HAI 2023
4 - 7 December • Gothenburg • Sweden
11th INTERNATIONAL CONFERENCE
on HUMAN-AGENT INTERACTION

TELECOM Paris

IP PARIS

# Comparing a mentalist and an interactionist approach for trust analysis in Human-Robot Interaction

Marc Hulcelle - Télécom Paris
Giovanna Varni - University of Trento
Nicolas Rollet - Télécom Paris
Chloé Clavel - Télécom Paris

# Introduction

2
main
perspectives

- Trust was mostly studied in Human-Robot Interaction (HRI) from a Psychological *mentalist* perspective
  - Mostly focused on robot-related factors that impact user's trust towards the robot

- Some adopt an *interactionist* perspective as described by Interactionist Sociology theories
  - Entirely focused on the user's display of trust

# Research Question

Can we identify criteria to differentiate both approaches based on their theoretical framework and trust assessment tools' methodologies ?

# Mentalist approach

Trust is defined as a user's mental-state. Definition from Rousseau [1] :

"**psychological state** comprising the intention to **accept vulnerability** based upon **positive expectations** of the intentions or behavior of another"

Trust is therefore a combination of :

- User's mental projection of the robot's capabilities (Cognitive trust)
- An affective response to these (Affective trust) [2]

[1] D. M. Rousseau, S. B. Sitkin, and C. Camerer. Not So Different After All : a Cross-Discipline View of Trust. Academy of Management Review 23, 3 (1998), 393–4
[2] Kristin E. Schaefer. The Perception and Measurement Of Human-robot Trust. Doctoral Dissertation. University of Central Florida, Orlando (2013)

# Mentalist approach

Trust is mostly measured through questionnaires filled by *users themselves* at the *beginning* and *end* of an interaction :

- Interpersonal Trust Scale [3]
- Negative Attitude towards the Robot Scale [4]
- Robot Trust Scale [5]
- Godspeed questionnaire [6]

Trust can also be measured through physiological sensors (e.g. EEG)

[3] J.B. Rotter. A new scale for the measurement of interpersonal trust. Journal of personality (1967)
[4] D.S. Syrdal, K. Dautenhahn, K. Koay, and M. Walters. The Negative Attitudes towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study. Proceedings of AISB09 (2009)
[5] K.E. Schaefer. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". Robust Intelligence and Trust in Autonomous Systems (2016)
[6] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics (2009)

# Interactionist approach

Trust is a result of the **state of the interaction**, and is oriented towards both the **content** and the **format** of the interaction, defined as [7] :

"form of affiliation and credit characterized by **a set of behaviors** that are intentional or not, expressive or propositional"

Observable at different bases :
- the robot's capacity to maintain a fluid and progressive interaction
- its skill in accomplishing a specific action at a given moment
- its knowledge

[7] M. Hulcelle, G. Varni, N. Rollet, and C. Clavel. TURIN: A coding system for Trust in hUman Robot INteraction. In 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)

# Experimental Study

We conducted a comparative annotation study of two measurement tools

- Dataset: First two phases of all 10 interactions of the Vernissage dataset [8]
- Tools : Robot Trust Scale (Mentalist) / TURIN (Interactionist)
- Procedure : Annotate 10 seconds long segments TOTAL : 180 segments
  5 experts A, B, C, D, and E annotated using one tool

[8] D.B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J.M. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, and D. Gatica-Perez The vernissage corpus: A conversational Human-Robot-Interaction dataset. ACM/IEEE International Conference on Human-Robot Interaction (2013

# Experimental Study

| | Interactions 1→ 5 | Interactions 6 → 10 |
|---|---|---|
| Expert A | RTS / TURIN | RTS / TURIN |
| Expert B | TURIN | |
| Expert C | | TURIN |
| Expert D | RTS | |
| Expert E | | RTS |

# Experimental Study

We used the Robot Trust Scale reduced to its 14 items.

"What % of the time will the robot..." (11 Likert scale)

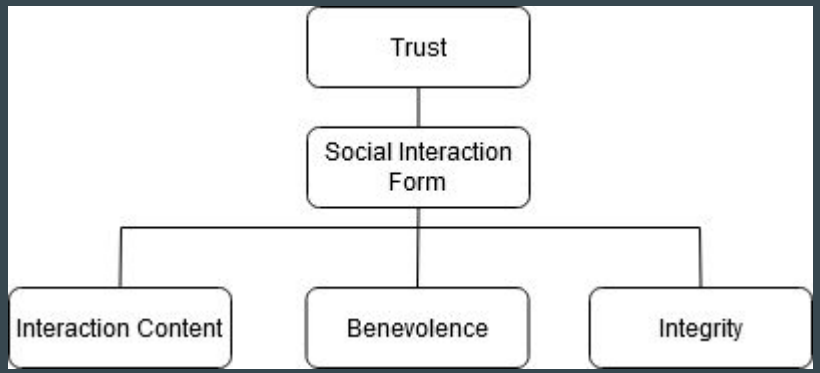
| | | |
|---|---|---|
| Have errors | Provide appropriate information | Be unresponsive |
| Malfunction | Communicate with people | Provide feedback |
| Function successfully | Act consistently | Be reliable |
| Be predictable | Be dependable | Meet the needs of the mission |

We dropped "perform exactly as instructed" and "follow directions" items

# Experimental Study

Trust in hUman Robot INteraction (TURIN) coding scheme :

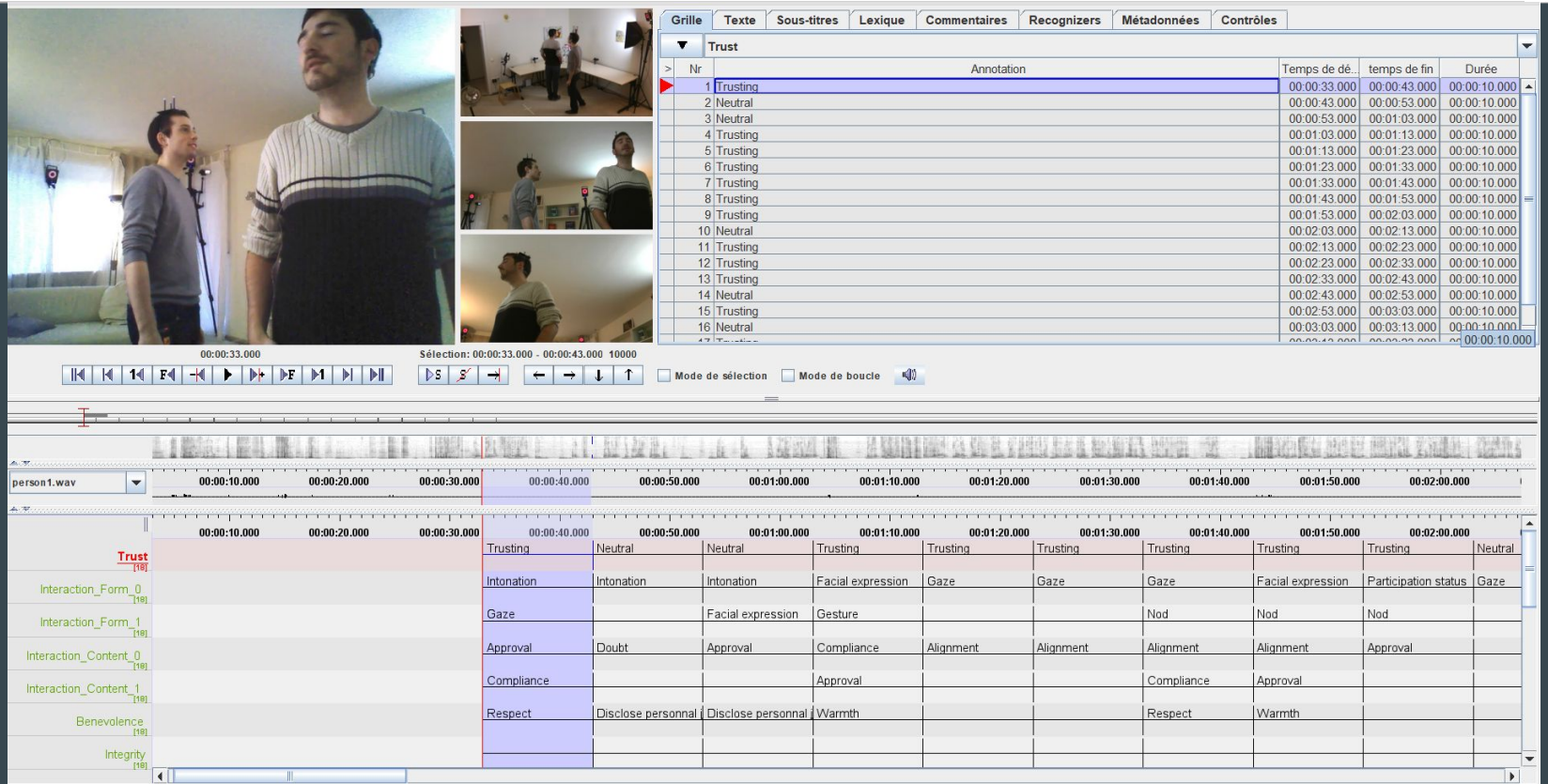
- At most 2 items from “Social Interaction Form”
- At most 2 items from “Interaction Content”
- At most 1 for “Benevolence”
- At most 1 for “Integrity”




| Social Interaction Form | Interaction Content |
|---|---|
| Gaze | Compliance |
| Facial expression | Cooperation |
| Nod | Alignment |
| Gesture | Approval |
| Phrasing | Out-of-context comment |
| Intonation | **Trusting only** |
| F-formation* | Joke |
| Speaking turn | **Mistrusting only** |
| Repetition | Doubt |
| Participation status* | |
| **Benevolence** | **Integrity** |
| Respect | Honesty |
| Personal info disclosure | Responsibility |
| Warmth | Promise |
| | **Mistrusting only** |
| | Manipulation |

*Group-specific item

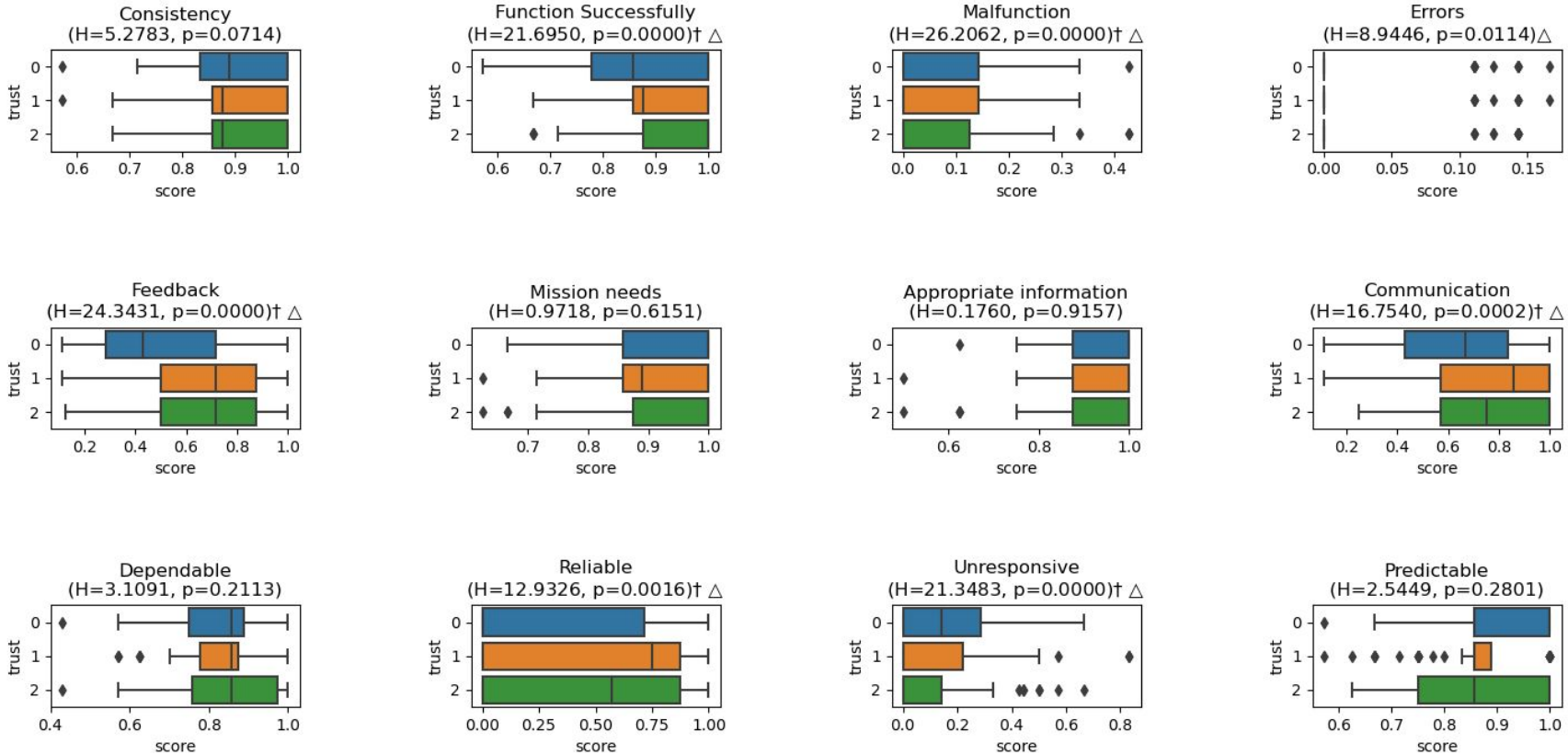
# Experimental Study



# Results

- We compute the global trust score for each segment by averaging all items' score
- For each annotator, we rescale the global trust score through a min-max scaling
- We aggregate segments according to their assigned TURIN label
- We compare the RTS global score distribution between different TURIN labels

# Results


Consistency
(H=5.2783, p=0.0714)
Function Successfully
(H=21.6950, p=0.0000)† △
Malfunction
(H=26.2062, p=0.0000)† △
Errors
(H=8.9446, p=0.0114)△
Feedback
(H=24.3431, p=0.0000)† △
Mission needs
(H=0.9718, p=0.6151)
Appropriate information
(H=0.1760, p=0.9157)
Communication
(H=16.7540, p=0.0002)† △
Dependable
(H=3.1091, p=0.2113)
Reliable
(H=12.9326, p=0.0016)† △
Unresponsive
(H=21.3483, p=0.0000)† △
Predictable
(H=2.5449, p=0.2801)
trust
score

# Results

We identified 4 criteria that differentiates both approaches to analyze trust:

- Time-framing
- Orientation
- Generalization
- Scalability

| | Time-framing | | | Orientation | | Generalization | | Scalability | |
|---|---|---|---|---|---|---|---|---|---|
| | BU | ST | EI | Data-driven | Theoretical-framework-driven | Specific | Generic | Individual | Group |
| **Mentalist** | | | X | | X | | X | X | X |
| **Interactionist** | X | X | | X | | X | | X | X |

**Table 1: Summary of the comparison of the mentalist and interacionist approach based on 4 criteria.**
**BU: Behavioral Unity. ST: Speaking Turns. EI: Entire interaction.**

QUESTIONS ?