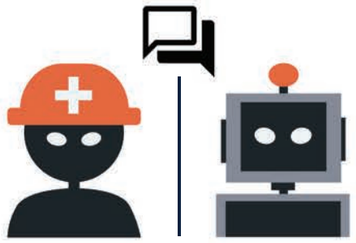
Tailoring Agent Explanations According to Human Performance on Human-AI Teamwork

1 Human-Agent Teaming

Interdependent Task Completion through Teamwork & Communication



Teleoperated Human Agent | Autonomous AI Agent

2 Urban Search-and-Rescue **Collaborative** Task Scenario

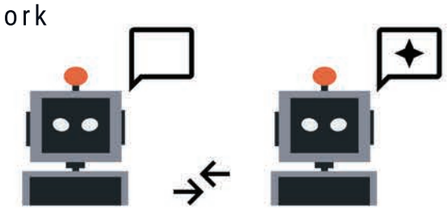
- Individual performance leads to an **unsuccessful** mission
- Critically injured victims must be rescued **collaboratively**
- Obstacles can (and sometimes must) be removed faster **together**
- **Efficient communication** is key to save all victims on time

Figure 1: MATRX Urban Search-and-Rescue Environment [matrx-software.com]

3

''people assign human-like traits to artificial agents, people will expect explanations using the **same conceptual framework** used to explain human behaviour'' [1, p. 2].

Human communication is a **complicated framework depending on many factors** including context, people involved, emotions etc, The main goal of **eXplainable AI** is to make AI systems **more understandable** to humans, making them **personalized and user-aware** [2, 3].

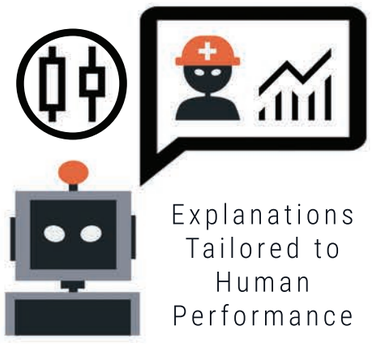


Basic Explanations of the AI Agent | Tailored Explanations of the AI Agent

4

Human Performance Metrics:
- Communication Performance
- Collaborative Performance
- Overall Score Performance

Tailoring Explanations:
- Changing the length and level of detail
- Adding motivational remarks
- Adding predictive outcome
- Adding reminders and tips



Explanations Tailored to Human Performance

Human Performance is a **Spectrum**

Worst Performance ⟷ **Best Performance**

Research Question and Methodology

How can an agent model and use **human performance** to tailor explanations?

creation of an internal model
↓
adaptation of explanations
↓
comparison with baseline AI
↓
evaluation of collaboration

5 Methodology

- 2-player Urban Search-and-Rescue Game
- 30 participants, where half of the participants teamed up with the Baseline Agent and the rest teamed up with the Tailored Agent

Baseline Agent → For every decision, providing suggestions with **random** length and detailed explanations

Providing suggestions with explanations **based on performance**, with additional tips ← **Tailored Agent**

- One tutorial and one full game is played, **objective game metrics** are logged after each game
- After the experiment, **subjective metrics** including, workload, trust, explanation satisfaction and collaboration fluency are measured

6

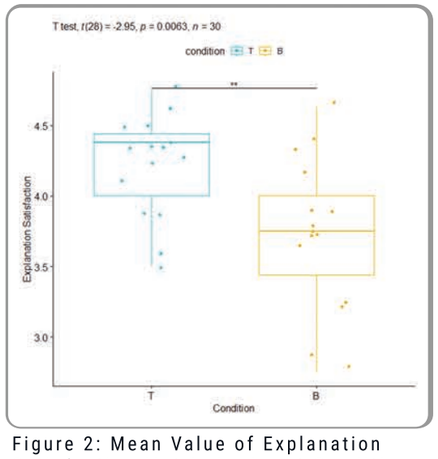


Figure 2: Mean Value of Explanation Satisfaction Metric by Agent Condition

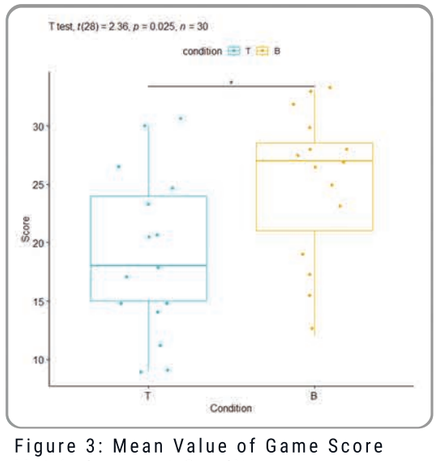


Figure 3: Mean Value of Game Score Metric by Agent Condition

7 Conclusion

- **No significant difference** was found between trust, workload and most of the performance metrics.

- Performance-tailored explanations were **more appreciated** by the user.

- Collaboration fluency **'improvement'** showed significant differences in favour of the tailored agent.

- Explanations were **satisfactory and appreciated,** however, due to the time pressure, they created a separate **challenge**.

- Long and frequent messages created **information overload** and **decreased** the overall **collaborative performance**.

References

[1] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artificial Intelligence, vol. 267, pp. 1–38, 2019.
[2] D. Gunning and D. Aha, "DARPA's Explainable Artificial Intelligence (XAI) Program," AI Magazine, vol. 40, no. 2, pp. 44–58, 6 2019.
[3] S. T. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable Agents and Robots: Results from a Systematic Literature Review," in 18th International Conference on Autonomous Agents and Multiagent Systems, 2019.

Can Parlar (c.parlar@student.tudelft.nl)
Responsible Professor: Myrthe Tielman | Supervisor: Ruben Verhagen

CSE3000 Research Project
Bachelor's Thesis | 16/05/2022

TU Delft Delft University of Technology