Tailoring User-Aware Agent Explanations to Properly Align Human Trust

Marin Vogel
m.vogel-3@student.tudelft.nl
Supervisor: R.S. Verhagen
Responsible Professor: M. Tielman
TU Delft

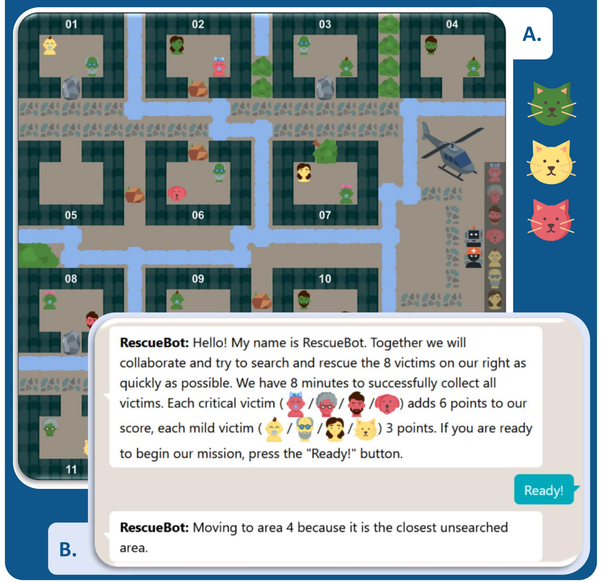
1. Theoretical Background

- Human agent teaming (HAT)
- Aligning human trust [1]
- Explainable Artificial Intelligence (XAI)
- User-Aware tailoring [2]

2. Research Question

How can an agent tailor its explanations to align human trust properly?

3. The Task



4. Suggestions

- The participant remains in charge
- Agent will only suggest what to do
- Suggestions are paired with different explanation types

5. Agent Design

Baseline Agent — *Capable of solving the task through collaboration*

Modelling Human Trust — *Suggestions Followed: trust ↑*
Suggestions Ignored: trust ↓

Tailoring Explanations — *↑ Trust: fewer explanations*
↓ Trust: more explanations

*I suggest to **continue** searching instead of removing the rock: 8/9 rescuers would decide the same, because the distance between us is large. If we had found more than 1 critical victim, I would have suggested to remove rock. Select your decision using the buttons "Remove" or "Continue".*

Low trust explanation

High trust explanation

*I suggest to **continue** searching instead of removing the rock. Select your decision using the buttons "Remove" or "Continue".*

6. Evaluation Metrics

Subjective

Collaboration Fluency [3]
Explanation satisfaction [5]
Workload [4]
Trust [5]

Objective

Completeness
Suggestions ignored
Agent moves
Score
Trust

7. Results

Correlation between subjective and objective trust — *Correlation: 0,194* *P-value: 0,3421*

C. Correlation Tests

variables	subjective correlation	subjective p-value	objective correlation	objective p-value
Collaboration fluency	0,58	0,00	0,01	0,98
Explanation satisfaction	0,70	0,00	0,01	0,98
Subjective workload	0,09	0,66	0,39	0,52
Suggestions ignored	-0,09	0,67	-0,88	0,00
Score	0,23	0,27	-0,09	0,66
Completeness	0,24	0,24	-0,08	0,69

D. T-Tests

variables	mean base	mean trust	mean difference	p-value	method
Explanation satisfaction	3,74	3,96	0,22	0,12	Wilcoxon
Completeness	0.70	0,63	-0,08	0,21	Wilcoxon
Score	25,00	22,36	-2,64	0,34	Student
Subjective trust	3,46	3,71	0,25	0,34	Welch
Collaboration fluency	4,94	5,11	0,17	0,55	Student
Subjective workload	48,95	46,21	-2,73	0.66	Student
Agent moves	290,47	278,55	-11,92	0,68	Student
Objective trust	0,40	0,42	0,01	0,87	Student
Suggestions ignored	0,29	0,29	0,00	0,97	Student

8. Discussion

- Results indicate no statistically significant difference between baseline and trust agent
- Therefore, the hypothesis is **rejected**
- Cause of the rejection does not lie in the metrics or number of experiments
- There are two possible solutions to this question

Flawed Assumption

- The method of tailoring explanations is built on a flawed assumption
- It performs as good as the baseline agent and therefore rejects the hypothesis
- Such an error in the agents' design causes the hypothesis to have failed
- Further research into tailoring explanations to human trust would be required.

Information Overload

- Feedback received indicated an information overload which caused the participants to skip the essential tailored explanations.
- This data was gathered incorrectly but would cause the hypothesis to be inconclusive
- Further research into the usage of suggestions in the current MATRX setup would be required.

References

[1] M. Johnson and A. Vera, "No ai is an island: The case for teaming intelligence," AI Magazine, vol. 40, pp. 16–28, 3 2019.
[2] S. T. Anjomshoae, A. Najjar, D. Calvaresi, and K. Fr¨amling, "Explainable agents and robots: Results from a systematic literature review," 2019.
[3] G. Hoffman, "Evaluating fluency in human-robot collaboration, " IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, vol. 49, 2019.
[4] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): results of empirical and theoretical research," Advances in Psychology, vol. 52, pp. 139–183, 1 1988.
[5] R. Hoffman, S. Mueller, G. Klein, and J. Litman, "Measuring trust in the xai context," Michigan Tech Publications, vol. PsyArXiv Preprints, 11 2021.