THE INFLUENCE OF ARTIFICIAL TRUST COMMUNICATION IN HUMAN-AGENT TEAMS

Author: Răzvan Loghin <r.loghin@student.tudelft.nl>
Supervisor: Carolina Ferreira Gomes Centeio Jorge
Responsible Professor: Myrthe Tielman

Introduction

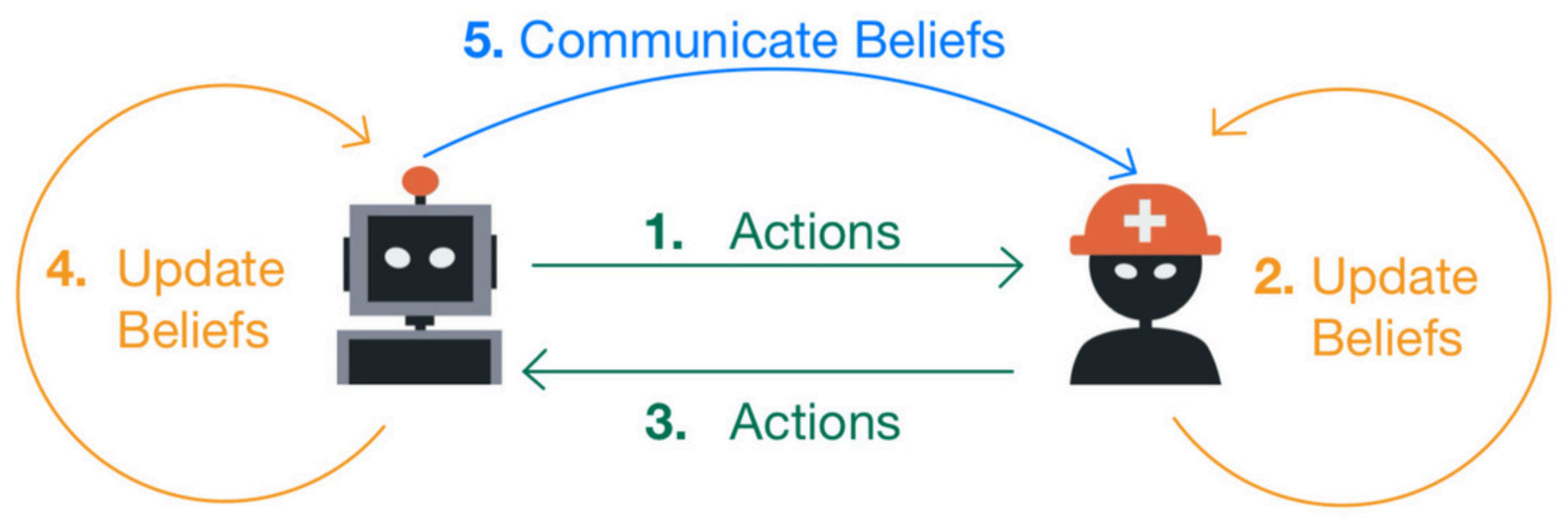


Figure 1: Influence of communicating trust-based beleifs in human-AI team dynamic

- **Collaborative AI:** Integration of AI systems alongside humans, leveraging both strengths and limitations **[1]**.

Background

- **Human-AI Teams** [2][3]:
  - Based on mutual trust and transparency
  - Superior performance compared to human-only or AI-only teams
- Mutual trust: A composite of **natural** and **artificial trust [4][5].**
- **Lack of representation** of **artificial trust** in current literature, with most models focusing only on natural trust **[5]**.

Motivation

- Effective communication of AI's reasoning and decisions is essential for building trust in AI systems [2] [6].
- **Research Gap:** The need for understanding how different communication types affect trust dynamics and human-AI team performance.
- **Trust** and **satisfaction** correlated with team effectiveness, positive team dynamics, and outcomes **[2][7]**.

Research Question

How does a **textual summary of changes (justification)** of the **mental model of the agent's trust** in the human teammate affect the **human teammate's trust** in the agent and **overall satisfaction**?

Why textual summary of changes?

- Excessive communication can distract and overwhelm humans **[8]**.
- Textual representation prevents misunderstanding of information **[9]**.

Sub-Questions

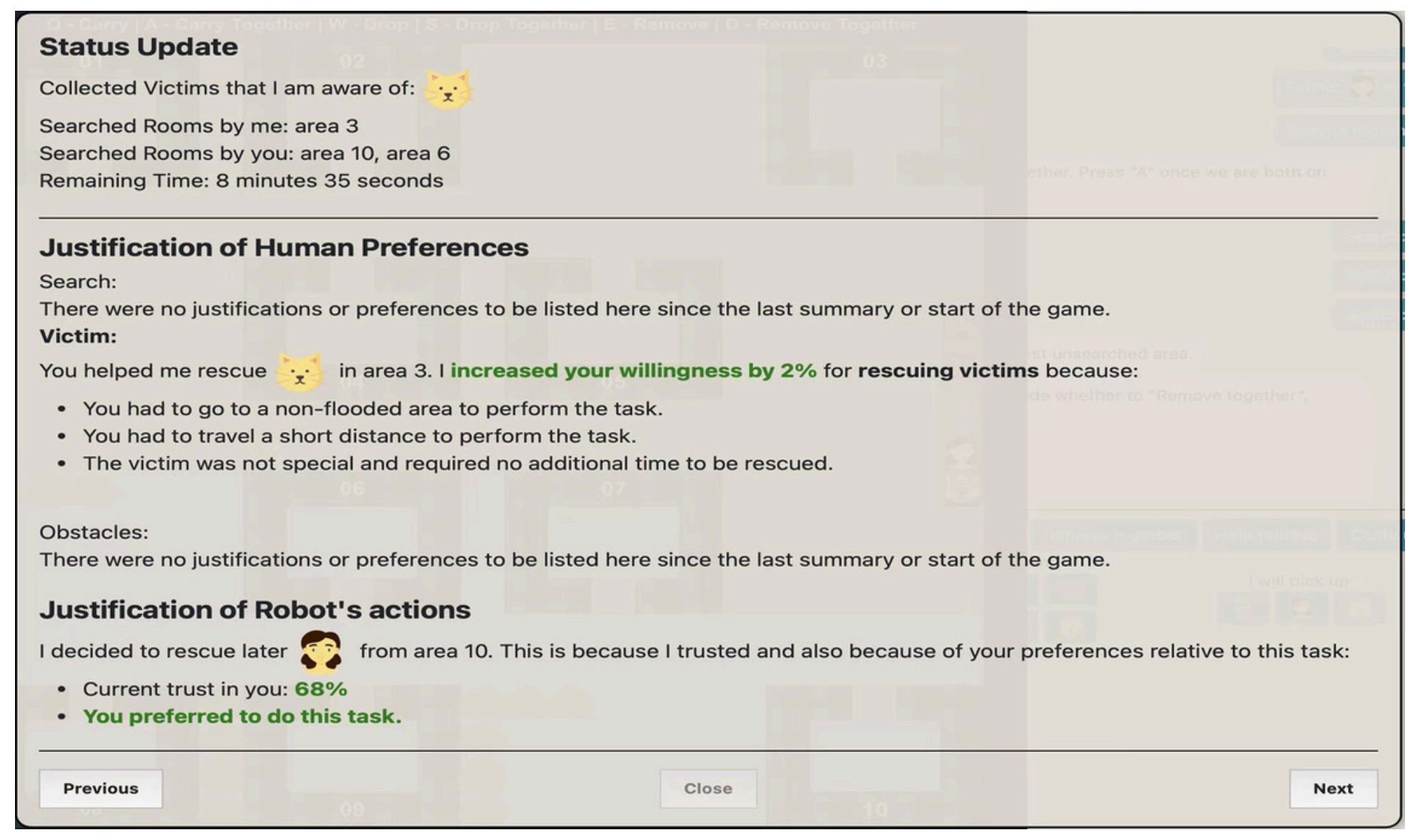
- **SQ1:** How can a textual summary of changes of the mental model of an AI agent be developed to effectively transmit artificial trust to human teammates?
- **SQ2:** What is the impact of the developed communication method on natural trust and overall satisfaction? (Figure 1)

Communication strategy

- **Type:** textual summary of changes
- Appears centrally on the screen and **pauses the game** until is closed.
- **3 summaries** generated at predetermined intervals based on the game's time progress and the number of victims rescued (Figure 6).

Information: displayed across 3 screens (example in Figure 2)

- **Status Update:** overview of current game state (victims, time, searched rooms).
- **Actions Impact on Trust:** human actions and their generated trust updates.
- **Justification of Human Preferences:** human actions related to preference factors.
- **Justification of Robot's Actions:** AI decisions influenced by behavior adaptation.
- **Trust and Confidence Levels:** with changes since the last summary.



Status Update

Collected Victims that I am aware of:

Searched Rooms by me: area 3
Searched Rooms by you: area 10, area 6
Remaining Time: 8 minutes 35 seconds

Justification of Human Preferences

Search:
There were no justifications or preferences to be listed here since the last summary or start of the game.
Victim:
You helped me rescue in area 3. I **increased your willingness by 2%** for **rescuing victims** because:

- You had to go to a non-flooded area to perform the task.
- You had to travel a short distance to perform the task.
- The victim was not special and required no additional time to be rescued.

Obstacles:
There were no justifications or preferences to be listed here since the last summary or start of the game.

Justification of Robot's actions

I decided to rescue later from area 10. This is because I trusted and also because of your preferences relative to this task:

- Current trust in you: **68%**
- **You preferred to do this task.**

Figure 2: Communication screen containing **Status Update, Justification of Human Preference,** and **Justification of Robot's Actions**

Method

User Study: in between, controlled experiment

- **Baseline group** (no communication): n = 28
- With **Communication group**: n = 28

Task: Urban Search and Rescue **(USAR)** mission (Figure 3)

Variables:

- **Independent Variable:** The presence or absence of the textual summary communication method.
- **Dependent Variables:** Trust and overall satisfaction of the human teammates.

Measurements:

- **Subjective:**
  - Trust and satisfaction were measured using established scales measured with Likert scales.
  - Optional open-ended questions for qualitative data.
- **Objective:** Compliance, Communication rate, Task success rate, Interaction frequency, and Task completion time.

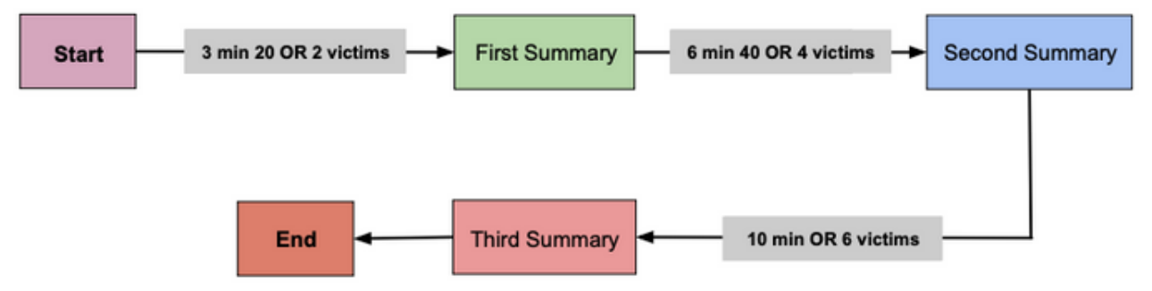


Figure 6: Generation of summaries

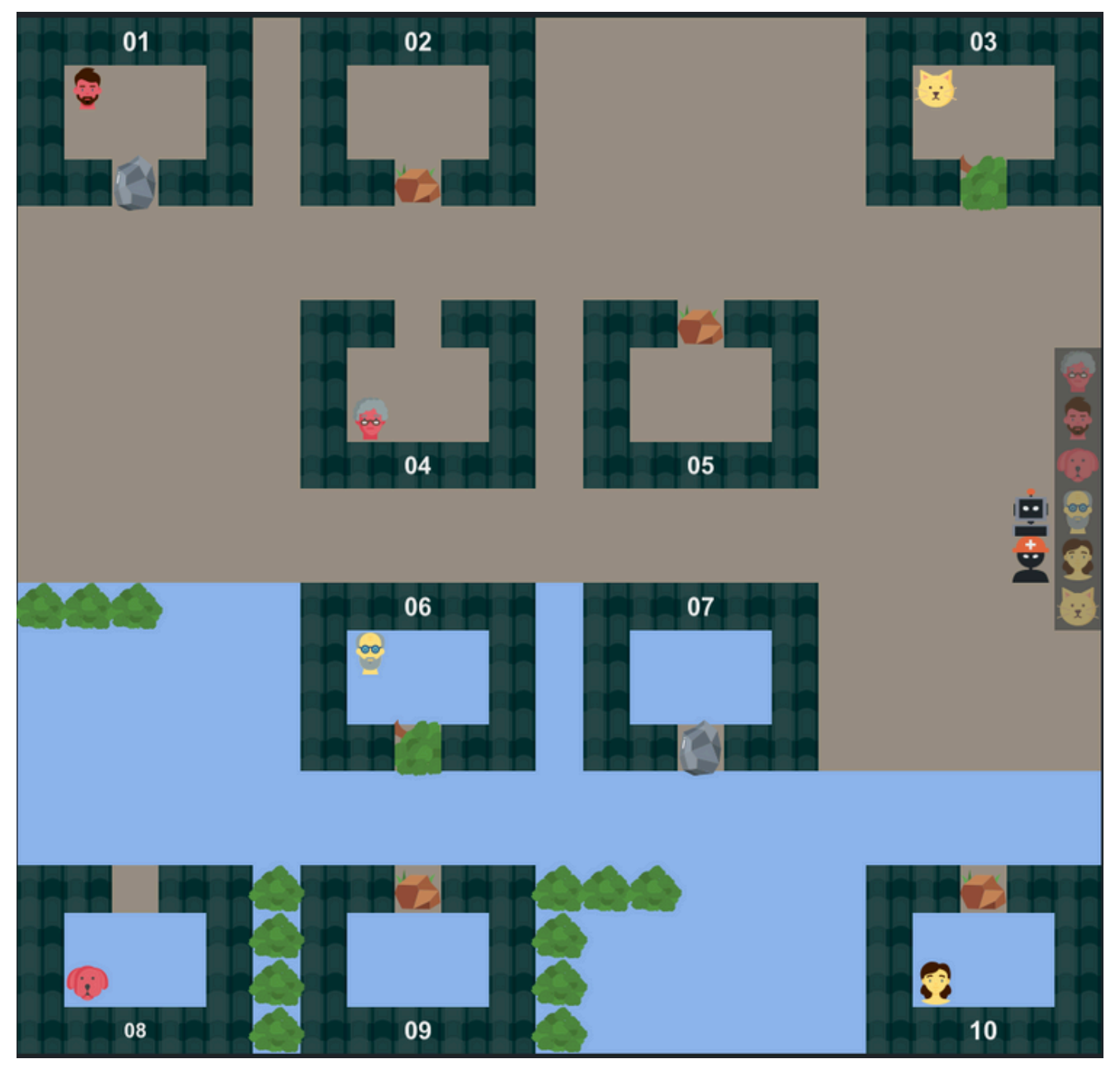


Figure 3: MATRX USAR Environment

Results

Data Processing:

- Translate Likert scale to numerical scale 1 to 5.
- Reliability consistency using **Cronbach's alpha.**
- Assessing normality of each measurement using **Shapiro-Wilk tests.**

Objective Results: only Task success rate has significant difference between the two groups (Figure 5)

Measurement	Test	p-value
Compliance	Mann-Whitney U	0.097
Ratio of joint actions	t-test	0.213
No. of human messages	Mann-Whitney U	0.347
Task success rate*	Mann-Whitney U	0.0027
Total task time (ticks)	t-test	0.053

Figure 5: Comparison of Baseline and Communication groups across objective measurements.

Subjective Results: Mann-Whitney U tests for both trust and satisfaction

- Communication group had **significantly higher trust levels** than the baseline group U = 185.0, p = 0.0012, *supporting H1*.
- Communication group had **significantly higher satisfaction** than the baseline group U = 177.0, p = 0.0007, *supporting H2*.

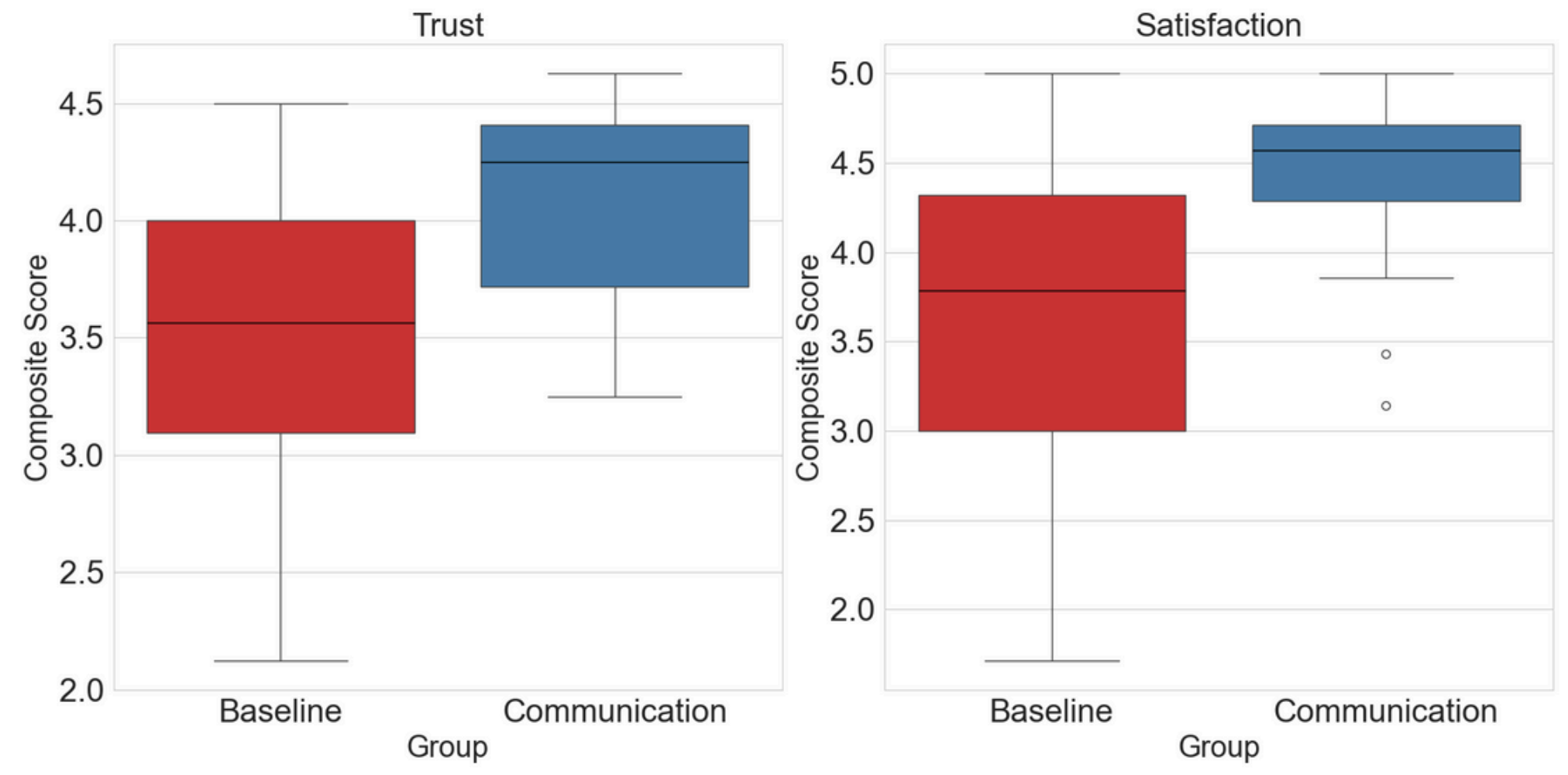


Figure 4: Box Plots comparing composite scores for trust and satisfacrion

Artificial Trust Mechanism

Mental Model

- Trust is influenced by the nature of the task and is **context-dependent [10].**
- Actions divided into 3 categories: **Search**, **Remove**, and **Rescue.**
- **Competence** and **Willingness** per task category (Formula 1)[11].

$$\begin{aligned}\mathbf{T} &= \{T_{\text{Search}}, T_{\text{Remove}}, T_{\text{Rescue}}\} \\ &= \{(C_{\text{Search}}, W_{\text{Search}}), (C_{\text{Remove}}, W_{\text{Remove}}), \\ &\quad (C_{\text{Rescue}}, W_{\text{Rescue}})\}\end{aligned}$$

Formula 1: Trust representation

Preference Integration: chosen heuristics (Figure 3)

- **Flooded areas (f):** blue areas, slow down the human agent.
- **Special victims (v):** additional rescue time.
- **Distance (s):** human agents favor nearer tasks.

Preference factor (p): adjusts the willingness based on task preference:

$$\mathbf{p(a)} = \frac{w_f \cdot \mathbf{f(a)} + w_d \cdot \mathbf{d(a)} + w_v \cdot \mathbf{v(a)}}{w_f + w_d + w_v}$$

Trust Update

- Trust values dynamically update after each human action.
- Updates depend on task workload and criticality.
- 3 update thresholds {±0.1, ±0.2, ±0.4}.
- **Formula:** $T_c(\text{new}) = (C_c(\text{old}) + \Delta C,\ W_c(\text{old}) + \Delta W + p_U(a))$

Behaviour Adaptation

- **Confidence:** AI's certainty in human trustworthiness
- AI **decides probabilistically** to trust human actions, considering willingness, competence and confidence in human

Discussion & Conclusion

Trust and Satisfaction:

- The textual summary increased significantly the natural trust and satisfaction levels **(SQ2)**.
- **Transparency and Explanations:** help calibrate human trust and enhance understanding of AI decisions **[2]**.
- **Qualitative Feedback:** Participants in the communication group appreciated the regular updates of the artificial trust levels.
  - Participant 11 Communication Group: *"It also makes me want to perform better, when seeing I am not trustworthy enough."*

Performance Metrics:

- **Task success rate** was higher in the Communication Group, other metrics showed no significant difference.
- Performance differences could be **influenced by operating systems** used by baseline participants.
- Performance logs could be influenced more by **individual user performance**.

Limitations:

- **Disparity in the background of participants** (Computer Science affiliation, experience with MATRX Software) could influence results
- **Sample homogeneity** and **number of participants**
- **Low Cronbach's alpha** in the Communication Group's trust data indicates poor internal consistency

Future Work:

- **Diversify** the participants' pool by conducting more experiments.
- Exploration of possible correlation of **confounding variables** (gaming experience, operating system used, familiarity with MATRIX Software) with reported trust and satisfaction levels.
- **Choose Trust and Satisfaction scales** that guarantee better internal consistency.
- **Compare** the textual summary method with other communication methods.

Conclusion:

- Human-centered textual summary communication builds human trust and satisfaction **(SQ2)**.
- Cornerstone for **future research** and **design of Explainable AI**.

References

[1] - Carolina Centeio Jorge, Myrthe L Tielman, and Catholijn M Jonker. "Artificial trust as a tool in human-AI teams". In: 2022 17th ACM/IEEE Interna-
tional Conference on Human-Robot Interaction (HRI).IEEE. 2022, pp. 1155-1157.
[2] - Joseph B Lyons. "Being transparent about transparency: A model for human-robot interaction". In:2013 AAAI Spring Symposium Series. 2013.
[3] - Carolina Centeio Jorge, Catholijn M Jonker, and Myrthe L Tielman. "How should an AI trust its human teammates? Exploring possible cues of artificial
trust". In: ACM Transactions on Interactive Intelligent Systems 14.1 (2024), pp. 1-26.
[4] - C Centeio Jorge et al. "Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams". In: 22nd International Trust Workshop 2021. 2021.
[5] - Hebert Azevedo-Sa et al. "A unified bi-directional model for natural and artificial trust in human-robot collaboration". In: IEEE robotics and automation letters 6.3 (2021), pp. 5913-5920.
[6] - Anthony R Selkowitz et al. "Displaying information to support transparency for autonomous platforms".In: Advances in Human Factors in Robots and Unmanned Systems: Proceedings of the AHFE 2016 International Conference on Human Factors in Robots and Unmanned Systems, July 27-31, 2016, Walt Disney World®, Florida, USA. Springer. 2016, pp. 161-173.
[7] - Sami Abuhaimed and Sandip Sen. "Human Satisfaction in Ad Hoc Human-Agent Teams". In: International Conference on Human-Computer Interaction. Springer. 2023, pp. 207-219.
[8] - Marin Le Guillou, Laurent Pr´evot, and Bruno Berberian. "Trusting artificial agents: Communication trumps performance". In: AAMAS 2023. 2023.
[9] - Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. "Visual, textual or hybrid: the effect of user expertise on different explanations". In: 26th international conference on intelligent user interfaces. 2021, pp. 109-119.
[10] - Carolina Centeio Jorge et al. "Appropriate context-dependent artificial trust in human-machine teamwork". In: Putting AI in the Critical Loop. Elsevier, 2024, pp. 41-60.
[11] - Carolina Centeio Jorge, Myrthe L Tielman, and Catholijn M Jonker. "Assessing artificial trust in human-agent teams: a conceptual model". In: Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents. 2022, pp. 1-3.