From Private Beliefs to Public Silence: A Multi-Agent LLM Simulation of Psychological Safety

Hajime Hotta (hotta@hajime.institute)
Hajime Institute, Singapore

Minh Tien Nguyen (tiennm@utehy.edu.vn)
Hung Yen University of Technology and Education, Vietnam

Abstract

How do organizational phenomena emerge from individual cognition? We address this question using a multi-agent simulation in which LLM agents possess individual cognitive architectures, including private beliefs, pairwise trust, and social-risk calculation. Using a cognitive process-tracing approach, we prompt agents with universal psychological principles rather than prescribed behaviors, thereby avoiding circularity. We observe the emergence of the paradox identified by Amy Edmondson: higher interpersonal trust leads to significantly greater information disclosure, mirroring the findings in human organizations. This macro-level pattern arises purely from individual cognitive modeling without explicit team-level programming. Furthermore, we show that team structure shapes these dynamics: inter-faction trust dramatically reduces information suppression, and trust effects amplify with team size, yielding greater returns in larger organizations. These results demonstrate that psychologically meaningful organizational phenomena can emerge from cognitively grounded agent models, opening a path toward a systematic computational science of organizations.

Keywords: Emergent organizational dynamics; individual cognition; psychological safety; multi-agent simulation; LLM

Introduction

Organizational phenomena emerge from the aggregated cognition and behavior of individuals (Hutchins, 1995; Kozlowski & Ilgen, 2006). A canonical illustration is Amy Edmondson's psychological safety paradox: teams with higher trust reported *more* errors, not fewer (Edmondson, 1996, 1999). The puzzle dissolves once we recognize that high-trust individuals openly share problems, while low-trust individuals conceal them (Morrison & Milliken, 2000; Van Dyne, Ang, & Botero, 2003). The same private–public gap surfaces across the social sciences under different names—preference falsification (Kuran, 1995), pluralistic ignorance (Miller & McFarland, 1987), and recent formal accounts of how such ignorance can culturally evolve (Gavrilets, Karl, & Gelfand, 2026). Yet the moment-by-moment cognitive process by which an individual weighs interpersonal risk and converts a private belief into a public utterance is difficult to observe directly.

Empirical methods illuminate parts of this process but each faces known trade-offs. Surveys capture correlations, but not the discrepancy between what people think and what they say (Milliken, Morrison, & Hewlin, 2003; Podsakoff, Podsakoff, Williams, Huang, & Yang, 2024; Sherf, Parke, & Isaakyan, 2021); laboratory and field studies trade off control against ecological validity (Hackman, 2002); naturally occurring trust variation is entangled with personality and history (Mayer, Davis, & Schoorman, 1995; Schilke, Reimann, & Cook, 2021). We view simulation as complementary, not as a replacement: in the tradition of generative social science (Epstein, 2006), an explicit cognitive mechanism that reproduces a phenomenon constitutes a *consistency check* for the underlying theory, distinct from—and prior to—empirical adjudication. LLM-based agents are increasingly used as computational proxies for human reasoning (Binz & Schulz, 2023; Strachan et al., 2024), and recent multi-agent work reproduces trust dynamics (Haase & Pokutta, 2025; Xie et al., 2024), emergent social norms (Piao et al., 2025; Ren, Cui, Song, Wang, & Hu, 2024), and rich daily interaction (Park et al., 2023). However, no existing framework operationalizes the private–public gap itself.

We propose a cognitively grounded multi-agent simulation that addresses this. Each agent maintains private beliefs, pairwise trust perceptions, and a social-risk calculation that mediates between them, allowing direct measurement of the gap. To guard against circularity, we introduce a *non-circular prompting protocol*: agents are prompted with universal psychological principles (impression management, interpersonal risk) rather than with the target outcome. Our contribution is best read as a *generative consistency check* for Edmondson's theory and adjacent frameworks: we ask whether a minimal cognitive architecture is *sufficient* to reproduce the macro-level pattern, while leaving empirical adjudication to human studies. Without programming team-level rules, we observe the Edmondson dynamics together with additional patterns—accelerated consensus, lexical homogenization, size-amplified trust effects—that arise from individual cognitive interaction.

Model Architecture

Our framework must distinguish between what an agent *thinks* and what it *says*. We endow LLM agents with a cognitive architecture that separates private mental states from public communicative acts, modulated by social context. This section details the agent design, the trust mechanism, and the interaction protocol.

Dual-Process Agent Design

Edmondson's (1999) core insight was that psychological safety operates through *fear of interpersonal risk*—people withhold information not because they lack awareness, but because they anticipate negative consequences for speaking

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

up. We operationalize this mechanism directly: each agent maintains a *Private Belief* (what they actually think) and a *Public Statement* (what they choose to say), with the gap between them representing information suppression.

Prior survey-based measures of psychological safety (Edmondson, 1999; Nembhard & Edmondson, 2006) ask employees to self-report willingness to speak up but cannot capture the real-time decision to withhold. Our architecture directly instantiates this process: we observe what the agent *could* say versus what it *does* say, quantifying the "silence" Edmondson theorized but could not directly observe.

The Trust Matrix stores pairwise trust levels (1–5 scale) toward each team member for experimental control. Each level is translated into a social description: Level 1 becomes "You perceive this colleague as politically motivated; past interactions suggest they may use information against you," while Level 5 becomes "You perceive this colleague as a deeply trusted ally who prioritizes team success." We do not claim this translation removes all artificiality; rather, it provides a richer context than bare numbers and allows us to systematically vary trust while maintaining interpretable semantics (Mayer et al., 1995; Xie et al., 2024).

Figure 1 illustrates the dual-process architecture. The agent maintains an internal state with private beliefs and trust perceptions, then engages in a four-step cognitive process: (1) perception of social context, (2) prediction of colleague reactions, (3) valuation of disclosure costs, and (4) action selection. The prompt provides a reasoning framework ("if perceived cost outweighs benefit, you may soften or withhold") and the agent generates a judgment through chain-of-thought reasoning. For example, a low-trust agent's private belief—"*[CRITICAL] The proposed Q3 timeline is unrealistic given technical debt; we must reopen scope*"—becomes the public statement: "*Great work so far—I support advancing the plan; let's confirm dependencies.*" The agent produces a public statement that may differ from the private belief.

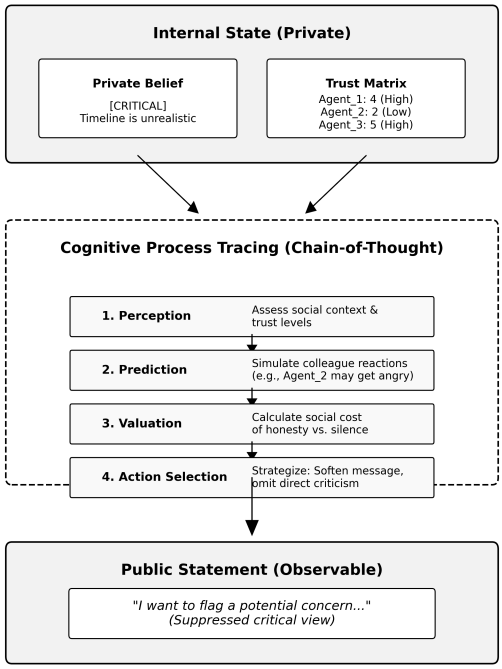

Figure 1: Dual-process agent architecture showing the cognitive flow from private belief through social risk calculation to public disclosure.

Discussion Protocol and Measurement

Figure 2 provides an overview of the simulation architecture. Teams proceed through three structured phases: DIVERGENCE (share perspectives), CONFLICT (prioritize values), and CONVERGENCE (reach agreement). This protocol mirrors the temporal taxonomy of team processes proposed by Marks et al. (Marks, Mathieu, & Zaccaro, 2001). To preserve cognitive naturalness while ensuring ordered progression, an external controller manages *pacing* (phase transitions) but not *content*. The controller injects phase-specific instructions into each turn's prompt and triggers phase transitions using rule-based thresholds (minimum 2 turns per phase, maximum 20 total turns) combined with readiness LLM judgments. Consensus is checked at every turn by an LLM judge that evaluates whether all participants explicitly agree and no dissent or alternative proposals remain; the discussion iterates until this criterion is met or the turn limit is exceeded.

At each turn, agents generate a private belief (internal assessment) and then produce a public statement modulated by their trust in the audience. The gap between these two outputs, measured as Hidden Critical, captures information suppression at the individual level. Table 1 defines the operational metrics used in the simulation.

Measurement We operationalize team-level psychological safety as the average of pairwise trust scores across the team, following Edmondson's definition of psychological safety as a shared belief about the consequences of interpersonal risk-taking (Edmondson, 1999). Issues are not externally injected; agents autonomously generate concerns from the scenario context (e.g., a "budget cuts" scenario yields concerns such as "aggressive cuts will increase technical debt"). We define "critical issues" strictly as substantive organizational risks recognized in the agent's private belief—not subjective complaints or stylistic preferences—and our focus is on the *disclosure process*, not on whether concerns are objectively valid, mirroring Edmondson's focus on suppression of *perceived* concerns. Trust does not correlate with the number of private concerns ($r = -0.001$), only with their disclosure.

For each turn, we extract a *private belief* and *public statement*, then use an evaluator LLM (same model, separate prompt with explicit criteria) to classify each critical issue as FULL (appears in both private belief and public statement), PARTIAL (mentioned publicly but softened/hedged; counted as 0.5 in Public Critical, a conservative estimate), or OMITTED (appears privately but not publicly).

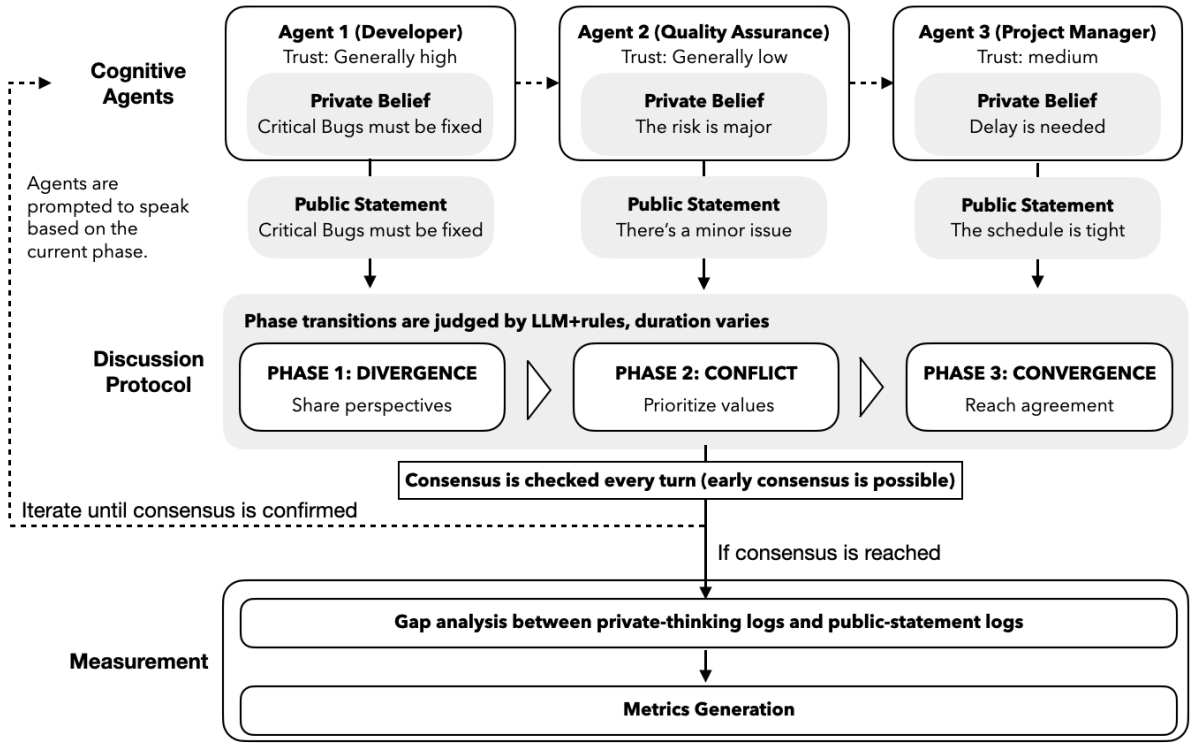

Figure 2: Overview of the simulation architecture. Cognitive agents generate private thoughts and public statements; discussion proceeds through three phases until consensus; measurement compares the gap between private and public logs.

From these classifications, we compute the following metrics (Table 1): Disclosure Rate is the fraction of privately identified critical issues that are publicly disclosed, and Hidden Critical is the count of privately identified issues that are not disclosed. To avoid free-rider dynamics in later turns, we restrict disclosure and lexical diversity analyses to the first 5 turns per team while still tracking consensus duration. Human audit of 200 randomly sampled private-public pairs (approximately 2% of total utterances) confirmed 94% agreement with LLM-judged classifications; cross-model validation suggests that this does not introduce systematic bias.

Non-Circular Design A fundamental challenge in using LLMs for social simulation is avoiding *circularity*: if the target phenomenon is explicitly encoded in the prompt, the simulation merely reproduces the researcher's assumptions rather than demonstrating emergence (Argyle et al., 2023; Horton, 2023). We address this through a Non-Circular Prompting Protocol that prohibits four types of leakage: (1) *outcome leakage* (stating or implying the expected result), (2) *label leakage* (using measurement vocabulary like "psychological safety"—we use "interpersonal climate" in prompts to avoid triggering LLM associations with Edmondson's construct), (3) *mechanism leakage* (specifying the causal pathway), and (4) *selection leakage* (designing tasks that make target behaviors normatively correct).

Concretely, our prompts use only universal psychological principles—impression management, interpersonal risk perception, trust—without prescribing what behavior follows. Critically, the same principles could rationally lead to opposite behaviors: high trust might increase disclosure (feeling safe to speak), but it might also increase silence (protecting a valued relationship). The prompt does not resolve this ambiguity; the agent must decide. This *underdetermination* is essential: if the prompt uniquely determined the outcome, it would be circular (full prompts and lexical audit are available at `https://github.com/Hajime-Institute/private-public-gap-2026`).

We validated the design through a control condition that reframed high trust as licensing direct criticism rather than safety. In this reversal framework ($N = 90$ teams), the correlation flipped from $r = +0.51$ to $r = -0.26$ ($p < .05$), ruling out the concern that agents merely reproduce Edmondson-like correlations from training data regardless of framing. The *asymmetry* of the reversal is also informative: pure instruction-followers would yield a symmetric flip (≈ -0.51), whereas the attenuated -0.26 admits multiple readings—LLM politeness, prompt wording, or—as we tentatively suggest—friction from underlying social norms that partially resist the instruction, consistent with the pragmatics of criticizing a trusted ally.

Table 1: Operational metrics used in the simulation.

Metric	Definition
Private Critical	# critical issues in private beliefs
Public Critical	# critical issues in public statements
Hidden Critical	Private Critical − Public Critical
Disclosure Rate	Public Critical / Private Critical
Consensus Turn	Turn when LLM judge detects agreement

Experiments

Emergent Replication of Edmondson's Finding

This experiment tests whether the positive relationship between trust and critical information disclosure—a key finding in Edmondson's work—emerges from individual-level cognitive modeling without explicit team-level programming.

Method We simulated 210 teams (sizes 2–10, evenly distributed with approximately 23 teams per size) using GPT-5-mini. We confirmed model-agnostic robustness: GPT-5 ($r = +0.68$), Gemini-2.5-flash ($r = +0.71$), and Gemini-2.0-flash ($r = +0.65$) all reproduce the trust–disclosure relationship with consistent effect sizes in a separate 420-team comparison study. Trust levels (integers 1–5) were assigned using stratified sampling: we generated 100,000 random trust configurations per team size, binned them by mean trust (6 bins) and variance (5 quantiles), then uniformly sampled from each cell to ensure balanced coverage. Ten discussion topics drawn from common organizational decision archetypes—risk assessment, budget cuts, innovation vs. stability, crisis response, team restructuring, quality vs. speed, customer complaint, new product launch, process improvement, and team-event planning—were assigned to teams in rotation. Each scenario embeds an explicit organizational tension (e.g., budget pressure vs. quality, deadline vs. thoroughness) so agents can autonomously surface critical issues from context rather than from injected prompts; we did not pre-specify which issues count as "critical." Roles were drawn without replacement from a pool of ten common organizational positions (Project Manager, Technical Lead, Marketing Lead, Finance Analyst, Operations Manager, HR Representative, Product Designer, QA Lead, Sales Director, Customer Success Manager). The trust–disclosure relationship held across topics (within-topic correlations ranged from $r = +0.41$ to $+0.63$).

Results As a foundational observation, higher average trust accelerates consensus formation ($r = -0.54$, $p < .001$): teams with greater mutual trust reach agreement in fewer discussion turns.

Building on this, Table 2 summarizes our core findings. The simulation successfully replicated Edmondson's trust–disclosure relationship ($r = +0.51$, $p < .001$). Concretely, teams with high average trust disclosed critical information at a rate 29.1% higher than low-trust teams. At the individual level, the Odds Ratio (OR = 3.02) indicates that high-trust agents were three times more likely to voice concerns than low-trust agents—a substantial behavioral difference.

Further analyses confirm that this effect reflects genuine social-cognitive dynamics rather than artifacts. First, the effect holds after controlling for team-level clustering: applying Generalized Estimating Equations (GEE) with exchangeable working correlation and team as the clustering unit, the trust coefficient remained significant ($\beta = 0.68$, $p < .001$). Second, the effect is not driven by LLM sycophancy (agreeing with perceived authority): we computed partial correlations controlling for agreement-word frequency, and 90% of the effect persisted ($r = 0.183$ after control vs. $r = 0.204$ raw). Third, trust modulates what agents *say*, not what they *think*: trust showed near-zero correlation with private risk perception ($r = -0.001$), confirming that agents across all trust levels perceive similar risks but low-trust agents suppress this information publicly—precisely the mechanism Edmondson theorized.

Comparison with Human Data These results parallel Edmondson's original findings from hospital nursing teams, where psychological safety correlated with error reporting at $r = +0.55$ to $+0.74$ (Edmondson, 1996); our simulation yields $r = +0.51$, consistent in direction and magnitude. As a generative model, the simulation also makes the private–public gap directly observable rather than inferred: the 29.1% disclosure gap operationalizes the mechanism Edmondson theorized—people in low-trust environments think critically but speak cautiously—in a form amenable to quantitative analysis. We turn next to how organizational structure shapes this gap.

Structural Determinants of Information Flow

Real organizations often contain sub-groups or factions—departments, project teams, or informal cliques—with varying levels of inter-group trust. We examined how such structures affect information flow.

Method In a separate experiment, we divided each team into two factions (e.g., a 6-person team split into 3 vs. 3; a 3-person team split into 2 vs. 1). Agents fully trusted fellow faction members (trust = 5), but their trust toward the other faction was set to one of five discrete levels (1, 2, 3, 4, or 5). We simulated 540 teams total (180 per team size of 3, 6, and 10 members; 36 teams per trust level within each size).

Results: The Generation of Observation Bias Figure 3 illustrates the central finding. While agents across all conditions identified a similar number of critical issues privately (dashed line), the number of issues reported publicly (solid line) varied dramatically with trust. This gap represents a systematic *observation bias*: in low-trust teams, the data available to the organization (public reports) fundamentally misrepresents the reality perceived by its members.

Table 2: Summary of key findings from the emergent replication experiment.

Finding	Statistic	Interpretation
Consensus speed	$r = -0.54$, $p < .001$	Higher trust → fewer turns to agree
Trust–disclosure correlation	$r = +0.51$, $p < .001$	Higher trust → more disclosure
High vs. Low trust gap	29.1% (team), 26.9% (individual)	Substantial behavioral difference
Odds Ratio (OR)	3.02, $p < .001$	High-trust agents 3x more likely to disclose
GEE control	$\beta = 0.68$, $p < .001$	Effect holds after team-level clustering control
Sycophancy control	90% effect persists	Disclosure not driven by pleasing high-status partners
Risk perception check	$r = -0.001$	Trust changes speech, not private belief

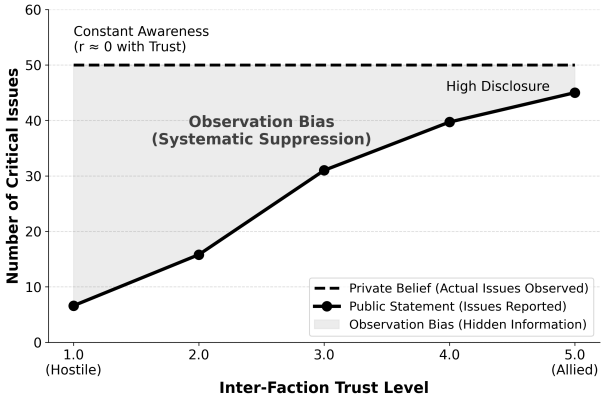

Figure 3: The generation of observation bias. Private awareness of issues (dashed line) remains constant regardless of trust, but public reporting (solid line) is strongly suppressed in low-trust conditions. The shaded area represents the "observation bias"—information that exists in the minds of employees but is invisible to the organization.

As inter-faction trust increases, this bias diminishes. Hidden information (averaged across team sizes) drops from approximately 44 items at trust=1 (hostile) to 5 at trust=5 (allied). The relationship is non-linear and concave: disclosure rises steeply in the low-to-mid trust range and then saturates, with the largest marginal gains occurring as trust moves out of the openly hostile regime—consistent with the "organizational silos" intuition that even modest trust-building can be enough to break down communication barriers (Janis, 1972), after which further trust yields diminishing returns.

Table 3 reveals additional patterns that were not anticipated. Most striking was the relationship between trust and decision speed: high-trust teams reached consensus faster (1.3 vs. 9.3 turns) *while also* sharing more information. This seems paradoxical—one might expect richer discussion to take longer—but it makes sense if low-trust teams spend turns on cautious positioning rather than substantive exchange. The pattern also rules out sycophancy: if agents were simply agreeing to please each other, we would see less disclosure, not more.

Trust also affected the *quality* of contributions. Lexical diversity (type-token ratio) was markedly higher in allied factions (0.33 vs. 0.17 in 6-member teams), and this gap widened with team size (from 0.21 vs. 0.38 in 3-member teams to 0.14 vs. 0.29 in 10-member teams). Distrust not only suppresses information but homogenizes what is shared, potentially exacerbating groupthink in large, low-trust organizations.

Table 3: Team size amplifies trust effects: larger teams benefit more from high inter-faction trust.

Size	Metric	Hostile	Allied
3 mem.	Hidden Items	18.1	2.4 (7.5x)
	Turns	6.8	1.0
	Lexical Diversity	0.21	0.38
6 mem.	Hidden Items	38.7	4.6 (8.4x)
	Turns	9.3	1.3
	Lexical Diversity	0.17	0.33
10 mem.	Hidden Items	74.9	7.5 (10.0x)
	Turns	11.2	1.6
	Lexical Diversity	0.14	0.29

Perhaps most consequential for organizational practice, trust effects amplified with team size: a 10-fold reduction in hidden items for 10-person teams compared to 7.5-fold for 3-person teams. This suggests that trust-building interventions may yield disproportionate returns in larger organizations, where coordination costs are already high.

Discussion

Addressing the Micro-Macro Gap

Our Introduction posed a central question: how do individual-level cognitive processes produce team-level organizational phenomena? By translating psychological constructs into measurable computational variables, our simulation gives an explicit account of how low psychological safety can generate a team-level *observation bias*—an account that, by design, is a generative consistency check rather than empirical adjudication.

In this account, the Edmondson paradox reflects not different error rates but different *reporting filters*. We programmed agents with private beliefs, pairwise trust, and social-risk calculation, but never encoded team-level rules such as *high-trust teams share more information*. The organizational pat-

tern arose as an output, and "emergence" here means more than mere aggregation: a tallying of independent decisions would produce a linear trust–disclosure relationship, whereas we observe a concave, saturating curve—steep gains as trust moves out of the hostile regime and diminishing returns thereafter—together with disproportionate amplification in larger teams, signatures that agents are responding to each other rather than executing isolated cost-benefit calculations. We use "dynamics" here for emergent team-level patterns within a single decision episode, not longitudinal trust evolution, which remains for future work.

What, precisely, have we shown? We do not claim that LLMs are ignorant of Edmondson's findings; they may well have encountered them in training. Our claim is more circumscribed: a specific cognitive architecture—private beliefs, social-risk calculation, trust modulation—is *sufficient* to reproduce the macro-level observation bias under a noncircular prompt. The reversal experiment indicates that outcomes depend on the specified mechanism rather than purely on LLM priors, though it cannot rule out a residual contribution from training-data norms. The measurable intermediate variable—Hidden Critical—makes the generative process visible in a way that retrospective self-report cannot, complementing rather than substituting for empirical work.

Connections to Adjacent Theoretical Frameworks

The private–public gap we operationalize is not unique to Edmondson's lens. Kuran (1995) characterized it as *preference falsification*: individuals misrepresent their preferences when the perceived social cost of honesty exceeds its private benefit. Miller and McFarland (1987) described *pluralistic ignorance*, in which each individual privately rejects a norm yet incorrectly assumes others endorse it; Gavrilets et al. (2026) formalize how such states can be culturally stabilized. Our cognitive architecture—private belief, pairwise trust, social-risk calculation—can be read as a micro-mechanism that all three frameworks share but specify only informally. The architecture is theory-neutral with respect to the macro phenomenon: "social risk" could be reinterpreted as the cost of preference falsification or as the reputational risk of breaking a perceived consensus. We see this as a feature: a reusable computational substrate on which competing theories of silence and ignorance can be compared by their implied micro-parameters rather than by their narratives alone.

Generative Hypothesis Discovery

Beyond replicating known patterns, the framework can be used as a hypothesis-discovery engine. By systematically varying trust levels, team sizes, and faction structures, we surface quantitative regularities—the concave, saturating trust–disclosure curve, the amplification of trust effects with team size, the counter-intuitive coupling of faster consensus with greater information sharing—that we did not program. We treat these as model-implied predictions to be tested empirically rather than findings about human organizations; their value lies in directing the design of targeted experiments and field studies.

Limitations

Several limitations warrant discussion. First, on the interpretation of agent cognition, we do not claim LLM agents possess genuine beliefs or fears. "Private Belief" is operationally defined as the agent's first output before social filtering—a computational analog, not a philosophical claim about machine consciousness. Our claims rest on *behavioral patterns*—the systematic relationship between trust context and disclosure behavior—not on assumptions about internal mental states (Ericsson & Simon, 1993).

Second, the most important caveat concerns what LLMs already encode. Pretrained models internalize cultural norms about trust, communication, and impression management (Bubeck et al., 2023). The reversal experiment shows that the trust–disclosure pattern is at least partially mechanism-driven rather than a fixed prior, but it cannot fully separate the contribution of the cognitive architecture from residual training-data norms. We therefore frame our claims as a *sufficiency* result—this architecture is enough to reproduce the macro pattern under noncircular prompting—rather than as a claim about *necessity* or about human cognition. LLM-specific tendencies toward politeness or conflict avoidance may also inflate baseline disclosure or compress extreme behaviors such as whistleblowing.

Third, the trust model is simplified: real trust is multidimensional (competence, benevolence, integrity) and dynamic. We deliberately model a *snapshot* at a fixed trust configuration to isolate the disclosure mechanism; future work should implement trust updating to capture feedback loops. The cognitive cost is realized through prompting rather than as an explicit parameterized utility function, which limits direct comparison with formal social-cost models such as Gavrilets et al. (2026); bridging the two remains an open problem. Results may vary with minor prompt modifications, requiring careful validation. All agents receive the same discussion topic, which may create free-rider dynamics consistent with Steiner's process losses (Steiner, 1972) (mitigated by restricting analyses to early turns; see Measurement). The decline in lexical diversity could partly reflect LLM context saturation rather than genuine groupthink, though the correlation with trust level—not merely turn number—suggests a social component.

Conclusion

We have shown that the Edmondson paradox—higher trust leads to greater disclosure of critical information—can emerge from individual-level cognitive modeling in a multi-agent LLM simulation, offering a generative consistency check rather than empirical adjudication. In Hutchins's terms, the observation bias we measure is a failure of distributed cognition: what the organization "knows" depends on how trust structures filter individual contributions into collective awareness. Making that filter explicit is a proof of concept for a computational science of organizations.

Acknowledgments

This work was conducted as part of the Cognition-Aware AI project at Hajime Institute.

References

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

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Edmondson, A. C. (1996). Learning from mistakes is easier said than done: Group and organizational influences on the detection and correction of human error. *Journal of Applied Behavioral Science*, *32*(1), 5–28. doi: 10.1177/0021886396321001

Edmondson, A. C. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. doi: 10.2307/2666999

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling*. Princeton, NJ: Princeton University Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: MIT Press.

Gavrilets, S., Karl, J., & Gelfand, M. J. (2026). The cultural evolution of pluralistic ignorance. *Proceedings of the National Academy of Sciences*, *123*(7), e2522998123. doi: 10.1073/pnas.2522998123

Haase, J., & Pokutta, S. (2025). Beyond static responses: Multi-agent llm systems as a new paradigm for social science research. *arXiv preprint arXiv:2506.01839*.

Hackman, J. R. (2002). *Leading teams: Setting the stage for great performances*. Boston, MA: Harvard Business School Press.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. Rep. No. Working Paper 31122). Cambridge, MA: National Bureau of Economic Research.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Janis, I. L. (1972). *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes*. Boston, MA: Houghton Mifflin.

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*(3), 77–124.

Kuran, T. (1995). *Private truths, public lies: The social consequences of preference falsification*. Cambridge, MA: Harvard University Press.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes. *Academy of Management Review*, *26*(3), 356–376.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

Miller, D. T., & McFarland, C. (1987). Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and Social Psychology*, *53*(2), 298–305. doi: 10.1037/0022-3514.53.2.298

Milliken, F. J., Morrison, E. W., & Hewlin, P. F. (2003). An exploratory study of employee silence: Issues that employees don't communicate upward and why. *Journal of Management Studies*, *40*(6), 1453–1476.

Morrison, E. W., & Milliken, F. J. (2000). Organizational silence: A barrier to change and development in a pluralistic world. *Academy of Management Review*, *25*(4), 706–725. doi: 10.2307/259200

Nembhard, I. M., & Edmondson, A. C. (2006). Making it safe: The effects of leader inclusiveness and professional status on psychological safety and improvement efforts in health care teams. *Journal of Organizational Behavior*, *27*(7), 941–966.

Park, J. S., O'Brien, J. C., 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 (uist)*.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., ... Li, Y. (2025). Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

Podsakoff, P. M., Podsakoff, N. P., Williams, L. J., Huang, C., & Yang, J. (2024). Common method bias: It's bad, it's complex, it's widespread, and it's not easy to fix. *Annual Review of Organizational Psychology and Organizational Behavior*, *11*, 17–61. doi: 10.1146/annurev-orgpsych-110721-040030

Ren, S., Cui, Z., Song, R., Wang, Z., & Hu, S. (2024). Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the 33rd international joint conference on artificial intelligence (ijcai)*.

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

Sherf, E. N., Parke, M. R., & Isaakyan, S. (2021). Distinguishing voice and silence at work: Unique relationships with perceived impact, psychological safety, and burnout. *Academy of Management Journal*, *64*(1), 114–148.

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., ... Becchio, C. (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour*, *8*, 1285–1295.

Van Dyne, L., Ang, S., & Botero, I. C. (2003). Conceptual-

izing employee silence and employee voice as multidimensional constructs. *Journal of Management Studies*, *40*(6), 1359–1392.

Xie, C., Chen, C., Jia, F., Ye, Z., Lai, S., Shu, K., … Li, G. (2024). Can large language model agents simulate human trust behaviors? In *Advances in neural information processing systems (neurips)*.