# 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**
Supervisors: M. Tielman, C. Ning
CSE3000 · TU Delft · June 2026

### WHY THIS MATTERS

AI agents make errors — users question their *intentions*, not just their skills

Trust questionnaires are post-hoc — they miss *when* and *why* trust shifts

CTA captures real-time reasoning as it happens

### RESEARCH QUESTION

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

**SQ1** What morality markers appear in CTA?

**SQ2** Which events trigger morality reasoning?

**SQ3** How do CTA markers compare with questionnaires?

### SQ1 — WHAT MARKERS APPEAR?

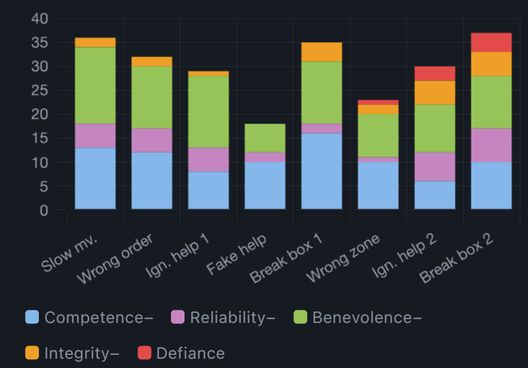
**Morality trust expressed through cooperation language, not abstract ethics**


**175** MORALITY CODES

**208** PERFORMANCE CODES

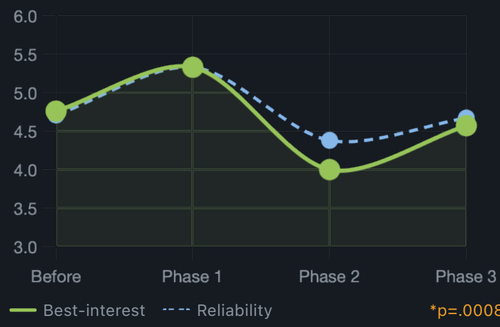
### SQ2 — WHICH EVENTS TRIGGER DISTRUST?

**Stronger distrust builds only after repeated failures**




### SQ3 — TRUST RATINGS OVER TIME

**Best-interest trust dropped significantly after Phase 2**




### SQ3 — CTA VS QUESTIONNAIRE SCORES

**CTA aligns with in-game ratings, not post-hoc MDMT**

**ρ=.60**
Morality+ CTA
↔ In-game rating
p = .004

**ns**
CTA markers
↔ MDMT morality
no correlation

CTA & pop-ups share temporal frame — both during interaction.
MDMT asks for abstract recall after the fact.

### KEY INSIGHT — TWO LAYERS OF MORALITY TRUST

**Trust operates at different abstraction levels depending on how and when you measure it**

**DURING INTERACTION**
"He ignored me"
"Great team"
"Does not care"
Cooperative language

⟺

**AFTER INTERACTION (MDMT)**
"Ethical"
"Principled"
"Sincere"
Abstract moral vocabulary

**1** **Morality trust = cooperative language.** Users say "he ignored me," not "he's unethical" — task-level framing, not abstract ethics.

**2** **Distrust escalates with repetition.** Early failures → competence doubt. Accumulation → suspicion of intent + defiance.

**3** **CTA complements questionnaires.** Real-time CTA (ρ=.60 with in-game ratings) captures what post-hoc MDMT cannot.