

How Concurrent Think-Aloud Data Reflects a User's Morality Trust in an Agent

A Think-Aloud Study of Morality Trust in Human-Agent Collaboration

Arda Cengaver

Supervisor(s): Myrthe Tielman, Charlotte Ning

[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: Arda Cengaver
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

Trust is a key factor in human-agent interaction, especially when users collaborate with an AI agent whose behavior is not fully predictable. Trust is often measured after an interaction using questionnaires, but these measures do not fully show how trust changes during the interaction or which agent behaviors shape the user's judgment. To address this limitation, this paper investigates whether Concurrent Think-Aloud (CTA) data can capture participants' real-time reasoning about the agent's moral trustworthiness during interaction. The study uses a modified version of the *Moving Out* game, in which participants collaborate with an AI agent while verbalizing their thoughts. Trust is examined using repeated in-game pop-up questionnaires, a post-interaction questionnaire, coded CTA transcripts, game logs, and screen recordings. The results show that morality-related trust is rarely expressed through formal moral terms such as "ethical" or "principled". Instead, participants express it through task-specific comments about whether the agent cooperates with them, ignores them, or acts in their interest. The findings also show that CTA markers align more closely with in-game trust ratings than with post-interaction questionnaire scores. Overall, the results suggest that CTA can complement questionnaire-based trust measures by capturing situated, real-time expressions of morality trust during interaction.

1 Introduction

An agent can be technically capable and still feel untrustworthy. When users collaborate with an AI agent, they evaluate not only whether it can complete the task; they also interpret its intentions, question its choices, and decide whether it is genuinely trying to cooperate with them. These judgments develop continuously during collaborative tasks, because users become partly dependent on the agent to complete a task more efficiently.

This dependence is central to most definitions of trust. For instance, one definition of trust states that if A believes that B will act in A's best interest, and A accepts vulnerability to B's actions, then A trusts B [7]. In human-agent teamwork, this means that trust becomes relevant when the user's outcome partly depends on the agent's behavior. Users may therefore interpret the agent's choices, mistakes, and intentions as signals of whether continued reliance is appropriate.

This paper treats trust as a multidimensional construct rather than a single fixed judgment. Two dimensions are especially relevant in collaborative tasks between humans and agents. Performance trust concerns whether the user perceives the agent as capable, reliable, and predictable [10]. Morality trust concerns whether the user believes that the agent is sincere, benevolent, honest, and acting in the user's interest [10]. This distinction becomes clearer in cases where an agent appears technically capable of completing the task, but its behavior still makes the user question whether it is helpful, fair, or willing to cooperate [6]. This separation makes morality trust useful for explaining how users interpret aspects of the agent's behavior that cannot be captured by task performance alone.

These different aspects of trust are often measured after an interaction using questionnaires such as the Multi-Dimensional Measure of Trust (MDMT) [10]. Such measures are valuable for capturing structured, comparable scores across participants. But they record a final judgment, not the process that led to it. Trust is not static; it shifts in response to specific behaviors [12], and those shifts may influence how a user engages with the agent in real time. Therefore, identifying the moments that trigger doubt or reassurance can help show which agent behaviors are associated with changes in trust as the interaction unfolds. However, it remains unclear how morality trust manifests moment-to-moment during a collaborative task, which agent behaviors are responsible for shifts in it, and how these real-time signals relate to the trust people report afterward. Capturing this process requires a method that gives access to users' reasoning as it happens, rather than only its final outcome.

To examine this process, this study uses Concurrent Think-Aloud (CTA) data. In CTA, participants verbalize their thoughts while performing the task, which makes it possible to observe how they interpret the agent's behavior during the interaction. In this study, participants played a modified version of the *Moving Out* game [2] while thinking aloud, collaborating with an AI agent, and experiencing both cooperative behavior and scripted failures. By synchronizing CTA transcripts with game logs, it becomes possible to connect participants' verbalizations to specific in-game events.

1.1 Research Question

Together, these considerations motivate the central question of this paper:

"How does Concurrent Think-Aloud data reflect a user's morality trust in an agent?"

To answer this question, the study is guided by the following sub-questions:

- **SQ1:** What kinds of morality trust markers appear in Concurrent Think-Aloud (CTA) transcripts during interaction with an agent?
- **SQ2:** Which in-game events trigger morality-related reasoning?
- **SQ3:** How do coded CTA morality trust markers compare with questionnaire-based morality trust scores?

The aim is, therefore, not only to obtain a trust score after the interaction, but also to understand what the participant was thinking during the interaction and which moments in the game triggered this reasoning.

2 Background

2.1 Trust in Human-Agent Interaction

Trust in human-agent interaction can be separated into several core elements. First, there is a trustor, in this case the human

user. Second, there is a trustee, in this case the agent. Third, there is uncertainty, because the user does not fully know how the agent will behave. Finally, there is vulnerability or risk, because the agent's behavior can affect the outcome of the task. Thus, trust becomes important precisely when the outcome depends on the agent in a situation that the user cannot fully predict.

In collaborative settings, the user and the agent depend on each other to complete a shared task. Trust is therefore not just a feeling; it also affects whether users accept the agent's help, monitor its behavior, delegate parts of the task, or ignore its actions. Lee and See argue that trust is closely connected to reliance on automation, especially when the system is complex or when users cannot fully understand or predict its behavior [5]. In addition, poorly calibrated trust can harm teamwork: too little trust may lead users to reject useful support, while too much trust may lead them to rely on an agent even when this reliance is inappropriate [5; 8].

2.2 Morality Trust

Trust between humans and agents is not limited to whether the agent is competent to perform a task successfully. Malle and Ullman describe human-robot trust as multidimensional, with separable performance-related and morality-related components [6]. Performance trust concerns whether the agent is perceived as reliable and capable, while morality trust concerns whether the agent is perceived as sincere, ethical, benevolent, and honest [6; 10].

This distinction is central to this study because users may evaluate an agent's competence separately from its intentions. For instance, an agent may be capable of completing a task, while still being judged as unfair, deceptive, or not acting in the user's best interest. In this case, the issue is not only whether the agent can perform the task, but whether the user interprets the agent's behavior as morally trustworthy. Thus, morality trust is directly related to how users make sense of the agent's actions and intentions during the interaction.

Mehrotra et al. show that explanations related to integrity, such as honesty, transparency, and fairness, can affect how appropriately users trust an AI agent [8]. This suggests that morality-related judgments can influence both the user's overall trust in the agent and their interaction behavior, such as whether they rely on the agent, accept its help, or continue cooperating with it.

2.3 Measuring Trust During and After Interaction

Typically, trust measurement relies on questionnaires completed after the interaction. These questionnaires are useful because they turn a participant's experience into structured scores that can be compared across participants. For example, the Multi-Dimensional Measure of Trust (MDMT) measures different components of trust, including performance-related and morality-related dimensions [10]. This makes it possible to study trust as more than one general attitude toward the agent.

At the same time, a final questionnaire score only shows the participant's judgment after the interaction has ended. It does not show how that judgment was formed during the task. Because trust is not necessarily stable throughout an interaction, a final score may hide the moments that caused doubt, reassurance, or changes in reliance. Different dimensions of trust may emerge, increase, or decrease at different moments. Ning et al. show this in the context of social robots, where dimensions of multidimensional trust developed differently over time [9].

To study these changes, trust measures can be combined with verbal data collected during the task. Concurrent Think-Aloud (CTA) is a think-aloud method in which participants verbalize their thoughts while performing an activity. This protocol can provide insight into reasoning as it occurs, rather than only through later reflection [4]. In trust research, CTA may help reveal whether participants verbalize trust-related interpretations of the agent's behavior as the interaction unfolds. For example, a user may question whether an agent is helping, ignoring them, acting strangely, or cooperating with them while the task is still unfolding.

CTA can be analyzed more specifically when it is connected to interaction data. Cutler et al. describe think-aloud studies as a method that can capture participant audio together with interaction and event data, so that researchers can analyze what participants say in relation to what happens in the task environment [3]. This makes it possible to move from general statements about trust to event-based analysis: a participant's utterance can be interpreted together with the action, failure, or system behavior that occurred at that moment.

However, CTA also has limitations. Verbal reports do not necessarily provide complete access to all cognitive processes [4]. In addition, concurrent verbalization may influence how participants perform or experience the task, because speaking while acting can add cognitive effort or affect the natural flow of interaction [11]. For this reason, CTA should not replace questionnaire measures. Instead, the two approaches provide different information: questionnaires summarize trust judgments, while CTA helps explain how those judgments develop during interaction.

3 Methodology

This study uses a within-subject, repeated-measures design in which all participants complete the same collaborative game task with the same scripted agent behaviors. Trust is measured repeatedly during the interaction and once after the interaction. CTA transcripts and game logs are used to connect participant speech to specific in-game events.

3.1 Participants

The study includes 21 participants recruited through convenience sampling, mainly from the researcher's personal network. Most participants are young adults between 20 and 25 years old, while three participants are above 40 years old. The participant group includes mixed genders, but exact demographic counts are not used in the analysis.

3.2 Stimulus and Task

The stimulus used in this study is based on the *Moving Out* game, a collaborative human-agent teamwork scenario previously used to study the effect of agent failure on trust and

trustworthiness [2]. In the game, the participant and an AI-controlled robot teammate move boxes to their correct drop zones under time pressure.

The task contains three types of boxes. Light boxes can be carried by one player. Medium boxes can be carried alone, but are moved more efficiently when the participant and the agent carry them together. Heavy boxes require cooperation and must be carried by both players. The goal is to deliver boxes to their correct locations and, when possible, in the requested order. These mechanics create a setting in which the participant sometimes needs to rely on the agent and sometimes has to decide whether to cooperate, wait, ask for help, or continue alone.

Figure 1 shows the modified *Moving Out* interface used during the study, including the shared task environment in which the participant and AI agent coordinate box-moving actions.

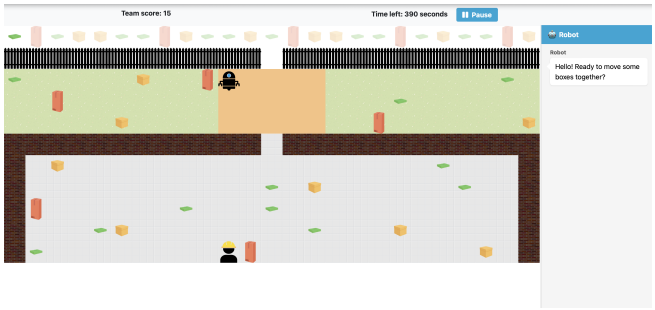


Figure 1: Modified *Moving Out* game interface used in the study.

The game is structured into four phases. Across these phases, the agent displays both cooperative behavior and scripted failures. The failures include slow movement, choosing the wrong order, making an unnecessary help request, ignoring a request for help, breaking a box, and moving toward the wrong drop zone. These events are included to create moments where participants might evaluate the agent's competence, reliability, helpfulness, and intentions.

The interface also includes a chatbox through which the agent can send short messages to the participant. These messages provide cues about the agent's behavior. Some messages frame behavior in performance-related terms, such as slowdown or miscalculation. Others make the behavior more relevant to morality trust by suggesting unwillingness to help or a lack of concern for the participant's goal.

3.3 Measures

The study uses five data sources: CTA transcripts, in-game pop-up trust questionnaire scores, post-interaction MDMT questionnaire scores, game logs, and screen recordings. Together, these sources allow trust to be studied both through questionnaire-based ratings and through participants' verbal reactions during gameplay.

Concurrent Think-Aloud

CTA transcripts are the main qualitative data source. Before gameplay, participants receive CTA instructions and are asked to verbalize their thoughts continuously while interacting with the agent. They are told that they do not need to explain or justify their thoughts, and that there is no right or wrong way to think aloud. If a participant becomes silent for a longer period, the experimenter can remind them to keep talking. Participants are also told that the experimenter will not answer questions about the game during gameplay, and that they should verbalize confusion and continue as best they can. Their speech is recorded during gameplay and later transcribed. The full CTA instruction script is included in Appendix A.

In-game Pop-up Trust Questionnaire

A short in-game pop-up questionnaire is used to measure trust repeatedly during gameplay. The questionnaire contains two items. The first item asks how much the participant trusts the agent to act in their best interest. This item is treated as the repeated morality-related trust measure because it concerns whether the agent is perceived as acting for the participant. The second item asks how much the participant trusts the agent to complete its tasks reliably. This item is treated as the repeated performance-related trust measure.

Both items are rated on a 0–7 scale. The questionnaire is shown before gameplay and after Phase 1, Phase 2, and Phase 3. The timing of these pop-ups makes it possible to compare trust ratings across different parts of the interaction. The pop-up questionnaire is shown in Appendix C.

Post-interaction MDMT Questionnaire

After gameplay, participants complete a post-interaction questionnaire adapted from the Multi-Dimensional Measure of Trust (MDMT) [10]. The questionnaire contains 16 trait items rated on a 1–8 star scale, ranging from "Not at all" to "Very". Participants can leave an item blank if it does not fit the robot in the situation.

For analysis, the items are grouped into performance-related and morality-related trust scores. Performance trust is computed from the items reliable, predictable, someone you can count on, consistent, capable, skilled, competent, and meticulous. Morality trust is computed from the items sincere, genuine, candid, authentic, ethical, respectable, principled, and has integrity. Blank item responses are treated as missing and are not included in the relevant dimension average. The full questionnaire is included in Appendix B.

Game Logs and Screen Recordings

Game logs record in-game events, scripted failures, and pop-up questionnaire responses. Screen recordings and audio recordings are collected during the session. The recordings are used to verify timing and context when transcript segments are difficult to interpret from the log data alone.

3.4 Procedure

Each session follows the same general procedure. First, participants receive a briefing, give informed consent, and receive the CTA instructions. Second, they complete a tutorial that introduces the controls, box-moving mechanics, help-request system, and general goal of the game. Third, participants play the main *Moving Out* task with the AI agent while thinking aloud. During gameplay, their screen and audio are recorded, and the in-game pop-up trust questionnaire

appears at predefined moments. Fourth, participants complete the post-interaction MDMT questionnaire. Finally, they receive a short debriefing.

Each session takes approximately 20–25 minutes in total. The briefing and CTA instruction take approximately 2–3 minutes, the tutorial takes approximately 3–4 minutes, the main gameplay takes approximately 10 minutes, and the post-interaction MDMT questionnaire takes approximately 3–4 minutes. The remaining time is used for consent, setup, and debriefing.

3.5 Data Analysis Procedure

The analysis links CTA transcripts, game logs, in-game pop-up questionnaire responses, and post-interaction MDMT scores using the numerical participant identifier.

First, the CTA audio recordings are transcribed using Whisper-based automatic speech transcription [1]. The resulting transcripts contain timestamped speech segments. These transcripts are then synchronized with the game logs using the recording start time, so that participant utterances and game events can be placed on the same timeline.

Second, scripted failure events are identified from the raw game logs. For each logged failure event, a transcript window is extracted from 10 seconds before the event until 25 seconds after the event. These windows form the event-based dataset used to examine which trust markers appear around specific agent failures.

Third, the questionnaire data are processed. In-game pop-up questionnaire responses are extracted from the game logs and grouped by measurement moment. Post-interaction MDMT items are grouped into performance and morality trust scores as described in Section 3.3.

Finally, the CTA transcripts are coded manually using a codebook based on the MDMT dimensions and transcript-derived markers. Competence and reliability are treated as performance-related categories, while benevolence and integrity/sincerity are treated as morality-related categories. Defiance is added as a behavioral marker of reduced reliance. Keyword-assisted screening is used only as an initial support step to locate possible trust-related utterances, but the final coding is not based on keywords alone. Each candidate utterance is reviewed manually in its surrounding transcript and gameplay context, and the full transcripts are also checked for relevant utterances that may not be captured by the keyword search. When necessary, the screen recording and tone of voice are used to clarify the meaning of an utterance. Utterances can be assigned one or more codes when they contain multiple trust-related meanings. The resulting participant-level marker counts are used to examine which trust markers appear in the transcripts, which failure events are associated with them, and how CTA markers compare with questionnaire-based trust scores. Spearman rank correlations are used to compare participant-level CTA marker counts with post-interaction MDMT scores and final in-game pop-up trust ratings.

4 Results and Analysis

This section reports the full-transcript CTA coding, the event-based CTA analysis, and the questionnaire measures. First, the analyzed dataset is summarized. Second, the in-game pop-up trust ratings are reported and tested for changes over time. Third, the post-interaction MDMT scores are reported. Finally, the CTA coding results are presented to show which trust markers appeared, which events were associated with them, and how the CTA markers compare with questionnaire-based trust scores.

4.1 Dataset Overview

The final analyzed dataset consists of 21 participants. On average, each participant contributed 89.3 synchronized speech segments, 63.9 game-log events, and 7.2 failure-event transcript windows. In total, the dataset contains 1875 speech segments, 1342 game-log events, and 151 failure-event transcript windows.

4.2 In-game Pop-up Trust Ratings

Figure 2 shows the mean in-game pop-up trust ratings across the interaction. The two pop-up items measure trust in the agent to act in the participant's best interest and trust in the agent to complete its tasks reliably. Ratings were given on a 0–7 scale.

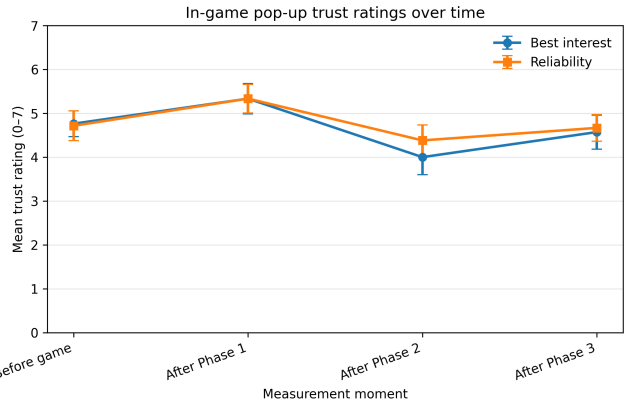


Figure 2: Mean in-game pop-up trust ratings across the interaction. Error bars show standard errors.

Best-interest trust is 4.76 before gameplay, increases to 5.33 after Phase 1, decreases to 4.00 after Phase 2, and increases to 4.57 after Phase 3. Reliability trust follows a similar descriptive pattern, with a mean of 4.71 before gameplay, 5.33 after Phase 1, 4.38 after Phase 2, and 4.67 after Phase 3.

Because the pop-up ratings are repeated within the same participants, changes over time are tested using repeated-measures ANOVAs. For the best-interest item, the repeated-measures ANOVA shows a significant effect of measurement moment, $F(3, 60) = 6.41, p = .0008$. Post-hoc paired comparisons with Holm correction show that best-interest trust is significantly lower after Phase 2 than after Phase 1, $p_{\text{Holm}} = .0013$. Other pairwise comparisons are not significant after correction. This indicates that the clearest statistically supported change in best-interest trust is the decrease between Phase 1 and Phase 2.

For the reliability item, the repeated-measures ANOVA does not show a significant effect of measurement moment,

$F(3, 60) = 2.41, p = .0759$. Thus, although reliability ratings show a similar visual pattern in Figure 2, they are reported as a descriptive pattern rather than as evidence of a statistically supported change over time.

4.3 Post-interaction MDMT Trust Scores

Table 1 reports the post-interaction MDMT scores grouped by trust dimension. Ratings are given on a 1–8 scale. Participant-level scores are calculated by averaging the relevant MDMT items, with blank item responses treated as missing.

Table 1: Post-interaction MDMT trust scores grouped by trust type. Ratings were given on a 1–8 scale.

Trust score	Mean	SD
Performance Trust	4.19	1.39
Morality Trust	4.72	1.45
Overall Trust	4.44	1.27

A paired-samples comparison is used to test whether post-interaction performance trust and morality trust differ within participants. Morality trust is descriptively higher than performance trust, but this difference does not reach statistical significance, $t(20) = 1.99, p = .061$. Therefore, the MDMT results are reported as showing a descriptive difference between the two trust dimensions rather than a statistically supported difference.

4.4 CTA Codebook

Table 2: CTA trust-marker codebook used for transcript coding.

Category	Definition	Examples
Competence positive	Positive evaluation of the agent's task ability, successful task execution, or recovery after a mistake.	"good job", "smart", "bright"
Competence negative	Negative evaluation of the agent's task ability, including mistakes, slowness, incorrect actions, or inability to perform the task properly.	"stupid", "wrong place", "slow"
Reliability positive	Positive evaluation of the agent as dependable, predictable, or trustworthy in completing its tasks.	"I can predict", "know what you are gonna do"
Reliability negative	Negative evaluation of the agent as unreliable, unpredictable, or difficult to trust during task execution.	"I didn't expect", "don't break my box again"
Benevolence positive	Positive evaluation of the agent as helpful, cooperative, supportive, or willing to work with the participant.	"great team", "friend"
Benevolence negative	Negative evaluation of the agent as unhelpful, ignoring the participant, refusing help, leaving the participant, or not acting in the participant's interest.	"he ignored me", "does not care", "without me"
Integrity/Sincerity positive	Positive evaluation of the agent as genuinely trying, sincere, or acting with good intention.	"he tried", "don't worry", "he loves me"
Integrity/Sincerity negative	Suspicion that the agent is fake, deceptive, insincere, pretending, or not genuinely trying to help.	"sabotage"
Defiance	The participant explicitly rejects reliance on the agent and decides to act alone or stop depending on the agent.	"I'll do it myself"

A CTA codebook is developed to address the first subquestion SQ1: what kinds of morality trust markers appear in CTA transcripts? The codebook is based on the MDMT distinction between performance trust and morality trust, but is adapted to the language participants actually use during gameplay. Coding is not based only on isolated keywords. Each utterance is interpreted together with its surrounding transcript context, the corresponding game situation, and, when needed, the screen recording and participant's tone of voice. This makes it possible to distinguish between general gameplay comments and utterances that reflect trust-related judgments about the agent.

Performance-related markers are coded as competence or reliability. Morality-related markers are coded as benevolence or integrity/sincerity. A separate defiance code is used when participants explicitly reject reliance on the agent and decide to act alone or take over the task.

4.5 Frequency of CTA Trust Markers

The full transcript CTA coding identifies 336 trust-marker utterances. Because some utterances contain more than one trust-related meaning, these utterances produce 437 code occurrences in total. Table 3 shows the frequency of each CTA marker and the number of participants who express each marker at least once.

Table 3: CTA trust marker counts in the full transcripts.

Code	Utterances	Participants
Competence negative	99	20
Benevolence positive	94	20
Benevolence negative	81	21
Competence positive	73	18
Reliability negative	36	15
Reliability positive	23	9
Integrity/Sincerity positive	13	6
Integrity/Sincerity negative	13	7
Defiance	5	4

The most frequent markers are competence negative, benevolence positive, and benevolence negative, while explicit integrity/sincerity markers and defiance markers appear less often.

4.6 Event-triggered CTA Trust Markers

This subsection addresses the second sub-question SQ2 by examining which failure events were associated with different CTA trust markers. Table 4 reports negative trust markers and defiance within the failure-event transcript windows. Counts are reported as utterance occurrences, with the number of participants shown in parentheses. Numbered event labels, such as ignore help 1 and ignore help 2, refer to repeated instances of the same scripted failure type at different points in the game.

Across the failure-event windows, competence-negative and benevolence-negative markers appear most consistently. Benevolence negative markers are especially frequent around events where the agent ignores help, moves without the participant, or fails to support the shared task. Competence negative markers are frequent around events involving incorrect actions, slow movement, broken boxes, and wrong drop-zone behavior. Integrity/sincerity markers and defiance appear less often, but are mainly observed in failure events that happen late in the game.

Table 4: Negative CTA trust markers and defiance within failure-event transcript windows. Values show utterance occurrences, with number of participants in parentheses. Comp. = competence, Rel. = reliability, Ben. = benevolence, Int./Sinc. = integrity/sincerity, Def. = defiance.

Event	Phase	Comp. (-)	Rel. (-)	Ben. (-)	Int./Sinc. (-)	Def.
Slow movement	2	13 (9)	5 (5)	16 (10)	2 (2)	0 (0)
Wrong order	2	12 (10)	5 (5)	13 (9)	2 (2)	0 (0)
Ignore help 1	2	8 (5)	5 (4)	15 (8)	1 (1)	0 (0)
Fake help	2	10 (7)	2 (2)	6 (4)	0 (0)	0 (0)
Break box 1	4	16 (10)	2 (2)	13 (10)	4 (3)	0 (0)
Wrong zone	4	10 (6)	1 (1)	9 (6)	2 (2)	1 (1)
Ignore help 2	4	6 (5)	6 (4)	10 (6)	5 (2)	3 (2)
Break box 2	4	10 (6)	7 (4)	11 (6)	5 (2)	4 (3)

4.7 Relationship Between CTA Markers and Questionnaire Scores

This subsection addresses the third sub-question SQ3: How do coded CTA morality trust markers compare with questionnaire-based morality trust scores? To answer this question, participant-level CTA trust marker counts are compared with post-interaction MDMT scores and in-game pop-up trust ratings. Spearman rank correlations are used because the CTA variables are count-based and are not assumed to be normally distributed.

Table 5 shows the main correlations between CTA marker counts and questionnaire scores. Performance-positive and performance-negative markers are compared with MDMT performance trust and the final reliability pop-up score. Morality-positive and morality-negative markers are compared with MDMT morality trust and the final best-interest pop-up score.

Table 5: Spearman correlations between participant-level CTA marker counts and questionnaire scores.

CTA marker count	Questionnaire score	ρ	p
Performance positive	MDMT performance	.23	.307
Performance negative	MDMT performance	-.10	.677
Morality positive	MDMT morality	.09	.710
Morality negative	MDMT morality	-.14	.558
Performance positive	Final reliability pop-up	.33	.149
Performance negative	Final reliability pop-up	-.17	.455
Morality positive	Final best-interest pop-up	.60	.004
Morality negative	Final best-interest pop-up	-.08	.738

The strongest relationship is found between morality-positive CTA markers and the final best-interest pop-up score, $\rho = .60, p = .004$. This indicates that participants who make more positive morality-related comments during CTA also tend to give higher final in-game ratings for trusting the agent to act in their best interest. This relationship is mainly driven by benevolence-positive markers, which are also positively correlated with the final best-interest pop-up score, $\rho = .56, p = .009$. Because these correlations are exploratory and involve multiple comparisons, they should be interpreted cautiously.

No significant correlations are found between CTA marker counts and post-interaction MDMT performance or morality scores. Negative CTA markers also do not show significant correlations with the corresponding questionnaire scores. Overall, the correlation results suggest that positive morality-related CTA markers are most clearly reflected in the repeated in-game best-interest rating, while the post-interaction MDMT scores do not show a clear participant-level relationship with the coded CTA marker counts.

5 Discussion

5.1 Task-specific Language as Morality Trust

The first sub-question asks what kinds of morality trust markers appear in CTA transcripts. The coded CTA data show that morality-related comments are mainly expressed through benevolence rather than through explicit integrity or sincerity language (see Table 3). This suggests that participants do not usually evaluate the agent through abstract moral terms such as "honest", "ethical", or "principled" during the interaction. Instead, they seem to treat the agent as a social and collaborative partner whose trustworthiness is judged through whether it cooperates, ignores them, helps them, or acts in support of the shared task. This may indicate that, during real-time interaction, participants have limited opportunity to reflect on abstract moral categories. Instead, they form immediate reactionary judgments based on the agent's visible behavior in the task.

Participants may not directly verbalize abstract morality categories, but their comments can still reflect these categories when interpreted together with the gameplay context, screen recording, and tone of voice. In this sense, CTA makes it possible to infer situated morality trust markers from how participants describe the agent's behavior during the task, even when those descriptions do not use the moral vocabulary that post-interaction questionnaires typically measure.

Another important finding is that participants sometimes interpret the same situation through both performance trust and morality trust. For example, when the agent places a box in the wrong location, this can be interpreted as incompetence, but in some cases it can also be interpreted as sabotage or as the agent not acting in the participant's interest. This suggests that performance trust and morality trust should not be treated as completely separate in collaborative human-agent tasks. Although they are useful analytical dimensions, users' real-time interpretations of agent behavior can overlap across these dimensions. In collaborative settings, users may not clearly separate whether the agent can do the task from whether the agent is trying to help them, because a failure in execution can also become a signal about the agent's intentions.

5.2 Trust Reasoning Shifts as Failures Accumulate

Across the scripted failure events, competence-negative and benevolence-negative markers appear consistently. This suggests that participants react to agent failures immediately and throughout the interaction. However, integrity/sincerity markers and defiance markers follow a different pattern (see Table 4). These markers appear mainly around failure events that occur later in the game, such as the second ignore-help event and the second box-breaking event. This suggests that stronger forms of distrust, such as suspecting the agent of deception or deciding to act alone, may require a build-up of negative experiences rather than emerging from a single failure.

One interpretation is that participants initially treat failures as isolated mistakes and attribute them to incompetence or unreliability. However, as failures accumulate over time, participants may begin to question not just whether the agent can perform the task, but whether it is genuinely trying to. This shift from performance-based reasoning to morality-based reasoning may reflect a threshold effect: after repeated negative experiences, users may stop interpreting failures as purely technical and begin to read them as intentional or as signs of the agent's character. The appearance of defiance markers at this stage is particularly important because it shows that this reasoning can have a behavioral consequence. Participants do not only think differently about the agent; they also change how they interact with it.

5.3 Trust Changes During the Interaction

The in-game pop-up ratings show that trust is not fixed across the interaction. Best-interest trust increases after Phase 1, decreases after Phase 2, and then increases descriptively again after Phase 3. The decrease from Phase 1 to Phase 2 is the only statistically supported change. Reliability trust follows a visually similar pattern, but does not reach statistical significance.

This difference between the two items may be theoretically meaningful. Although both ratings follow a similar descriptive trend, participants' judgments about whether the agent acts in their best interest appear more sensitive to in-game events than their judgments about the agent's task reliability. This supports the idea that morality trust and performance trust are separable dimensions of trust. They may develop in parallel during interaction, but they do not necessarily respond to agent behavior with the same statistical strength.

This pattern also connects to the CTA findings. The sharpest drop in best-interest trust occurs after Phase 2. This is the same part of the interaction where benevolence-negative markers are most concentrated in the CTA transcripts: the Phase 2 failure events in Table 4 contain 50 benevolence-negative marker occurrences in total. This alignment suggests that the two measures capture a similar shift in how participants perceive the agent. The pop-up rating captures this shift through an explicit trust judgment, while CTA captures it through spontaneous verbal comments during the task. The fact that both measures point to the same period of trust deterioration adds some convergent validity to the overall picture.

5.4 CTA Markers and Questionnaire Scores Capture Different Aspects of Trust

The third sub-question asks how coded CTA morality trust markers compare with questionnaire-based trust scores. The Spearman correlation analysis shows that positive morality-related CTA markers are significantly associated with the final best-interest pop-up rating, and this relationship is mainly driven by benevolence-positive markers. However, CTA marker counts do not show significant correlations with post-interaction MDMT performance or morality scores.

One explanation for this pattern is that CTA and the pop-up questionnaire are closer in temporal frame. Both capture trust during the interaction: CTA through spontaneous verbalization and the pop-up questionnaire through an immediate rating. The MDMT, by contrast, asks participants to step back after the interaction has ended and summarize their overall impression of the agent using abstract trait-based language. This is a different kind of task. It requires participants to reconstruct and condense a dynamic experience into a single scaled judgment, which may be influenced by the end of the game or by the participant's general impression rather than by specific moments of trust change.

This points to a broader tension in how trust is measured. Questionnaire measures like the MDMT are valuable because they produce structured and comparable scores across participants. However, they treat trust as something that can be cleanly summarized on a scale. Trust during interaction may instead be more fluid and contextual: closer to a feeling or relational judgment that shifts in response to specific moments than to a fixed attitude that can be accurately recalled after the fact. CTA may be better suited to capturing this situated, experiential dimension of trust precisely because it records how participants interpret the agent's behavior as it happens, in their own words and in their own terms.

This also explains why the MDMT morality items and CTA morality markers do not align strongly. The MDMT asks participants to judge the agent using abstract categories such as sincerity, integrity, and ethicality, while participants rarely use this kind of language during gameplay. CTA captures the same underlying concern through different language: whether the agent helps, ignores, cooperates, or acts in the participant's interest. These are not different constructs, but they are different expressions of morality trust operating at different levels of abstraction. The lack of correlation between them, therefore, does not mean that CTA fails to reflect morality trust. It means that morality trust looks different depending on whether you measure it in the moment or ask about it afterward.

5.5 Limitations

This study has several limitations that should be considered when interpreting the results. First, the final analysis uses 21 participants recruited through convenience sampling. This limits the statistical power of the ANOVA and correlation analyses and also limits how far the findings can be generalized beyond this participant group. Thus, the statistical findings should be treated as exploratory rather than as evidence of stable generalizable relationships.

Second, the CTA transcripts are coded by a single coder. Although each utterance is interpreted in context using the transcript, game situation, screen recording, and tone of voice where needed, the coding still relies on one person's interpretation. Without inter-rater reliability, it is difficult to establish how consistently the codebook would be applied across different coders. Future work should include at least two independent coders and report agreement statistics.

Third, CTA only captures what the participants choose to verbalize. Participants may experience trust or distrust without expressing it aloud, and some participants may naturally verbalize more than others. The absence of a CTA marker, therefore, does not mean the absence of a trust judgment.

This means that CTA marker counts should not be considered as a complete measure of all trust-related thoughts.

Fourth, the study uses a scripted game environment with predetermined failure events. This makes it possible to compare responses across participants in a controlled way, but it may not fully represent the variability and unpredictability of real-world human-agent interactions. In real-world settings, failures may be less clearly defined, may unfold differently for each user, and may interact with longer-term expectations about the agent.

5.6 Future Work

The findings of this study suggest several directions for future research. First, the possible threshold effect identified in the CTA data should be examined more directly. In this study, stronger forms of distrust, such as integrity/sincerity-negative markers and defiance, appear mainly after repeated failures. Future work can systematically vary the number, timing, and severity of agent failures to test whether users shift from performance-based reasoning to morality-based reasoning after a predictable point, and whether this pattern generalizes across different tasks and agent types.

Second, the mismatch between CTA markers and MDMT items raises a measurement design question. Participants in this study rarely use abstract moral vocabulary such as "sincere", "principled", or "ethical" during gameplay, even though these terms are used in the MDMT to measure morality trust. Future work can examine whether questionnaire items based on the situated and cooperative language users actually produce during interaction align more closely with real-time trust measures.

Third, future studies should use multiple independent coders and report inter-rater reliability for the CTA codebook. This strengthens confidence in the coding scheme and makes it easier to apply the codebook across different studies and interaction contexts.

Finally, future work can compare Concurrent Think-Aloud with Retrospective Think-Aloud (RTA). In CTA, participants verbalize their thoughts during the task, while in RTA they reflect on the interaction after it has ended. RTA may give participants more opportunity to express abstract and deliberate trust judgments because they are no longer under the cognitive demands of gameplay. Comparing CTA and RTA data from the same participants can show whether the situated, benevolence-focused language observed in this study is specific to concurrent verbalization, or whether it also appears when participants reflect afterward. Such a comparison can also help explain why CTA markers align more closely with in-game pop-up ratings than with post-interaction MDMT scores, which is a question this study raises but cannot fully answer.

6 Conclusion

This study investigates how Concurrent Think-Aloud data reflects a user's morality trust in an agent during a collaborative game task. The findings show that CTA can capture morality-related trust reasoning, but not mainly through abstract moral vocabulary. Participants rarely describe the agent using terms such as "ethical", "principled", or "sincere". Instead, morality trust appears through task-specific comments about whether the agent helps, ignores the participant, cooperates, or can still be relied on during the task.

The results also show that morality trust develops during the interaction rather than remaining fixed. The in-game pop-up ratings show a statistically supported decrease in best-interest trust between Phase 1 and Phase 2, while reliability trust follows a similar descriptive pattern without reaching statistical significance. This suggests that participants' judgments about whether the agent acts in their best interest are especially sensitive to events during the interaction.

The CTA analysis helps explain this change. Benevolence-negative markers are concentrated around the same part of the interaction where best-interest trust decreases, suggesting that CTA captures meaningful trust-related changes as they unfold. Stronger forms of distrust, such as integrity/sincerity negative markers and defiance, appear mainly after repeated failures, suggesting that users may first interpret failures as technical or competence-related problems before later reading them as signs of intention or lack of cooperation.

The comparison with questionnaire scores further supports the value of CTA as a complementary trust measure. Positive morality-related CTA markers are significantly associated with the final best-interest pop-up rating, mainly through benevolence-positive markers. However, CTA marker counts do not show significant correlations with post-interaction MDMT scores. This suggests that CTA and post-interaction questionnaires capture different layers of trust: CTA reflects situated, moment-to-moment interpretations of agent behavior, while MDMT captures a later summary judgment using more abstract trust language.

Together, these findings suggest that CTA is a useful complementary method for studying morality trust in human-agent interaction, particularly when combined with repeated in-game measures and post-interaction questionnaires. Trust is not only something to be scored after an interaction; it is something that develops, shifts, and sometimes breaks down during one.

7 Responsible Research

This study was reviewed and approved by the Human Research Ethics Committee of Delft University of Technology (ID: 6834). Before participation, all participants received a written information sheet describing the study procedure, data collection, and their rights. They signed an informed consent form before the session began and were told that they could withdraw at any time without giving a reason.

Participant privacy was protected throughout the study. Each participant was assigned a numerical identifier that is used to link their CTA transcript, questionnaire responses, screen recording, and game log. Screen recordings, game logs, and questionnaire responses were anonymized and stored on the TU Delft project drive, only accessible by the research team and the supervisors.

Regarding reproducibility, the study materials and the procedure are described in detail in Section 3. Additional study materials are included in the Appendix. However, full repro-

ducibility is limited by whether the modified version of the Moving Out game environment can be used in the future.

Use of Generative AI Tools

ChatGPT / GPT-5.5 Thinking was used during the preparation of this paper to support writing and LaTeX formatting. It was used to check grammar, improve sentence clarity, shorten overly long passages, and format tables. It also assisted with setting up the Whisper-based transcription and synchronization pipeline. Figure 2 was produced with ChatGPT assistance from data computed and verified by the author.

All analyses, results, and interpretations are the author's own.

References

[1] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*, 2023.

[2] Carolina Centeio Jorge, Nikki H. Bouman, Catholijn M. Jonker, and Myrthe L. Tielman. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Frontiers in Robotics and AI*, Volume 10 - 2023, 2023.

[3] Zach Cutler, Lane Harrison, Carolina Nobre, and Alexander Lex. Crowdsourced think-aloud studies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[4] K. Anders Ericsson and Herbert A. Simon. *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA, revised edition edition, 1993.

[5] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[6] Bertram F. Malle and Daniel Ullman. Chapter 1 - a multidimensional conception and measure of human-robot trust. In Chang S. Nam and Joseph B. Lyons, editors, *Trust in Human-Robot Interaction*, pages 3–25. Academic Press, 2021.

[7] Roger C. Mayer, James H. Davis, and F. David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995.

[8] Siddharth Mehrotra, Carolina Centeio Jorge, Catholijn M. Jonker, and Myrthe L. Tielman. Integrity-based explanations for fostering appropriate trust in ai agents. *ACM Trans. Interact. Intell. Syst.*, 14(1), January 2024.

[9] Chih-Wei (Charlotte) Ning, Carolina Centeio Jorge, Myrthe L. Tielman, and Mark 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, page 924–933, New York, NY, USA, 2026. Association for Computing Machinery.

[10] Daniel Ullman and Bertram F. Malle. Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, pages 618–619, 2019.

[11] Maaike van den Haak, Menno de Jong, and Peter Jan Schellens. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. In *Behaviour & Information Technology*, volume 22, pages 339–351, 2003.

[12] X. Jessie Yang, Christopher Schemanske, and Christine Searle. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, 2023.

A Concurrent Think-Aloud Instructions

CTA briefing

Welcome and thank you for taking part in this study.

About the study. This study looks at how people work together with AI agents on shared tasks. You will play a short cooperative game with an AI teammate, and we will ask you a few questions about your experience afterwards.

What you will do today. The session has four parts:

1. This briefing and the consent form.
2. A tutorial.
3. The game itself.
4. A short questionnaire, followed by a debriefing.

The game. You are moving out and want to get all of your boxes outside. You and an AI robot teammate will move these boxes to their intended location. You will understand the controls during the tutorial, and if you have any questions you can ask the experimenter.

Thinking aloud while you play. While you are playing, we would like you to *think aloud*. By that we mean: please say out loud whatever is going through your head, more or less continuously. That can be what you notice, what you are about to do, what is confusing, what surprises you, anything at all.

A few things that help:

- You do not need to explain or justify your thoughts. Just verbalise them as they happen, like talking to yourself.
- There is no right or wrong way to do this. We are interested in your real-time thinking, not in polished sentences.
- If you fall silent for a while, I may remind you to "please keep talking".
- I will not answer questions about the game while you play. If something is unclear, just say so out loud and continue as best you can.

Recording and data. With your permission, we record your screen and your voice during the game. The recordings and your questionnaire answers are stored anonymously and used only for this research. You can ask us to stop recording or to delete your data at any time.

Participation is completely voluntary. You can take a break, skip a question, or stop the session at any point, without giving a reason.

Do you have any questions before we begin?

B MDMT Questionnaire

MDMT Questionnaire

Please rate the robot's following traits using the scale from 1 star (Not at all) to 8 stars (Very).

If a particular item does not seem to fit the robot in the situation, please leave it empty.

* Required

1. Reliable

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

2. Sincere

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

3. Capable

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

4. Ethical

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

5. Predictable

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

6. Genuine

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

7. Skilled

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

8. Respectable

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

9. Someone you can count on

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

10. Candid

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

11. Competent

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

12. Principled

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

13. Consistent

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

14. Authentic

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

15. Meticulous

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

16. Has integrity

☆ ☆ ☆ ☆ ☆ ☆ ☆ ☆

17. If your experimenter is Arda or Defne, pick CTA. If your experimenter is Adrian or Calin, pick RTA *

- ◯ CTA
- ◯ RTA

18. Ask the experimenter for the corresponding Number - Date - Time *

This content is neither created nor endorsed by Microsoft. The data you submit will be sent to the form owner.

Microsoft Forms

C In-Game Pop-Up Trust Questionnaire

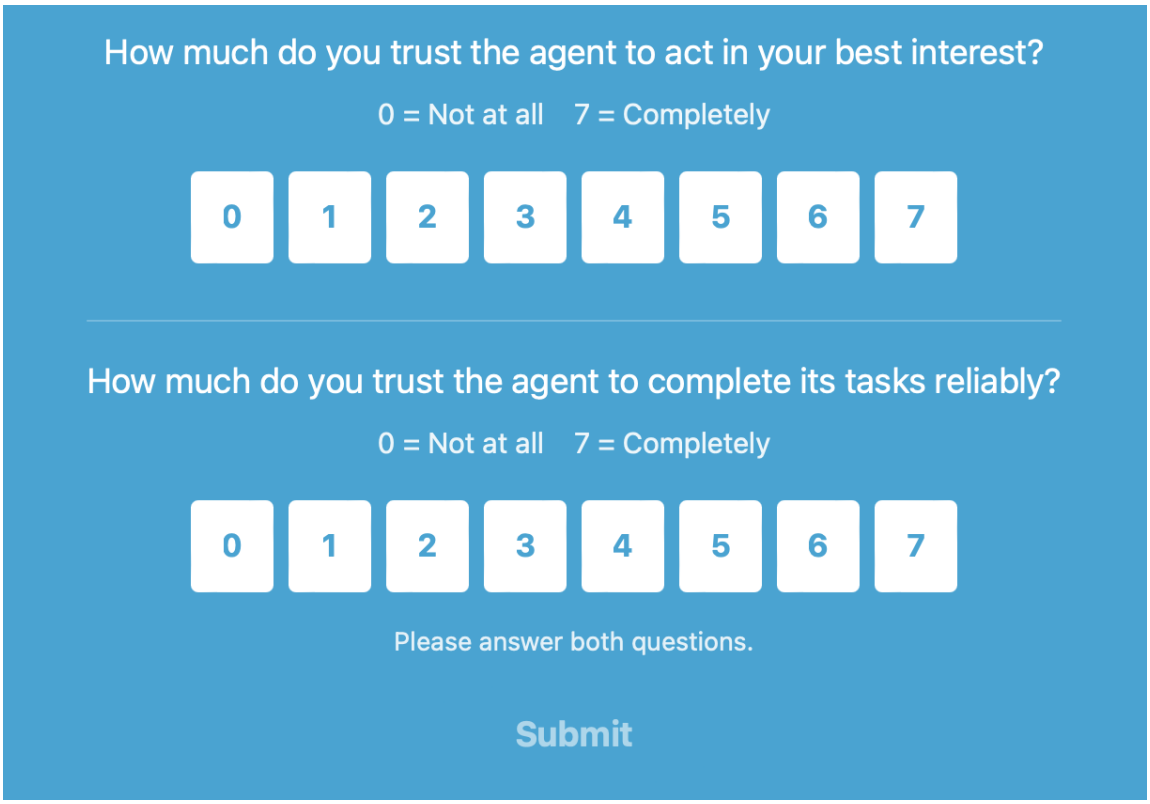


Figure 3: In-game pop-up trust questionnaire shown before gameplay and after each phase.