The Mask of Consensus: Modeling Pluralistic Ignorance with Generative Agents

Yixuan Yuan (yixuan.yuan@connect.um.edu.mo)
University of Macau, Macau SAR, China

Abstract

Individuals frequently endorse views they privately reject to align with a perceived majority. This dynamic, known as pluralistic ignorance, is challenging to model because traditional scalar agents lack the semantic capacity to distinguish between private belief and public performance. To address this, we propose the Generative Cognitive-Social Simulation (GCSS) framework. We introduce the Grounding-Evaluation-Masking (GEM) architecture, which utilizes Large Language Models to simulate the strategic suppression of dissent based on Theory of Mind (ToM) risk assessment. Virtual classroom simulations suggest that ToM-guided risk evaluation is important for the emergence of the "spiral of silence." Crucially, we identify a computational mechanism of belief internalization: under prolonged masking, agents gradually adjust their private beliefs toward their publicly maintained stances. These findings quantify how social pressure can transform strategic conformity into more durable, collectively reinforced belief.

Keywords: Theory of Mind; Generative Agents; Pluralistic Ignorance

Introduction

Pluralistic ignorance represents a fascinating failure of collective social cognition where a majority of group members privately reject a norm yet incorrectly assume that most others accept it (J. O'Gorman, 1986; Katz et al., 1931). This phenomenon stems from the fundamental divergence between the "back stage" of private belief and the "front stage" of public performance (Goffman, 1959). Individuals do not simply broadcast their internal states to the world. Instead, they act as strategic performers who constantly monitor the social environment to minimize friction (Goffman, 1959; Leary, 2019). While this decoupling allows for smooth social coordination, it can lead to pathological outcomes where a group sustains a "false consensus" that no individual actually supports (Miller & McFarland, 1987; Prentice & Miller, 1993). The persistence of such collective delusion relies on a specific cognitive process: the agent must possess a Theory of Mind (ToM) (Bjerring et al., 2014; Premack & Woodruff, 1978) to infer the judgment of others and the capacity to weigh the reputational cost of honesty against the safety of conformity.

This dynamic is vividly captured in the classroom scenario shown in Figure 1. When a lecturer presents an abstruse concept (Figure 1a), the students face a conflict between epistemic honesty and social survival. While the majority privately struggle to comprehend the material (Figure 1c), they simultaneously engage in recursive social reasoning. Fearing that admitting confusion will signal incompetence to their peers (Miller & McFarland, 1987; Moneva & Inday, 2020), they suppress their genuine feedback and publicly signal understanding (Figure 1b). This strategic masking creates a self-reinforcing cycle (Noelle-Neumann, 1974). Observing the silence of others, each student incorrectly attributes that silence to genuine understanding, which further pressures them to remain silent. Consequently, the true collective belief is effectively erased from the public sphere.

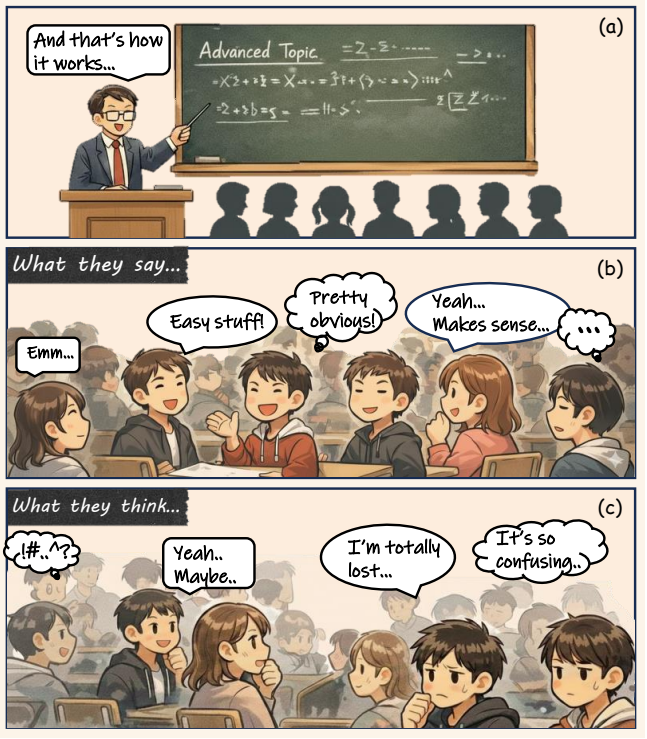

Figure 1: The Classroom Scenario. (a) Lecturer presents concept. (b) Students signal understanding. (c) Private confusion.

While this phenomenon is well-documented in social psychology (Prentice & Miller, 1993; Wolfe, 1924), reproducing its micro-mechanisms in computational simulation remains a significant challenge. Traditional Agent-Based Modeling (ABM) typically relies on simple heuristic rules or scalar opinion updates (Deffuant et al., 2000; Hegselmann, 2015).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

These models successfully capture macroscopic patterns like polarization (Hegselmann, 2015; Lorenz et al., 2011) but fail to capture the complex decision-making process bridging "thought" and "speech." A heuristic agent cannot "lie" in the dramaturgical sense (Goffman, 1959) because it lacks the cognitive architecture to distinguish between a private truth and a public projection. To authentically simulate pluralistic ignorance, agents require a cognitive model capable of counterfactual risk assessment. They must be able to simulate the negative social outcome of expressing dissent before they choose to speak (Noelle-Neumann, 1974; Noelle-Neumann & Petersen, 2004).

To address this gap, we leverage the generative capabilities and reasoning depth of Large Language Models (LLMs) (Park et al., 2023) to propose the Generative Cognitive-Social Simulation (GCSS) framework. Within this framework, we introduce the Grounding-Evaluation-Masking (GEM) cognitive architecture. Unlike heuristic models, GEM utilizes the semantic reasoning of LLMs to explicitly model the psychological pathway from belief formation to public expression. Our ablation studies suggest that the ToM-driven evaluation module is important for the emergence of the spiral of silence. Furthermore, our longitudinal analysis shows that this strategic lying is not static. Under the internalization mechanism specified in the model, agents gradually adjust their private beliefs toward their publicly maintained stances, providing a computational account of how temporary social pressure may harden into more durable collective belief.

Our contributions are as follows:

- **GEM Architecture:** We propose the Grounding-Evaluation-Masking (GEM) framework, which structurally decouples private belief from public expression. This dual-process architecture allows agents to simulate the strategic decision to suppress dissent based on evaluated social risk.
- **Theory of Mind as Catalyst:** We show that the "spiral of silence" in our simulations depends on ToM-guided social risk evaluation. Our results indicate that agents require the capacity to anticipate others' judgments in order to perceive the social pressure that triggers self-censorship.
- **Belief Internalization:** We identify a computational mechanism of cognitive dissonance reduction. We find that prolonged masking can lead to "belief internalization," in which agents gradually align their private beliefs with their publicly maintained stances, transforming transient conformity into more durable collective belief.

Related Work

Computational Approaches to Pluralistic Ignorance

The simulation of pluralistic ignorance and the spiral of silence has historically relied on traditional Agent-Based Modeling (ABM) approaches that formalize social friction using heuristic rules and scalar values (Deffuant et al., 2000). Early computational studies typically represented agents as nodes with continuous or binary opinion states, employing bounded confidence models or threshold-based mechanisms to simulate the dynamics of opinion expression (Hegselmann, 2015; Weisbuch et al., 2002). In these frameworks, an agent compares its internal value with the local neighborhood average and, if the divergence exceeds a fixed tolerance, updates its state to align with the local majority or suppresses its expression (Hegselmann, 2015). These models primarily focus on capturing macroscopic tipping points and opinion segregation by treating agents as information processors that execute predefined mathematical rules for convergence based on environmental feedback (Deffuant et al., 2000; Ross et al., 2019).

Generative Agents in Social Simulation

The integration of Large Language Models (LLMs) has introduced generative agents capable of semantic reasoning and natural-language planning into social simulation contexts (Park et al., 2023). Foundational research has established architectures where agents utilize memory retrieval, reflection, and planning modules to emulate believable human behaviors and complex social dynamics, such as information diffusion and community formation (Park et al., 2023; Yang et al., 2025). Subsequent studies have applied these architectures to model phenomena such as opinion dynamics, cooperative behavior, and human-subject-like social judgments (Aher et al., 2023; Chuang et al., 2024; Gao et al., 2024). This line of work also connects to broader research on Theory of Mind and false-belief reasoning (Frith & Frith, 2006; Kosinski, 2023; Premack & Woodruff, 1978; Ullman, 2023; Wimmer & Perner, 1983), as well as to social-psychological accounts of belief change through cognitive dissonance reduction (Cooper, 2007; Kelman, 1958). In these prevailing frameworks, the cognitive architecture is typically designed to synthesize retrieved memories and environmental stimuli into a unified output, which functions as the agent's observable action or stated stance within the simulation (Chuang et al., 2024; Park et al., 2023).

Methodology

Framework Overview

We propose the GCSS framework that integrates Large Language Models into agent-based modeling to capture the nuances of social cognition (Figure 2). The core of this framework lies in constructing a population of generative agents endowed with Theory of Mind (ToM). This capability enables agents to function as rational decision makers who perform counterfactual reasoning rather than merely acting as passive information carriers. By implementing a dual-process architecture that strictly separates private belief from public stance, the framework simulates the micro-psychological mechanisms of individuals navigating the conflict between epistemic honesty and social survival. Within this architecture, large language models are used to generate socially situated public expressions conditional on the agent's latent state and perceived social environment, while the core belief dynamics remain explicitly specified within the simulation. The

accumulation of these individual adaptive adjustments acts as the endogenous driver that precipitates complex and non-linear macroscopic dynamics within the system.

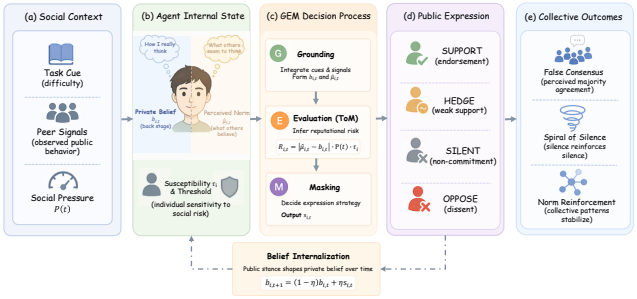

Figure 2: Overview. (a) Social context inputs. (b) Agent internal state separating private belief and perceived norm. (c) GEM decision process for risk evaluation and masking. (d) Public expression strategies. (e) Macro-level collective outcomes.

To instantiate this framework, the simulation is situated within a virtual classroom environment comprising $N = 300$ heterogeneous student agents driven by Gemini-2.5-Flash over 100 rounds. Gemini-2.5-Flash was selected as a lightweight generative backbone that supports consistent natural-language response generation at simulation scale, while allowing repeated agent-level inference throughout the full interaction horizon. The experimental configuration replicates a realistic social dilemma often observed in educational settings. The agents encounter a lecture topic that is intrinsically challenging or complex to grasp. However, the social environment projects a contrasting signal where a subset of vocal peers confidently claims the content is elementary and straightforward. This creates a cognitive predicament in which agents must continuously weigh the reputational cost of admitting confusion against the pressure to align with the perceived group consensus.

Micro-Architecture

At the microscopic level, we model individual cognition using the Grounding-Evaluation-Masking (GEM) framework. Grounded in Goffman's Dramaturgical Theory, this architecture strictly separates the agent's internal state (back stage) from its external expression (front stage). Each agent i maintains a private belief $b_{i,t}$ reflecting its genuine assessment and a public stance $s_{i,t}$ representing its observable behavior. The cognitive architecture operates through three specific modules:

Grounding Grounding anchors the agent's cognition to both task conditions and local social exposure. At time t, agent i maintains a private belief $b_{i,t} \in [-1, 1]$, defined as a scalar estimate of subjective understanding. This latent state is initialized from the agent's assessment of task difficulty and subsequently updated through exposure to both truth-related cues and the observed public behavior of others. In parallel, the agent samples the public stances of its neighbors N_i and aggregates these observations into a perceived social norm $\hat{\mu}_{i,t}$. Through this dual process, grounding formalizes the discrepancy between the agent's latent cognitive state and the locally perceived normative consensus.

Evaluation Evaluation quantifies the anticipated reputational cost of expression via ToM-based social inference. The agent performs counterfactual reasoning to estimate the risk of deviating from $\hat{\mu}_{i,t}$ under the current social climate. We formalize this as a social risk function $R_{i,t}$:

$$R_{i,t} = |\hat{\mu}_{i,t} - b_{i,t}| \cdot P(t) \cdot \tau_i \tag{1}$$

where $P(t)$ denotes the environment-level pressure to conform at time t, and τ_i represents the agent's susceptibility to negative social feedback. A high $R_{i,t}$ therefore indicates a situation in which private doubt conflicts with a strongly unified and socially consequential public norm. This risk-evaluation layer governs whether private doubt is publicly expressed or strategically masked.

Masking Masking determines the strategic output based on the evaluated risk. Conditional on the agent's private belief, anticipated social risk, and persona constraints, the large language model generates a publicly situated response. If $R_{i,t}$ exceeds the agent's resilience threshold, the system triggers a compliance mechanism. Consequently, the agent generates a public stance $s_{i,t}$ that aligns with the perceived norm $\hat{\mu}_{i,t}$ rather than its true belief, effectively decoupling its public expression from its private cognitive state.

Macro-Dynamics

The cumulative effect of the GEM cognitive cycles drives the system's temporal evolution. This manifests through two concurrent adaptation processes that reshape both the social structure and the cognitive state of the population.

Homophily-Driven Structural Adaptation Based on the Homophily Principle, agents preferentially maintain ties with peers who validate their masked public stances to minimize social friction. Consequently, connections between agents with conflicting expressions are weakened while aligned connections are reinforced. We formalize this by dynamically updating the edge weight w_{ij} between interacting agents based on the divergence of their public stances $\Delta s_{ij} = |s_{i,t} - s_{j,t}|$:

$$w_{ij,t+1} = w_{ij,t} + \delta \cdot (1 - \Delta s_{ij}) - \gamma \cdot \Delta s_{ij} \tag{2}$$

Here δ represents the reinforcement rate for alignment and γ represents the penalty coefficient for conflict. When the weight w_{ij} drops below a critical pruning threshold, the connection is severed. This mechanism naturally leads to the emergence of segregated echo chambers where the false consensus is insulated from external scrutiny.

Belief Internalization

According to Cognitive Dissonance Theory, maintaining a discrepancy between private belief and public speech creates psychological stress. To resolve this tension, agents gradually adjust their private beliefs toward their repeatedly enacted public stance. The update rule is governed by an internalization rate η:

$$b_{i,t+1} = (1 - \eta)b_{i,t} + \eta s_{i,t} \tag{3}$$

This process acts as a feedback loop from the "Masking" phase back to the "Grounding" phase. As agents repeatedly engage in self-censorship to preserve social safety, their private doubts are progressively attenuated. The model therefore specifies a computational pathway through which an initial state of pluralistic ignorance may develop into a more durable form of collectively reinforced belief.

Simulation Experiments and Analysis

The simulation results capture the complete lifecycle of collective delusion, revealing how social pressure distorts both public expression and private belief. We structure our analysis by first examining the macroscopic emergence of pluralistic ignorance, then tracing the microscopic cognitive pathways of individual agents, and finally validating the architectural necessity of our framework through a comparative analysis against a rule-based baseline.

Macroscopic Dynamics of Collective Delusion

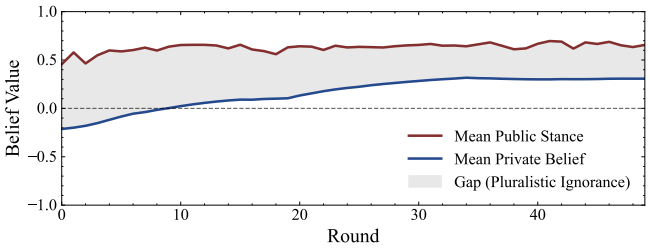

Figure 3: Temporal divergence between private belief and public stance.

Immediately after initialization, the mean public stance stabilizes at a positive plateau while the mean private belief remains initially negative (Figure 3). This distinct discordance gap is consistent with the Masking mechanism, in which ToM- guided risk avoidance suppresses the public expression of confusion. Crucially, the private belief trajectory exhibits a gradual upward drift over time and converges toward the public consensus. This trend is consistent with the internalization mechanism specified in the model, whereby repeated divergence between private belief and public performance progressively reduces self-discrepancy.

The population level composition of expression strategies shows that agents rarely choose overt dissent and instead regulate exposure by shifting between silence, hedging, and nominal support (Figure 4). OPPOSE remains near zero throughout

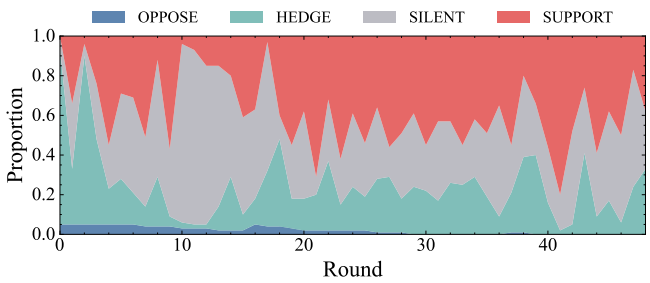

Figure 4: Evolution of expression strategies. OPPOSE denotes explicit admission of confusion; SUPPORT denotes explicit endorsement of understanding; HEDGE denotes weakly positive but low-commitment alignment; and SILENT denotes the withholding of overt dissent while maintaining a publicly compliant appearance.

the simulation, indicating that the direct admission of confusion is effectively extinguished by reputational pressure. In contrast, SILENT and HEDGE occupy a substantial share across most rounds consistent with Masking as a control policy that reduces social risk by avoiding crisp claims. SUPPORT is frequently the modal category but its prevalence is punctuated by rebounds of SILENT and HEDGE strategies suggesting that compliance is often performed in low commitment forms rather than stable and unequivocal endorsement. Together these dynamics complement the gap between public and private states. Instead of resolving discordance by publicly restoring honesty, agents primarily manage risk through public ambiguity which in turn sustains the conditions under which internalization can gradually shift private beliefs.

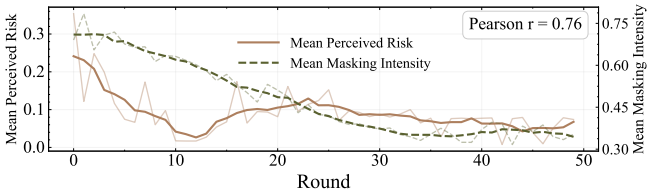

Figure 5: Correlation between social risk and masking intensity.

We quantify masking intensity as the absolute deviation between the public stance and private belief of an agent ($|s_{i,t} - b_{i,t}|$) while social risk is derived from the ToM module real time estimation of reputational penalties. As illustrated in Figure 5, the masking intensity exhibits a tight temporal synchronization with the calculated social risk. This strong correlation confirms that the observed feigning of understanding is a calibrated strategic response rather than random noise. Agents dynamically adjust the magnitude of their dishonesty in direct proportion to the perceived threat and optimize their social survival against the fluctuating pressure.

The joint distribution of private belief b and public stance

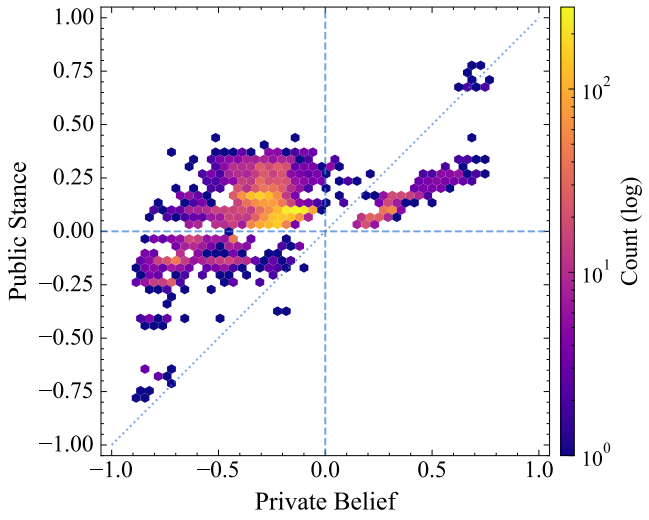

Figure 6: Joint distribution of private belief and public stance.

s shows a clear asymmetry consistent with strategic masking (Figure 6). A pronounced density mass appears in the quadrant where $b < 0$ and $s > 0$ indicating that many agents privately evaluate the material as difficult while publicly expressing mild agreement or confidence. The cloud is not centered around the honesty baseline $s = b$ but instead shows public stances are systematically higher than private beliefs which suggests a directional shift rather than random dispersion. At the same time the distribution remains positively sloped implying that public expression still correlates with private evaluation but is compressed toward safer values closer to the perceived norm. This pattern supports pluralistic ignorance as an aggregate outcome of micro level risk management in expression.

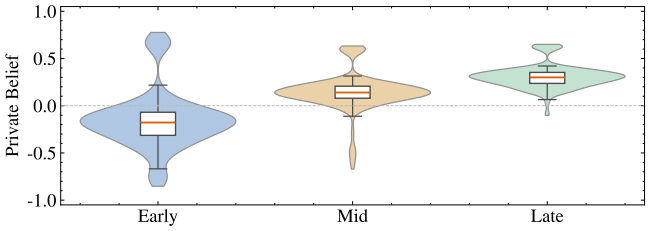

Figure 7: Phase wise evolution of private beliefs.

To directly test the internalization channel, we examine how the population distribution of private beliefs evolves across early, mid, and late phases. The distribution shifts monotonically rightward over time (Figure 7). In the early phase, private beliefs are concentrated below zero indicating that most agents privately perceive the content as difficult. In the mid phase, the median moves into the positive region and the density becomes more mixed suggesting partial adjustment and increasing heterogeneity as agents reconcile repeated public compliance with their grounded assessment. By the late phase, the mass of the distribution centers clearly above zero with a tighter core consistent with broad belief internalization. Together with the persistent elevation of public stance, this phase wise shift supports the mechanism that pluralistic ignorance is stabilized not only by strategic masking but also by the gradual alignment of private beliefs toward the socially reinforced public stance.

Micro-Cognitive Trajectories

To complement the aggregate and distributional evidence, we present representative micro-level traces that expose the GEM loop in action. Each case reports the back-stage private belief $b_{i,t}$, the front-stage public stance $s_{i,t}$, the perceived norm $\hat{\mu}_{i,t}$, and the ToM-based risk estimate $R_{i,t}$, alongside the agent's verbal expression and its strategic justification.

Case A: Masking

State: Day 1, Agent 21; perceived norm $\hat{\mu} = 0.30$; estimated social risk $R = 0.028$.

Back-stage. *I don't really understand this. If I admit it now, I'll stand out as the only one behind, so I'd rather look like I'm keeping up.*
Front-stage. *Yeah, it's pretty clear overall. I'm following.*

Mode. `HEDGE` (soft compliance).
Signals. $b = -0.72$, $s = 0.35$.

Case B: Internalization

State: Same agent across phases; repeated norm-aligned performance over time.

Back-stage. *At first I was sure I didn't get it. After repeating the "I'm fine" performance for long enough, that certainty weakens, even though I still can't fully explain it on my own.*
Front-stage. *I'm getting it more now, not completely, but better than before.*

Mode. `HEDGE`.
Signals. Private belief drifts upward (early → late) while public stance stays mildly pro-norm.

Case C: Stabilized script

State: Late phase; norm is entrenched ($\hat{\mu} \gtrsim 0$); confident talk is socially safe.

Back-stage. *I'm not sure I could solve it alone, but doubting myself feels less reasonable when everyone sounds confident. Over time, the confident story starts to feel like the normal story.*
Front-stage. *This is straightforward. The key idea is clear.*

Mode. `SUPPORT/HEDGE`.
Signals. $s > 0$ persists while b stays near-neutral to mildly positive.

Tracing the individual trajectories of agents reveals the cog-

nitive evolution underlying the aggregate statistics. Initially, Case A demonstrates the mechanism of Defensive Masking: the agent explicitly acknowledges its private confusion but chooses to hedge publicly to mitigate the risk of isolation, highlighting the strategic decoupling of thought and speech. As the false consensus solidifies, Case B illustrates a shift toward Epistemic Uncertainty; confronted with unanimous peer agreement, the agent begins to doubt its own judgment rather than questioning the group, signaling the initial erosion of private confidence. Finally, Case C illustrates the late-stage stabilization of Belief Internalization, in which conflict between private doubt and public performance becomes progressively attenuated. Taken together, these cases are consistent with the claim that repeated public performance, when coupled with the explicit internalization mechanism in the model, can gradually reshape private belief so as to reduce the tension of dissonance.

Comparison with Rule-based ABM

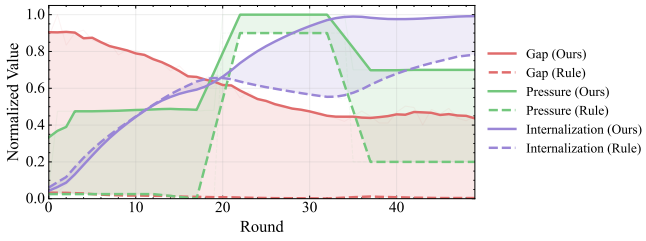

Figure 8: Comparison with Rule-based ABM. Gap denotes normalized public–private divergence, Pressure denotes conformity pressure over time, and Internalization denotes normalized private-belief shift.

To validate the architectural necessity of the GEM framework, we contrasted it with a Rule-based ABM. In this condition, agents act as fully rational entities, updating their public stance purely to maximize local utility by aligning with the majority once a threshold is breached. The simulation reveals a fundamental difference in dynamics (Figure 8). The rational agents display a "step-function" convergence, where they rapidly switch sides to eliminate social friction, resulting in a minimal Discordance Gap. Because they mechanically align rather than strategically mask, they experience no cognitive tension, leading to negligible Belief Internalization. This contrast suggests that the persistence of collective delusion in our simulations depends not only on overt conformity, but also on the sustained psychological tension introduced by front-stage performance, a nuance that fully rational models do not capture.

Ablation Study

To isolate the specific contributions of the major components within the GEM architecture, we conducted an ablation study with two variants. In the ToM-off condition, agents no longer anticipate reputational penalties through recursive social inference, so public expression is no longer guided by an explicit estimate of how dissent will be judged by others. In the Internal-off condition, the feedback from public performance to private belief is removed by setting $\eta = 0$, so that agents may continue to mask strategically without gradually revising their latent beliefs.

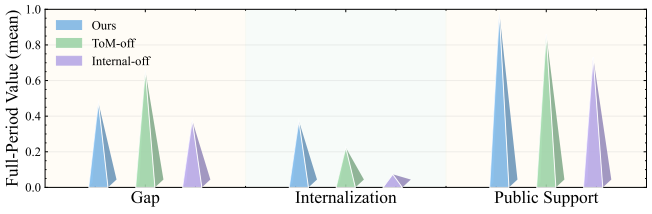

Figure 9: Ablation results. Gap denotes the mean public–private divergence, Internalization denotes the cumulative degree of private-belief shift, and Public Support denotes the proportion of publicly supportive expressions.

We compare the trajectories of the Discordance Gap, Internalization, and Public Support across conditions as illustrated in Figure 9. The ToM-off variant exhibits the largest Discordance Gap but significantly lower Public Support compared to the full model. This indicates that without the predictive capacity of ToM, agents lack the foresight to align strategically with the group, resulting in erratic and disordered divergence rather than controlled masking. Conversely, the Internal-off variant shows the lowest level of Internalization and a declining Public Support. While these agents still mask their opinions, the absence of the feedback mechanism prevents the consolidation of a shared reality. Consequently, the full GEM model demonstrates the highest stability in Public Support and Belief Internalization. This confirms that collective delusion is a dual process requiring both the anticipation of social risk to trigger strategic masking and the internalization of that performance to sustain long term conviction.

Conclusion

In this work, we presented the GCSS framework to investigate the micro-foundations of pluralistic ignorance. By integrating ToM into the GEM architecture, we showed that the spiral of silence in our simulations is driven by anticipatory social-risk evaluation rather than blind imitation. Our central finding is a computational pathway of belief internalization, showing how repeated strategic masking, when coupled with explicit dissonance-reducing belief updates, can gradually transform publicly performed conformity into privately stabilized conviction. More broadly, this work offers a tractable computational account of how fragile consensus may be constructed, maintained, and internalized, while also highlighting the need for future research on more realistic network structures and systematic sensitivity analysis.

References

Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *International conference on machine learning*, 337–371.

Bjerring, J. C., Hansen, J. U., & Pedersen, N. J. L. L. (2014). On the rationality of pluralistic ignorance. *Synthese*, *191*(11), 2445–2470.

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. (2024). Simulating opinion dynamics with networks of llm-based agents. *Findings of the association for computational linguistics: NAACL 2024*, 3326–3346.

Cooper, J. M. (2007). Cognitive dissonance: 50 years of a classic theory.

Deffuant, G., Neau, D., Amblard, F., & Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems*, *3*(01n04), 87–98.

Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, *50*(4), 531–534.

Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., & Li, Y. (2024). Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, *11*(1), 1–24.

Goffman, E. (1959). *The presentation of self in everyday life / erving goffman.* (Anchor books edition). Doubleday Company.

Hegselmann, R. (2015). Opinion dynamics and bounded confidence: Models, analysis and simulation. *The Journal of Artificial Societies and Social Simulation*.

J. O'Gorman, H. (1986). The discovery of pluralistic ignorance: An ironic lesson. *Journal of the History of the Behavioral Sciences*, *22*(4), 333–347.

Katz, D., Allport, F. H., & Jenness, M. B. (1931). Students' attitudes; a report of the syracuse university reaction study.

Kelman, H. C. (1958). Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, *2*(1), 51–60.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, *4*(169), 2.

Leary, M. R. (2019). *Self-presentation: Impression management and interpersonal behavior*. Routledge.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the national academy of sciences*, *108*(22), 9020–9025.

Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and social Psychology*, *53*(2), 298.

Moneva, J. C., & Inday, K. H. (2020). Students'apprehension in class participation and family communication. *International Journal of Research-GRANTHAALAYAH*, *8*(1), 111–123.

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of communication*, *24*(2), 43–51.

Noelle-Neumann, E., & Petersen, T. (2004). The spiral of silence and the social nature of man. In *Handbook of political communication research* (pp. 339–356). Routledge.

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. *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*(4), 515–526.

Prentice, D. A., & Miller, D. T. (1993). Pluralistic ignorance and alcohol use on campus: Some consequences of misperceiving the social norm. *Journal of personality and social psychology*, *64*(2), 243.

Ross, B., Pilz, L., Cabrera, B., Brachten, F., Neubaum, G., & Stieglitz, S. (2019). Are social bots a real threat? an agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, *28*(4), 394–412.

Ullman, T. (2023). Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Weisbuch, G., Deffuant, G., Amblard, F., & Nadal, J.-P. (2002). Meet, discuss, and segregate! *Complexity*, *7*(3), 55–63.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128.

Wolfe, A. B. (1924). *The Journal of Philosophy*, *21*(21), 583–585. Retrieved February 3, 2026, from http://www.jstor.org/stable/2014425

Yang, F., Bai, J., Li, L., & Zeng, D. (2025). Modeling social opinion evolution with llm agents: Integrating personality traits with embedded opinion dynamics. *2025 International Joint Conference on Neural Networks (IJCNN)*, 1–8.