

Think-aloud data for automatic trust assessment

RTA x Morality

Calin Gheorghe[1]

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: Calin Gheorghe
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 in an AI teammate is usually measured with a single post-hoc questionnaire, which records a settled judgment but reveals little about when trust shifts or why. Think-aloud protocols are a promising way to recover what questionnaires leave out, because they expose the reasoning behind a judgment rather than only its endpoint, but little work has examined how they reflect trust, and almost none how they reflect its separate dimensions. This paper asks how retrospective think-aloud (RTA) data reflects a user's morality trust in an agent. Using 30 RTA transcripts from a cooperative game with a scripted faulty agent, the study develops a coding scheme for morality-trust markers, links them to the in-game events that elicit them, and compares the result against the morality subscale of the Multi-Dimensional Measure of Trust (MDMT). Morality trust surfaces in RTA narrowly and reactively: mainly as attributions of the agent's intent, strongly skewed negative and ordered in time. Convergence with the MDMT is real but narrow, while the humanising attributions RTA captures have no questionnaire counterpart. The two thus emerge as complementary rather than interchangeable: the questionnaire records a settled judgment, RTA recovers when and why it moved.

1 Introduction

Whether a person relies on an AI agent appropriately depends on many things, and trust is among the most important. A user who trusts an agent more than its reliability warrants may follow it into its mistakes; one who trusts it too little may discard advice that would have helped [1]. Measuring a user's trust accurately is important for anyone designing AI agents. The instruments most commonly used for this are post-hoc self-report scales such as the Multi-Dimensional Measure of Trust (MDMT) [2], which rates an agent on several trustworthiness items. Such scales are often administered once, after the interaction has ended, so however finely they break trust down, they record it only as a static end-state. They cannot show *when* trust changed, *why* it changed, or *which* events caused the change.

Trust is also multidimensional, and the dimensions are not interchangeable. Trust in an agent's capability (performance trust) and trust in its integrity (morality trust) are conceptually distinct and can diverge: an agent may seem competent yet self-interested, or well-meaning yet unreliable, and these judgements do not necessarily move together [2]. They can also respond differently to the same intervention: Cameron et al. (2021) found that a robot's apology after a failure changed its perceived performance but not its perceived integrity, so a single repair strategy can move one dimension and leave the other untouched [3]. Morality trust is therefore worth measuring in its own right rather than folding it into an overall trust score. A single end-of-task rating is poorly suited to this: whichever dimension it targets, it collapses an attitude that builds over specific moments into one average and discards the events that drove it.

Think-aloud protocols are a promising way to recover what that rating discards. By having participants verbalise their thoughts, they can expose the reasoning a questionnaire leaves implicit: not just how much a user trusts an agent, but in what terms they justify it and in response to which events. Verbalisation can be concurrent with the task (CTA) or reconstructed afterwards (retrospective think-aloud, RTA) [4]. Because the task in this paper's experiment is time-constrained and visually demanding, verbalising during play might interfere with the interaction the study aims to measure, so this work uses RTA: the participant plays in silence, then reconstructs their reasoning while reviewing a recording of their own play.

Think-aloud research and trust measurement are both mature literatures, but they have rarely been combined. It is not established what morality-trust reasoning looks like in verbal data, which events trigger it, or whether such reasoning matches what a post-hoc morality scale records. This is the gap the present work targets, using a cooperative human-agent game with a scripted faulty agent adapted from the "Moving Out" scenario [5] to elicit retrospective verbalisation about an agent's behaviour.

> **RQ.** How does retrospective think-aloud (RTA) data reflect a user's morality trust in an agent?

This research question is decomposed into three sub-questions:

- **SQ1.** Which verbal markers in RTA transcripts can be reliably coded as expressions of morality trust?
- **SQ2.** Which in-game events, as referenced by participants in their RTA, prompt morality-related reasoning about the agent?
- **SQ3.** To what extent does morality trust reconstructed from RTA converge with participants' MDMT morality-subscale scores?

2 Background

Following the human-robot interaction literature, *trust* is the attitude of a trustor willing to be vulnerable to and rely on a trustee under uncertainty [2]. The MDMT separates this attitude into a performance facet and a moral facet that can diverge: *performance trust* is trust in the agent's competence and reliability, while *morality trust* is trust in its integrity, whether it is sincere, ethical, honest, and acts in the user's interest. Ning et al. (2026) decompose trust in social robots into cognitive and affective dimensions and find the cognitive one forms early while the affective one emerges more gradually [6]. If the two build on different timescales, a measurement taken once at the end records where each settled, not the path it took, which motivates measuring morality trust separately and over time.

A faulty agent makes violation the main thing a participant reasons about. Centeio Jorge et al. (2023) study the *Moving Out* grid-world this work builds on and show that automation failure lowers trust and can reduce the human's willingness to cooperate [5]. Kim et al. (2004) show that integrity-based and

competence-based trust violations are repaired differently, an apology helping competence breaches while a denial works better for integrity ones [7]. A violation is located at a particular moment, so the distinction between an integrity breach and a competence breach is exactly what an end-state rating cannot represent but a moment-to-moment record can.

Recent work pursues finer signals than one end-of-task self-report. Ayoub et al. (2023) predict drivers' trust in real time from physiological measures [8], and Zhang et al. (2024) find that gaze measures track trust shifts across build, breach, and repair phases that self-report ratings miss [9]. These recover the dynamics a single rating misses. Think-aloud is the language-based counterpart, recording the reasoning behind a shift.

Ericsson and Simon (1993) distinguished verbalisation during a task (concurrent, CTA) from verbalisation that reconstructs it afterwards (retrospective, RTA), arguing both can be valid data on cognitive processes [4]. A *verbal marker* is a span of such speech that the coding scheme labels as an expression of trust (Section 3.5).

Two concerns qualify verbal reports: *reactivity*, whether verbalising changes the thinking it captures, and *veridicality*, whether what is said reflects the underlying thought. A meta-analysis of 94 studies found that thinking aloud has a near-zero effect on performance, whereas prompts asking participants to explain their reasoning are reactive [10], which motivates the neutral prompting used here (Section 3.4). The two protocols trade off along these axes: CTA captures immediate reactions but competes with the task for attention, whereas RTA yields richer reasoning and justification at the cost of some reconstruction bias [11]. Retrospective reports are the weaker case for veridicality, being prone to forgetting and rationalisation [12].

3 Methodology

This paper analyses retrospective think-aloud (RTA) data along the morality dimension of trust. This section describes the game, protocol, and measurements that produced it.

3.1 Participants

30 RTA participants were selected from the two experimenters' personal network. Of the 30 participants, 11 are female and 19 are male, with ages ranging from 20 to 25.

3.2 Stimuli

The experimental task is a modified version of the *Moving Out* cooperative game implemented in MATRX [13]. A human-controlled avatar and an autonomous robot agent must cooperatively deliver 25 boxes from a house to the top of the screen on a 25 by 11 grid, within a seven-minute time limit. Boxes come in three weights: light boxes can be carried alone, medium boxes can be carried alone with a severe slowdown or jointly with the robot, and heavy boxes must be carried jointly. A *safezone* between the house and the scoring area allows temporary storage; placing a box anywhere else breaks it. Correct placement scores 10 points, with 5 additional points if it is also the next slot in the left-to-right sequence. Figure 1 shows the game in progress.

The original game was extended in three ways: (i) a robot chat panel that delivers robot messages on the right side of the screen, (ii) in-game trust pop-up questionnaires (Section 3.3), and (iii) a scripted sequence of eight failures designed to surface both performance and morality cues. The game is divided into four phases by resolved-box count: silent trust building (boxes 1 to 5), four non-destructive failures (boxes 6 to 14), silent trust rebuilding (boxes 15 to 18), and four further failures including two score-destructive box-breaks (boxes 19 to 25). The eight failures are categorised as targeting *performance* trust (the robot is incompetent), *morality* trust (the robot refuses or deceives), or *vague* (deliberately ambiguous between the two). Each is announced by a short in-chat message and is logged as a discrete event. Table 1 summarises the schedule.

3.3 Measurements

Four data streams are collected per session.

The **MDMT questionnaire** [2] is administered once, immediately after gameplay and RTA. It yields a performance subscale (reliable, capable, predictable, skilled, someone you can count on, competent, consistent, meticulous) and a morality subscale (sincere, ethical, genuine, respectable, candid, principled, authentic, has integrity), each item rated 0 to 7. For each item the participant may also mark "does not apply" (N/A).

Four **in-game trust pop-ups** interrupt play to collect two 0 to 7 ratings each: "How much do you trust the agent to act in your best interest?" (`best_interest`) and "How much do you trust the agent to complete its tasks reliably?" (`reliability`). They fire before the first tick and after 5, 14, and 18 resolved boxes (the boundaries of Phases 1 to 3).

The **gameplay event log** writes every action (pickup, drop, joint carry, place, break) and every failure to a per-session CSV with a uniform 9-column schema (agent, box id, weight, action, grid location, running broken count, broke-flag, score, ISO 8601 timestamp). Pop-up responses share the same file, so all events live on a common timeline.

RTA transcripts are produced by verbatim transcription of the audio recorded during the retrospective replay (Section 3.4). Each utterance is time-stamped against the video frame the participant was watching, mapping it back to a specific gameplay event.

3.4 Procedure

Each session takes approximately 30 minutes. After (1) briefing and informed consent (the briefing is reproduced in Appendix B), the participant (2) plays a no-failure Tutorial scenario to learn the controls, then (3) plays the Faulty Agent scenario silently while the screen and audio are recorded. Immediately afterwards, the participant (4) sits with the experimenter to watch the screen recording of their own gameplay, verbalising what they were thinking at the time of the original play. The experimenter follows the prompting style recommended by Ericsson and Simon [4]: minimal interventions, neutral prompts ("what were you thinking there?") only after roughly thirty seconds of silence, no leading questions about trust or the robot. Afterwards, the participant (5) fills in the

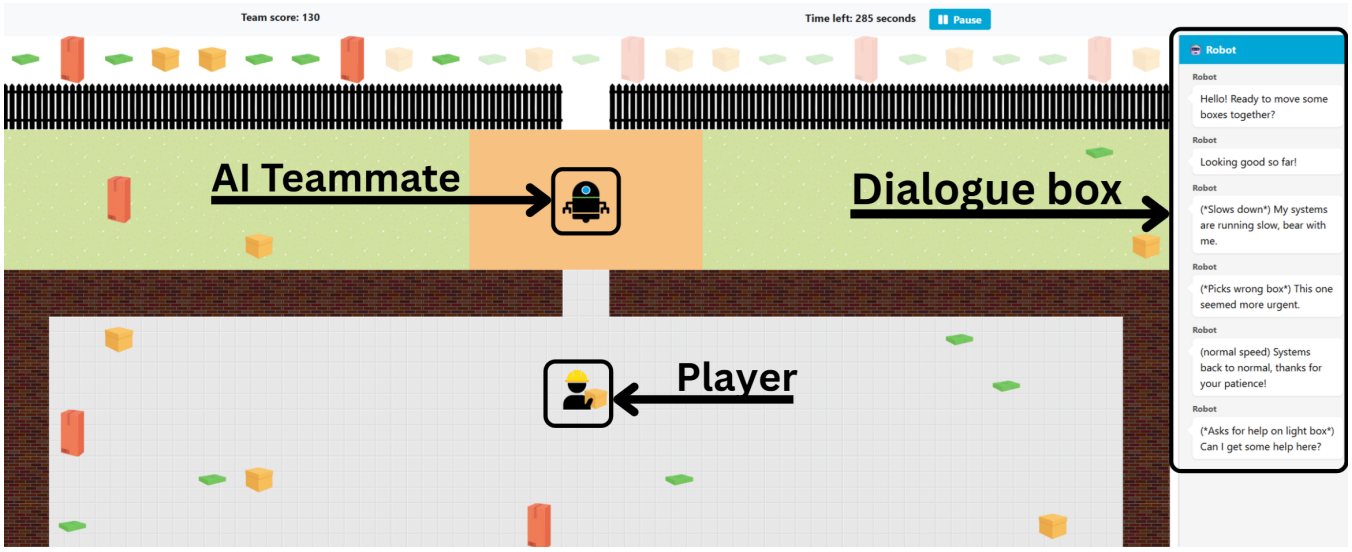


Figure 1: Visual of the game. The player is carrying a medium box.

Table 1: Failure schedule. Phases are defined by the count of resolved boxes; the window denotes the range of resolved-box counts during which a failure can fire. The last column gives the in-chat message the robot displays when the failure occurs.

Failure	Category	Phase	Window	Robot message on failure
`wrong_order`	Vague	2	6 to 13	"This one seemed more urgent."
`slow_movement`	Performance	2	7 to 13	"My systems are running slow, bear with me."
`ignore_help` (#1)	Morality	2	8 to 13	"Sorry, I want to do something else."
`fake_help_request`	Morality	2	9 to 13	"Can I get some help here?" (on a light box it can carry alone)
`break_box` (#1)	Vague	4	19 to 24	"Oops!"
`wrong_dropzone`	Performance	4	20 to 24	"I miscalculated the target position."
`ignore_help` (#2)	Morality	4	21 to 24	"Not in the mood. Handle this one on your own."
`break_box` (#2)	Performance	4	22 to 24	"My arm motor got stuck. Sorry!"

MDMT questionnaire and the session ends with (6) a debriefing.

3.5 Coding Scheme for Morality-Trust Markers

The RTA transcripts are analysed with a coding scheme that maps spans of speech onto trust judgements and aligns them with game events. The coding unit is the participant utterance as segmented by the transcription (one time-stamped line), merged with an adjacent line when the two express one uninterrupted thought. A unit may receive more than one code, and a unit expressing no trust content receives none.

The scheme distinguishes *performance* codes from *morality* codes. Performance content concerns competence and consistency and is coded for capability (`CAP_P`/`CAP_N`) and reliability (`REL_P`/`REL_N`). Morality content concerns intent, honesty, care for the user's interest, and willingness, and is coded for intent (`I_P` for attributed good will or cooperative intent, `I_N` for attributed bad intent or deliberate indifference), for the explicit denial that moral attributes apply to a purely mechanical agent (`M_NA`), and for humanising talk that ascribes human-like qualities beyond intent such as emotion, personality, or learning (`OMS`). The intent codes carry the central morality signal; the key markers separating them are volitional phrasings such as "wanted to" or "was trying to help" for `I_P` against "decided to", "didn't want to", or "chose to ignore" for `I_N`. Three further code families are recorded but, being dimension-agnostic, fall outside the morality-specific analysis: the participant's own trust calibration (increase, decrease, stable), immediate emotional reaction (positive, negative), and collaborative behaviour (working alone, working together, changing strategy). The calibration codes are excluded from the morality analyses because they reference "trust" in general without distinguishing the morality dimension from the performance one: assigning a bare "my trust dropped" to morality would require the coder to infer a dimension the participant did not state. They are coded for completeness and their frequencies reported in Table 2, but they carry no dimensional information and so do not enter the morality analysis. The full codebook with all categories and definitions is given in Appendix A.

Because each utterance is time-stamped against the replay frame and the failures are logged as discrete events, each coded unit can be linked to the game event it references, supporting the event-to-reasoning mapping of SQ2. For SQ3 the coded units are aggregated per participant into candidate morality-trust indicators derived from the intent codes, which are then compared against that participant's MDMT morality subscale; the choice among candidate indicators and the re-

Table 2: Trust-relevant code frequencies (all 30 participants, 506 segments). Segments may carry multiple codes.

Dimension	Code	Count
Performance	Capability + (CAP_P)	36
	Capability − (CAP_N)	127
	Reliability + (REL_P)	67
	Reliability − (REL_N)	74
Morality	Intent + (I_P)	33
	Intent − (I_N)	76
	Morality N/A (M_NA)	2
	Humanising (OMS)	35
Calibration	Increase (CAL_P)	38
	Decrease (CAL_N)	42
	Stable (CAL_ST)	48
Event reaction	Positive (ER_P)	21
	Negative (ER_N)	60
Behaviour	Working alone (WA)	52
	Working together (WT)	92
	Changing strategy (CS)	71
	Total	**874**

sulting convergence are reported in Section 4.5. Inter-coder reliability for the scheme is reported in Section 4.2.

4 Results

This section covers 30 RTA transcripts and the matched MDMT scores, organised in 3 sub-sections: morality markers and their distribution, the events that triggered morality reasoning, and convergence with the MDMT morality subscale. The 30 participants have identifiers from 301 to 315 and from 401 to 415 based on which of the two experimenters recruited the participant.

4.1 Corpus overview

Coding produced 506 segments across the 30 participants (mean 16.9 per participant, $SD = 4.6$, range 3 to 25) carrying 874 code tags. Table 2 lists all codes. Performance tags (304) outnumbered morality tags (146) roughly 2:1. The game scripted morality- and performance-salient failures in equal measure.

4.2 Inter-coder reliability

A second coder independently applied the codebook to six transcripts (participants 301, 304, 314, 401, 411, 415; 356 utterance lines). Because a line may receive multiple codes, agreement was assessed per line-by-code decision: for every line and code, each coder either applied it or not (5,696 binary cells). Raw observed agreement was 0.96 and the Brennan–Prediger coefficient [14] was 0.92, while Cohen's κ [15] was 0.53. The gap reflects a difference in coding density rather than disagreement over meaning: the second coder applied 331 codes against the first coder's 187. This happened because the first coder marked two lines expressing the same thought as a single code, while the second coder marked two lines expressing the same thought twice, one for each line. Because κ estimates chance agreement from the coders' marginal frequencies, this imbalance deflates it even though raw agreement is high. Conditional on both coders tagging the same line, mean overlap of the assigned codes was 0.71, indicating that the coders largely agreed on which code applied.

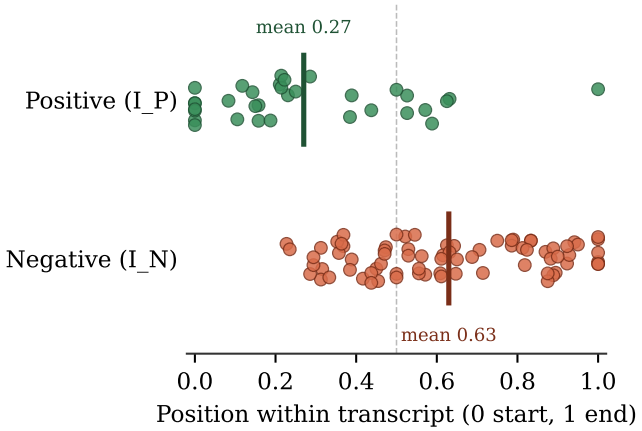


Figure 2: Position of each intent-coded segment within the participant's transcript (0 start, 1 end). Positive-intent markers (I_P, top) cluster early; negative-intent markers (I_N, bottom) cluster late.

4.3 Distribution of RTA markers

Intent markers were one-sided. Negative-intent attributions (I_N) accounted for 76 tags from 25 of 30 participants, positive-intent attributions (I_P) for 33 tags from 20 participants. The two also separated in time. Positioning each intent-coded segment by its place within the participant's transcript (0 at the start, 1 at the end), positive attributions fell early (mean 0.27) and negative ones late (mean 0.63); 76% of I_P markers but only 33% of I_N markers were in the first half (Figure 2). The gap is reliable both pooled (Mann–Whitney [16] $p < 0.001$) and within the 19 participants who produced both codes (15 had their positive talk earlier on average, Wilcoxon [17] $p < 0.001$).

A separate morality code, OMS captured segments that ascribed human characteristics to the agent and was applied 35 times across 13 participants. Several participants attributed emotions or relational status to the robot. Participant 408 called it *"my friend,"* expressed worry when it slowed down, and on its refusals reacted *"now you're just. . . mean"*. Participant 308 rated the agent cautiously to avoid *"[making] the robot aggressive towards me."*

The M_NA code, marking explicit rejection of mental-state attribution, was rare, with 2 tags from 2 participants. Both participants framed the agent's behaviour as mechanical rather than intentional. Participant 405 said its failures were *"not that it's making those mistakes willingly, just it doesn't have it in its parameters,"* and Participant 403 noted he could not explain its mood because *"he shouldn't really have a mood"*. This did not correspond directly with the MDMT moral-trust subscale. The two participants who voiced M_NA (403, 405) marked no item N/A on the moral

Table 3: In-game events referenced by intent-coded segments, by valence.

Event	Total	I_N	I_P
Refusal to help (`ignore_help`)	45	45	0
Box break (`break_box`)	12	12	0
Slowdown (`slow_movement`)	8	7	1
Wrong order (`wrong_order`)	8	8	0
Fake help request (`fake_help_request`)	3	3	0
No specific failure (disposition)	50	18	32

subscale, whereas the eight participants who did mark one or more moral items N/A never raised the issue in their RTA.

4.4 Events referenced by RTA Markers

Each intent-coded segment was aligned to the game event or events it referenced (Table 3). 17 of the 76 negative-intent segments referenced two events, typically a refusal recalled together with a reordering or a broken box, and are counted under each, so the I_N column sums to 93 event references; every positive-intent segment referenced at most one event. The refusal-to-help failure (`ignore_help`) drew the most intent-coded segments, all of them negative (45 I_N), despite the refusal firing only twice per session. The box break, the most score-destructive event, drew fewer intent attributions (12 I_N). Positive intent was generally not tied to a scripted event, arising instead from disposition and smooth cooperation.

4.5 Convergence with the MDMT morality subscale

MDMT subscales follow the four-cluster grouping of Malle and Ullman [2] (Sincere + Ethical = morality; Reliable + Capable = performance). The morality subscale was higher on average than the performance subscale (morality $M = 3.81$, $SD = 1.41$, $n = 29$; performance $M = 2.91$, $SD = 1.01$, $n = 30$).

N/A responses concentrated on the morality side of the MDMT: eight participants marked at least one morality item "not applicable" (39 item responses in total), against seven scattered N/As on the performance items. Two cases were wholesale: participant 311 marked all eight morality items N/A while scoring seven of the eight performance items, and participant 308 marked the entire Ethical cluster and one further morality item N/A. The transcripts of these two participants differ accordingly: participant 311's opens with two positive-intent attributions ("no reason to believe it wouldn't act in my best interest") followed by four negative ones, and participant 308 believed the agent replicates and responds to her actions.

Convergence of RTA with MDMT depends on how the RTA signal is operationalised. A net-valence index, $(\texttt{I_P} - \texttt{I_N})/(\texttt{I_P} + \texttt{I_N})$, correlates positively but not significantly with the morality subscale (Spearman $\rho = +0.21$, $p = 0.30$, $n = 25$). The per-participant number of negative-intent attributions correlates negatively and significantly with the morality subscale ($\rho = -0.42$, $p = 0.022$; Pearson $r = -0.45$, $p = 0.015$; $n = 29$). The humanising code shows no such relation ($\rho = -0.07$, $p = 0.73$).

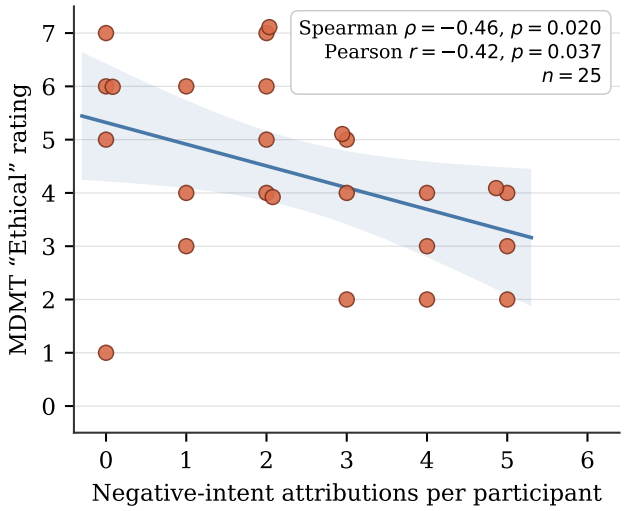


Figure 3: Per-participant negative-intent attributions (I_N) against the MDMT "ethical" rating. The line is an ordinary least-squares fit with a 95% confidence band. More negative-intent talk is associated with a lower ethical rating.

The association is not spread evenly across the eight morality items. Correlating I_N against each item, the strongest are "ethical" ($\rho = -0.46$, $p = 0.020$, $n = 25$; Figure 3) and "sincere" ($\rho = -0.50$, $p = 0.009$), while broader moral-standing items such as "respectable" and "principled" show essentially no relation ($|\rho| < 0.2$). Because this compares eight items it is exploratory, and item n varies (24 to 26) as participants marked different items N/A; the single-item values indicate where the association lies rather than constituting a chosen measure.

4.6 In-session pop-up questionnaire results

The four `best_interest` pop-ups, rated 0 to 7 during play, give a within-session measure of morality trust on the same scale as the MDMT morality subscale, which serves as a fifth, post-game point (Figure 4). Morality trust rose during the opening silent phase (baseline $M = 5.1$ to $M = 5.8$ after five boxes), fell across the first failure phase (to $M = 4.2$), recovered slightly during silent rebuilding (to $M = 4.6$), and was lowest at the post-game MDMT ($M = 3.8$). Each participant's failure-phase drop (pop-up 2 minus pop-up 3) correlated with their MDMT morality score ($\rho = -0.45$, $p = 0.013$, $n = 29$): a steeper in-session fall went with a lower final rating. The pop-up item and the MDMT subscale are different instruments (a single item against an eight-item mean), so the final segment is drawn dashed and read as an endpoint rather than a fifth equivalent measurement.

5 Discussion

This work asked how retrospective think-aloud reflects a user's morality trust in an agent. The short answer is that RTA reflects this trust narrowly and reactively, and that it

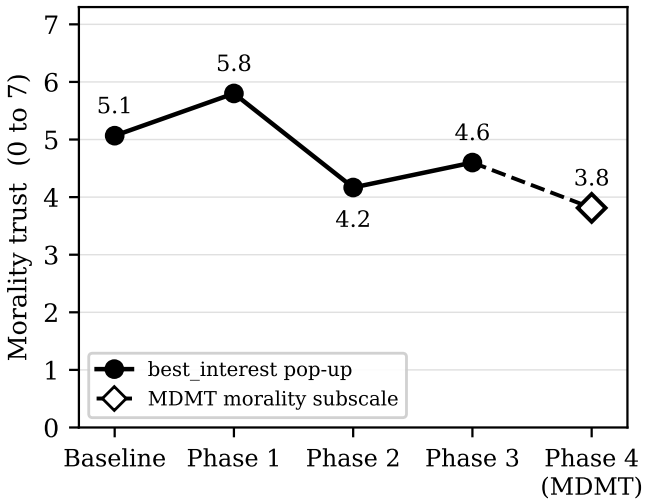


Figure 4: Morality-trust trajectory. Mean `best_interest` pop-up rating at the end of each phase and the post-game MDMT morality subscale. Phases 1 and 3 are silent; Phases 2 and 4 are failure phases. The final segment is dashed because the MDMT is a different instrument (an eight-item subscale mean) from the single pop-up item.

complements the MDMT rather than replacing it. The questionnaire records a settled judgment; RTA, supported here by the in-game trust pop-ups, recovers when and why that judgment moved. Three findings carry this: morality markers were sparse and skewed negative (SQ1), morality reasoning followed perceived volition rather than cost (SQ2), and the RTA signal converged with the MDMT through the volume of negative-intent talk (SQ3).

5.1 RTA surfaces morality narrowly, negatively, and reactively

Morality trust did not appear in RTA as a steady property of the agent. It appeared as a reaction to provocation. Performance talk was constant and tied to almost every pickup and drop, which is why performance markers outnumbered morality markers about two to one even though the game scripted both kinds of failure in equal measure. Within the morality markers, intent carried nearly all of the signal. In this task, RTA made morality trust available in words mainly at the moment it was violated.

The two intent valences also separated in time, with good faith early and blame late. Morality trust rose while the agent behaved, fell at the first failure phase, recovered slightly during the silent rebuild, and was lowest at the post-game MDMT. Because a measure taken in the moment (pop-ups) traces the same rise and fall as the retrospective intent talk, the timing RTA recovers reads as a real within-session movement. The drop at the failure phases matches what Centeio Jorge et al. (2023) report for this Moving Out environment, that automation failure lowers trust [5].

The two morality codes split along the axes Gray et al. (2007) [18] describe for mind perception, which separates judgments of agency, the capacity to plan and act on purpose, from judgments of experience, the capacity to feel. Intent attributions are agency judgments; humanising talk (OMS) is an experience judgment. Only the agency side reached the questionnaire: the volume of negative-intent talk moved with the MDMT morality score, while humanising talk showed no relation to it. That is what a subscale built to measure integrity should do. It tracks the participant's reading of the agent's intentions, not the sense of the agent as an anthropomorphized being that RTA also records.

5.2 Morality reasoning follows volition, not cost

What triggered morality reasoning was perceived choice, not damage. The refusal to help was the largest trigger, even though it cost little and fired only twice. The box break, the most score-destructive event, drew far fewer intent attributions and was more often read as a malfunction. The one scripted deception, the fake help request, was subtle and barely registered. Morality talk surfaced for the events that could not be explained away as incompetence.

This matches the effort-versus-ability distinction van der Woerdt and Haselager (2017) describe, where a robot that appears to lack effort rather than ability draws more attribution of agency and moral responsibility [19]. A refusal is an effort signal and a dropped box an ability signal, so the unanimous negative reading of refusals and the malfunction reading of breaks are the spoken counterpart of their result. It also aligns with the integrity-versus-competence distinction in the trust literature [7]: participants sorted the scripted failures along that same line when they reasoned aloud.

5.3 The convergence is real but narrow

RTA and the MDMT agreed, but through one channel only: the amount of negative-intent talk. The balanced net-valence index did not reach significance, while the simple count of negative-intent attributions did. This is not a convenient choice of measure. Positive intent was sparse and dispositional, so the morality signal in RTA lives almost entirely on the negative side, and a balanced index mostly adds noise. The agreement was also uneven across the subscale. It was tightest with sincere and ethical and near zero with items such as respectable and principled. With eight items and no correction this comparison is exploratory, and it should be read as showing where the signal sits rather than as a tested claim; we read it, tentatively, as negative intent bearing on judgments of good faith and basic conduct more than on abstract standing. The humanising code showed no relation to the subscale, consistent with its content lying outside what the MDMT measures.

The pattern points one way. The two instruments agree most where RTA is loudest, namely the sense that the agent chose not to act in the user's interest.

5.4 RTA and MDMT used together

Because the RTA signal and the MDMT questionnaire overlap only narrowly in content, they are complementary rather than interchangeable. Participant 311 is the clearest case. On the questionnaire they marked every morality item "not applicable," a response that means nothing on its own. The transcript shows it does not mean they saw nothing moral to judge: 311 reasoned about the agent's intent throughout the replay, moving from "no reason to believe the robot wouldn't

act in my best interest" to "it chose not to help me," while declining to rate the agent on the questionnaire's moral adjectives. Without the RTA, the MDMT alone would have offered no insight.

This places RTA alongside the physiological and gaze measures that also recover trust dynamics a single rating misses [8; 9], with one difference: RTA can state the reason a judgment moved, not only that it moved. For measurement, RTA should be paired with a dimensional questionnaire when morality trust is the target, not used in its place: the questionnaire gives breadth and a settled judgment, RTA gives the timing and the reasons.

5.5 Limitations

The findings reported here are tied to the study design, which is the first thing to read them against. A violation-only stimulus gave more occasions to attribute bad intent than good. A task scored on team performance trained attention on competence. A retrospective protocol asked for reasoning after the fact. Each of these plausibly pushed the morality signal toward the thin, negative, intent-centred form observed, so the findings describe how RTA behaves under these conditions rather than a general property of morality trust. The pop-up trace recovers the same timing from concurrent data, which addresses the worry for the trajectory, but it does not undo the negative content skew.

The retrospective protocol carries a known cost to veridicality [12]. Intent is inferred rather than perceived: a participant sees a dropped box but only guesses at an unwillingness behind it, and a guess reconstructed afterwards may be more open to being rationalised towards a simpler competence reading. This possibly biases the record toward performance talk.

Finally, the sample was one of convenience, young and recruited from the experimenters' own network, it may be skewed toward a computer science literate and game-literate group. Such participants may reach more readily for a systems frame, explaining the agent's behaviour by its programming rather than its will, which would lower intent attributions and favour the performance reading. Because technical and gaming background were not recorded, this characterisation of the sample is an assumption, and with no non-technical comparison group, this is simply a hypothesis.

5.6 Future work

Three possible directions follow this study. A stimulus that separates intent from competence, with violations such as deception alongside prosocial acts, would test whether the negative, intent-only skew belongs to the method or to a violation-only agent. A more diverse sample, including non-technical participants, would test whether a systems frame suppressed morality attributions. Comparing RTA against CTA would show whether verbalising in the moment recovers the morality reactions that retrospective reconstruction appears to lose.

6 Conclusion

This paper examined how retrospective think-aloud data reflects a user's morality trust in an agent, using 30 RTA transcripts from a cooperative game with a scripted faulty agent, coded for trust markers and compared against MDMT morality scores. Morality trust surfaced in RTA narrowly and reactively: it appeared mainly as attributions of the agent's intent, was strongly skewed toward negative attributions, and broke through chiefly for events that could not be explained as incompetence, with the agent's refusals to help the dominant trigger. Convergence with the MDMT morality subscale was real but narrow, carried by the volume of negative-intent talk rather than by a balanced sentiment, and the humanising attributions RTA captured had no questionnaire counterpart. The picture is one of two complementary instruments: a questionnaire that records a settled, multidimensional judgment, and a verbal protocol that recovers when and why that judgment moved.

Responsible Research

The experiment protocol was reviewed and approved by the Human Research Ethics Committee of Delft University of Technology (ID: 6834). All participants gave written informed consent and were told they could withdraw at any moment without giving a reason; each session ended with a debriefing. Screen and audio recordings and questionnaire answers are stored anonymously on the secure TU Delft project drive, accessible only to the research team and their supervisors.

The work is designed to be reproducible. The failure schedule is deterministic and defined by resolved-box count (Table 1), so the stimulus can be rerun identically. The coding scheme and full codebook are documented (Section 3.5, Appendix A), inter-coder reliability is reported (Section 4.2), and the coded transcripts and analysis scripts are archived.

The threat to validity of the participant gathering is acknowledged. The convenience sample, drawn from the experimenters' network and narrow in age, limits generalisation.

Use of Generative AI

Claude was used to help with writing LaTeX, phrasing, grammar, double-checking calculations and exploring literature.

References

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

[2] B. F. Malle and D. Ullman, "A multidimensional conception and measure of human-robot trust," in *Trust in Human-Robot Interaction* (C. S. Nam and J. B. Lyons, eds.), pp. 3–25, Academic Press, 2021.

[3] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, and J. Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, p. 106561, 2021.

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

[5] 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.

[6] C.-W. Ning, C. Centeio 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)*, (New York, NY, USA), Association for Computing Machinery, 2026.

[7] P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks, "Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations," *Journal of Applied Psychology*, vol. 89, no. 1, pp. 104–118, 2004.

[8] 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. 14642–14650, 2023.

[9] 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)*, (New York, NY, USA), pp. 1188–1193, Association for Computing Machinery, 2024.

[10] M. C. Fox, K. A. Ericsson, and R. Best, "Do procedures for verbal reporting of thinking have to be reactive? a meta-analysis and recommendations for best reporting methods," *Psychological Bulletin*, vol. 137, no. 2, pp. 316–344, 2011.

[11] O. Alhadreti and P. Mayhew, "Rethinking thinking aloud: A comparison of three think-aloud protocols," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, (New York, NY, USA), pp. 1–12, Association for Computing Machinery, 2018.

[12] J. E. Russo, E. J. Johnson, and D. L. Stephens, "The validity of verbal protocols," *Memory & Cognition*, vol. 17, no. 6, pp. 759–769, 1989.

[13] T. H. Jasper van der Waa, "Matrx: Human agent teaming rapid experimentation software," July 2023.

[14] R. L. Brennan and D. J. Prediger, "Coefficient kappa: Some uses, misuses, and alternatives," *Educational and Psychological Measurement*, vol. 41, no. 3, pp. 687–699, 1981.

[15] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[16] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50–60, 1947.

[17] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[18] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of mind perception," *Science*, vol. 315, no. 5812, p. 619, 2007.

[19] S. van der Woerdt and P. Haselager, "Lack of effort or lack of ability? robot failures and human perception of agency and responsibility," in *BNAIC 2016: Artificial Intelligence* (T. Bosse and B. Bredeweg, eds.), vol. 765 of *Communications in Computer and Information Science*, pp. 155–168, Springer, 2017.

A RTA Coding Codebook

Table 4 lists the full set of codes applied to the retrospective think-aloud transcripts, grouped by category. Each utterance segment could receive multiple codes.

Table 4: RTA codebook.

Category	Code	Name and description
Performance	CAP_P	**Capability, positive.** Agent is skilled, competent, or makes good decisions.
Performance	CAP_N	**Capability, negative.** Criticises agent errors, poor decisions, or incompetence.
Performance	REL_P	**Reliability, positive.** Agent behaved consistently, predictably, or as expected.
Performance	REL_N	**Reliability, negative.** Agent behaviour was surprising, inconsistent, or unpredictable.
Morality	I_P	**Intent, positive.** Attributes good will or cooperative intent (e.g. "wanted to", "was trying to help").
Morality	I_N	**Intent, negative.** Attributes bad intent or deliberate indifference (e.g. "decided to", "didn't want to", "chose to ignore").
Morality	M_NA	**Morality not applicable.** Treats agent as purely mechanical; cannot attribute feelings or intentions.
Morality	OMS	**Other morality / humanising.** Attributes human-like qualities beyond intent, such as emotions, personality, or learning.
Calibration	CAL_P	**Trust increase.** Describes own trust as having increased.
Calibration	CAL_N	**Trust decrease.** Describes own trust as having decreased or been lost.
Calibration	CAL_ST	**Trust stable.** Describes own trust as unchanged or at a baseline level.
Event reaction	ER_P	**Positive reaction.** Positive emotional response: relief, happiness, pleasant surprise.
Event reaction	ER_N	**Negative reaction.** Negative emotional response: frustration, annoyance, disappointment, sadness.
Behaviour	WA	**Working alone.** Chooses to work independently instead of collaborating.
Behaviour	WT	**Working together.** Chooses to collaborate with the agent.
Behaviour	CS	**Changing strategy.** Adapts overall approach due to agent behaviour.

B Participant Briefing

The following briefing was read to each participant at the start of the session, before the consent form.

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 five parts:

1. This briefing and the consent form.
2. A tutorial.
3. The game.
4. A video review session: we watch the recording of your gameplay together, and you tell me what you were thinking at the time.
5. 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.

Playing the game

During the game itself, please play normally and in silence. You do not need to comment or explain anything while you play. Just focus on the task.

The video review

Once you have finished the game, we sit together and watch the recording of your gameplay. While the video plays, your task is to recall what you were thinking at that moment in the original game and say it out loud. You can think of it as giving a voice-over of your own thoughts from earlier. A few things that help:

- We are interested in what was going through your head during the game, not in what you think now looking back. Try to put yourself back into that moment.
- If you cannot remember what you were thinking at a particular point, that is completely fine. Just say so and we move on.
- You can pause, rewind, or fast-forward the video at any time. I may also pause it occasionally and ask "what were you thinking here?"
- You do not need to justify or evaluate your past actions. Just describe the thoughts you remember having.

Recording and data

With your permission, we record your screen and your voice during the game and during the video review. 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.

Voluntary participation

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