Beyond Idealized Cooperation: Epistemic Noise in Collaborative Problem Solving

Videep Venkatesha
Colorado State University

Ethan Seefried
Colorado State University

Changsoo Jung
Colorado State University

Nathaniel Blanchard
Colorado State University

{Videep.Venkatesha, Ethan.Seefried, Changsoo.Jung, Nathaniel.Blanchard} @colostate.edu

ABSTRACT

AI agents that support collaborative learning must model the phenomena that drive productive group work, including communicative acts, belief formation, trust dynamics, and reasoning quality, as they unfold in real time. Existing collaborative datasets capture idealized cooperation under shared epistemic goals, but the dynamics pervasive in real classrooms, such as misconceptions, unreliable contributions, overconfident errors, and their downstream effects on group reasoning, remain largely absent and unlabeled. We propose social deduction games, specifically Secret Hitler, as a proxy testbed where collaborative problem solving and controlled injection of unreliable information co-occur with researcher-accessible ground truth. The game's role structure produces contributions that are functionally equivalent to classroom misconceptions and misdirection, but with known provenance, enabling traceable study of how unreliable information propagates through groups. We present a multi-layer annotation framework that captures communicative acts using a validated persuasion strategy taxonomy, player-reported belief states yielding first- and second-order Theory of Mind data, and post-hoc epistemic vigilance assessments of whether belief updates were warranted given available evidence. We pair this framework with multi-person egocentric sensing using Meta Aria research glasses alongside exocentric recording, providing synchronized capture of gaze, attention, and social signals from every participant's perspective. Pilot data from two 6-player, 9-round sessions demonstrates that failed epistemic vigilance, belief fragmentation, and traceable misinformation effects emerge naturally and are captured by the framework.

Keywords

epistemic vigilance, collaborative problem solving, multimodal learning analytics, Theory of Mind, social deduction games, belief dynamics

Videep Venkatesha, Ethan Seefried, Changsoo Jung, and Nathaniel Blanchard. Beyond Idealized Cooperation: Epistemic Noise in Collaborative Problem Solving. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 600–604. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039891

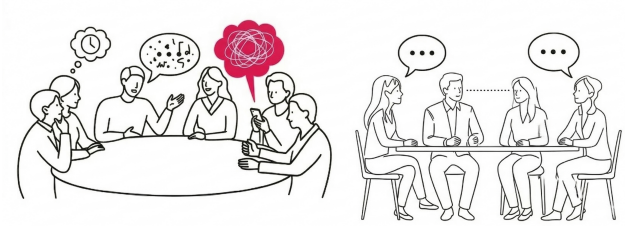

Figure 1: Real collaboration (left) involves heterogeneous contributions, competing attention, and participants who may be disengaged or misinformed—dynamics pervasive in collaboration but largely absent from controlled datasets. Idealized settings (right) capture only cooperative interaction. Our testbed produces the dynamics on the left with the ground truth needed to study them.

1. INTRODUCTION

AI agents that facilitate collaborative learning must model the phenomena that drive productive group work as they unfold: the *communicative acts* members perform, the *beliefs* they hold and revise, the *reasoning quality* with which they evaluate one another's contributions, and the *trust dynamics* that determine whose voice carries weight [2, 15, 18]. Developing models that capture these phenomena requires extensive iteration on sensing, annotation, and inference that would consume scarce instructional time if conducted in classrooms. Instead, we need proxy testbeds: controlled environments where groups collaborate naturally, producing multimodal signals (speech, gaze, gesture, posture), but where data collection serves system development rather than student learning.

Existing testbeds such as the DELI corpus [9] and Physics Playground [18] capture multi-party deliberation and collaborative problem solving (CPS), but under shared epistemic goals where all participants work toward the right answer. Real collaboration is messier: students unknowingly spread misconceptions, assert incorrect solutions with unwarranted confidence, freeload, or dominate discussions unproductively [4]. These dynamics of *unreliable contributions*—not intentional sabotage, but the natural heterogeneity of contributor quality—are pervasive in classrooms [10, 14] yet rare in existing datasets.

We propose social deduction games, specifically *Secret Hitler*,

as a proxy testbed that fills this gap. Players are secretly assigned to cooperative or disruptive teams and engage in structured rounds of discussion, voting, and policy enactment. For the majority of players, the game *is* collaborative problem solving: pooling incomplete evidence, evaluating reliability of claims, and coordinating decisions under genuine uncertainty about who is trustworthy [15, 18]. What makes the setting valuable is that a minority of players are structurally incentivized to introduce unreliable information—spreading misinformation, asserting false claims with confidence, and undermining group reasoning. These contributions are *functionally equivalent* to the misconceptions, overconfident errors, and unproductive dominance documented in classroom groups [3]: the collaborative challenge they pose is the same, even though the underlying intent differs. Crucially, every piece of unreliable information has known provenance (role assignments provide ground truth), and its effects on group beliefs are traceable through our annotation framework.

We introduce a multi-layer annotation framework targeting progressively deeper social-cognitive constructs. Layer 1 classifies communicative acts using a validated persuasion strategy taxonomy [11]. Layer 2 collects player-reported belief states after each round, yielding first- and second-order Theory of Mind data grounded in Baker et al.'s [2] framework. Layer 3 infers epistemic vigilance post-hoc [16], the capacity to evaluate communicated information for source reliability and content plausibility rather than accepting it uncritically, assessing whether belief updates were warranted given available evidence. The cascading structure of communicative act → belief update → vigilance assessment—captures the trajectory from speech to social-cognitive outcome that a classroom agent must track. We pair this framework with multimodal sensing using Meta Aria egocentric research glasses [7], instrumenting every player to capture synchronized video, eye tracking, and inertial data from first-person perspectives.

Our contributions are:

- A multi-layer annotation framework integrating communicative act classification, player-reported belief states (first- and second-order Theory of Mind), and post-hoc epistemic vigilance inference for coding group social-cognitive dynamics.
- A positioning of Secret Hitler as a CPS environment with controlled injection of unreliable information with known provenance, addressing a gap left by datasets capturing only idealized cooperation.
- Pilot data from two 6-player sessions demonstrating that failed vigilance, belief fragmentation, and traceable misinformation effects emerge naturally and are captured by the framework.

2. SECRET HITLER AS A CPS TESTBED

In Secret Hitler, 5 to 10 players are secretly assigned roles as Liberals or Fascists, with one Fascist designated as Hitler. Each round, a President nominates a Chancellor; the group votes to approve or reject; and if approved, the government enacts a policy from a constrained deck. Liberals win by enacting five Liberal policies or assassinating Hitler; Fascists win by enacting six Fascist policies or electing Hitler as Chancellor.

For the majority of players (Liberals), the game *is* collaborative problem solving: they must externalize private information, build shared representations of the game state, and negotiate coordinated action, all without knowing who their allies are. This maps onto the joint problem space [6, 15, 19] operationalized as "constructing shared knowledge" in the General Competency Model [18]. Critically, the game enforces the information asymmetry that the hidden profile literature identifies as the hardest part of real collaboration [12, 17], a challenge most CPS testbeds sidestep. Discussion phases require players to surface disagreements, provide evidence, and converge on action [3, 8], instantiating the argumentative structure that predicts decision quality [9, 13, 20].

What makes the setting distinctive is not that it introduces dynamics foreign to CPS, but that it makes *visible and labeled* the kinds of unreliable contributions that pervade real collaboration but are rarely captured in research settings. Barron [4] documented how classroom collaborations fail when members push incorrect solutions with unwarranted confidence or dismiss valid contributions. Students regularly spread misconceptions, sometimes unknowingly, sometimes through carelessness and the group's ability to detect and correct these contributions determines whether collaboration succeeds [4]. These dynamics are pervasive [10, 14] yet absent from controlled CPS datasets designed to elicit cooperation under shared goals [18]. Secret Hitler produces functionally equivalent dynamics through its role structure: Fascist players introduce misinformation, assert false claims with confidence, and selectively withhold evidence. The collaborative challenge this poses of forcing the group to evaluating contributions of uncertain reliability, pooling incomplete evidence, and reaching decisions despite misleading information, is the same challenge groups face whenever collaboration involves distributed knowledge and heterogeneous contributor quality. We do not claim that game players are cognitively equivalent to struggling students; we claim that the epistemic problem is shared. The methodological advantage is that the ground truth is known: role assignments let us track who introduced unreliable information, who accepted it (Layer 2), and whether that acceptance was warranted (Layer 3). In a classroom, the same dynamics occur but leave no traceable record.

The testbed also provides structured temporal episodes (nomination → discussion → vote → reveal), rich multimodal interaction in a facetoface small group, and repeatability across sessions. Our annotation framework (Section 3) captures what existing deliberation datasets cannot: the downstream effects of unreliable contributions on *individual* beliefs, traced from specific communicative acts to specific belief changes in specific individuals.

3. MULTI-LAYER ANNOTATION FRAMEWORK

3.1 Multi-Layer Annotation Framework

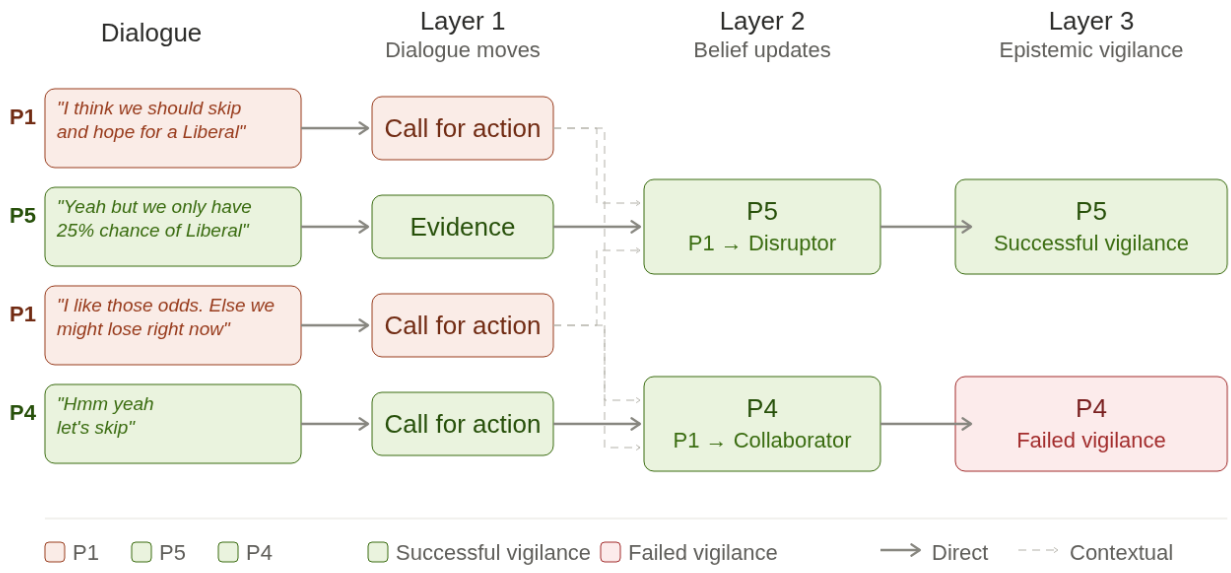

Figure 2: Three layer annotation of a discussion phase. Utterances are classified by communicative function (Layer 1), linked to belief updates (Layer 2), and assessed against ground truth as successful or failed epistemic vigilance (Layer 3). Color indicates true role (red = adversarial, green = collaborative; hidden during play).

The framework is hierarchical: each layer targets a progressively deeper dimension of group social-cognitive dynamics. Layers 2 and 3 operationalize two cognitive science frameworks: Baker et al.'s [2] Bayesian Theory of Mind and Sperber et al.'s [16] epistemic vigilance which have been influential but not previously applied to multimodal behavioral annotation in group settings.

3.2 Layer 1: Communicative Acts

We adopt the six-category taxonomy from Lai et al. [11], annotated at the utterance level: *Identity Declaration*, *Accusation*, *Interrogation*, *Call for Action*, *Defense*, and *Evidence*. This layer captures what communicative act a speaker performs, the information an agent needs to classify conversational moves in real time. Lai et al. report Krippendorff's $\alpha > 0.6$ across all categories, establishing feasibility for reliable annotation.

3.3 Layer 2: Player-Reported Belief States

At the end of each round, players report two types of beliefs:

First-order ToM. Each player records their belief about every other player's role (Liberal, Fascist, or uncertain). Since researchers know the true assignments, these reports yield calibration data: how accurately does each player's belief state track reality as evidence accumulates?

Second-order ToM. Each player records what they think others believe about *their own* role. Comparing Player A's second-order report against Player B's actual first-order report yields a direct measure of Theory of Mind accuracy [2].

These belief reports serve as supervision signals for training models that infer beliefs from observable behavior—gaze patterns, speech acts, voting behavior, and postural cues—since a classroom agent cannot ask students to report beliefs directly.

3.4 Layer 3: Epistemic Vigilance

This layer is inferred post-hoc using Layers 1 and 2 together with ground-truth role assignments [16]. For each communicative act targeting a specific player, we examine the receiver's belief trajectory:

- **Successful vigilance**: belief moved toward truth (e.g., correctly increasing suspicion of a Fascist after inconsistent behavior)
- **Failed vigilance**: belief moved away from truth (e.g., persuaded by a misleading claim, or failing to update when evidence warranted it)

Figure 2 illustrates this cascading structure with an excerpt from a pilot session. P1 (true role: Disruptive) advocates for an irrational move of skipping a turn and hoping for a favorable outcome. P5 correctly objects by presenting evidence that the odds are poor, updating their belief that P1 is a Disruptor showing successful vigilance. P4, however, accepts P1's reasoning and agrees to skip, maintaining the belief that P1 is a trustworthy Collaborator despite the weak justification: failed vigilance. The same exchange thus produces both outcomes in different receivers, and the frame-

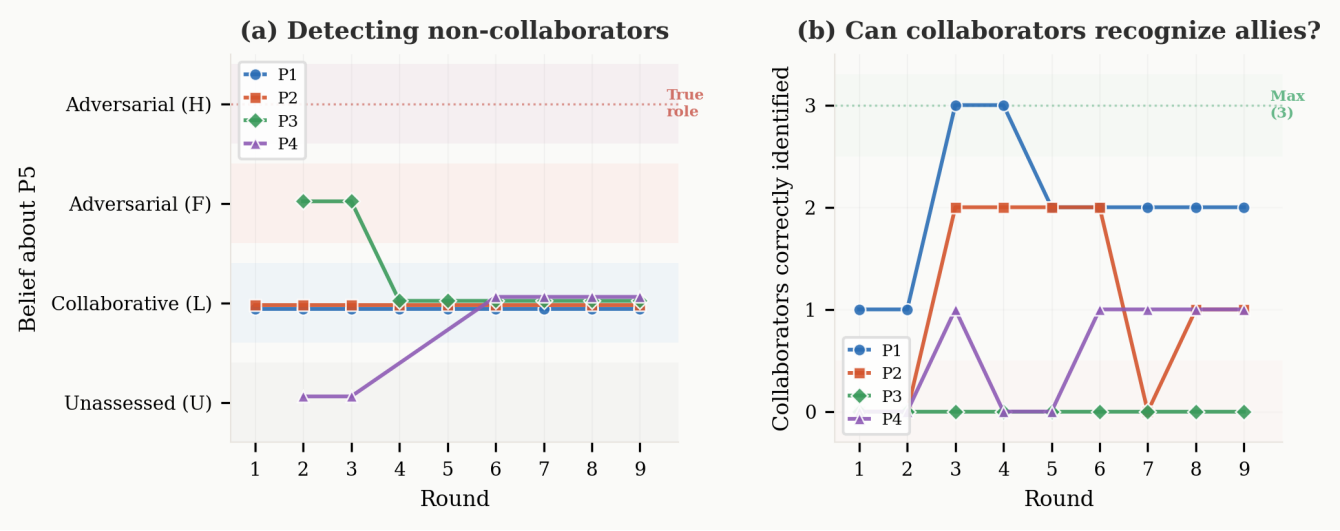

Figure 3: **(a)** Four collaborative players' beliefs about P5 (true role: Disruptive) across rounds. No collaborator ever correctly classified P5. P3 briefly suspected P5 (Rounds 2–3) but reversed by Round 4—a belief update *away* from truth, illustrating failed vigilance. **(b)** Number of fellow collaborators each player correctly identified per round (max = 3).

work captures exactly where and why the group's reasoning diverged.

4. SENSING AND DATA CAPTURE

All players wear Meta Aria Gen 1 research glasses [7], capturing egocentric RGB video, eye-tracking data, and IMU signals. Group audio is recorded via the glasses' microphones, and an overhead camera provides an exocentric reference view. Game state is logged by a facilitator.

Most multimodal learning analytics systems rely on exocentric sensing—ceiling-mounted cameras and external microphones [1]. While scalable, these setups cannot reliably recover participant-relative signals: whether someone is looking *at you*, directing a statement toward you, or attending to the current speaker. Our testbed instruments *every player* with egocentric glasses, producing synchronized capture from all participants' viewpoints. This enables gaze-derived features aligned with our annotation layers: identifying each player's attention target at utterance onset (linking Layer 1 acts to perceptual targets), characterizing gaze sequences during deceptive vs. truthful communication [5], and computing group-level attention divergence across all players at each timestep. Both egocentric and exocentric streams are captured simultaneously, enabling empirical comparison of whether first-person signals carry more information about the social-cognitive constructs in Layers 1–3.

5. PILOT DATA

To demonstrate that the target phenomena emerge naturally, we report observations from a pilot session (6 players, 9 rounds, 2 Disruptive/Adversarial and 4 Collaborative). All players wore Aria glasses and completed belief reports after each round.

Detecting Failed Vigilance. Figure 3(a) tracks collaborative players' evolving beliefs about P5, whose true role was Disruptive. Despite 9 rounds of discussion and evidence, no collaborator correctly identified P5. P3 initially suspected P5 (Rounds 2 and 3) but reversed toward Collaborative by Round 4, constituting *failed vigilance* in our Layer 3 framework: P3 possessed a partially correct signal but abandoned it, likely in response to P5's communicative acts (Layer 1) that undermined the suspicion.

Active but Counterproductive Participation. Figure 3(b) reveals a complementary pattern. P3 correctly identified zero allies across all 9 rounds, placing every collaborator in the Disruptive category—the profile of a group member who is actively reasoning but systematically misdirected. This is analogous to a student who is engaged but whose contributions fragment rather than strengthen shared understanding. Distinguishing active-but-counterproductive participation from constructive collaboration is precisely the capability our Layer 3 assessment operationalizes. P1, by contrast, peaked at identifying all three allies (Rounds 3–4) before declining, suggesting group-level belief fragmentation as unreliable information accumulated.

6. DISCUSSION AND CONCLUSION

The pilot data, though limited to a single session, demonstrates two properties. First, the multi-layer framework produces *linked annotations*: specific communicative acts traced to specific belief changes in specific individuals, assessed for epistemic warrant which is a cascading structure not available in existing CPS [18] or deliberation [9] datasets. Second, the testbed produces unreliable contributions with known ground truth, enabling study of how misinformation propagates through groups and what behavioral signals distinguish listeners who detect it from those who do not.

We frame the testbed as addressing a prior question: is it feasible to infer social-cognitive constructs from multimodal behavioral signals at all? The signals captured here are nat-

uralistic with overlapping speech, occlusions, and rapid gaze shifts, but with clean supervision unavailable in educational settings. We do not claim models trained on game data will deploy directly in classrooms; what transfers is the inference machinery and the insight that proxy testbeds with clean supervision can accelerate model development impractical to iterate on in instructional settings.

7. LIMITATIONS

The pilot comes from a single 6-player session; we do not claim generalizability. Player-reported beliefs are subject to social desirability and retrospective rationalization. Layer 3 assessments depend on binary ground truth cleaner than any classroom would provide. The game's competitive structure may elicit more deliberate reasoning than typical collaboration, and the sensing setup requires research-grade hardware whose value is in model development, not deployment.

8. ACKNOWLEDGMENTS

This material is based in part upon work supported by the U.S. National Science Foundation (NSF) under award DRL 2454151 (Institute for Student-AI Teaming). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

9. REFERENCES

[1] K. Ahuja, D. Kim, F. Xhakaj, V. Varga, A. Xie, S. Zhang, J. E. Townsend, C. Harrison, A. Ogan, and Y. Agarwal. Edusense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(3):1–26, 2019.

[2] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):0064, 2017.

[3] B. Barron. Achieving coordination in collaborative problem-solving groups. *The journal of the learning sciences*, 9(4):403–436, 2000.

[4] B. Barron. When smart groups fail. *The journal of the learning sciences*, 12(3):307–359, 2003.

[5] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, 44(4):1079–1100, 2012.

[6] P. Dillenbourg and D. Traum. Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. *The Journal of the Learning Sciences*, 15(1):121–151, 2006.

[7] J. Engel, K. Somasundaram, M. Goesele, A. Sun, A. Gamino, A. Turner, A. Talattof, A. Yuan, B. Souti, B. Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

[8] Y. Hayashi. The power of a "maverick" in collaborative problem solving: An experimental investigation of individual perspective-taking within a group. *Cognitive science*, 42:69–104, 2018.

[9] G. Karadzhov, T. Stafford, and A. Vlachos. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25, 2023.

[10] S. J. Karau and K. D. Williams. Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, 65(4):681–706, 1993.

[11] B. Lai, H. Zhang, M. Liu, A. Pariani, F. Ryan, W. Jia, S. A. Hayati, J. Rehg, and D. Yang. Werewolf among us: Multimodal resources for modeling persuasion behaviors in social deduction games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6570–6588, 2023.

[12] L. Lu, Y. C. Yuan, and P. L. McLeod. Twenty-five years of hidden profiles in group decision making: A meta-analysis. *Personality and Social Psychology Review*, 16(1):54–75, 2012.

[13] D. Moshman and M. Geil. Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, 4(3):231–248, 1998.

[14] S. L. Piezon and W. D. Ferree. Perceptions of social loafing in online learning groups: A study of public university and us naval war college students. *International Review of Research in Open and Distributed Learning*, 9(2):1–17, 2008.

[15] J. Roschelle and S. D. Teasley. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer, 1995.

[16] D. Sperber, F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson. Epistemic vigilance. *Mind & language*, 25(4):359–393, 2010.

[17] G. Stasser and W. Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6):1467, 1985.

[18] C. Sun, V. J. Shute, A. Stewart, J. Yonehiro, N. Duran, and S. D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672, 2020.

[19] S. D. Teasley, F. Fischer, A. Weinberger, K. Stegmann, P. Dillenbourg, M. Kapur, and M. Chi. Cognitive convergence in collaborative learning. In *Proceedings of the 8th International Conference for the Learning Sciences*, volume 3, pages 360–367. 2008.

[20] S. E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2 edition, 2003.