CSE3000 RESEARCH PROJECT
June 2026

Retrospective Think-Aloud as a Window into Performance Trust

A Human-Agent Collaborative Game Study

Adrian Todorov · Prof.: Dr. Myrthe Tielman · Supervisor: Charlotte Ning
A.E.Todorov@student.tudelft.nl

1 BACKGROUND

- Trust in AI agents evolves dynamically but most measures capture only a static end-state score
- MDMT questionnaire separates performance trust (reliable, capable) from morality trust (sincere, ethical)
- No prior work has directly tested whether RTA reflects performance trust specifically

2 RESEARCH QUESTION

"How does retrospective think-aloud (RTA) data reflect a user's performance trust in an agent?"

SQ 1: What themes appear in RTA when users express high vs. low performance trust?
SQ 2: Which game events prompt performance trust reasoning in RTA?
SQ 3: How well do RTA verbal expressions align with MDMT performance subscale scores?

3 STUDY DESIGN & METHODOLOGY

STUDY DESIGN

- 30 participants; mixed academic and non-academic backgrounds
- Sessions run in-person or remotely; standardized protocol
- Single shared game scenario for all participants

MEASUREMENTS

- MDMT - 16-item; Reliable + Capable subscales = performance trust; In-session pop-up ratings – performance item
- RTA transcripts - audio verbalisations during gameplay replay; Audio recordings
- Game logs - timestamped record of scripted failure onsets; Screen recordings

PROCEDURE — RTA CONDITION

Briefing & Consent → Play game (silent) → RTA → MDMT

DESIGN RATIONALE

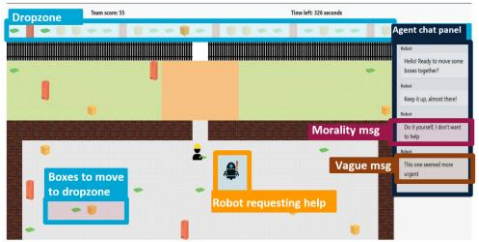
- Scripted failures create controlled, repeatable trust-relevant moments
- Three message framings let the study examine how attribution type affects trust language
- Performance trust is a persistent attitude - less vulnerable to RTA memory decay

REFERENCES

[1] Lee & See (2004). Trust in automation. Human Factors
[2] Malle & Ullman (2021). MDMT. Trust in HRI
[3] Alhadreti & Mayhew (2018). Rethinking TA. CHI
[4] Cutler et al. (2025). Crowdsourced TA. CHI
[5] Pathan et al. (2025). CTA vs RTA. ICCE
[6] Centeio Jorge et al. (2023). Automation failure. Front. Robotics AI

TU Delft

4 THE GAME: MOVING OUT (MODIFIED)



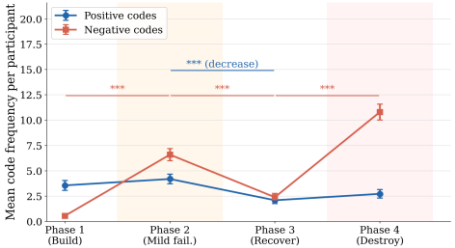
Example failure type	Agent message (example, not all are given)	Message Type
Wrong dropzone	"I miscalculated the target position"	Performance
Break box	"Oops!"	Vague
Ignore call for help	"Do it yourself, I don't want to help"	Morality

The game is a human-agent collaborative box-moving task. The AI agent has 8 scripted failures during the game, each paired with a performance, morality-framed, or vague chat message (6 types of failures that happen). The session is divided into 4 phases. Phases 1 and 3 are trust-building; Phase 2 introduces moderate failures; Phase 4 introduces the most serious ones. Participants rate their performance trust at baseline and after each of Phases 1-3 (T0-T3).

5 RESULTS

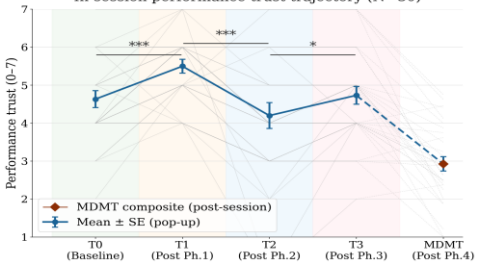
A 16-code scheme (5 categories: Performance, Morality, Calibration, Event Reaction, Behaviour) was applied to all 30 transcripts

RTA trust-relevant language across game phases (N=30)



- RTA negative codes track the scripted trust trajectory; all phase transitions $p < .001$
- Phase 4 peak: 10.8 codes/participant vs. 0.57 in Phase 1

In-session performance trust trajectory (N=30)



- Pop-up ratings recovered fully by T3, yet RTA negative codes remained 4× Phase 1 level
- MDMT composite (M=2.93) fell well below T3 (4.73)

6 DISCUSSION

- Pop-up ratings recovered fully by T3, yet RTA negative codes remained 4× Phase 1 level - verbal processing lags behind explicit ratings
- Aggregate RTA frequency does not predict MDMT scores - diagnostic value lies in timing, not total volume
- RTA consistently surfaced evaluative language, causal attributions, and explicit trust-update moments - performance codes dominated, with RELIABILITY_NEG and CAPABILITY_NEG being the two most frequent
- Explicit attribution pre-empts participant reasoning rather than amplifying it - performance-framed messages suppressed verbalisation rather than prompting it; Vague failures produced significantly more language
- Performance language dominated even after morality-framed failures - participants defaulted to a capability lens regardless of framing

LIMITATIONS

- N=30; single game context limits generalisability
- Memory decay for minor events; single session limits longitudinal claims

> NEXT STEPS

- *Time-weighted or phase-specific RTA metrics as MDMT predictors*
- *Event-level trust probes to test whether less verbalisation means less trust impact*
- *Integrating RTA-style debriefs into iterative agent design pipelines*