Think-aloud data for automatic trust assessment

RTA x Morality

Calin Gheorghe
Supervisors: Myrthe Tielman, Charlotte Ning | EEMCS, Delft University of Technology

1 The problem

Trust in an AI teammate is usually measured once, after the task, with a single questionnaire.

That records a settled end state, not **when** trust moved or **why** it moved.

Trust is also multidimensional. Performance trust (is it capable?) and morality trust (does it act in my interest?) can diverge.

Retrospective Think-Aloud (RTA) asks a participant to recall what they were thinking during an interaction

Using RTA to measure morality trust may recover when and why trust changed.

2 Research question

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

SQ1. Which verbal markers can be reliably coded as morality trust?

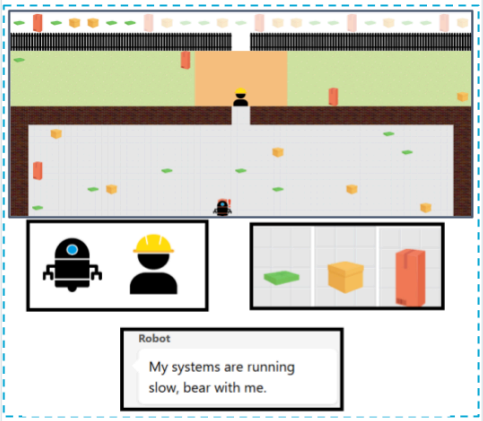
SQ2. Which in-game events trigger morality reasoning about the agent?

SQ3. Does it converge with the MDMT morality subscale?

3 The Experiment

A cooperative Moving Out game (MATRX): a participant and a scripted faulty agent move 25 boxes within a 7 minute limit.

Box weights, a safe zone, scoring, and a robot chat panel.



8 scripted failures across 4 phases (build, fail, rebuild, fail), each accompanied by a message to code the failure as a morality one, performance one, or interpretable, vague one.

30 participants: Briefing -> Tutorial -> Game -> RTA verbalisation -> MDMT questionnaire -> Debriefing

4 Measuring and coding

4 data streams are collected:

1. RTA Utterances are coded for intent: **I_P** (positive intent), **I_N** (bad intent), plus a humanising code (OMS).
2. MDMT questionnaire answers
3. Mid-game pop-up questionnaire answers
4. Game logs tracking when events happened

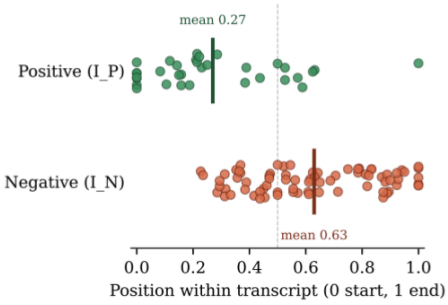
Dimension	Code
Performance	Capability + (CAP_P) Capability − (CAP_N) Reliability + (REL_P) Reliability − (REL_N)
Morality	Intent + (I_P) Intent − (I_N) Morality N/A (M_NA) Humanising (OMS)

5 Finding 1: narrow and negative

Morality surfaced narrowly, negatively, and reactively.

It appeared mainly at the moment trust was violated, not as a steady trait of the agent.

76 negative intent vs 33 positive



6 Finding 2: volition, not cost

What triggered morality talk was perceived choice, not damage.

The refusal to help drew the most blame, despite firing only twice and costing little.

Refusal: 45 negative Box break: 12

The box break, the most destructive event, drew far fewer attributions and was read as a malfunction.

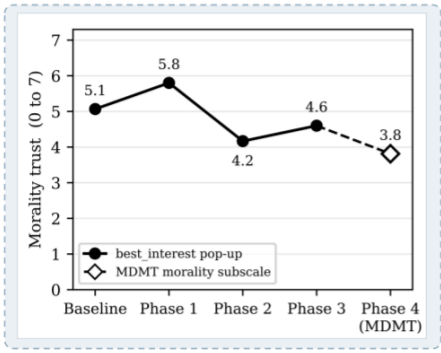
Event	Total	I_N	I_P
Refusal to help (ignore_help)	45	45	0
Box break (break_box)	12	12	0
Slowdown (slow_movement)	8	7	1
Wrong order (wrong_order)	8	8	0
Fake help request (fake_help_request)	3	3	0
No specific failure (disposition)	50	18	32

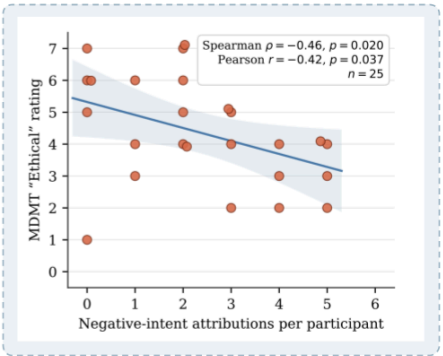
7 Finding 3: convergence with MDMT

The volume of negative intent talk tracked the MDMT morality score.

rho = -0.42 (p = 0.022)

A balanced sentiment index did not, and the humanising talk had no questionnaire counterpart. Agreement was tightest on sincere and ethical.





8 Takeaway

The questionnaire records a settled judgment. RTA recovers when and why it moved.

The two are complementary, not interchangeable. MDMT recovers a settled judgement across dimensions, RTA recovers when and why trust changed.

Example: Participant 311 marked every morality item not applicable, yet reasoned about the robot's intent all through the replay. Without RTA the questionnaire showed nothing.

9 Limitations & Future work

Violation only stimulus. Performance scored task. Retrospective protocol. Convenience sample.

Future works can try a similar experiment with:

1. a stimulus that elicits more diverse morality responses
2. a larger and more generalised participant sample
3. Concurrent think-aloud instead of retrospective