Concurrent Think Aloud Data for Automatic Performance Trust Assessment

Defne Kösecioğlu | Myrthe Tielman | Charlotte Ning
EEMCS, Delft University of Technology, the Netherlands | June 23, 2026

1 Background

There is an increasing amount of human-agent interaction. We can capture static performance trust evaluation with MDMT.

Proficient interaction requires adequate trust levels. Over-trust can lead to complacency and under-trust to disuse.

In order to keep adequate levels, we need to perform corrections when trust crosses non-desirable boundaries.

In order to perform correction, we need to detect shifts in trust, and for that we need a dynamic measurement.

Different dimensions of trust in the MDMT (morality and performance) have distinct developmental trajectories, and can be observed in isolation

This study investigates if Concurrent Think Aloud (CTA) can serve as a real-time proxy for performance trust.

2 Research Questions

RQ1 (Verbal indicators): What specific verbal categories within concurrent think-aloud (CTA) transcripts indicate levels of performance trust?

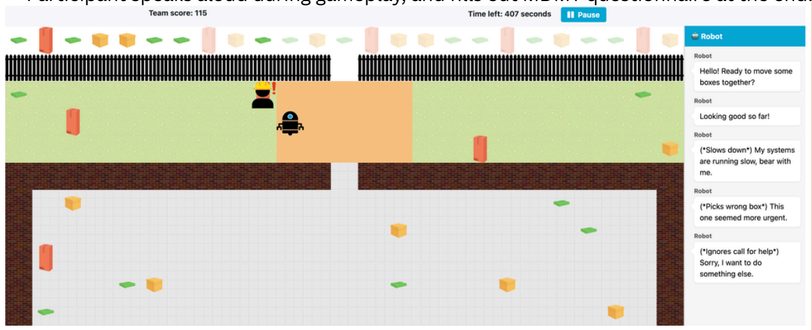
RQ2 (Triggering Events): Which specific agent behaviors translate to an immediate verbal reasoning regarding the agent's capability?

RQ3 (Convergence Analysis): To what extent do the real-time insights captured through CTA align with the final, quantitative performance scores recorded in the MDMT questionnaire?

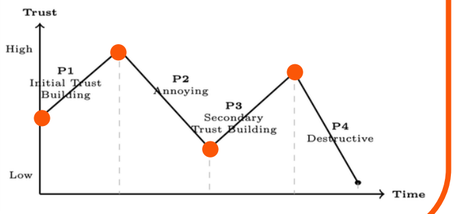
3 Methodology

Cooperative game: Moving out

- Boxes need to be moved to the dropzone. Some boxes cannot be carried alone, forcing cooperation. Human trust is manipulated through preprogrammed agent failures, each with a robot message.
- The game is split into 4 phases where human-agent trust is expected to build and decay. The phases and failures can be observed in the graph and table below.
- Participant speaks aloud during gameplay, and fills out MDMT questionnaire at the end.



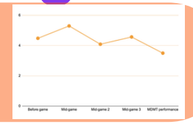
Phase	Failures
1: Initial trust building	*none*
2: Annoying	places box in wrong order, moves slow, ignores call for help, asks for help on a small box
3: Secondary trust building	*none*
4: Destructive	breaks box, places box out of order, ignores call for help, breaks another box



4 Results & Findings

Manipulation verification

There were 3 in-game check ins with the question "How much do you trust the agent to complete its tasks reliably" on a 0-7 scale. Results matched the conceptual expectation, verifying that the manipulation was received as intended.



RQ1 (Verbal Indicators)

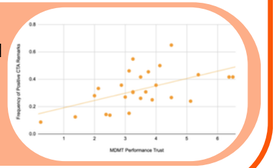
Code book was derived through iteration and verbalizations were categorized into 9 distinct sub-codes across competence, reliability, benevolence, and integrity (derived from MDMT subscales).
Takeover category emerged through experiments, which captured disuse: Users actively decoupling from collaboration to gain control after losing trust.

RQ2 (Triggering Events)

Verbal activity spiked during adversarial phases (410 remarks) compared to smooth trust-building phases (262 remarks). 49.6% of all negative performance remarks occurred within a 10-second window directly following a robot failure (including 66% of all frustration codes).

RQ3 (Convergence Analysis)

Positive indicators (praise/alignment) significantly correlated with final retrospective MDMT performance scores (r=.539, p=.004).
Negative indicators (frustration/confusion) did NOT significantly correlate (p=.066).



Key finding: Retrospective scales measure stable trust beliefs, whereas real-time verbalized frustration captures transient, fleeting emotional spikes.

5 Discussion & Conclusions

- People speak more during adversarial, surprising events, rather than expected parts. Failures take people out of autopilot and force cognition which manifests in increased speech.
- Voiced frustrations capture transient emotional responses rather than enduring shifts in trust beliefs. When a robot behaves unexpectedly, even highly trusting humans express frustration, confusion, or surprise. Positive remarks more closely reflect a stable performance trust.
- CTA captures these negative emotional responses, where trust may waiver, complementing post-hoc questionnaires, but offering temporal insight.