Modeling Epistemic Vigilance in Collaborative Problem Solving: A Multimodal Approach Using Social Deduction as a Proxy Testbed

Videep Venkatesha
Colorado State University

Nathaniel Blanchard
Colorado State University

ABSTRACT

AI agents that support collaborative learning must detect reasoning failures as they unfold: misconceptions propagating unchecked, weak claims accepted without scrutiny, and participants reasoning actively but reaching wrong conclusions. Building such detection models requires datasets where these failures occur and are labeled and yet existing collaborative datasets capture only idealized cooperation under shared epistemic goals. This paper presents a doctoral research program addressing this gap through three integrated contributions: (1) a multi-layer annotation framework that traces communicative acts through belief updates to epistemic vigilance assessments, operationalizing constructs from cognitive science for behavioral annotation; (2) a proxy testbed using social deduction games where unreliable information enters group reasoning with known provenance, enabling traceable study of how epistemic failures propagate; and (3) multi-person egocentric sensing that captures participant-relative signals unavailable to traditional exocentric setups. This work extends my prior research on propositional extraction, common ground tracking, deliberation chain modeling, and automated annotation in collaborative settings, building toward an integrated system that can monitor the quality of group reasoning in real time. Pilot data from a 6-player session demonstrates that failed 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, belief dynamics

1. PROBLEM AND STATE OF THE ART

Collaborative problem solving (CPS) is a core competency in education that involves sharing information, building common ground, reasoning about evidence, forming and revising beliefs, and coordinating action through communication [15, 18]. For collaboration to be productive, participants must evaluate each other's contributions, weigh evidence, update beliefs, and coordinate decisions often under genuine uncertainty, especially when there is no single clear answer and the task is open-ended. The quality of this evaluation process, what Sperber et al. [16] term epistemic vigilance, is what separates productive collaboration from groups that converge on misconceptions or fragment into incompatible understandings.

The challenge for educational data mining is that existing collaborative datasets overwhelmingly capture idealized cooperation. Corpora such as the DELI dataset [8] and frameworks like the General Competency Model [18] operationalize CPS under shared epistemic goals, where all participants work toward a correct answer. Real classrooms are messier where students unknowingly spread misconceptions, assert incorrect solutions with unwarranted confidence, freeload, or dominate discussions unproductively [3]. These dynamics of unreliable contributions, the natural heterogeneity of contributor quality, are pervasive in educational settings [9, 14] yet rare and unlabeled in controlled datasets. Without data where epistemic failures occur with known ground truth, building detection models is intractable.

A second gap concerns sensing modality. Current multimodal learning analytics for collaboration rely predominantly on exocentric sensing with overhead cameras and external microphones [1]. Egocentric sensing offers the potential to capture participant-relative signals that exocentric setups cannot reliably recover: whether a listener is attending to the speaker during a critical claim, or scanning others' reactions before updating a belief. Whether these first-person signals carry additional information about the social-cognitive constructs relevant to CPS, and epistemic vigilance in particular, is an open empirical question that this work aims to investigate.

1.1 Prior Work: Building Toward Epistemic Monitoring

My doctoral research has progressively built the computational infrastructure needed to model collaborative reasoning. This trajectory establishes both the technical foundations and the motivating gaps that the current work addresses.

Extracting what groups discuss. I developed methods for extracting task-relevant propositions from natural speech in collaborative settings, using cross-encoder architectures

Videep Venkatesha, and Nathaniel Blanchard. Modeling Epistemic Vigilance in Collaborative Problem Solving: A Multimodal Approach Using Social Deduction as a Proxy Testbed. 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. 865–869. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21039976

adapted from coreference resolution [22]. This work established that the semantic content of group dialogue can be recovered automatically, even from noisy overlapping speech, and was later extended to handle automated transcription with minimal degradation [21]. These methods became a core component of TRACE, a real-time multimodal common ground tracking system that integrates speech, gesture, gaze, and object detection to model the shared beliefs of collaborating groups as they emerge [19].

Modeling how discussions evolve. I contributed to work on linking deliberation chains in collaborative dialogues, modeling the causal structure from earlier utterances to probing questions that drive group reasoning forward [13]. This work formalized the sequential structure of collaborative discourse, showing that probing interventions and their causal antecedents can be automatically clustered using joint learning frameworks.

Understanding the limits of automation. Most recently, I evaluated 13 LLMs across six model families on the task of annotating pedagogical moves in educational dialogue [20]. A key finding was that models achieve moderate agreement on content-focused categories with clear surface markers but fail on categories defined by pedagogical function, precisely the categories (like CPS facets) that require understanding speaker intent and social dynamics rather than surface text patterns.

The gap this work addresses. Collectively, this prior work can track *what* groups discuss and *whether* they reach agreement, but it cannot assess *whether the reasoning that led there was sound*. A group may converge on a shared belief through careful evidence evaluation or through uncritical acceptance of a confident but wrong claim. Current systems, including those I have helped built, treat convergence as success without evaluating the epistemic quality of the path. The present work addresses this gap directly, introducing a framework and testbed for studying how groups evaluate (or fail to evaluate) the reliability of contributions.

2. THEORETICAL FRAMING AND METHODOLOGY

2.1 Theoretical Foundations

The research integrates three theoretical frameworks. *Epistemic vigilance* [16] provides the overarching lens: the cognitive mechanism for evaluating incoming information by assessing source reliability and content plausibility rather than accepting claims uncritically. In collaborative learning, failed epistemic vigilance manifests as misconceptions spreading through groups, weak arguments being accepted without challenge, and overconfident errors going undetected.

Bayesian Theory of Mind [2] provides the belief-modeling framework, formalizing how observers infer others' mental states from observed behavior. In my framework, players report first-order beliefs (what they believe about others' roles) and second-order beliefs (what they think others believe about them), grounding Theory of Mind measurement in self-report data that can serve as supervision signals for computational models.

Argumentative Theory of Reasoning [12] motivates the focus on persuasion dynamics. Mercier and Sperber argue that reasoning quality in collaborative settings depends critically on the group's argumentative practices: whether claims are challenged, whether evidence is demanded, and whether the group evaluates contributions on epistemic merit rather than social dynamics.

2.2 Testbed: Social Deduction as a CPS Environment

I use the social deduction game Secret Hitler as a proxy testbed where players are secretly assigned cooperative or adversarial roles 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 [15, 18]. A minority of players are structurally incentivized to introduce unreliable information such as asserting false claims with confidence, and undermining group reasoning. These contributions are functionally equivalent to the misconceptions and overconfident errors in collaboration. The epistemic challenge is shared, even though the underlying intent differs. Crucially, every piece of unreliable information has known provenance (role assignments provide ground truth), and the game enforces the information asymmetry that the hidden profile literature identifies as the hardest aspect of real collaboration [11, 17].

2.3 Multi-Layer Annotation Framework

The framework is hierarchical, with each layer targeting a progressively deeper dimension of group social-cognitive dynamics.

Layer 1: Communicative Acts. Each utterance is classified using the six-category persuasion strategy taxonomy from Lai et al. [10]: Identity Declaration, Accusation, Interrogation, Call for Action, Defense, and Evidence (Krippendorff's $\alpha > 0.6$). This layer draws on my prior work in propositional extraction [22] and epistemic move classification [19], extending those methods to a setting where communicative acts include deliberate misinformation. Categories like Interrogation and Defense parallel the questioning and justification behaviors that characterize productive student reasoning, the moments where a collaborator challenges a claim or defends their position with evidence, rather than accepting contributions uncritically.

Layer 2: Player-Reported Belief States. After each round, players report first-order beliefs about every other player's role (yielding calibration data against ground truth) and second-order beliefs about what others believe about them (enabling direct measurement of Theory of Mind accuracy [2]). These reports serve as supervision signals for models that must eventually infer beliefs from observable behavior alone. In the classroom parallel, this captures whether a collaborator can identify when someone else is not contributing productively to the group, whether through providing misleading information, dominating the discussion, or steering the group in an unproductive direction.

Layer 3: Epistemic Vigilance. Inferred post-hoc using Lay-

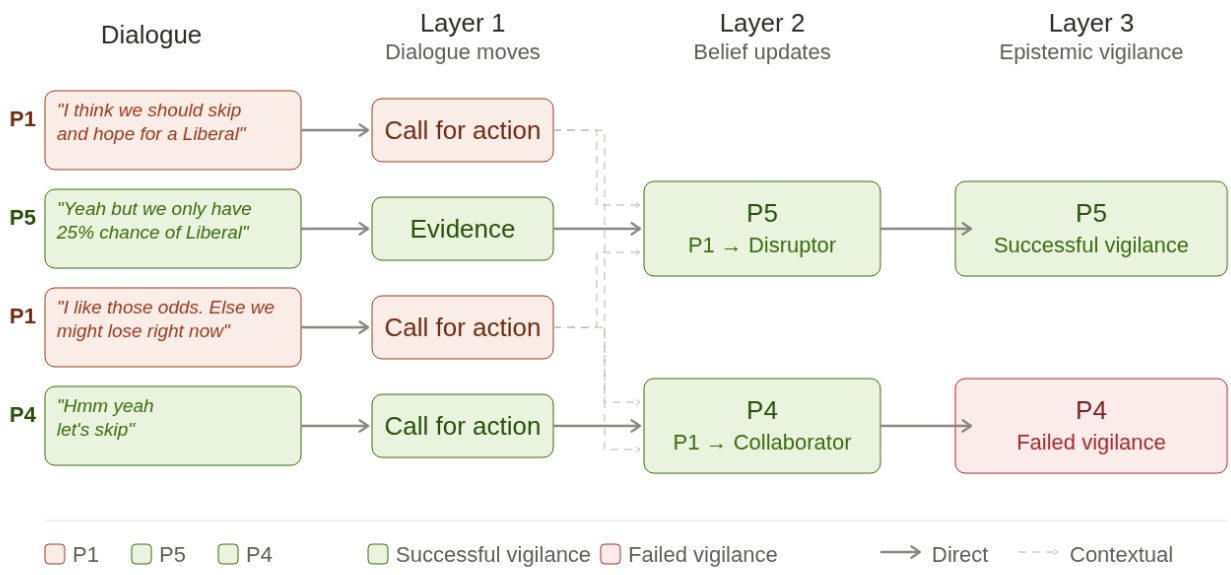

Figure 1: Three-layer annotation of a discussion phase. Utterances are classified by communicative function (Layer 1), linked to belief updates (Layer 2), and assessed against ground truth as successful or failed epistemic vigilance (Layer 3). Color indicates true role (red = adversarial, green = collaborative;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

ers 1 and 2 together with ground-truth role assignments [16]. For each communicative act, we assess whether the receiver's belief update was warranted: *successful vigilance* (belief moved toward truth) or *failed vigilance* (belief moved away from truth). The cascading structure of communicative act → belief update → vigilance assessment aims to mirrors what a classroom agent must track, extending the common ground tracking paradigm from my prior work [19]. This layer captures not just whether students can identify unreliable contributions, but whether they actually succeeded in doing so, and by examining which communicative behaviors (Layer 1) precede successful vs. failed vigilance, we can study what kinds of group dynamics lead collaboration to succeed or break down.

Figure 1 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 framework captures exactly where and why the group's reasoning diverged.

2.4 Sensing Infrastructure

All players wear Meta Aria Gen 1 research glasses [5], capturing egocentric RGB video, eye-tracking data, and IMU signals. An overhead camera provides a synchronized exocentric reference view, producing $n + 1$ synchronized video streams per session wehre n is the numebr of players. This multi-person egocentric configuration enables gaze-derived features aligned with our annotation layers: gaze target identity at utterance onset, scanpath structure [4] during successful vs. failed vigilance episodes, and group-level attention divergence computed across all players which is a measure unavailable in single-viewpoint datasets [7, 6].

2.5 Research Trajectory

My research proceeds in three phases. *Phase 1 (current): Framework design and pilot validation.* I have designed the annotation framework and conducted pilot sessions to verify that target phenomena emerge naturally. *Phase 2: Data collection and annotation.* I will conduct additional sessions to build a dataset sufficient for computational modeling, adapting the annotation methods from my propositional extraction and deliberation chain work [22, 13] to this new setting. My evaluation of LLM annotation capabilities [20] will inform which layers can be partially automated and which require human coding. *Phase 3: Computational modeling.* Using the annotated multimodal data, I will train models that predict Layer 3 vigilance outcomes from observable behavioral features, systematically comparing egocentric-only, exocentric-only, and fused feature sets, building on the multimodal fusion architecture of TRACE [19].

3. PROGRESS TO DATE

Pilot data from two 6-player, 9-round session demonstrates that the target phenomena emerge naturally. Figure 2(a) tracks four collaborative players' evolving beliefs about P5, whose true role was Hitler/Disruptive. Despite 9 rounds

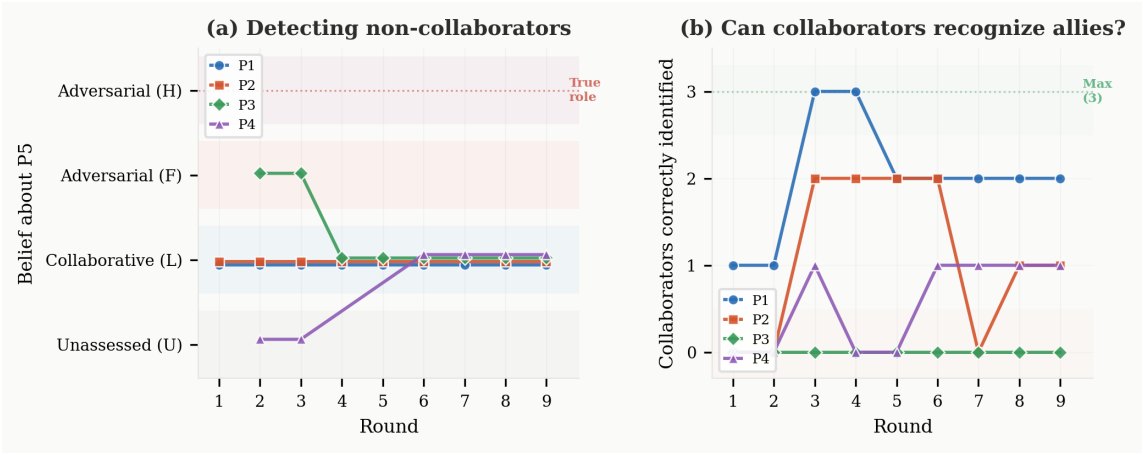

Figure 2: (a) Four collaborative players' beliefs about P5 (true role: Disruptive) across rounds. No collaborator ever correctly classified P5, illustrating failed vigilance. (b) Number of fellow collaborators each player correctly identified per round (max = 3).

of discussion and evidence, no collaborator ever correctly identified P5. P3 initially suspected P5 (Rounds 2–3) but reversed toward a collaborative classification by Round 4, constituting failed vigilance: P3 possessed a partially correct signal but abandoned it, likely in response to P5's communicative acts. Figure 2(b) reveals a complementary pattern where P3 correctly identified zero allies across all 9 rounds, placing every collaborator in the adversarial category which is the profile of a group member who is actively reasoning but systematically misdirected, analogous to a student who is engaged but whose contributions fragment rather than strengthen shared understanding [3]. Distinguishing active-but-counterproductive participation from constructive collaboration is precisely the capability the 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.

4. EXPECTED CONTRIBUTIONS AND IMPACT

For learning sciences. The annotation framework provides a structured methodology for studying how unreliable information propagates through collaborative groupsa phenomenon widely documented [3, 9] but rarely studied with the temporal granularity and ground truth this testbed provides. The operationalization of epistemic vigilance as a measurable construct in group behavioral data offers a new lens for understanding when and why collaborative reasoning fails.

For computer science. The testbed and dataset enable development of multimodal models for detecting reasoning failures in group settings. The multi-person egocentric sensing configuration provides a template for capturing participant-relative signals in small-group research, and the systematic comparison of egocentric vs. exocentric features will inform sensing design for future learning analytics systems.

For the broader EDM/AIED community. This work addresses a fundamental bottleneck: the scarcity of labeled data capturing collaborative dynamics beyond idealized cooperation. By providing a proxy testbed where epistemic failures occur frequently, with known ground truth, and with rich multimodal capture, the research accelerates model development that would be impractical to iterate on in instructional settings [18]. Combined with my prior work on propositional extraction, common ground tracking, and automated annotation evaluation, this research will produce an integrated pipeline: from sensing raw multimodal signals, through extracting propositional content, to assessing the epistemic quality of group reasoning—the full trajectory a collaboration monitoring agent would need to traverse.

5. 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. When smart groups fail. *The journal of the learning sciences*, 12(3):307–359, 2003.

[4] 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.

[5] 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.

[6] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding

skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024.

[7] K. Grauman, M. Wray, A. Fragomeni, J. P. Munro, W. Price, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, et al. Around the world in 3,000 hours of egocentric video. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[8] 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.

[9] 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, 1993.

[10] 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.

[11] 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.

[12] H. Mercier and D. Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

[13] A. Nath, V. Venkatesha, M. Bradford, A. Chelle, A. C. Youngren, C. Mabrey, N. Blanchard, and N. Krishnaswamy. "any other thoughts, hedgehog?" linking deliberation chains in collaborative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314, 2024.

[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] H. VanderHoeven, B. Bhalla, I. Khebour, A. C. Youngren, V. Venkatesha, M. Bradford, J. Fitzgerald, C. Mabrey, J. Tu, Y. Zhu, et al. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50, 2025.

[20] V. Venkatesha, S. Anindho, E. Seefried, and N. Blanchard. Using LLMs to annotate pedagogical moves: You know what i mean? In *Proceedings of the 27th International Conference on Artificial Intelligence in Education (AIED 2026)*. Springer, 2026. Accepted for publication.

[21] V. Venkatesha, M. Bradford, and N. Blanchard. Dude, where's my utterance? evaluating the effects of automatic segmentation and transcription on cps detection. In *International Conference on Artificial Intelligence in Education*, pages 144–151. Springer, 2025.

[22] V. Venkatesha, A. Nath, I. Khebour, A. Chelle, M. Bradford, J. Tu, J. Pustejovsky, N. Blanchard, and N. Krishnaswamy. Propositional extraction from natural speech in small group collaborative tasks. *International Educational Data Mining Society*, 2024.