Modeling Epistemic Vigilance in Collaborative Problem Solving

A Multimodal Approach Using Social Deduction as a Proxy Testbed (non archival)

Videep Venkatesha · Ethan Seefried · Changsoo Jung · Nathaniel Blanchard

Colorado State University

Videep's website

IMPACT lab

CVPR JUNE 3-7, 2026 DENVER COLORADO

The Problem

In collaboration, people exercise epistemic vigilance: evaluating others' claims before accepting them. When it fails, misinformation spreads, weak claims pass unchecked, errors go unchallenged. An AI agent monitoring collaboration needs to detect these failures as they happen.

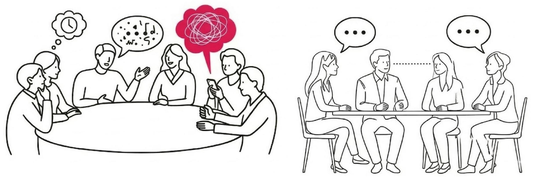

Figure 1: Real collaboration is messy with unverified claims, misconceptions, divided attention (left). Existing datasets capture only the idealized version (right).

GAP 1: Labeled epistemic failures are missing

Existing datasets capture idealized cooperation. No ground truth for who introduced misinformation or who accepted it.

GAP 2: Participant-relative signals are missed

Overhead cameras can miss who someone is looking at, who they're addressing, or who they're attending to.

OUR CONTRIBUTION

A multimodal testbed pairing multi-person egocentric sensing with a structured environment that produces labeled epistemic vigilance failures with known provenance.

Testbed

WHY A SOCIAL DEDUCTION GAME?

Hidden roles make adversarial players inject unreliable information by design, equivalent to misconceptions and overconfident errors in real groups, but with known provenance for every claim.

Multimodal Sensing

MULTI-PERSON EGOCENTRIC CAPTURE

Every player wears Meta Aria Gen 1 research glasses (egocentric RGB, gaze, IMU, audio) plus an overhead camera, yielding n + 1 synchronized streams.

Figure 2: Exocentric: overhead view captures the group but loses participant-relative attention.

Figure 3: Egocentric view reveals participant-relative attention that overhead cameras may not recover.

WHY THIS MATTERS

Overhead cameras suffer from occlusion and cannot distinguish fine-grained attention targets. Receiver-perspective signals like directed eye contact, mutual gaze, gaze toward an ally during an accusation are available only egocentrically.

Three Layer Annotation Framework

Layer 1: Dialogue moves. Utterances tagged with a 6-category persuasion taxonomy.
Layer 2: Belief states. Player-reported 1st and 2nd order beliefs each round.
Layer 3: Epistemic vigilance. Belief updates scored against ground-truth roles.

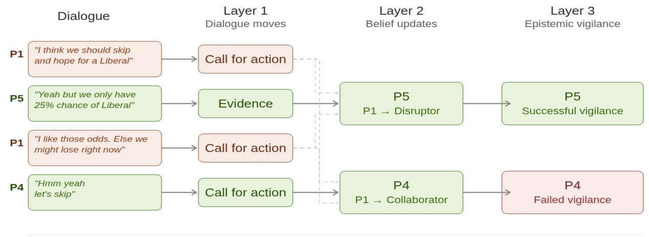

Figure 4: P1 (adversarial) calls for a risky action. P5 challenges with evidence and flags P1 as a disruptor: successful vigilance. P4 accepts P1's call and treats them as a collaborator: failed vigilance.

Pilot Observations

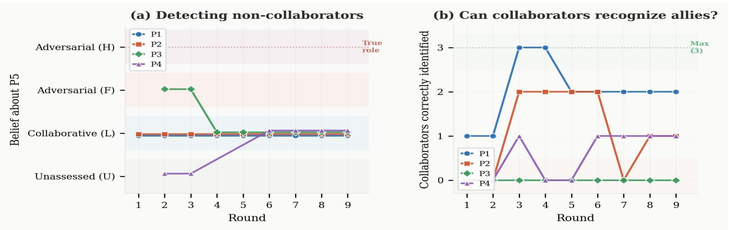

Figure 5: (a) Collaborators' beliefs about P5 (true: Disruptive) across 9 rounds; none ever classified P5 correctly. (b) Allies correctly identified per round (max = 3).

Conclusion

When epistemic vigilance fails, groups commit to wrong answers, and AI agents monitoring collaboration need to detect it. We pair a social deduction game with multi-person egocentric sensing to produce labeled vigilance failures with known provenance. Next: scale annotation and train multimodal models to predict them.