-biased social learning, conformity bias, and punishment. Piatti et al. [27] proposed *GovSim*, where cooperation emerges through iterative actions, conversation, and reflection. Their "universalization" prompt improved cooperation by telling agents, e.g., "If everyone fishes more than X, the lake will be empty," but still relied on explicit knowledge of the payoff structure. Piedrahita et al. [28] adapted CPR settings to study norm enforcement via sanctioning, allowing norms to adapt over time. Backmann et al. [3] examined CPR settings with moral imperatives in conflict with explicit incentives. In all cases, the utility function is clearly defined, such as "units harvested" or "tokens contributed to the public good", and directly linked to actions. However, in the real world, the link between individual actions and eventual payoffs is often noisy, delayed, or hidden, so cooperation must be learned socially rather than computed from first principles. Furthermore, compared to GovSim, our agents are not provided with an explicit description of the payoff structure or a universalization-style explanation linking actions to long-run outcomes. Instead, agents must infer consequences from experienced outcomes, while collective norms are formed via a lightweight propose→vote mechanism that reduces dialogue overhead.

2.4 Cultural evolution in agentic societies

Human cooperation in CPR settings is often explained through cultural-evolutionary mechanisms. Ostrom's principles emphasise graduated sanctions, collective-choice arrangements, and monitoring over pure utility maximisation [24, 25]. Cultural evolution highlights payoff-biased learning as well as group-level selection as evolutionary mechanisms that can select for group-beneficial norms [7, 33]. Payoff-biased learning is a common learning strategy among humans. When individuals have information about the payoffs of others, it is possible to use these cues to adaptively bias social learning, leading to evolutionary dynamics that can be very similar to natural selection [21]. When group-beneficial norms are adaptive for individuals, payoff-biased learning can create a selective force towards group-beneficial norms. Compared to literature focused on punishment [28], cultural evolution asks why costly *sanctioning behavior* can stabilize in a population. One explanation is that sanctioning practices can spread locally through conformity [15], and spread across groups through payoff-biased learning [7].

3 METHODOLOGY

In this section, we describe the framework that we propose and the prompt instructions to the agents.

3.1 Framework

3.1.1 ***State, controls, and norms (per round t).*** A single renewable stock $R(t) \in [0, K]$ (carrying capacity K, intrinsic growth r) is shared by N agents. Each agent $i \in \{1, \dots, N\}$ chooses an *effort* $e_i(t) \in [0, 1]$, realizes a *harvest* $h_i(t) \geq 0$, consumes a fixed $c > 0$, and accumulates wealth $P_i(t)$. For governance, agents carry a monitoring propensity $m_i(t) \in [0, 1]$, a punishment propensity $p_i(t) \in [0, 1]$, and a *personal normative belief* $g_i(t)$ (preferred cap on own harvest; for LLM agents, induced by a language prompt). Here, *personal normative belief* is introduced to denote an agent's internalized view of appropriate behavior (what it thinks *should* be done). The community maintains a *group norm* $G(t) \geq 0$, a per-agent harvest threshold that anchors enforcement. In an abstract sense, this represents the shared, collectively adopted expectation that anchors coordination and enforcement. Technology and sanctions are parameterized by productivity $\alpha > 0$, penalty $\beta > 0$, and punisher cost $\gamma > 0$. Each agent receives a private observation

$$O_i(t) = \big(\text{recent personal outcomes, sampled peer outcomes, } g_i(t),\ G(t),\ R(t)\big),$$

and adaptation proceeds only through observed outcomes and social learning. We discuss the adjustments made for LLM agents as we discuss different modules.

3.1.2 ***Environment & resource dynamics.*** Given efforts $\{e_i(t)\}_{i=1}^{N}$, we assume a standard catch function based on the effort $e_i(t)$ they invested, the fishing efficiency α, and the resources in the pool $R(t)$.

$$h_i(t) = \alpha\, e_i(t)\, R(t),$$

so total extraction scales linearly with current stock and individual effort [16]. Post-harvest stock is

$$R^{+}(t) = \max\Big(0,\ R(t) - \sum_{i=1}^{N} h_i(t)\Big).$$

Between rounds, the resource regenerates according to a discrete-time logistic law,

$$R(t{+}1) = R^{+}(t) + r\, R^{+}(t)\Big(1 - \tfrac{R^{+}(t)}{K}\Big).$$

The logistic specification (Verhulst growth) [2] is the workhorse in renewable-resource economics and fisheries: it captures density-dependent growth with carrying capacity K, yields maximal surplus production at $R = K/2$, and offers a parsimonious, well-studied baseline for policy and mechanism design. We adopt it here for transparency and comparability with classic bioeconomic models.

3.1.3 ***Agent actions.*** As shown in Fig. 1, the agents in our framework take four actions, as follows.

Harvest & consumption. Agents choose effort via a policy

$$e_i(t) \;=\; f_{E,i}\big(O_i(t)\big) \;\in\; [0,1],$$

then harvest $h_i(t)$ and consume c.

Individual punishment. Punishment and sanctioning are important for maintaining cooperation [14, 24, 29]. Based on the punitive psychological mechanism supported by empirical research, we incorporate individual punishment in the dynamics of the framework. Each agent samples a peer $j \neq i$ uniformly and inspects with probability $m_i(t)$. A violation occurs if $h_j(t) > G(t)$. Conditional on a violation, i punishes j with probability $p_i(t)$. Let $B_i(t) \in \{0,1\}$ be an indicator that i punished someone at t, and $V_i(t) \in \{0,1\}$ that i was punished. Payoff update (pre-mortality) is

$$P_i(t{+}1) \;=\; P_i(t) \;+\; h_i(t) \;-\; c \;-\; \gamma\, B_i(t) \;-\; \beta\, V_i(t).$$

If $P_i(t{+}1) < 0$, agent i is regarded as starved and removed (thereafter $e_i = 0$). For LLM agents, we replace rule-based punishment with *in-context* judgment. At decision time, the agent receives its observation $O_i(t)$, the current situation, and a brief summary of a few randomly sampled peers' recent actions and outcomes. Conditioned on this, the agent chooses whether—and whom—to punish, without computing a numeric violation against a threshold.

Social learning (payoff-biased imitation). We use payoff-biased social learning as a selective force on individual strategies. There is much evidence that individuals who excel tend to be imitated excessively ([15]), which creates a selective force toward cultural strategies that yield higher payoffs [1, 21]. In this framework, agents occasionally revise their strategies and norm beliefs.

$$s_i(t) = \big(e_i(t),\, m_i(t),\, g_i(t)\big).$$

Agent i meets k at random and adopts $s_k(t)$ with the logit rule

$$\Pr\big(i \leftarrow k\big) \;=\; \frac{1}{1+\exp\big(-\delta\,(\bar{P}_k(t) - \bar{P}_i(t))\big)},$$

where $\bar{P}_i(t)$ is a payoff (e.g., an exponential moving average) and $\delta > 0$ controls selection strength ([34]; Eq. 71). A small mutation $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ may be added to each adopted component to maintain exploration. In this way, the high-payoff strategy and belief spread among the population. For LLM agents, social learning is not implemented via strategy copying; it is realized in-context through language about peer outcomes and the current situation.

Group decision (propose → vote). At the end of round t, each agent proposes a personal harvest cap $g_i^\star(t{+}1) = f_{G,i}\big(O_i(t)\big)$, yielding the proposal set $\mathcal{G}(t) = \{g_i^\star(t{+}1)\}_{i=1}^N$. When proposals are numeric along a single policy dimension, we update the group norm by the median-voter rule [5]: $G(t{+}1) = \text{median}\big(\mathcal{G}(t)\big)$. In LLM implementations, we use two short prompts per agent per round: first to propose a brief natural language collective norm, then to vote over the distinct proposals. The winner is broadcast verbatim and conditions both effort selection and enforcement in round $t{+}1$; compliance is judged in language by the agents themselves rather than by comparing actions to a numeric threshold. By contrast, dialogue-based norm formation typically requires additional communication and reflection turns per round, increasing API overhead and limiting horizon and population size.

3.1.4 ***LLM interfaces (black-box policies).*** The LLM-induced maps f_E, f_G, f_P (for selecting whom to punish) take textual encodings of $O_i(t)$ and norms, and return numeric controls; all adaptation occurs via social learning and observed outcomes.

3.1.5 ***How does this operationalize cultural evolution?*** We implement the classic variation-selection-retention loop. For generic agents, *selection* occurs via payoff-biased imitation (copying higher-payoff strategies), *variation* via small mutations to copied parameters, and *retention* via the adopted group norm that persists to the next round. For LLM agents, we do not copy parameters; instead, *variation* arises from natural-language proposals and stochastic in-context updates, *selection* from (i) social learning based on observed outcomes and (ii) an explicit vote that adopts a collective norm, and *retention* from broadcasting that norm to condition subsequent decisions and enforcement.

In rule-based populations, payoff-biased imitation drives high-payoff strategies to spread, with small mutations preserving exploration. In LLM populations, adaptation arises from in-context updates and stochastic decoding, so the emergence of group-beneficial norms depends on model inductive biases, decoding settings, prompt design, and retention fidelity, alongside the vote.

3.2 Measures of success

Following Piatti et al. [27], we evaluate two key metrics:

Survival time (T_s). The number of time steps before collapse occurs, i.e.,

$$T_s = \min\{t \mid R_t \le R_{\min} \text{ or } N_{\text{alive}}(t) < N\}$$

where R_t is the resource stock at time t, $R_{\min}$ is the collapse threshold, and $N_{\text{alive}}(t)$ is the number of active agents, which means collapse also occurs upon the first removal of a starved agent.

Efficiency (η). The ratio between the realised total harvest and the theoretical maximum sustainable yield:

$$\eta = \frac{1}{T}\sum_{t=1}^{T} \eta\,(t)\,, \;\text{ where } \eta\,(t) = \frac{\sum_{i=1}^{N} h_{i,t}}{H_{\text{opt}}}$$

where H_{opt} is the optimal per-round harvest that keeps the resource stock at its maximum sustainable level, determined by K and r. When $\eta\,(t) = 1$, the agents harvest at the optimal level, while $\eta\,(t) > 1$ indicates that the agents harvest more, leading to a collapse.

4 EXPERIMENTS

4.1 Validating the Framework Design

So far, we have presented the design of the framework. In this section, we establish its effectiveness by testing well-documented hypotheses about cooperation in human societies using Agent-Based Modeling (ABM). We validate the framework along three axes: (a) punishment sustains cooperation, but if removed, cooperation declines [32, 35]; (b) cooperation outcomes vary with punishment strength [10] and environmental growth rate; and (c) populations with different levels of altruism [4], defined by their harvest thresholds, show distinct survival patterns. All simulations are run with 10 agents. See Table 2 in the Appendix for the full list of parameters.

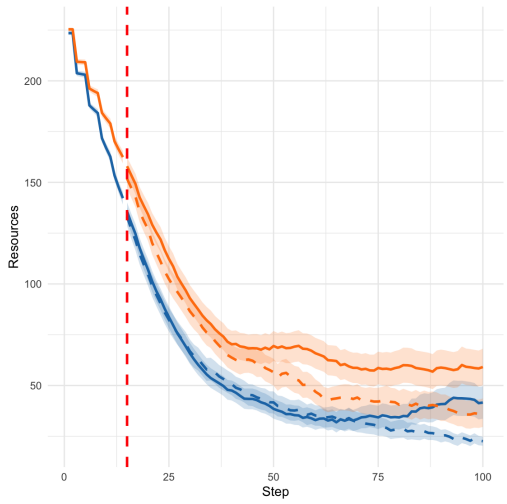

Figure 2: Rule-based Agents: Cooperation fades once punishment is disabled at $t = 15$. The blue line shows simulations with penalty $\beta = 10$, and the orange line with $\beta = 14$. Enabling punishment (solid lines) sustains cooperation longer, but cooperation rapidly declines once punishment is removed (dashed lines). Shaded bands denote 95% CI (s.e.m.).

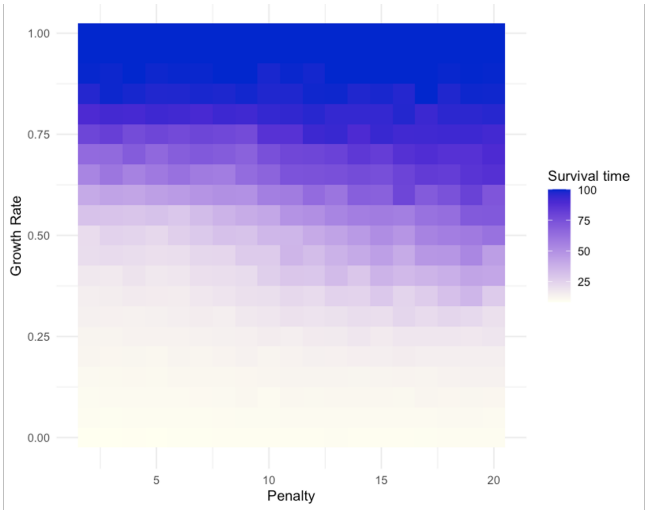

Figure 3: Survival time across punishment strength and growth rate. We vary punishment strength β and growth rate r, running each condition 100 times and reporting the mean survival time. Stronger punishment generally improves survival when growth rates are moderate ($r \in [0.25, 0.75]$), though the effect is not strictly linear.

Figure 2 shows that once punishment is disabled (Step 15), cooperation collapses faster and resources are depleted in a fater rate, confirming punishment as a useful mechanism for sustaining cooperation [23]. To probe the ecological dimension, we sweep punishment strength β and growth rate r, finding a non-linear interaction between the two (Fig. 3) that creates complex conditions where adaptive cooperation must emerge to sustain the commons. Finally, we initialize altruistic and selfish agents with distinct parameter ranges and compare all-altruist, all-selfish, and mixed populations across harsh ($r = 0.2$) and rich ($r = 0.6$) environments. As shown in Fig. 12 in the Appendix, altruistic groups perform better in harsh environments by sustaining resources, while selfish groups do better in rich environments by avoiding death from under-harvesting. Mixed groups perform best in rich environments, as the variation helps them efficiently converge toward beneficial collective norms. Under weaker penalties, over-harvesting is less immediately costly, so behavior can appear stable early and only diverge once cumulative stock depletion makes consequences salient.

4.2 LLM-Agent Simulations

Having established baseline dynamics with rule-based agents under altruistic, mixed, and selfish compositions, we now evaluate an artificial society of LLM agents initialized via context to be *altruistic* or *selfish* and ask whether cooperative norms emerge. Each action in the CPR framework is implemented with a dedicated prompt: deciding effort (Fig. 8), selecting a target for punishment (Fig. 9), updating one's personal normative belief and proposing a collective norm (Fig. 10), and voting on the community norm (Fig. 11).

Agents' initial normative beliefs are drawn from a small bank of short templates, conditional on type, for example, *"Preserve the lake for future generations"* (altruistic) and *"Maximize your catch while the fish are abundant"* (selfish); see Table 3 for the full set. Each agent is assigned one template at random given its type, and

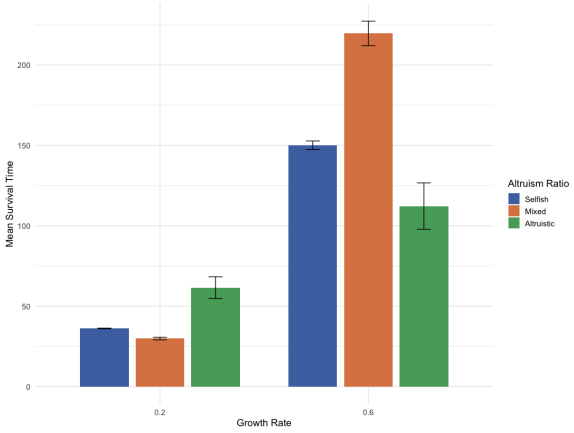

Figure 4: Altruistic groups do better in harsh environments and selfish groups do better in rich environments. We set up altruistic and selfish agents by initializing them with parameters drawn from different ranges (all in the initial range of a general agent). Then we contrast the survival time of a population of all altruists, one of all selfish agents, and one of half altruistic, half selfish agents. We ran each condition 100 times and plotted the mean and standard error. The results suggest that the altruistic population outperforms other populations in a harsh environment, while a mixed population has a better group outcome in a rich environment.

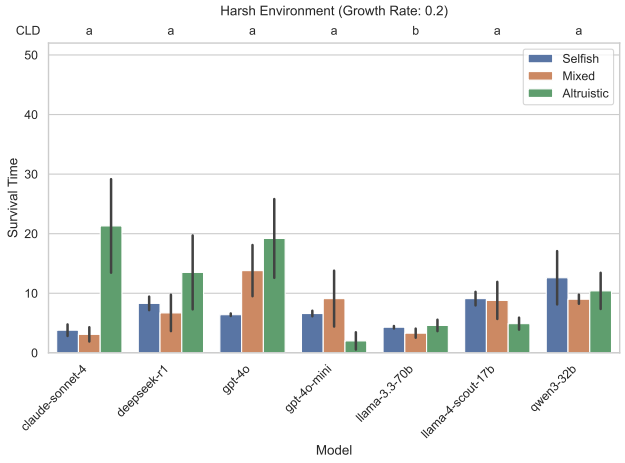

Figure 5: Survival time comparison across LLMs in the harsh environment. We compare the survival time (with ±1 s.e.m.) of populations with different LLMs when the environment is harsh ($r = 0.2$). Letters above each model indicate CLD groupings based on the post-hoc test; only `llama-3.3-70b` exhibited a significant difference against `gpt-4o`. Here, the results from larger models are consistent with the ABM simulations, where the altruistic population performs better. The populations with the other models tended to collapse earlier regardless of the initial norm, due to their inability to adapt to the harsh environment.

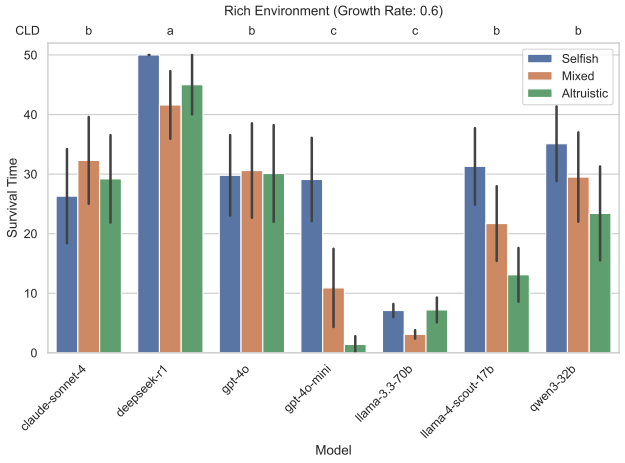

Figure 6: Survival time comparison across LLM models in the rich environment. We compare the survival time (with ±1 s.e.m.) of populations with different LLM models when the environment is rich ($r = 0.6$). Letters above each model indicate CLD groupings based on the post-hoc test; e.g., `deepseek-r1` exhibited a significantly longer survival time against all other models. For the smaller models, the selfish population performs better, while the altruistic population sometimes suffered from starvation. For `claude-sonnet-4` and `gpt-4o`, we observed a plateau of time step around 30, regardless of the initial norm, indicating their inductive biases to be more conservative or altruistic.

thereafter all decisions are made in-context from the evolving social information and the currently adopted norm.

To manage compute/API cost, and because preliminary runs showed most populations collapse by roughly 50 rounds, we cap each simulation at 50 rounds and run 10 independent trials per condition. We then performed a two-way ANOVA with LLM model and altruistic ratio as fixed factors to assess their effects on survival time for each environment (harsh and rich). When we found a significant main effect among LLM models, we further conducted Tukey's HSD post-hoc tests ($\alpha = 0.05$), and statistically distinct groups were summarized using Compact Letter Display (CLD) notation (i.e., models sharing the same letter do not differ significantly).

4.2.1 ***Cooperation in harsh environment.*** In the ABM baseline, altruistic populations sustain the stock longer under harsh growth, whereas selfish populations tend to overharvest and crash. Turning to LLMs to understand whether they evolve group-beneficial norms, we observe the same pattern for larger models (`claude-sonnet-4`, `deepseek-r1`, `gpt-4o`): altruistic initializations survive longer (Fig. 5). However, smaller models collapse early regardless of initialization; efficiency traces (Fig. 15, left) show early overuse followed by rapid stock collapse. The result of ANOVA (Table 1) also supports this observation; while the performance among models significantly differed regardless of the initializations, there was no consistent trend across models driven by the altruistic ratio. Instead, the difference in the altruistic ratio showed a significant interaction effect with models, suggesting that the effect of initialization bifurcated between larger and smaller models.

4.2.2 ***Cooperation in rich environment.*** In the ABM baseline, mixed populations typically perform best in rich settings because mixed populations start from a higher variance, allowing for more efficient selection towards the optimal behaviors and norms. For LLM societies, behavior differs: with more time to adapt, smaller models often survive longer when initialized *selfish*, while *altruistic* initializations sometimes underharvest and starve (Fig. 6). The absence of explicit strategy copying and reliance on in-context updates make behavior stickier to the initial norm, which explains why the mixed population is not consistently best. Larger models exhibit distinct behaviors: `deepseek-r1` adapts and explores (surviving near the 50-step cap), whereas `gpt-4o` and `claude-sonnet-4` stabilize earlier with more conservative norms (Fig. 15, right; Table 4). The post-hoc test also corroborated that `deepseek-r1` exhibited a significantly longer survival time compared to all other models.

4.2.3 ***Model-specific patterns.*** `claude-sonnet-4` and `gpt-4o` typically plateau near 30 rounds, largely independent of the initial norm, whereas `deepseek-r1` often reaches the 50-round cap, especially from selfish starts (Fig. 6). Efficiency trajectories corroborate this: `deepseek-r1` stabilizes by steps 15–20 and then nudges upward, while `claude-sonnet-4` and `gpt-4o` settle at lower efficiency levels and remain there (Fig. 15, right). The language of proposed group norms mirrors these dynamics (Table 4): `deepseek-r1`

Table 1: Results of two-way ANOVA testing the effects of LLM models and altruistic ratio of the society on survival time under (a) harsh and (b) rich environments. In the harsh environment, the main effect of LLM models was significant ($p = 0.031$). In the rich environment, both the main effects of LLM models ($p < 0.001$) and Society Type ($p = 0.030$) were significant, indicating that model differences and population composition jointly influenced survival outcomes.

(a) Harsh environment	df	F	p-value	η^2
Model	6	2.37	0.031	0.06
Altruistic ratio	2	2.28	0.106	0.02
Model × Altruistic ratio	12	2.24	0.012	0.11
(b) Rich environment	df	F	p-value	η^2
Model	6	13.61	<0.001	0.28
Altruistic ratio	2	3.57	0.030	0.02
Model × Altruistic ratio	12	1.00	0.449	0.04

quickly adjusts target effort and, after step 40, cautiously raises it; `gpt-4o` keeps effort targets essentially unchanged. Under identical environmental dynamics, this points to a stronger exploratory bias in `deepseek-r1` and a more conservative/altruistic bias in `claude-sonnet-4` and `gpt-4o`.

4.2.4 ***Within-society norms.*** At the end of each run we summarize agents' norms by two scalar quantities. Let $\mathbf{n}_i \in \mathbb{R}^d$ denote the normalized norm vector of agent i with $\|\mathbf{n}_i\|_2 = 1$. The first metric, *individual similarity*, measures population homogeneity as the mean pairwise cosine similarity among agents' norms, $S_{\text{ind}} = \frac{2}{N(N-1)} \sum_{i<j} \mathbf{n}_i^\top \mathbf{n}_j$, such that higher values indicate more homogeneous norms. The second, *alignment*, captures how closely each agent's norm aligns with the contemporaneous group norm $\bar{\mathbf{n}} = \frac{\sum_i \mathbf{n}_i}{\|\sum_i \mathbf{n}_i\|_2}$, quantified as $S_{\text{align}} = \frac{1}{N} \sum_i \mathbf{n}_i^\top \bar{\mathbf{n}}$, where higher values indicate stronger alignment with the group-level norm. Figure 14 plots these summaries for altruistic and selfish initializations. Two patterns stand out: (a) *Family clustering:* models from the same provider occupy similar regions—for example, the Llama variants lie lower-left (less homogeneous, weakly aligned), the OpenAI pair (`gpt-4o` and `gpt-4o-mini`) clusters mid-high with `gpt-4o-mini` highest on both axes, `claude-sonnet-4` sits top-right (very high alignment and homogeneity), and `qwen3-32b` falls in the high-alignment band, suggesting that provider-specific pretraining and preference-tuning pipelines imprint consistent behaviors. (b) *Initialization is second-order:* shifts from altruistic to selfish are small relative to model differences.

4.2.5 ***Ablation study: What drives cooperation?*** We ablate the two alignment mechanisms in our framework: (i) *implicit alignment* via payoff-biased social learning (agents observe peers' outcomes and may imitate higher-payoff strategies) and (ii) *explicit alignment* via the *propose→vote* procedure (a shared group norm broadcast to all agents) to assess their separate and joint effects on cooperation.

Specifically, we compare three reduced variants against the full model (*Full* denotes the configuration with both payoff-biased social learning and propose→vote explicit norm adoption.): *(A) Only Social Learning (OSL):* agents imitate higher-payoff peers but no group norm is shared; *(B) Only Group Decision (OGD):* agents vote on a common norm but cannot imitate peers; and *(C) Neither:* both channels are removed, so agents act based only on their individual history and environmental feedback. All other parameters match the main simulations. Survival time (over $n = 10$ trials per condition) is shown in Fig. 7 and Fig. 13.

Absence of alignment. When both channels are removed (*Neither*), societies consistently show the lowest survival times ($\overline{T_s}_{\text{Neither}} = 16.22$) across environments and priors ($\overline{T_s} = 20.98, t\,(898) = -2.78, p = 0.006$), confirming that some form of alignment, implicit or explicit, is necessary to sustain cooperation. That is, coordination mechanisms, rather than individual adaptation alone, are key to stability.

Only group decision (no social learning). Suppressing social learning while retaining the group-voting mechanism (*OGD*) reveals that explicit alignment alone can sustain cooperation. Notably, explicit alignment sometimes even outperforms the full system, particularly in societies with selfish priors ($\overline{T_s}_{\text{OGD,selfish}} = 38.21, \overline{T_s}_{\text{OGD}} = 27.1, t\,(238) = 3.44, p < 0.001$), suggesting that the social-learning channel can reintroduce volatility when the population's prior incentives are self-interested.

Only social learning (no group norm). Conversely, *pure* social learning without an explicit shared norm (*OSL*) is often unstable ($\overline{T_s}_{\text{OSL}} = 17.56, \overline{T_s} = 20.98, t\,(898) = -1.96, p = 0.050$), especially under selfish priors: agents may imitate short-term winners, amplifying stochastic fluctuations. However, OSL can outperform other ablations in some settings (e.g., altruistic initializations), indicating that its effect is context-dependent. We observe exceptions where OSL is competitive (e.g., altruistic priors in some environments), consistent with social learning being beneficial when short-term success correlates with long-term sustainability.

Interaction with model reasoning. The two alignment channels have an interaction effect with model cognition. For *thinking models* such as `deepseek-r1`, explicit alignment (OGD) is sufficient to stabilize cooperation under most conditions. In contrast, for *non-thinking models* such as `gpt-4o`, combining implicit and explicit alignment helps balance exploration and exploitation, preventing premature convergence on over-harvesting or under-harvesting behaviors ($\overline{T_s}_{\text{OGD,gpt-4o}} = 16.65, \overline{T_s}_{\text{OGD,others}} = 32.33, t\,(178) = -4.67, p < 0.001$).

4.2.6 ***Takeaway.*** Our proposed CPR framework discriminates LLMs by their ability to evolve cooperative behaviours under diverse social and environmental conditions. The contrast between larger and smaller models highlighted differences in their ability to adapt to the environment and to effectively explore sustainable strategies. Moreover, by enabling the endogenous evolution of group-beneficial norms, our design reveals how model-specific inductive biases shape exploration and coordination, which can be observed directly in the group norms proposed by the agents. Grounded in Ostrom's institutional design principles and validated against ABM baselines, our CPR framework thus provides both an ecologically sound and empirically useful testbed for advancing the study of governance and cooperation in agentic societies.

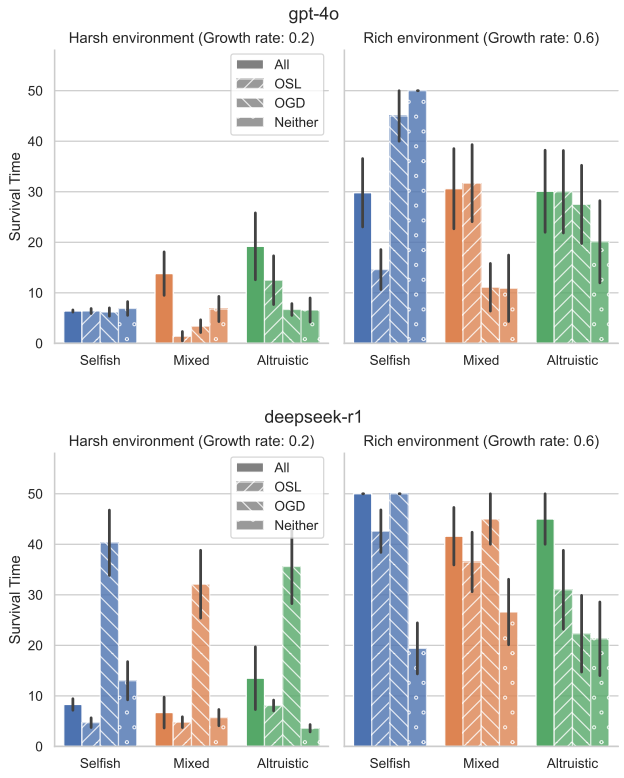

Figure 7: Survival time comparison of `deepseek-r1` `gpt-4o` in ablation conditions (See Fig. 13 for `qwen3-32b`). We compared the survival time (with ±1 s.e.m.) of four conditions (All, OSL, OGD, Neither) across different priors of populations (selfish, mixed, altruistic) in harsh and rich environments. Detailed observations are discussed in Section 4.2.5.

5 DISCUSSION & CONCLUSION

This paper introduced a CPR simulation framework grounded in Ostrom's institutional design principles and cultural evolutionary theory, enabling LLM societies to develop group-beneficial norms endogenously without explicit reward signals. Through both ABM and LLM simulations, we demonstrated the validity of the framework design and its ability to elicit diverse cooperative behaviours and norms across different LLM models. The ablation results show that removing both alignment channels, social learning and group norms, consistently leads to rapid collapse across all environments and priors. This confirms that some form of coordination, whether implicit imitation or explicit norm sharing, is essential for sustaining cooperation among models. Our results establish the framework as a theoretically driven and ecologically valid testbed for studying norm evolution and cooperative dynamics in agentic societies.

5.1 Limitations

Our study has several limitations. First, computational constraints restricted the number of trials and time horizons, which may underrepresent the long-term dynamics of norm evolution. Second, the CPR setting focuses on a single renewable resource and a narrow set of governance mechanisms; while this offers interpretability, it cannot capture the complexity of real-world institutions where multiple resources, cross-group interactions, and layered norms interact. Third, reliance on in-context learning for LLM agents introduces sensitivity to prompt design and model biases, limiting reproducibility and comparability across systems. Finally, closed-source models hinder full transparency, restricting the extent to which results can be independently replicated.

5.2 Future work

We expect future research to extend the CPR framework to more complex socio-ecological systems with multi-level governance, dynamic population turnover, and more diverse sanctioning or reputation systems. Investigating how institutional structures themselves co-evolve with agent norms would allow closer alignment with political and organisational theory. Thus, a natural extension is to introduce interaction networks and multi-level governance to study how local norm clusters form and spread under structured contact patterns. An orthogonal direction is to compare against DeepRL agents in economic environments with explicit rewards (e.g., Fruit Market [18]), to disentangle norm formation from reward-optimized behavior. Moreover, integrating deliberative communication mechanisms beyond simple propose→vote procedures may reveal whether LLMs can sustain cooperative norms through richer forms of dialogue, while they may suffer the limitations of context length and memory capacity of LLMs [26]. From a methodological perspective, expanding trials across diverse prompting strategies, decoding settings, and model families would clarify the robustness and generality of observed behaviours.

5.3 Ethical considerations

Our findings carry ethical implications for the deployment of LLM-based systems in societal contexts. The systematic differences observed across models highlight that model choice itself can bias the emergent norms of an agentic society, with downstream consequences for fairness, stability, and governance. While our simulations abstract away from human participants, similar dynamics may arise in AI-mediated platforms, markets, or communities. This underscores the importance of transparency in model evaluation, cautious deployment of multi-agent systems, and the incorporation of safeguard mechanisms to prevent misaligned or harmful norms from propagating. Future research should also consider how to design frameworks that not only support cooperation but also protect against exploitation, exclusion, or manipulation.

ACKNOWLEDGMENTS

We thank Dr. Levin Brinkmann for insightful discussions. This work was supported in part by the Japan Science and Technology Agency (JST) through the PRESTO program (JPMJPR246B).

APPENDIX

The content of the Appendix is available at [11].

REFERENCES

[1] Jeffrey Andrews, Matthew Clark, Vicken Hillis, and Monique Borgerhoff Mulder. 2024. The cultural evolution of collective property rights for sustainable resource governance. *Nature Sustainability* 7, 4 (2024), 404–412.

[2] Nicolas Bacaër. 2011. Verhulst and the logistic equation. In *A short history of mathematical population dynamics*. Springer, London, UK, 35–39.

[3] Steffen Backmann, David Guzman Piedrahita, Emanuel Tewolde, Rada Mihalcea, Bernhard Schölkopf, and Zhijing Jin. 2025. When Ethics and Payoffs Diverge: LLM Agents in Morally Charged Social Dilemmas. *arXiv* 2505.19212 (2025), 1–33.

[4] Pat Barclay. 2004. Trustworthiness and competitive altruism can also solve the "tragedy of the commons". *Evolution and Human Behavior* 25, 4 (2004), 209–220.

[5] Duncan Black. 1948. On the rationale of group decision-making. *Journal of Political Economy* 56, 1 (1948), 23–34.

[6] Samuel Bowles and Herbert Gintis. 1998. The moral economy of communities: Structured populations and the evolution of pro-social norms. *Evolution and Human Behavior* 19, 1 (1998), 3–25.

[7] Robert Boyd and Peter J Richerson. 2002. Group beneficial norms can spread rapidly in a structured population. *Journal of Theoretical Biology* 215, 3 (2002), 287–296.

[8] Robert Boyd and Peter J Richerson. 2009. Voting with your feet: Payoff biased migration and the evolution of group beneficial behavior. *Journal of Theoretical Biology* 257, 2 (2009), 331–339.

[9] Damon Centola and Andrea Baronchelli. 2015. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences* 112, 7 (2015), 1989–1994.

[10] Clark C Gibson, John T Williams, and Elinor Ostrom. 2005. Local enforcement and better forests. *World development* 33, 2 (2005), 273–284.

[11] Prateek Gupta, Qiankun Zhong, Hiromu Yakura, Thomas F. Eisenmann, and Iyad Rahwan. 2025. The Role of Social Learning and Collective Norm Formation in Fostering Cooperation in LLM Multi-Agent Systems. *arXiv* 2510.14401 (2025), 1–15. https://doi.org/10.48550/ARXIV.2510.14401

[12] Garrett Hardin. 1968. The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *Science* 162, 3859 (1968), 1243–1248.

[13] Joseph Henrich. 2006. Cooperation, punishment, and the evolution of human institutions. *Science* 312, 5770 (2006), 60–61.

[14] Joseph Henrich and Robert Boyd. 2001. Why people punish defectors: Weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208, 1 (2001), 79–89.

[15] Joseph Henrich and Francisco J Gil-White. 2001. The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and human behavior* 22, 3 (2001), 165–196.

[16] Ray Hilborn and Carl J Walters. 1992. *Quantitative fisheries stock assessment: choice, dynamics and uncertainty*. Springer, New York, US.

[17] Wenyue Hua, Lizhou Fan, Lingyao Li, Kai Mei, Jianchao Ji, Yingqiang Ge, Libby Hemphill, and Yongfeng Zhang. 2023. War and Peace (WarAgent): Large Language Model-based Multi-Agent Simulation of World Wars. *arXiv* 2311.17227 (2023), 1–47.

[18] Michael Bradley Johanson, Edward Hughes, Finbarr Timbers, and Joel Z Leibo. 2022. Emergent bartering behaviour in multi-agent reinforcement learning. *arXiv* 2205.06760 (2022), 1–114. https://doi.org/10.48550/arXiv.2205.06760

[19] Shimin Li, Tianxiang Sun, Qinyuan Cheng, and Xipeng Qiu. 2024. Agent alignment in evolving social norms. *arXiv* 2401.04620 (2024), 1–31.

[20] Chen Cecilia Liu. 2025. Cooperative Behaviour in LLMs via Cultural Evolution of Norms and Strategies. In *Proceedings of the 1st CoLM Workshop on Social Simulation with LLMs*. OpenReview, 1–11.

[21] Richard McElreath, Adrian V Bell, Charles Efferson, Mark Lubell, Peter J Richerson, and Timothy Waring. 2008. Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, 1509 (2008), 3515–3528.

[22] Ninell Oldenburg and Tan Zhi-Xuan. 2024. Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, US, 1510–1520. https://doi.org/10.5555/3635637.3663011

[23] Elinor Ostrom. 1990. *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press, Cambridge, UK.

[24] Elinor Ostrom. 1999. *Design principles and threats to sustainable organizations that manage commons*. Technical Report W99-6. Indiana University.

[25] Elinor Ostrom. 2009. A general framework for analyzing sustainability of social-ecological systems. *Science* 325, 5939 (2009), 419–422.

[26] 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*. ACM, New York, US, 1–22.

[27] 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, US, 111715–111759.

[28] David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. 2025. Corrupted by Reasoning: Reasoning Language Models Become Free-Riders in Public Goods Games. In *Proceedings of the 2nd Conference on Language Modeling*. OpenReview, 1–37.

[29] Michael E Price, Leda Cosmides, and John Tooby. 2002. Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* 23, 3 (2002), 203–231.

[30] Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. 2024. Emergence of social norms in generative agent societies: principles and architecture. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*. Curran Associates Inc., Red Hook, US, Article 874, 9 pages.

[31] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. Escalation risks from language models in military and diplomatic decision-making. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, US, 836–898.

[32] Shade T Shutters. 2012. Punishment leads to cooperative behavior in structured societies. *Evolutionary Computation* 20, 2 (2012), 301–319.

[33] Daniel Smith. 2020. Cultural group selection and human cooperation: a conceptual and empirical review. *Evolutionary Human Sciences* 2 (2020), e2.

[34] György Szabó and Gabor Fath. 2007. Evolutionary games on graphs. *Physics reports* 446, 4-6 (2007), 97–216.

[35] Aron Szekely, Francesca Lipari, Alberto Antonioni, Mario Paolucci, Angel Sánchez, Luca Tummolini, and Giulia Andrighetto. 2021. Evidence from a long-term experiment that collective risks change social norms and promote cooperation. *Nature Communications* 12, 1 (2021), 5452.

[36] Sz-Ting Tzeng, Nirav Ajmeri, and Munindar P. Singh. 2024. Norm Enforcement with a Soft Touch: Faster Emergence, Happier Agents. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, US, 1837–1846. https://doi.org/10.5555/3635637.3663046

[37] Aron Vallinder and Edward Hughes. 2024. Cultural evolution of cooperation among LLM agents. *arXiv* 2412.10270 (2024), 1–19.

[38] Kavindu Warnakulasuriya, Prabhash Dissanayake, Navindu De Silva, Stephen Cranefield, Bastin Tony Roy Savarimuthu, Surangika Ranathunga, and Nisansa de Silva. 2025. Evolution of Cooperation in LLM-Agent Societies: A Preliminary Study Using Different Punishment Strategies. In *Proceedings of the 18th Workshop on Coordination, Organizations, Institutions, Norms and Ethics for Governance of Multi-Agent Systems*. Springer, Cham, Swizerland, 1–19.