Validating Generative Agent-Based Models of Social Norm Enforcement: From Replication to Novel Predictions

Logan Cross[1], **Nick Haber**[3], **Daniel L.K. Yamins**[1,2]
[1]Department of Computer Science, [2]Department of Psychology, [3]Graduate School of Education, Stanford University

Abstract

As large language models (LLMs) advance, there is growing interest in using them to simulate human social behavior through generative agent-based modeling (GABM). However, validating these models remains a key challenge. We present a systematic two-stage validation approach using social dilemma paradigms from psychological literature, first identifying the cognitive components necessary for LLM agents to reproduce known human behaviors in mixed-motive settings from two landmark papers, then using the validated architecture to simulate novel conditions. Our model comparison of different cognitive architectures shows that both persona-based individual differences and theory of mind capabilities are essential for replicating third-party punishment (TPP) as a costly signal of trustworthiness. For the second study on public goods games, this architecture is able to replicate an increase in cooperation from the spread of reputational information through gossip. However, an additional strategic component is necessary to replicate the additional boost in cooperation rates in the condition that allows both ostracism and gossip. We then test novel predictions for each paper with our validated generative agents. We find that TPP rates significantly drop in settings where punishment is anonymous, yet a substantial amount of TPP persists, suggesting that both reputational and intrinsic moral motivations play a role in this behavior. For the second paper, we introduce a novel intervention and see that open discussion periods before rounds of the public goods game further increase contributions, allowing groups to develop social norms for cooperation. This work provides a framework for validating generative agent models while demonstrating their potential to generate novel and testable insights into human social behavior.

Introduction

Advances in LLMs have created new possibilities for studying human social behavior through simulation. While agent-based modeling has long been used in the social sciences, the emergence of LLM-based agent enables a qualitatively new approach: generative agent-based modeling (GABM) (Vezhnevets et al., 2023). However, for GABMs to be useful tools for social science, we need rigorous frameworks to validate whether these agents actually capture relevant aspects of human behavior. A key challenge is determining whether agents' response distributions match those observed in human studies. In this paper we introduce a two-stage approach that prototypes GABMs to replicate and extend human studies.

1. **Model Validation:** First, we conduct an iterative model comparison to identify the minimal set of cognitive components sufficient to reproduce known behavioral effects. For theoretical parsimony, we begin with the simplest possible agent architecture and test new cognitive components only when simpler models fail to capture human behavior patterns. In addition, by varying the agent architectures and comparing their output to human data, we can test specific hypotheses about the cognitive mechanisms underlying social behavior.

2. **Novel Prediction Generation:** Second, we leverage validated agent architectures to explore counterfactual scenarios that would be slow or costly to implement in laboratory settings. By maintaining the validated cognitive components while varying environmental conditions, we can generate precise, quantitative predictions about human behavior in novel situations. These predictions can then be tested empirically, providing a rigorous test of the model's generalization capabilities while advancing our understanding of human social behavior.

Here, we present this validation approach with two case studies to examine the components necessary for GABMs to replicate key findings from experimental social dilemmas. These mixed-motive paradigms provide an ideal test case as they involve complex social cognition, including reputation management, moral judgment, theory of mind, and norm enforcement. First, we investigate third-party punishment (TPP) from Jordan et al. (2016), which demonstrated that third-party punishers are perceived as—and actually are—more trustworthy than non-punishers. Through careful comparisons of different model architectures built in Concordia, we show that both persona prompting and theory of mind reflection are necessary components for reproducing these effects. Second, we examine Feinberg et al. (2014)'s work on gossip and ostracism in promoting cooperation in a public goods game, finding that while our initial validated architecture from the TPP study provides a good foundation, an additional strategic component proves necessary to replicate a majority of the behavioral effects in this more complex social environment.

We then apply these validated model architectures to explore two distinct types of novel questions not investigated in the original papers. For the TPP study, we use the validated model for theoretical disambiguation, examining how punishment behavior differs between public and private settings. Our results quantify the relative contributions of reputational vs. intrinsic motivations, helping resolve the ongoing

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

theoretical debate about why people engage in costly third-party punishment. For the public goods study, we use the validated model to test a novel intervention—adding pre-round discussion periods—and find that this further improved cooperation rates, as groups could self-organize and establish shared norms for contribution. These findings demonstrate how GABM can help social scientists both evaluate conflicting cognitive theories and investigate potential interventions to improve collective welfare.

Related Work

Recent advances in large language models have enabled generative agent-based modeling (GABM) as a promising approach for simulating human behavior and social dynamics (Vezhnevets et al., 2023). Initial work by Park et al. (2023) demonstrated that LLM-based agents could produce believable simulations of human behavior, while our research focuses on validating such models against established behavioral effects at the population level. We build on the growing body of work exploring how cognitive components like theory of mind and individual differences can enhance agent behavior (Li et al., 2023; Chen et al., 2024).

Our validation targets two classic experimental economics paradigms. The Trust Game (Berg et al., 1995) measures trust and trustworthiness through sequential monetary exchanges, with third-party punishment extending this to include observers who can pay costs to sanction norm violations (Jordan et al., 2016). The Public Goods Game represents a fundamental paradigm for studying cooperation where participants face a social dilemma between personal gain and group welfare (Ledyard et al., 1994). These mixed-motive settings provide ideal test cases for validating social agents as they require sophisticated reasoning about reputation, norms, and strategic behavior.

Methods

Generative Agent Implementation

We implemented our experimental paradigms using Concordia (Vezhnevets et al., 2023), a framework for LLM-based agents that enables natural language interaction while maintaining grounded state variables for game decisions. Following a principle of parsimony, we began with minimal agent architectures and systematically added components only when simpler models failed to reproduce human behavioral patterns. This approach enabled precise identification of which cognitive capabilities are necessary for social behavior. We used GPT-4o as the base model for all architectures.

Our Base Agent Architecture includes fundamental components: an Observation Summary (processes information about current/past game states), Situation Assessment (prompts the LLM to assess the current situation), and Decision (generates decisions based on previous components). We then tested additional components: Persona (provides agents with distinct personalities), Theory of Mind Reflection (enables explicit reasoning about others' mental states), Strategy Reflection (explicitly considers long-term payoff maximization), and Emotion Reflection (considers emotional responses to situations). Our model comparisons systematically included or excluded these components to identify minimal sufficient architectures for replicating human behavior.

Additional Cognitive Components After testing the Base Architecture against human data and identifying gaps in behavioral replication, we implemented these social and personality components that make up what we call the **Social Architecture**. **Persona**: Provides agents with distinct personalities and behavioral tendencies. We asked Claude Sonnet 3.5 to generate 40 personas with varying demographics, occupations, and traits. **Personality Reflection**: Prompt displaying the persona and asking the LLM: "Based on this background and your past actions, how would you describe your approach to trust and fairness in economic decisions vs maximizing your own payoff?" **Theory of Mind Reflection**: Enables explicit reflection about others' mental states and intentions and how they may respond to your behavior. The output of the Observation Summary is included in the prompt such that the agent can reflect on what their partner has done in previous decisions.

To replicate Study 2, we additionally developed these components in our iterative agent development process to help the agent to value possible outcomes more precisely: **Strategy Reflection**: Explicitly prompts the LLM to think strategically about maximizing long-term earnings. **Emotion Reflection**: Prompts the agent to consider their emotional response to the situation: "How are you feeling emotionally about the current situation?"

Study 1: Third-party Punishment as a costly signal of trustworthiness

We based our first generative agent simulation on the third-party punishment game developed by (Jordan et al., 2016) which consists of two sequential stages designed to test whether costly punishment serves as a signal of trustworthiness. Here, both stages consist of the Trust game. The Trust game is a classic behavioral economics paradigm where one player (the Helper) can choose to send some portion of their endowment to another player (the Recipient), with the sent amount being tripled by the experimenter. Then, the Recipient decides how much of this tripled amount to return to the Helper. This creates a social dilemma that measures trust and trustworthiness: while both players can benefit from cooperation, the Helper may keep the endowment if they do not trust the Recipient, and the Recipient faces a temptation to defect by keeping the tripled amount.

In the first stage, one LLM agent acts as a third party, the Punisher/Signaller, watching two people play the Trust Game. The Trust Game consists of a Helper with an initial endowment of $10 that could potentially be shared with another player: the Recipient. In Stage 1, the Helper is programmed to give $0 in order to evaluate whether the Signaller punishes selfish behavior. The Signaller observes the Helper's decision

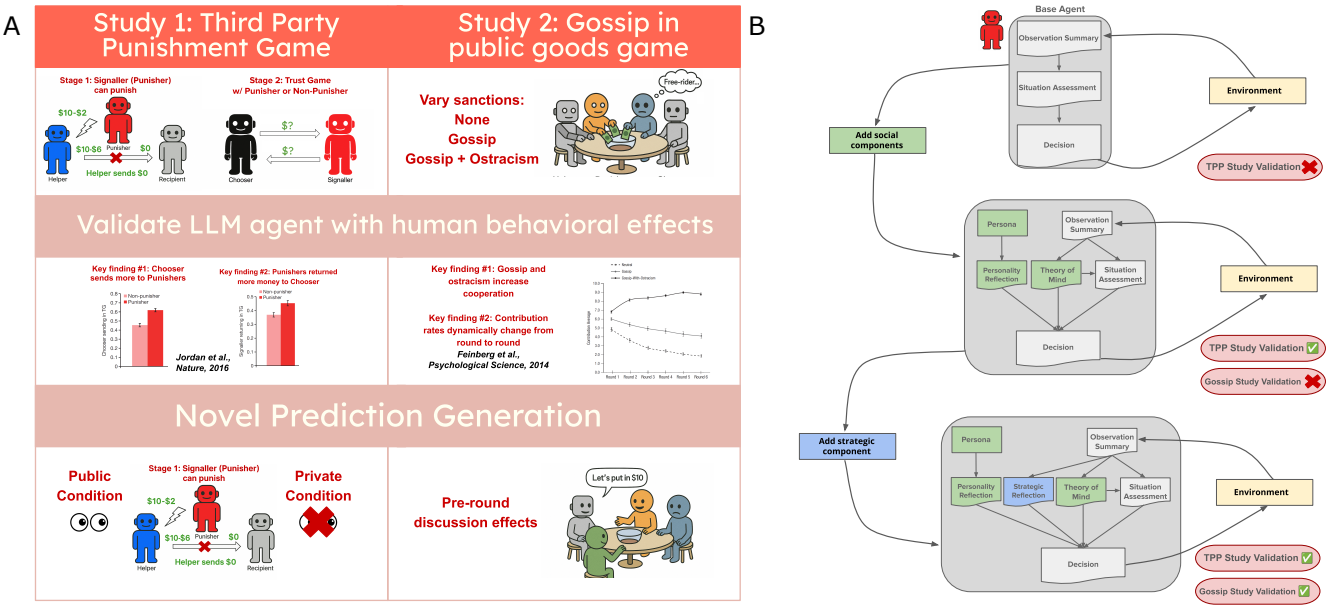

Figure 1: A. Two-stage approach of model validation and novel prediction generation. B. Iterative refinement of model architectures.

and can pay a personal cost ($2) to punish the Helper ($6 reduction of Helper's winnings). The key feature of this stage is that punishment is costly to the Punisher and provides no immediate direct benefit, as they are an unaffected third party.

In the second stage, a new participant (the Chooser) plays a trust game with either a punisher or non-punisher from Stage 1, knowing their previous punishment decision in the public condition (see below). The trust game proceeds as follows: 1. Chooser receives an endowment as the Helper and decides how much to send to the Signaller (Punisher in Stage 1). 2. Any amount sent is tripled by the experimenter. 3. Signaller, acting as the Recipient now, decides how much of the tripled amount to return to the Chooser

Public and Private Conditions. To disentangle reputational from intrinsic motivations in third-party punishment, we varied punishment decision observability across two conditions. In the public condition (replicating Jordan et al., 2016), the Signaller's punishment decision was explicitly communicated to the Stage 2 Chooser, and Signallers were informed their decision would be public. In the private condition, the Signaller's punishment decision remained unknown to the Chooser, and Signallers were told their decision would remain private. This manipulation removes potential reputational benefits while preserving all other aspects of the decision context.

Study 2: Gossip and Ostracism Promote Cooperation in Groups

We based our second generative agent simulation on the gossip and ostracism paradigm developed by Feinberg et al. (2014) which examines how the spread of reputational information through gossip facilitates cooperation and limits defection in groups. The study employed a public goods exercise where participants faced a social dilemma between maximizing personal gain and contributing to group welfare.

Public Goods Exercise The foundation of this study is the public goods exercise. In each round: 1. Participants in groups of 4 each receive an allotment of 10 points (worth 2.5¢ each in the study and $1 for our agents), 2. Each participant decides how many points to contribute to a group fund versus keep for themselves, 3. The total points contributed to the group fund are doubled and redistributed equally to all group members. This creates a social dilemma where individuals benefit most by free-riding on others' contributions. After each round, participants learned how much each group member had contributed and earned, were assigned to a group with different partners for the next round.

Experimental Conditions The study employed a repeated measures design where all participants played three distinct conditions, each for six rounds: **Basic** In the basic game, participants played the standard public goods exercise without modifications. **Gossip** In the gossip condition, after learning the results of each round, participants could send an anonymous gossip note about one of their current group members to that person's future interaction partners. This allowed reputational information to flow between groups despite no direct interaction between them. **Gossip With Ostracism** This condition added an ostracism mechanism. At the beginning of each round, after receiving any gossip notes, participants could anonymously vote to exclude one participant from playing in the upcoming round. If someone received

Figure 2: TPP Results: (A) Average amount sent by choosers to signallers in Stage 2. (B) Percentage returned by recipients in Stage 2. (C) Punishment rates in public versus private conditions.

two or more exclusion votes, they were ostracized and did not participate in that round, earning nothing. The remaining three participants played the public goods exercise with a reduced multiplier (1.5 instead of 2) to maintain proportionality of potential earnings.

Results

Study 1: Third-party punishment as a costly signal of trustworthiness

We first examined what are the minimal components needed in a generative agent model to reproduce key findings from Jordan et al. (2016)'s human study of third-party punishment shown in Figure 1: **Key Finding 1:** Choosers demonstrated greater trust in punishers by sending them more money in Stage 2. **Key Finding 2:** Punishers proved to be actually more trustworthy by returning a higher percentage of funds as Recipients in Stage 2.

By systematically varying agent components and comparing their outputs to human data, we can identify which cognitive mechanisms are necessary for reproducing these social behaviors.

Model Validation Key Finding 1: Punishment Perceived as a Signal of Trust. In the original human study, participants consistently demonstrated greater trust in individuals who engaged in costly third-party punishment. Initially, our analyses with the base agent and GPT4o base model showed that persona prompting is necessary to produce these crucial individual differences in punishment rate **??**. Without distinct personas, agents showed uniform punishment rates of 100%, failing to capture the natural variation seen in human studies, even with a relatively high LLM sampling temperature of 1.0. Thus, adding personas was the first crucial component necessary to model human behavior for this experimental paradigm.

With only a persona component, agents failed to discriminate between punishers and non-punishers in their trust decisions (t = -0.566, p = 0.572, N=100), sending similar amounts to both groups (Figure 2a). Analysis of agent reasoning revealed that without explicit prompting to consider their partner's past behavior, agents defaulted to uniform trust levels regardless of punishment history. This suggests that explicit partner modeling is necessary for translating punishment signals into trust decisions, mirroring theories about human social cognition.

Following our iterative process, we added a theory of mind component to our agent architecture. With this addition, choosers sent significantly more money (22% more) to Signallers who had previously engaged in punishment compared to those who did not (t = 6.08, $p < 0.001$, N = 400) (Figure 2a). This matches the effect observed in human studies, suggesting that our agents captured the key dynamics of social trust.

Key Finding 2: Punishment is a Signal of Actual Trustworthiness The second key finding from the human study was that punishers were actually more trustworthy when playing as the Recipient in the trust game. This relationship demonstrates that punishment isn't merely cheap talk but reflects genuine prosocial tendencies. The Social architecture (ToM + Persona) successfully reproduced this pattern: Signallers who had previously punished selfish behavior returned significantly more money to Choosers compared to non-punishers (t = 9.83, $p < 0.001$, N = 400) (Figure 2b). This behavioral pattern provided validation that our agents weren't simply learning to signal trustworthiness superficially, but were exhibiting internally consistent patterns of prosocial behavior across different contexts. This pattern persists in the Theory of Mind ablation (t = 6.95, $p < 0.001$, N=100), as punishers acted less selfishly as Recipients similarly to the Social architecture with ToM (Figure 2b). This suggests that individual differences implemented through personas are sufficient to drive behavioral consistency through the LLM's modeling of stable prosocial traits.

Novel Predictions: Public vs. Private Punishment Having validated our Social agent architecture against human behavioral patterns, we used the model to generate and test novel predictions about punishment behavior under public vs. private conditions. Our simulations revealed a significant difference in punishment rates between conditions Figure 2c. In the public condition, 64.8% of agents chose to punish the selfish behavior of the Helper in Stage 1, while this rate dropped to 48.9% in the private condition where punishment decisions were not observable ($\chi^2 = 18.68$, $p < 0.001$, N = 400). This 16 percentage point difference helps quantify the relative contributions of reputational versus intrinsic motivations for punishment. The 16% drop in punishment rates when decisions are private represents the portion of punishment motivated by reputational concerns. This suggests that the desire to signal trustworthiness to potential future interaction partners drives a significant portion, but not the majority, of punishment behavior. The substantial baseline rate of punishment in the private condition, where there are no reputational benefits, reveals a strong intrinsic motivation to enforce fairness norms in this task. This suggests that a sizable portion of punishment decisions are driven by genuine prosocial preferences rather than strategic concerns. These quantitative predictions about punishment rates in public versus private conditions offer clear hypotheses for future work with human participants.

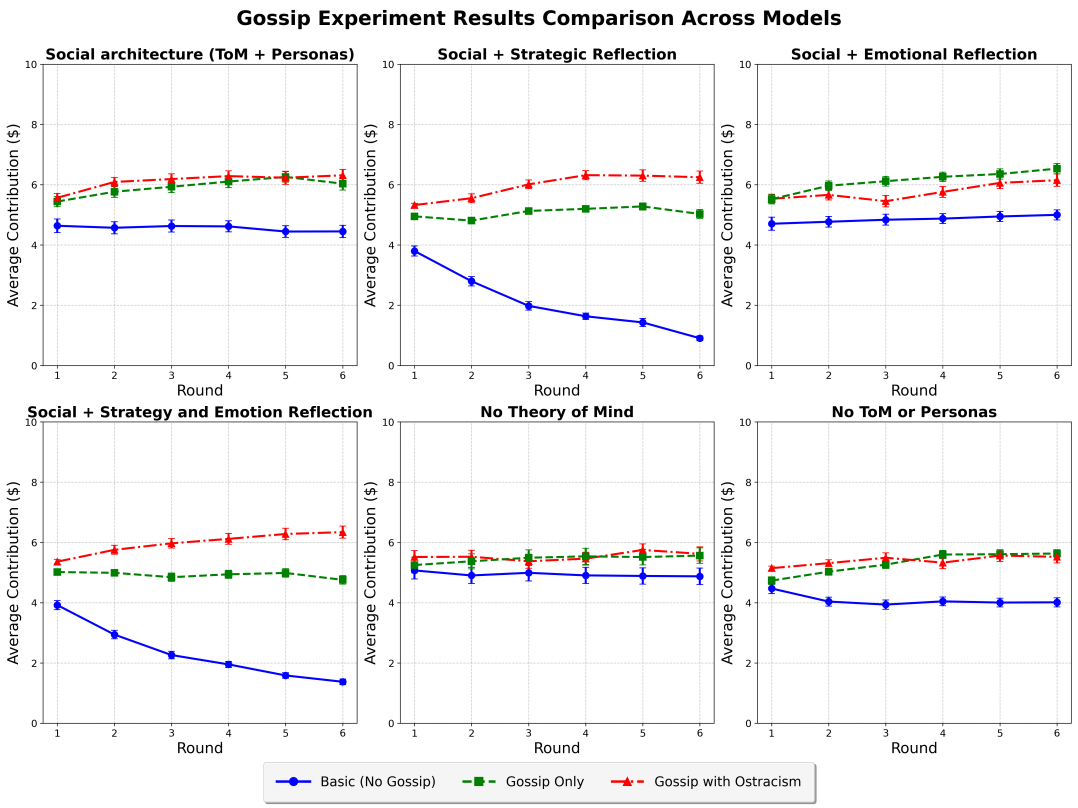

Figure 3: Gossip experiment results. Each line includes 5 experiments of 24 agents each.

Study 2: Gossip and Ostracism Promote Cooperation in Groups

Building on previous experiments, we investigated the minimal components needed to replicate human social dynamics in Feinberg et al. (2014). In this study, participants in groups of 4 played multiple rounds of a public goods game across three conditions. In the basic condition, participants played a standard public goods game. Two conditions allowed gossip, where participants could write notes about players to their future partners. In the gossip with ostracism condition, players could use this information to vote to ostracize non-cooperative members for a round. Researchers identified key behavioral patterns to replicate: 1. Higher contributions in the gossip + ostracism condition vs basic. 2. Higher contributions in the gossip condition vs basic. 3. Higher contributions in the gossip + ostracism condition vs gossip. 4. Increasing trend in contributions for gossip + ostracism by round. 5. Decreasing trend in contributions for gossip condition by round. 6. Decreasing trend in contributions for basic condition by round. The study demonstrates that gossip and ostracism mechanisms significantly increase group cooperation.

Model Validation We first tested the Social architecture (LLM agent with persona and ToM components) from our previous experiments. This model replicated 3/6 of the experiment effects previously highlighted (Figure 3). Contributions are significantly higher in both gossip and gossip+ostracism conditions compared to basic ($F(1,119) = 77.56$, $\eta^2 = 0.39$ for gossip vs. basic; $F(1,119) = 86.88$, $\eta^2 = .42$ for gossip+ostracism vs. basic). These results show the Social architecture captures reputational transfer effects that function as sanctioning mechanisms increasing overall contributions.

However, the Social architecture fails to replicate other crucial effects from the original study. It doesn't produce the temporal dynamics of declining contributions in the basic and gossip conditions. This effect, common in public goods games (Ledyard et al., 1994; Fischbacher et al., 2001), occurs as incentives lead to more free-riding behavior over time. This creates a downward spiral in the basic condition, and to a lesser degree in the gossip-only condition. Even though mutual cooperation yields higher payouts for everyone, free-riding becomes the dominant strategy without effective sanctioning mechanisms like ostracism.

Relatedly, the Social architecture fails to replicate the significant difference between the gossip and gossip + ostracism conditions. In humans, contributions were highest in the gossip and ostracism condition as it creates a social norm in which cooperation is advantageous and norm-breaking is costly. The explicit sanction of ostracism additionally causes a phase shift in temporal dynamics, making cooperation ratio-

nal and altering the game's equilibrium. Our model shows increasing contributions over time in both the gossip+ostracism condition and the gossip condition, leading to no significant difference between the groups (F(1, 119) = 1.55, $\eta^2 = .01$, p > 0.05).

We had two cognitively-grounded hypotheses for improving our model: a strategic reflection component to explicitly consider payoff maximization, and an emotional component to capture risk aversion and affective responses to cooperation/defection.

The strategic component significantly improved model performance (Figure 3). When added to the social architecture, it enabled clear differentiation between all three conditions, with highly significant pairwise comparisons between conditions. Contributions were highest for gossip-with-ostracism, and significantly higher than the gossip condition ($t(119) = -7.30$, $p < 0.001$), matching human data and showing that ostracism provides benefits beyond reputation alone. The Social + Strategic model also demonstrated appropriate temporal trends—decreasing contributions in the basic condition (slope: -0.5407, $p = 0.0017$), and increasing contributions in gossip-with-ostracism (slope: 0.205, $p = 0.014$)—closely matching human patterns. However, contributions in the gossip condition remained flat rather than decreasing (slope: 0.053, $p = 0.225$).

In contrast, the emotional reflection component did not improve replication. All conditions actually showed increasing linear trends, and the emotional component model showed higher contributions in the gossip condition compared to gossip-with-ostracism ($t(119) = 2.83$, $p = 0.005$), the opposite of what was observed in the human study. Combining both components yielded results similar to the strategic-only model, suggesting strategic reasoning is the critical cognitive component.

Ablation studies confirmed the importance of our base architecture components, with models lacking theory of mind or personas failing to replicate key behavioral patterns and condition differences. Importantly, our new architecture with strategic reflection maintained successful replication of the key findings from Study 1, providing additional validation that these cognitive architectures capture essential social dynamics across different experimental paradigms.

Novel Predictions: Pre-round Discussion Condition We next used the Social + Strategic Reflection model to explore a novel intervention: adding structured discussion periods before each round of the public goods game. This condition builds on the gossip-with-ostracism mechanism but adds a collective deliberation phase, allowing agents to coordinate their behavior and establish shared norms explicitly rather than only through indirect gossip channels.

Contributions in the discussion condition (M = 38.17, SD = 4.27) were significantly higher than in the standard gossip-with-ostracism condition (M = 35.76, SD = 6.64), t(119) = -3.52, $p < 0.001$), particularly for the first round before implicit coordination can occur. This demonstrates that this open discussion period further increases cooperation levels beyond the gossip and ostracism condition.

Discussion

This paper demonstrates a systematic approach to validating generative agent-based models through careful replication of behavioral effects from the psychological literature and generation of novel predictions. Using the TPP paradigm and gossip-ostracism studies, we identified key cognitive components necessary for reproducing human social behavior.

Our finding of higher punishment rates in public settings aligns with previous work on prosocial behavior under observation (Barclay, 2004; Milinski et al., 2002), supporting costly signaling theory (Jordan et al., 2016). However, the substantial punishment that persists in private settings suggests the presence of both reputational and intrinsic motivations for norm enforcement. Similarly, our finding that discussion periods enhance cooperation extends Ostrom's research on how communities establish self-organized norms to overcome social dilemmas (Ostrom, 1990).

While our validation approach provides multiple levels of evidence through replication and model comparison (Vezhnevets et al., 2023), significant limitations remain. We couldn't replicate all behavioral patterns (such as the decrease in contributions in the gossip condition), and our hand-designed prompts represent degrees of freedom that may limit generalization. The most immediate opportunity to strengthen this work would be empirical testing of our novel predictions with human participants, followed by refinement of our cognitive architectures.

This work provides a template for rigorous validation of generative agent-based models in social science. By combining replication, systematic component analysis, and novel prediction generation, these models can serve as both theoretical tools for understanding human behavior and practical tools for generating testable hypotheses. As LLMs advance, this validation approach will be crucial for establishing their scientific utility while maintaining rigorous standards of evidence, ultimately providing a scalable framework for studying dynamic social phenomena (Vezhnevets et al., 2023; Leibo et al., 2024).

Altogether, this work provides a template for rigorous validation of generative agent-based models in social science. By combining careful replication, systematic component analysis, and novel prediction generation, we demonstrate how these models can serve as both theoretical tools for understanding human behavior and practical tools for generating testable hypotheses. As large language models continue to advance, this kind of systematic validation approach will be crucial for establishing their scientific utility while maintaining rigorous standards of evidence. Most importantly, once validated and expanded, this new paradigm provides a scalable framework for studying dynamic social phenomena such as status, culture, and societal scale social dilemmas (Vezhnevets et al., 2023; Leibo et al., 2024).

References

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

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122–142.

Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., ... others (2024). From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological science*, *25*(3), 656–664.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics letters*, *71*(3), 397–404.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476.

Ledyard, J. O., et al. (1994). *Public goods: A survey of experimental research*. Division of the Humanities and Social Sciences, California Inst. of Technology.

Leibo, J. Z., Vezhnevets, A. S., Diaz, M., Agapiou, J. P., Cunningham, W. A., Sunehag, P., ... others (2024). A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*.

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). Camel: Communicative agents for" mind" exploration of large scale language model society.

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the 'tragedy of the commons'. *Nature*, *415*(6870), 424–426.

Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. *Cambridge University*.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1–22).

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