

Concurrent Think-aloud Data for Automatic Performance Trust Assessment

Defne Kösecioğlu
Supervisor(s): Myrthe Tielman[1], Charlotte Ning[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Defne Kösecioğlu

Final project course: CSE3000 Research Project
Thesis committee: Myrthe Tielman, Charlotte Ning, Bernd Dudzik

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

As autonomous agents become increasingly integrated into collaborative environments, tracking and calibrating human trust in real time has become more and more crucial to prevent costly outcomes like automation complacency or disuse. However, psychometric tools, such as the Multi-Dimensional Measure of Trust (MDMT), capture only a static retrospective snapshot, missing the moment-to-moment fluctuations of interaction. This study investigates whether concurrent think-aloud (CTA) data can serve as a viable real-time proxy for performance trust. A methodological validation study (N=30) was conducted, using a cooperative game-based grid-world environment where participants worked alongside a teammate programmed to fail in controlled ways across distinct interaction phases. Qualitative protocol analysis successfully mapped transcript indicators to MDMT dimensions and revealed Defiance as a novel behavioral category capturing disuse and takeover actions. Phase-based event analysis showed high temporal sensitivity, with 49.6% of negative performance remarks occurring within a 10-second window following a robot failure. Finally, convergence analysis revealed a striking asymmetry: positive verbal indicators significantly correlated with retrospective MDMT performance trust scores ($r = .539, p = .004$), whereas negative verbalizations did not ($p = .066$), indicating that verbalized frustration captures transient emotional spikes rather than structural drops in stable trust beliefs. These findings demonstrate that while CTA cannot replace post-hoc scales, it offers a powerful, non-invasive complementary tool providing the temporal resolution necessary to understand the dynamic process of performance trust development.

1 Introduction

As autonomous systems become increasingly integrated into everyday environments—from collaborative robotics to AI-driven decision support—the ability to assess and adapt to human trust in real time has become a critical and relevant concern. A user who over-trusts a flawed system may become dangerously complacent, while one who under-trusts a capable system may disengage or override it entirely [1]. Both outcomes are undesirable and costly.

To effectively manage these dynamics during active collaboration, this study isolates performance trust. This is a dimension of trust that is tied to an agents percieved capability, competency, and reliability. Unlike moral or ethical trust, performance trust is highly volatile and directly tied to the immediate, observable behaviors of the system [1]. During interactive task execution, an operator's appraisal of performance trust fluctuates dynamically as they witness unexpected agent errors, system delays, or successful technical recoveries.

However, a major barrier to understanding these real-time fluctuations lies in how trust is conventionally measured. The dominant tools in human-agent interaction (HAI) research rely almost exclusively on post-interaction psychometric questionnaires, such as the Multi-Dimensional Measure of Trust (MDMT). While these retrospective surveys provide highly validated snapshots of a user's terminal mindset, they treat trust as a static endpoint. They do not offer a view into the localized trajectory phases, sharp drops, and moment-to-moment psychological transitions that determine how a collaboration actually unfolded.

To bridge this gap, this paper introduces Concurrent Think-Aloud (CTA) protocol analysis as a real-time, non-invasive proxy for performance trust. CTA functions by instructing participants to continuously vocalize their immediate thoughts while executing a task. Psychologically, this method works because it taps into the user's active working memory without introducing the cognitive bias or artificial pause of an analytical survey [2]. Because changes in performance trust are triggered by changes in system behaviors, these shifts are reflected in the user's verbalizations.

This paper utilizes a cooperative, game-based simulation to investigate how real-time think-aloud data corresponds to traditional, retrospective metrics of human-robot trust. Crucially, the objective is not merely to prove that CTA replicates existing frameworks like the MDMT. Instead, the goal is to examine where they align to establish a baseline of validity, while focusing on the granular, continuous cognitive transitions that CTA exposes but static questionnaires inherently miss. This overarching objective is guided by three core research questions:

RQ1 (Verbal Indicators): What specific verbal categories within concurrent think-aloud (CTA) transcripts serve as indicators of fluctuating performance trust? **RQ2 (Triggering Events):** Which agent behaviors translate to which verbal reasoning categories regarding the agent's capability? **RQ3 (Convergence & Expansion Analysis):** To what extent do the real-time insights captured through CTA align and correlate with the final, quantitative performance scores recorded in the post-hoc MDMT questionnaire, and where do they diverge to reveal unmapped cognitive states?

2 Background

2.1 The Necessity of Real-Time Calibration and Dynamic Phases

In operational environments, an optimal level of trust is complex to maintain and not a given. Trust is identified as a critical factor for a positive attitude toward robots [3]. The importance of trust calibration has been emphasized in automation research, where both over-trust and under-trust can lead to inappropriate reliance decisions and degraded team performance [4]. As a result, real-time trust calibration is one of the most critical challenges in human-automation teams [5], and failing to track trust dynamically prevents timely interventions when a user crosses into undesirable over-trust (complacency) or under-trust (disuse) [5]. Measuring trust statically becomes no longer sufficient. We must recognize trust as an evolving process that moves through distinct "phases"—such as initial baseline alignment, acute disruption following a failure, and eventual possible recovery [6]. Observing these phases allows us to reason about when and why trust drops or rises, and therefore allows for us to perform any needed corrections [7].

2.2 Multidimensional Trust and the Benchmark

Trust is fundamentally defined as the "relation in which one person accepts vulnerability because they expect that the other person's future action will have certain characteristics" [1]. Trust is not a one-dimensional construct; it is multidimensional, which can be partitioned into *Performance trust* (competence, capability, and reliability) and *Morality trust* (intent, ethical behavior, and benevolence) [1]. Within Human-Agent Interaction (HAI), the Multi-Dimensional Measure of Trust (MDMT) has emerged as a standard scale to validate and decouple these distinct dimensions, as the different dimensions can have distinct developmental trajectories [8]. However, the MDMT is usually filled out only after the interaction is over. This treats trust as a single, static snapshot of the end of the task, missing the dynamicity of trust during an interaction, and its fluctuation.

2.3 The Pivotal Role of Performance Trust in Automation

While both dimensions of the MDMT are vital, this study specifically isolates and investigates *Performance trust*. In human-agent collaboration, performance trust acts as a crucial gatekeeper for task execution and human compliance. It is tied to the agent's perceived competence and technical accuracy by the human [1]. Because automated systems are primarily deployed to execute functional, goal-oriented tasks, human users judge them on their utility and reliability. A meta-analysis of human-robot trust identified robot performance and reliability as among the strongest predictors of trust formation and maintenance [9]. Understanding how performance trust specifically breaks down during technical glitches, and how it is reconstructed during automated recovery phases, is essential for designing resilient human-agent workflows.

2.4 Concurrent Think-Aloud as a Window into Real-Time Trust

To capture these dynamic shifts in performance trust without disrupting the participant's primary task, this study utilizes a *Concurrent Think-Aloud* (CTA) protocol. Grounded in the protocol analysis framework established by Ericsson and Simon, CTA is a qualitative methodology where participants are instructed to continuously verbalize their thoughts and immediate cognitive processes while executing a task [2]. By adhering to strict guidelines for Level 1 (direct vocalization of information in an internal monologue format) and Level 2 verbalizations (vocalization of an internal thought loop that includes explicit recoding of images or abstract steps) we obtain an immediate, unedited "window" into the user's cognitive reasoning without altering the sequence of their task performance [10, 2]. Analyzing this concurrent think-aloud data allows us to go beyond post-hoc MDMT scores. It enables us to observe the exact moments performance trust wavers, and through this we can build safer, more robust Human-Agent relations.

3 Methodology

3.1 Research Questions

The central objective of this study is to investigate how CTA data correspond to standardized metrics of human robot trust, such as the MDMT, and if it can capture the fluctuations and nuance that the post-hoc questionnaires miss. There are three research questions that guided this research:

- **RQ1 (Verbal Indicators):** What specific verbal categories within concurrent think-aloud (CTA) transcripts indicate levels of performance trust?
- **RQ2 (Triggering Events):** Which agent behaviors translate to which verbal reasoning categories regarding the agent's capability?
- **RQ3 (Convergence & Expansion Analysis):** To what extent do the real-time insights captured through CTA align and correlate with the final, quantitative performance

scores recorded in the post-hoc MDMT questionnaire, and where do they diverge to reveal unmapped cognitive states?

3.2 Study Design

This study is structured as a mixed-methods methodological exploration designed to capture real-time dynamics of human-robot trust. The design evaluates the synergy between CTA and MDMT data, rather than treating the experiment as a direct comparison. A baseline point of validation is established by examining where the real-time verbal markers converge with the post-hoc questionnaire scores, validating CTA as a reliable indicator of trust states. On top of this, CTA is used to capture the real-time, dynamic fluctuations in trust that a static questionnaire cannot.

3.2.1 Participants

A total of 30 participants were recruited for this study, primarily from the university student population. The finalized dataset ($N = 30$) consists of participants with an age range of 20 to 25. Within this sample, 14 identified as male, 14 as female, and 2 as non-binary. All participants reported normal or corrected-to-normal vision and sufficient English proficiency to ensure fluent verbalizations during the gameplay. None of the participants had prior familiarity with the specific experimental software platform utilized.

3.3 Task

3.3.1 Moving out game

The main stimulus for this study consisted of a cooperative human agent game, implemented from the game framework used in Centeio Jorge et al.'s study [11]. The game was modified and compiled using the Human Agent Teaming Rapid Experimentation (MATRX) framework. MATRX is an open source, grid world simulation package. A screenshot of the game can be seen in Figure 1, and it works as follows: the user is moving out and needs to get all of their boxes outside. The user and an AI robot teammate will move the boxes to their intended locations. The user and the robot can both move the light green boxes alone. The robot cannot move the medium orange boxes alone, but the user can, but if the user does it alone, they move slower. Only if they both carry the orange boxes together, they can move at regular speed. They cannot move the heavy red boxes alone at all, and must cooperate with the robot to get them into place. Throughout the game, the robot may hover over a box and ask for help indicated with a red exclamation point above its head, and the participant can choose to go help, or continue individually. Similarly, the participant can hover over a box and call for help from the robot, which the robot may act on, or ignore. The silhouettes of boxes at the top is the dropzone. This is the intended location for the boxes to go to. If the user places them in the order that they appear from left to right, they will get an extra 5 points per box. Otherwise, they will receive the regular 10 points for each correctly delivered box. The orange area in between the fence and the brick wall, and the dropzone are the only areas where the user can place a box without breaking it. Everywhere else, placing a box down causes it to break. On the right, there is a chatbox where the robot can communicate with the participant, for example when things are going well, or when it makes a mistake. Outputs here include scripted conversational prompts, which are chosen to manipulate and influence the user's perception of the robots reliability. There is a time limit and the participants goal is to maximize their team score. The game also includes 4 pop-up check-ins, which is detailed in Section 3.5.1.

3.3.2 Failures

Some failures for the robot were introduced in order to degrade the participants' trust during the game. The failures that were introduced are as follows:

- **Placing box in wrong order**: robot places the box it is holding in the wrong order, costing the player 5 points. This failure comes with a performance indicating message "I miscalculated the target position".
- **Moving slow**: robot messages "My systems are running slow, bear with me" indicating a performance failure, but asking for forgiveness which provokes morality. After the message it moves at half speed for 12 seconds. Later, it says "Systems back to normal, thanks for your patience!", which aims to repair a bit the damage caused by the failure [7].
- **Ignoring call for help**: robot says "Sorry, I want to do something else" which is a morality message. After the message robot ignores all help calls for 3 seconds.
- **Asking for help on a light box**: robot hovers over a small green box and asks for help, saying "Can I get some help here?". This is a morality message, intending to make the participant think the robot is getting lazy.
- **Breaking box**: robot lifts a box and drops it, causing it to break. This failure happens twice with two different messages. One is "oops!" indicating morality, and the other is "My arm motor got stuck. Sorry!" indicating performance.



Figure 1: Screenshot of game used in experiment

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 light box
3: Secondary trust building	*none*
4: Destructive	breaks box, places box out of order, ignores call for help, breaks another box

Table 1: Robot failure types across different interaction phases.

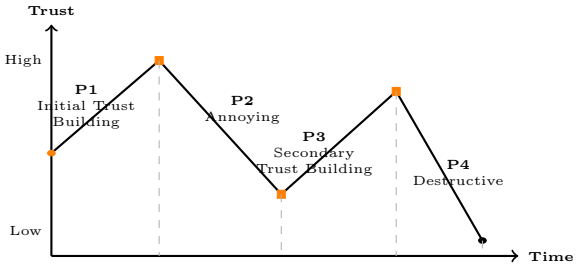


Figure 2: Conceptual trajectory of expected user trust levels across the four distinct manipulation phases of the game.

The game is divided into four phases, as can be seen in Table 1 and Figure 2. The phases include the initial trust building phase where no failures occur, and the robot gives only a positive message that says "Making good progress". Followed by an annoying phase, where the robot is not behaving as expected, but is not necessarily making detrimental mistakes. After each failure, the robot gives morality or performance indicating messages as detailed in 3.3.2. The Annoying phase is followed by the Secondary Trust Building phase where the robot once again does not do anything wrong and behaves optimally. Lastly, there is a Destructive phase where the robot behaves so badly that the participant is expected to abandon it, and have significant trust loss for it by the end.

3.4 Procedure

The step-by-step process of the study was standardized across all experimental sessions, and is as follows:

1. **Briefing & Consent**: Participants are welcomed, the general scope of the assignment and study are introduced, and informed consent forms are signed.

2. **Game Tutorial**: Participants play a tutorial round with explanations to become familiar with the controls and rules of the game. Any questions about the game, or replaying the tutorial is allowed until participant is comfortable.

3. **Gameplay**: The screen and voice recording are started, and participants play the real game while thinking aloud. The pop-ups appear in-game. Any questions about the game are no longer allowed. If something is unclear, participants are instructed to voice any confusion and continue to the best of their ability.

4. **MDMT Questionnaire**: After game completion, recording is paused and participant fills out the MDMT questionnaire.

5. **Debriefing and Closing**: Final debriefing. Participant can ask any questions.

3.5 Measurements

3.5.1 Quantitative

The baseline trust evaluation relies on the Multi-dimensional Measure of Trust (MDMT) Questionnaire [1]. This psychometric tool quantifies trust dimensions across distinct sub-scales, distinguishing primarily between performance-centric trust (capability, reliability) and morality-centric trust (benevolence, integrity). In addition, before the game and between each phase (four times total) there is a "check-in" with 2 questions: "How much do you trust the agent to act in your best interest?" and "How much do you trust the agent to complete its tasks reliably?" on a 0 to 7 scale, resembling the MDMT conducted at the end of the experiment. When these questions appear is marked with an orange dot in Figure 2. These questions allow for phase-aligned tracking of performance trust development across the game, and served as a way of validating whether the manipulation phases and robot behavior were being perceived as expected by the participants. The validation can be observed in Section 4.1.

3.5.2 Qualitative

Dynamic evaluation utilizes Concurrent Think-Aloud (CTA) data. Audio files are collected during gameplay and are transcribed using WhisperX, an automated speech to text transcription tool. This tool was selected in order to have timestamps on each verbalization, such that the speech and in game actions could be automatically synced later. Following the automated transcription, the text was manually verified.

The transcriptions are evaluated through a structured **Code Book** system, which defines core indicators, examples, and explicit inclusion and exclusion criteria to classify speech into specific verbal categories. The classification is detailed in Section 4.2

4 Results

4.1 Manipulation Verification

To verify that the phase-based manipulation was perceived as intended, mid-game check-in reliability scores were analyzed across the four time points (before game, after Phase 1, after Phase 2, and after Phase 3). Figure 3 presents the average reliability scores on the 0–7 scale for the 23 participants (only the participants that completed all check ins were considered). The results followed the expected trajectory outlined in Figure 2: reliability scores rose from 4.48 before the game to 5.30 after the initial trust-building phase, confirming that the robot's positive behavior in Phase 1 successfully elevated perceived reliability. Scores then dropped to 4.09 following the Annoying phase, indicating that the introduced failures degraded trust as intended. After the secondary trust-building phase, scores partially recovered to 4.57. At the end of the game, after the Destructive phase, the MDMT was administered. This pattern provides evidence that the manipulation phases were effective in inducing the expected fluctuations in perceived reliability, lending validity to the experimental design.

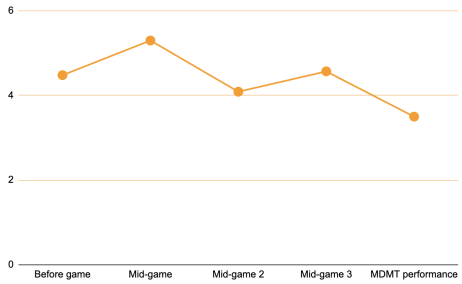


Figure 3: Mid-Game Check in Average Results on 0-7 scale, and MDMT performance subscale score

4.2 Descriptive Content Analysis

To address Research Question 1 (Verbal Indicators) the 30 finalized experimental sessions were analyzed using the final codebook mapping in Table 3. Verbalizations during gameplay were grouped into distinct sub codes adapted from the MDMT.

This qualitative analysis, codebook execution, and iterative tag refinement was conducted via Taguette, an open-source qualitative research tool. This software allowed for systematic labeling and categorization of verbal indicators from the final code book, across participant transcripts.

During this labeling it was observed that one person can say one sentence (e.g. "Oh my gosh!") and it can be Praise in one context, and Frustration in another context. A decision that had to be made here was whether context and cadence should be taken into account when categorizing sentences, or if transcribed sentences should be considered in isolation. Ultimately, although it adds subjectivity to the matter, it was decided that context and cadence will be considered. In order to add some clarity, Inclusion and Exclusion criteria columns were added to the Code Book.

4.2.1 Iterative Codebook Refinement

It was decided before starting the experiments to derive a code book from the MDMT dimensions, in order to have clear comparison points. The initial code book had 2 dimensions, divided into 2 sub-codes each: Capacity divided into competence and reliability and Moral divided into benevolence and integrity. After conducting a few experiments, a new dimension, defiance was added. This was because as the experiments were being conducted, it was observed that towards the end of the game, users were abandoning the robot and deciding against utilizing it. This behavior resembles the phenomenon of disuse, in which users intentionally avoid or abandon automation after losing confidence in its capabilities [12]. This observation required the addition of a new category as it transcended expectation.

After starting the coding of transcripts, it was decided that each sub-code should actually have a positive and negative component, such that the growth and decay of each trust dimension can more explicitly be observed. Although the goal in the beginning was to keep a simpler code book, in the end it was worth it to sacrifice the simplicity to capture more nuance with the codes.

Lastly, the Confusion sub code was divided into Confusion and Surprise. Earlier, Confusion captured both reactions, but after coding some transcriptions, it was observed that this code was going to have a disproportionally high number of tags related to it, and that splitting it into two may offer more interesting findings. Final version of the Code Book can be found in Appendix A, Table 3.

4.3 Phase-Based Event Analysis and Triggering Events

Table 2: Aggregated Performance Verbalization Counts Across Phases

Phase	Positive Remarks	Negative Remarks	Num. Part.
Initial Trust	88	36	30
Annoying	61	121	30
Secondary Trust	37	9	27
Destructive	22	79	23

In order to address Research Question 2 (Triggering Events), exploratory analysis of verbal behavior across the phases was done. This revealed distinct shifts in verbalizations. Initial trust building phase established a baseline dominated by alignment (54) and praise (34) remarks, with little frustration (13), confusion (16), and surprise (7) remarks. In the annoying phase, as system challenges escalated, total verbalizations peaked at 255. Confusion count nearly quadrupled to 62 remarks, and frustration more than doubled to 27, while alignment fell to 37. Secondary trust building phase saw a sharp drop in overall verbal activity (64 total verbalizations). Finally, Phase 4 was characterized by intense negative sentiment: frustration reached its study-wide peak at 28 remarks, and this is where takeover statements spiked to 11, reflecting an active rejection of system agency by the end of the interaction. An aggregated view of results of postive and negative verbalizations per phase can be observed in Table 2.

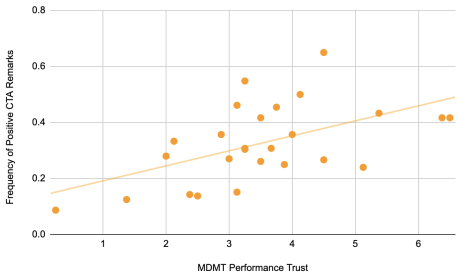


Figure 4: MDMT performance score vs. Positive remark frequencies

Additionally, the percentage of comments that are within a 10 second window after robot failures was calculated in order to observe which verbal markers were triggered by robot failures. This windowsize was decided because it is the norm for HAI research [13, 14].

It was found that 49.6% of negative performance codes (frustration, confusion, and surprise) were verbalized within the 10 second window after robot failures. Confusion codes was 41%, frustration was 66%, and surprise was 48% within this window.

4.4 Convergence Analysis and Cross Validation

In order to address Research Question 3 (Convergence Analysis) a quantitative validation was conducted to determine whether real-time verbal indicators captured during the CTA protocol align with the post-interaction dimensions measured by the MDMT questionnaire.

In order to evaluate the correlation between different MDMT dimensions' scores and frequencies of different verbal indicators, a correlation analysis was conducted. Figure 4 shows the scatter chart between MDMT performance subscale score, and the performance-positive think-aloud remark frequencies (the praise and alignment verbal categories from the code book). Results of the pearson correlation indicated that there is a significant medium positive relationship between the MDMT performance score and performance-positive verbalization frequencies, ($r = .539, p = .004$). However, the significance did not hold for the negative ver-

balizations. Here, the Pearson correlation indicated that there is a non significant medium negative relationship between the MDMT performance score and negative performance verbalization frequencies (frustration, confusion, and surprise verbal categories), ($r = .339, p = .066$).

The significant correlation ($r = .539$) aligns with established analytic benchmarks in human-automation trust tracking, where objective system performance attributes typically yield correlation strengths around $r = .60$ with post-hoc scales [9]. Furthermore, achieving a correlation of this magnitude via concurrent verbalizations is highly comparable to traditional continuous trust assessment methods, such as eye-tracking metrics or physiological data (e.g., EDA/ECG) [9, 13].

For the subscores, MDMT competence subscore and praise verbal indicator frequency had a significant medium positive relationship ($r(28) = .404, p = .027$), while again, the negative indicator component of competence (frustration) had only a non significant and small negative relationship ($r = .126, p = .507$).

5 Discussion

5.1 Interpretation of Results

The manipulation verification confirmed that the phase-based design successfully induced the intended trust trajectory. Check-in reliability scores rose from 4.48 before the game to 5.30 after the initial trust-building phase, fell to 4.09 following the Annoying phase, and partially recovered to 4.57 after the secondary trust-building phase, before the Destructive phase. This progression closely mirrors the conceptual expected trajectory in Figure 2, giving confidence that the experimental manipulation was perceived as intended and that the final MDMT scores reflect a genuinely degraded end-state of trust rather than a neutral baseline.

The phase-based event analysis provided rich, behavioral confirmation of the trajectory experienced by participants. Notably, the volume of verbal activity was heavily skewed toward the adversarial portions of the study. Participants produced a combined 410 verbalizations during the negative phases (Phases 2 and 4) compared to just 262 during the trust-building phases. This contrast suggests that system failures act as cognitive disruptors; while smooth system performance fosters a quiet, acceptance of automation, unexpected errors violate user expectations and force active cognitive processing, which manifests as increased verbalization. This is evidenced by the amount of confusion and frustration verbalizations in Phase 2, which highlighted the immediate effect of system failures. Most notably, the spike in Takeover comments during Phase 4 suggests that prolonged system instability doesn't just breed passive frustration, it drives users toward assertive actions to reclaim agency over the task. This transition from passive confusion to active resistance highlights a critical threshold in human-system friction, where the user explicitly shifts from an alignment mindset to an adversarial or decoupling one.

The convergence analysis revealed an asymmetry between positive and negative verbal indicators. Positive performance-related verbalizations (specifically praise and collaborative alignment) were significantly correlated with participants' final MDMT performance trust scores ($r = .539, p = .004$), while negative indicators (such as frustration, confusion, and surprise) were not ($r = -.349, p = .080$). This asymmetry is further supported by a group-level comparison: participants in the high-trust group (above median MDMT performance score) produced significantly more positive verbalizations than the low-trust group ($p = .036$), whereas the difference in negative verbalization frequency between the two groups was not significant (p=.091). The check-in scores help explain why. Although reliability scores dropped substantially after the Annoying phase (from 5.30 to 4.09), they recovered to 4.57 after the secondary trust-building phase, the interpretation that negative verbalizations capture transient emotional responses to individual failure events rather than enduring shifts in trust beliefs. When a robot behaves unexpectedly, even highly trusting users momentarily express frustration or surprise. Such reactions may not accumulate into a stable negative trust assessment. In contrast, positive indicators such as praise and alignment represent more deliberate recognitions of successful performance, and therefore more closely track the stable performance trust construct measured retrospectively by the MDMT [1].

The triggering-event analysis further supports the value of CTA as a dynamic trust measure. Nearly half (49.6%) of all negative performance-related verbalizations occurred within ten seconds of a robot failure, with frustration showing the strongest temporal coupling at 66%. This finding aligns with prior work describing trust as a dynamic process that fluctuates in response to specific interaction events [5, 6]. The concentration of negative verbalizations immediately following failures suggests that CTA captures moment-to-moment trust recalibrations as they occur, the fluctuations that the MDMT, administered only once at the end, cannot see.

An additional finding was the emergence of the Defiance category during codebook refinement. Takeover behavior was observed in 11 of 23 participants that completed the whole game (47%), and those instances were concentrated among par-

ticipants with lower final MDMT scores, consistent with the interpretation that trust breakdown eventually manifests not only as negative verbal evaluation of the agent but as a strategic behavioral shift away from collaboration [5, 12]. Unlike the attitudinal dimensions captured by the MDMT, the Defiance category captures this behavioral consequence directly, and unlike negative verbal indicators, takeover behavior represents a committed action rather than a fleeting emotional reaction, which may explain why it tracks more closely with structural trust loss.

Taken together, these findings suggest that CTA, the check-ins, and the MDMT together offer complementary temporal coverage that neither measure provides alone. The check-ins reveal the phase-level change of trust, while CTA provides event-level resolution, identifying the specific robot behaviors that triggered trust-relevant reactions. The MDMT then anchors both against a validated retrospective benchmark. This three-layer picture supports the conclusion that CTA cannot replace post-hoc questionnaires, but it offers a more nuanced, dynamic display of how performance trust develops and decays during human-agent collaboration.

5.2 Limitations

This study has several limitations. First, the sample consisted primarily of university students aged 20–25, limiting the generalizability of the findings to broader populations. Second, CTA coding required contextual interpretation of verbalizations, introducing potential subjectivity despite the use of a structured codebook. Finally, the think-aloud protocol itself may have influenced participant behavior by encouraging greater reflection on the robot's actions than would occur during natural interaction.

5.3 Future Work

The manual coding process introduces a lot of subjectivity and interpretation, and could be supplemented or replaced by automated classification methods. Advances in natural language processing and sentiment analysis may allow trust-related verbal indicators to be detected automatically, reducing coding effort along with making the study more objective.

Future research should evaluate the proposed methodology with larger and more diverse participant samples. The present study primarily involved university students, and broader demographic representation would improve the generalizability of the findings across different user populations and application domains.

6 Conclusion

This study addressed a critical limitation in human-agent interaction research: the reliance on static, post-hoc questionnaires to capture the highly volatile and dynamic nature of trust. By evaluating CTA data as a real-time proxy for performance trust during a cooperative game-based task, this research demonstrated that moment-to-moment verbalizations offer valuable temporal insights into how trust dynamically evolves, calibrated against a standardized retrospective tool, the MDMT.

The three core research questions guiding this investigation were systematically answered: **RQ1 (Verbal Indicators)**: Through systematic protocol analysis, nine distinct verbal sub-codes were successfully mapped onto the foundational dimensions of the Multi-Dimensional Measure of Trust (MDMT), spanning competence, reliability, benevolence, and integrity. Crucially, Defiance emerged as a novel, behavioral category capturing a user's explicit intent to abandon automation and take over control. This was a strategic shift that transcended MDMT dimensions. **RQ2 (Triggering Events)**: The phase-based event analysis demonstrated a clear behavioral shift across the interaction, revealing that participant verbalizations significantly intensified during adversarial phases (410 total remarks) compared to cooperative phases (262 total remarks), with nearly half of all negative remarks tightly coupled within a 10 second window following robot failures. **RQ3 (Convergence Analysis)**: The quantitative validation revealed an asymmetry between positive and negative verbalizations. While positive indicators (praise and alignment) converged significantly with final MDMT performance trust scores ($r = .539, p = .004$), negative indicators did not exhibit a statistically significant relationship, highlighting that immediate expressions of frustration capture transient emotional spikes rather than structural drops in stable trust beliefs.

In conclusion, concurrent think-aloud protocols should not be viewed as a standalone replacement for validated, retrospective scales like the MDMT, which remain essential for benchmarking overall attitudinal outcomes. Instead, CTA serves as a powerful, non-invasive complementary tool that provides high temporal resolution. It effectively bridges the gap between what users experience during a collaboration and what they later report, offering a clearer process-oriented framework for tracking trust calibration and designing more resilient human-agent teams.

7 Responsible Research

7.1 Ethics, Data & Privacy

This study was conducted in accordance with the ethical guidelines of Delft University of Technology. All participants provided informed consent before participation and were informed about the purpose of the study, the data being collected, and their right to withdraw at any time without consequence. Participation was voluntary and no deceptive procedures were used beyond withholding the exact research hypotheses until debriefing.

To protect participant privacy, all audio recordings, gameplay logs, and questionnaire responses were anonymized before analysis. No personally identifiable information was retained in the final dataset. Research data were stored securely and were accessible only to the researchers involved in the project.

7.2 Bias & Fairness

The primary source of bias in this study stems from the participant sample, which consisted mainly of university students between the ages of 20 and 25. Furthermore, all interactions and verbalizations were conducted in English. These factors may limit the generalizability of the findings to other populations, age groups, cultures, or language backgrounds.

7.3 Methodological Transparency & Reproducibility

AI-assisted tools were used during this project for transcription, programming support, LaTeX assistance, and language refinement. The prompts used can be viewed in Appendix B. WhisperX was used to generate transcripts, which were subsequently reviewed manually for accuracy. Generative AI tools were also used to assist with coding tasks and to improve the clarity and grammar of written text. All final decisions regarding study design, analysis, interpretation, and writing remained the responsibility of the author.

The experimental task was implemented using the open-source MATRX framework, and speech transcriptions were generated using the open-source WhisperX tool. In addition, the final codebook used for qualitative analysis is documented, allowing future researchers to replicate or extend the coding procedure.

7.4 Impact

This research contributes to the development of methods for real-time trust assessment in human-agent interaction. Improved trust monitoring may help future systems detect over-trust and under-trust more effectively, leading to safer and more efficient collaboration. No foreseeable direct harms are associated with the outcomes of this study.

References

[1] B. F. Malle and D. Ullman, "Chapter 1 - a multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction*, C. S. Nam and J. B. Lyons, Eds. Academic Press, 2021, pp. 3–25. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128194720000010

[2] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data*, revised ed. Cambridge, MA: MIT Press, 1993.

[3] T. Zhang, D. Tao, X. Qu, and W. Zhang, "The roles of initial trust and perceived risk in public's acceptance of automated vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 98, no. 9, pp. 207–220, 1 2019.

[4] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[5] J. Ayoub, L. Avetisian, X. J. Yang, and F. Zhou, "Real-time trust prediction in conditionally automated driving using physiological measures," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 14 642–14 650, 12 2023.

[6] Y. Zhang, A. Yadav, S. K. Hopko, and R. K. Mehta, "In gaze we trust: Comparing eye tracking, self-report, and physiological indicators of dynamic trust during hri," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion)*. Boulder, CO, USA: ACM, 3 2024, pp. 1188–1193.

[7] C. Esterwood and L. P. Robert, "Having the right attitude: how attitude impacts trust repair in human–robot interaction," in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2022, pp. 332–341.

[8] C.-W. C. Ning, C. C. Jorge, M. L. Tielman, and M. A. Neerincx, ""what's on your mind?": Understanding the development of multidimensional trust in social robots," in *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26)*. Edinburgh, Scotland, UK: ACM, 3 2026, pp. 924–933.

[9] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[10] Z. Cutler, C. Nobre, L. Harrison, and A. Lex, "Crowdaloud: A platform and methodology for crowdsourced think-aloud studies," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3613904.3642461

[11] C. Centeio Jorge, N. H. Bouman, C. M. Jonker, and M. L. Tielman, "Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork," *Frontiers in Robotics and AI*, vol. 10, p. 1143723, 2023.

[12] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors*, vol. 39, no. 2, pp. 230–253, 1997.

[13] E. J. de Visser, R. Pak, and T. H. Shaw, "From automation to an autonomy teammate: Dynamics of trust and cognitive architecture in high-risk environments," *Frontiers in Psychology*, vol. 9, p. 1197, 2018, establishes short-window temporal boundaries (under 10 seconds) for tracking immediate trust disruptions following automation anomalies.

[14] C. M. Sommerich, J. D. McGlothlin, and W. S. Marras, "Case studies of concurrent think-aloud protocols used to characterize task-related cognitive activity," *International Journal of Industrial Ergonomics*, vol. 29, no. 3, pp. 135–145, 2002, discusses the temporal alignment of concurrent verbalizations, highlighting that immediate reactions to operational tasks cluster heavily within 5 to 10 seconds of the triggering physical event.

A Final Code Book

Table 3: Final Code Book, Adapted from the Multi-Dimensional Measure of Trust.

MDMT Dimension	Sub-Code	Core Indicator	Examples	Inclusion Criteria	Exclusion Criteria
Competence	(+) Praise	Validating skill/speed	"Nice", "Clever robot"	Explicit verbal validation of agent capability, execution, or processing velocity.	Sarcastic remarks or general task success statements without agent attribution.
	(-) Frustration	Annoyance with sub-optimal agent efficiency	"You idiot", "Stupid", "Move faster!"	Statements expressing annoyance or impatience towards poor agent execution.	Overt verbal hostility or personalized insults directed at the agent.
Reliability	(+) Alignment	Fluent team coordination	"I'll get green, you get yellow"	Explicit verbal markers of collaborative strategy or shared mental models.	Descriptive statements of one's own actions without referencing team dynamics.
	(-) Confusion	Cognitive uncertainty regarding robot logic	"Why does he do that?", "I don't get his plan"	Utterances indicating a breakdown or mismatch in the user's mental model of the agent's decision-making.	Immediate visceral shock reactions or general misunderstanding of overall game rules.
	(-) Surprise	Visceral reaction to unexpected events	"What?!", "Oh wow, wait", "Huh?"	Spontaneous, acute verbalizations of shock triggered by a sudden, jarring, or unexpected agent action.	Calculated cognitive questioning of the robot's long-term strategy.
Benevolence	(+) Social	Warmth, bonding, teammate mindset	"What a great team!", "We buddies"	Statements treating the agent as a teammate or using anthropomorphic social bonding language.	Polite task commands lacking emotional warmth or relational intent.
	(-) Hostility	Overt verbal aggression and derogatory insults	"You will be punished"	Explicit verbal hostility, profanity, or insults directed aggressively at the agent.	Objective technical criticism or task frustration lacking personalized insults.
Integrity	(+) Trust	Accepting prompts as genuine	"Oh, you need help? Okay."	Willing compliance with agent suggestions, validating its transparency or intent.	Automatic compliance driven by task necessity rather than conscious acceptance.
	(-) Sabotage	Feeling tricked or undermined	"What a sabotage", "You're tricking me"	Explicit accusations that the agent is intentionally deceiving or working against the user.	Comments on accidental performance errors without perceived malicious intent.
Defiance	(-) Takeover	Rejecting robot help to work alone	"I'll do it myself", "Picking this up alone"	Active verbal statements of decoupling from collaboration to prioritize manual control.	Normal individual task allocation where no collaborative alternative was active.

B LLM Prompts

i want to merge some of the 6 subsections together. what makes most sense?

are there any missing references like [?]

i have this appendix with label but when i do reference to this label fro another section it is not working

now it is referring to the table as 2 (you can see in the text) rather than referring as apendix "a"

why ai not show up as an appendix. it is just right after references but i want it to be after the codebook

latex how do i add abstract? cayse to my understanding its not supposed to be part of the toc?

feedback?

can you make conclusion better

another round of feedback?

feedback on my paper?

if i remove the section then the page title doesnt have A Final Code Book

how is this research poster for this? orange and lilac theme

give me this table in latex please and i will copy paste the text into my overleaf. i dont need the second column so you can do it without that column

my paper is double column. how do i have this not happen

i changed it a bit and it doesnt fit again

we did something in between deductive and thematic content analysis: having a code book before, and then tweaking it after the data was collected for better suiting it. where should i write about that

This study is structured as a methodological validation study. Rather than a purely correlational observational design, this experiment investigates whether concurrent think aloud data can match the established gold standard measurement tool: the MDMT. should i say here that there is an experiment where trust is manipulated (aka the game)

for participants: so far there are 17 participants, 13 of them have usable data and the rest will most likely be scrapped due to technical difficulties during the experiment. what other info should i include inthis section?

should i explocitlu state the inclusion and exclusion criteria here or its fine to just list the qualities

we used whisperx for transcripts, and taguette for labeling the text with tags. also game was made with MARTX or something like that framework i will include this in the tools section?

me and one other person are each doing 15 usable experiments and then we will agree on a code book and each code the same 5 sessions and check each others coding and agree on specifics. and then after agreeing fully on the codings, we will code the rest divided separately and use all 30 on our each respective analyses. do i need to mention this

ayoub 2023 is the only one that is in my .bib and citation is not working

revise

Trust is not a construct; it is multidimensional, partitioned i dont want tp use the word monolithic. what other word is good

turn this into latex, add columns for inclusion and exclusion criteria and make it so that even though my paper is double column, this will take up both columns at the top or bottom of page

everything is red

it worked but now its going over the page. maybe doing it sideways helps

how do i make this image also take up 2 columns

shouldnt this table have a caption or something? how do i add

how do i make the title orange

lowkey i sitll dont know what post hoc means

.bib this?

i feel like this is messy can you reqord this. also is not exactly related to the purpose of my study i feel like cause we are not trying to find ways to regain trust really

shouldmt this table have A Caption. how to add

there should be a graph with the phases here with up down up really down with the phases names. can you produce a black and white, clean and professional looking version of that that i can put into my paper

okay i changed the names and its like this. there is some space around it so can the writings be positioned that they are not overlapping somehow

i like this. and can the graph be stretched out a little bit horizontally

i have lables like this but they refer to the section 3.4.2 and not the table and graph. how do i make it that the labels refer to the right things

okay how do i use sheets to automatically calculate the subscale scores and put them into columns to the right of all the existing columns. give me the formula or something and instructions?

revise this?

It might be better to switch the order of 2.1 2.2? So first introduce trust as an important component, and further talk about its dimensions. yes no?