Tailoring User-Aware Agent Explanations to Properly Align Human Trust

Marin Vogel
Supervisor: R.S. Verhagen
Responsible Professor: M.L. Tielman
EEMCS, Delft University of Technology, The Netherlands
22-6-2022

A Dissertation Submitted to EEMCS faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

Abstract

Aligning human trust to correspond with an agent's trustworthiness is an essential collaborative element within Human-Agent Teaming (HAT). Misalignment of trust could cause sub-optimal usage of the agent. Trust can be influenced by providing explanations which clarify the agent's actions. However, research often approaches explanations statically, making them not adjustable to real-time situations. In this research, we study the effectiveness of an agent capable of modelling human trust and tailoring explanations to influence it. We achieve this by modifying an existing HAT environment and setting up an experiment comparing a trust and baseline agent. Modelling human trust is calculated through the number of suggestions ignored. When the model estimates low trust, more explanation types are used during communication. Higher trust uses fewer explanation types in order to save time. However, the results indicate no difference between the baseline and trust agent rejecting the hypothesis. A potential cause for the rejection can be found in either a flaw in the agents' design or information overload.

1 Introduction

The research field of human-agent teaming (HAT) concerns artificial intelligence (AI) agents solving tasks through collaborating with humans [1, 2, 3]. One such task could be a search-and-rescue operation where an agent assists a doctor after a natural disaster. The doctor is still an expert in medical aid, but the agent's searching capabilities contribute to their collective welfare. As AI team members get increasingly introduced into new and complex environments, the necessity to develop their social intelligence also increases [4, 5].

The agent's social intelligence impacts mutual trust most. A human and agent will try to properly establish trust based on the others' trustworthiness [6]. Misalignment of the two factors causes sub-optimal usage of the agent; the consequences can vary from lowered performance to critical failures [7, 4, 8]. Continuing with the search-and-rescue example, the doctors' trust in the agent will quickly deteriorate if its promised searching capabilities underperform. This instance presents the misalignment problem; the doctors' trust was misaligned with the agents' actual trustworthiness.

Appropriately aligning trust and trustworthiness requires improvements to the agents' social intelligence. Implementing the method of explaining an agent's actions is one such factor that contributes to this required social skillset. The concept of explanations is explored in the research field of Explainable Artificial Intelligence (XAI), providing theories and different approaches to making agents understandable to humans. The result of humans understanding agents is that they can better predict and rely on their teammates [9, 4, 1]. Often these explanations are performed ad-hoc on black box models; these are not capable of tailoring explanations to specific users or situations[5].

Goal-driven XAI develops a better method by making agents intrinsically understandable. These agents do not necessarily have to compensate for their performance [10], a misunderstanding hindering the development of this field. Agents capable of adjusting their behaviour to humans are classified as user-aware. If the agent senses that the human does not trust it, the explanation can be altered.

The following research question is formed to understand explanation tailoring further: how can an agent tailor its explanations to align human trust properly? The experiment compares baseline and trust agents, measuring general performance and the effects of tailoring on human trust. It is hypothesised that the trust agent will perform better on both fronts; based on the fact that it provides an improved dynamic approach to its explanations.

The rest of this paper is organised as follows. Section 2 will provide related works. Section 3 will provide the methodology. Section 4 will justify this research. Section 5 will present the results. Section 6 discusses the outcomes. Section 7 concludes this paper.

2 Related Works

The research questions contains two main techniques: modelling human trust and tailoring explanations. Both have a vast amount of literature describing different strategies. This section will provide a generalised overview of each research field and explain the picked methods.

2.1 Trust and Explainability

Before explaining the different approaches to both fields, the terminology should be discussed. Both definitions are explored in this research paper's responsible scientific faculty and will therefore be used. Jorge [6] formalises trust as a belief about trustworthiness. How much a human trusts the agent is dependent on how trustworthy the agent seems. Verhagen [11] formalises the XAI field further into three different systems: incomprehensible, interpretable and understandable. Each system provides different attributes to which agents can be classified, simplifying the different terminology uses.

2.2 Modelling Human Trust

Research into modelling human trust provides different interdisciplinary approaches. The proposed modelling method will be based on simplified techniques. A good starting point is a method provided by Kaniarasu [12]. By requesting explicit feedback from the user during the experiment, it is able to estimate human trust. This model seems most intuitive but requires periodic interruptions of the participant's workflow. Floyd [13] proposes an inverse trust metric that looks at the agent's internal state. Human trust can thereby be calculated through assessing metrics such as agent performance and task completion. However, this method has only been experimented on within an agent-agent simulation, making it not as reliable.

The following methods of modelling human trust perform better but are outside this research's scope. Integrating these techniques would require more time or budget than is currently available. Guo [14] proposes a method through a

Bayesian inference approach. However, this method would require a preliminary experiment on which the model is based. Neubauer [15] measures human trust through facial expressions, which would require capturing and storing the participant's face. Ajenaghughrure [16] achieves modelling through an electroencephalogram (EEG). An impressive interdisciplinary method but including medical hardware is unachievable.

2.3 Tailoring Explanations

The explainability of agents inhabits the research field of goal-driven XAI; Anjomshoae [5] provides an extensive survey on the topic. Tailoring the explanations can be based on a combination of context- or user-aware modelling. One relevant conclusion of this research states that user-aware or context-user-aware agents are under-researched. Rudin [10] argues that this happens due to the misconception of a decrease in performance when deploying goal-driven XAI agents.

The different types of explanations are founded in the psychological research field. A classical contribution is that of Miller [17], introducing human explanation concepts to the XAI field. In turn, Hendrickx [18] provided more insights into applying the combination of statistical and counterfactual explanations.

3 Methodology & Experimental Setup

3.1 Search-and-Rescue Task

The participant is paired up with an agent in a virtual search-and-rescue operation visible in figure 1. The environment is developed in a specialised HAT research software tool called MATRX [19]. Through basic game mechanics, the participant is tasked to score as many points as possible. Points can be collected by saving the lives of victims; a distinction is made between the mild and severely injured, providing more points for the latter. Along the way, obstacles are presented that will hinder the participant's progress. The participant will have eight minutes to try and score 36 points.

Figure 1: MATRX environment at the start of an experiment

To enforce collaboration, specific tasks need to be performed together. Therefore, agent and human communications are exchanged through the chat messages visible in figure 2. Most obstacles, for instance, need assistance from the agent to gain access to the different rooms. Severely injured victims will also need to be saved together. The participant can request help for each possible action, to which the agent will comply. Interesting knowledge is also communicated through the interface, such as the location of mild victims. The full chat interface is visible in appendix A. The participant will remain in charge of the agent; at each decision point the human decides what action is performed. The agent will attempt to clarify the situation through suggestions.

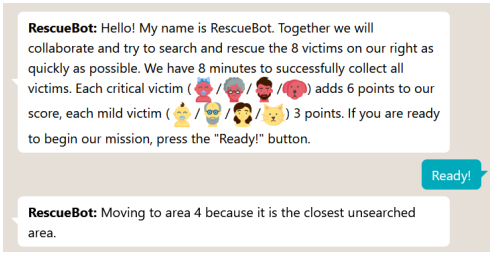

Figure 2: Chat messages exchanged between agent and human

3.2 Agent Design

The baseline agent is developed by the research faculty and capable of solving the search-and-rescue task through collaboration. It is a rule-based agent constantly deciding what needs to be done next. These decisions are based on weighing parameters such as remaining time, score and approximate distance to the human. Outgoing communication originates from these actions and will inform the human what it will do. Incoming knowledge updates from the participant get stored for potential later usage. Incoming requests are prioritised and will override its current plan.

The trust agent is built on the baseline, adding trust modelling and tailoring explanations. Modelling human trust is calculated through the number of suggestions followed or disregarded and maps to a value between 0 and 1. Suggestions disregarded have a more significant impact than those which are followed [17]. The research faculty provided the agent with a confidence level for each suggestion. When confidence in a suggestion is low, the agent will have a reduced change in trust. This allows for a more realistic approach; low confidence suggestions should not be treated the same as high confidence ones.

Tailoring explanations depends on the calculated trust value, which maps to four different so-called trust phases. Each phase adds the following explanation type from high to low: suggestion, confidence, feature explanation and counterexample. A detailed example can be seen in Appendix B. The baseline agent uses the same explanation dataset but picks one randomly for each suggestion. The following assumption was made during the agents' design: if trust is low more explanations are necessary to improve trust; if trust is high, fewer explanations are necessary to preserve time.

Using four different phases is assumed to be the sweet spot. The participants should see a noticeable change in agent explanations, but it also allows for more accessible measurements in the limited time window for the experiment. The mapping is achieved by equally dividing the provided range. The starting phase contains the following explanation types: suggestion, confidence and feature explanation.

3.3 Experimental Setup

The experimental setup involves human participants and has therefore been carefully planned out. They are first asked to fill in a consent form which conforms to TU Delft standard regulations. A preliminary questionnaire captures some general data such as age and gaming experience. The tutorial will explain all the basic game mechanics after which the game will begin. Once the eight minutes are up, the rest of the questionnaire is filled in.

3.4 Evaluation Metrics

The base agent provides various quantitative metrics ready to be used through its integrated loggings system. Interesting variables include score, suggestions ignored and agent moves. The trust agent adds its trust modelling value to the logging system, and from here on out, it will be called objective trust. The mean of each experiment's variable is calculated and stored after the eight minutes.

There are also four subjective quantitative metrics in the form of a questionnaire. These are included because they serve as a guaranteed metric. Each has been tried and tested within its paper and is considered reliable. Hoffmans' [20] fluency metric measures the general collaboration between agent and human. A different Hoffman [21] provides two great measures of explanation satisfaction and trust. The trust measurement will be called subjective trust. Lastly, to measure the workload, Harts' [22] NASA-TLX is used.

3.5 Analysis strategy

Two methods of statistical analysis will be used: correlation tests and two-sample t-tests. The correlation test can check whether specific agent metrics influence each other. For example, the hypothesis anticipates that the number of suggestions the participant ignored negatively influences objective and subjective trust. Two-sample t-tests will be used to analyse the differences between baseline and trust agents' variables. Do note that some differences will be present between objective and subjective metrics. Objective conveys a mean value over the eight minutes, whilst the subjective metrics are done after the experiment representing a single timeframe.

The primary research goal is focused on two things. How did the trust modelling perform, and did tailoring affect trust? Trust modelling can be studied through the correlation analysis between subjective and objective trust. Furthermore, comparing the baseline and trust agent with a comparison t-test could provide answers to the effects of tailoring. The t-test will look at factors such as suggestions ignored and subjective- and objective trust.

The analysis will also compare broader aspects through several t-tests. General performance will be analysed through score, completeness and subjective workload. Collaboration between agent and human will be examined through collaboration fluency. The quality of the explanations will be analysed through the explanation satisfaction questions.

4 Responsible Research

Reproducing this research has minimal random variables to it. The entire codebase which includes the base and trust agent is available through a publicly available fork of the faculties sub-repository [23]. The group of participants should be along the lines of our general data, with gaming experience being the most important factor. The only factor not accounted for is the decisions made by the participants. Different results can also occur due to this research's small number of participants. An important clarification is that no data was left out, in total 26 participants sat down for the experiment which have ended up in the results.

Furthermore, all data gathered during the experiment has been anonymised. An informed consent form was created along the general TU Delft guidelines with the approval of our supervisor and can be provided on request.

5 Results

In total, 26 experiments were conducted, of which 15 were on the baseline agent and 11 on the trust agent. The base agent tended to have more male participants than the trust agent. The average age of the base agent was more diverse, ranging from 18-34 whilst the trust agent spanned mainly the 25-34 age group. Both primarily consisted of participants with some college credits, the trust agent however did include almost each education level at least once. Gaming experience spanned equally over the two agents with both peaks on moderate and a lot of experience.

Comparing the baseline and trust agents is done through a method called two-sample t-tests. It checks whether the baseline agent means differ significantly from the trust agent mean. All assumptions required for a t-test are accounted for with the following options: Wilcox rank sum test with continuity correction, Welch two-sample t-test and Student two-sample t-test. All numbers are rounded to three digits.

5.1 Modelling Trust

Running a correlation calculation between objective and subjective trust presents a correlation value of 0.194 and a p-value of 0.3421, making them uncorrelated. Table 1 presents the pairwise correlation tests and corresponding p-values of interesting variables with subjective and objective trust. Subjective trust is moderately correlated with collaboration fluency and explanation satisfaction with a p-value of 0.00. Objective trust is only slightly correlated with explanation satisfaction with a p-value rounded down from 0.052. Objective trust is also highly correlated with suggestions ignored with a p-value of 0.00.

5.2 Performance and Collaboration

Measuring performance was done through a comparison t-test on the variables presented in table 2. The Student t-test was used on 'score' since it met all assumptions. The null

	sub corr	sub p-value	obj corr	obj p-value
collaboration fluency	0.58	0.00	0.01	0.98
explanation satisfaction	0.69	0.00	0.01	0.98
subjective workload	0.09	0.66	0.38	0.05
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

Table 1: Correlation results

hypothesis of equal means could not be rejected with a p-value of 0.336, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

The Wilcoxon rank sum test was used on 'completeness' since the data was not normally distributed. The null hypothesis of equal means could not be rejected with a p-value of 0.208, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

The Student t-test was used on 'agent moves' since it met all assumptions. The null hypothesis of equal means could not be rejected with a p-value of 0.676, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

The Student t-test was used on 'subjective workload since it met all assumptions. The null hypothesis of equal means could not be rejected with a p-value of 0.66, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

	mean diff	p-value	CI low	CI high	method
score	-2.636	0.336	-2.907	8.179	Student
completeness	-0.075	0.208	-1	-1	Wilcoxon
agent moves	-11.921	0.676	-46.211	70.054	Student
subjective workload	-2.733	0.66	-9.948	15.414	Student

Table 2: Performance results

Measuring collaboration was done through a comparison t-test on the variable presented in table 3. The Student t-test was used on 'collaboration fluency' since it met all assumptions. The null hypothesis of equal means could not be rejected with a p-value of 0.551, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

	mean diff	p-value	CI low	CI high	method
collaboration fluency	0.165	0.551	-0.729	0.399	Student

Table 3: Collaboration results

5.3 Trust and Explanations

Measuring trust was done through a comparison t-test on the variables presented in table 4. The Student t-test was used on 'objective trust' since it met all assumptions. The null hypothesis of equal means could not be rejected with a p-value of 0.865, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

The Welch t-test was used on 'subjective trust' since the data was normally distributed but differed in variances. The null hypothesis of equal means could not be rejected with a p-value of 0.339, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

The Student t-test was used on 'suggestions ignored' since it met all assumptions. The null hypothesis of equal means could not be rejected with a p-value of 0.974, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

	mean diff	p-value	CI low	CI high	method
objective trust	0.014	0.865	-0.187	0.158	Student
subjective trust	0.248	0.339	-0.785	0.289	Welch
suggestions ignored	-0.002	0.974	-0.136	0.141	Student

Table 4: Trust results

Measuring explanation satisfaction was done through a comparison t-test on the variable presented in table 5. The Wilcoxon rank sum test was used on 'explanation satisfaction' since the data was not normally distributed. The null hypothesis of equal means could not be rejected with a p-value of 0.124, resulting in the conclusion that the baseline agent and trust agent mean seem equal.

	mean diff	p-value	CI low	CI high	method
explanation satisfaction	0.221	0.124	-1	-1	Wilcoxon

Table 5: Explanation results

6 Discussion

The correlation results provide several insights into the agents' design. Most important is the performance of the human trust modelling estimate. With a correlation of 0.194, it underperforms the expectations. The subjective trust questionnaire is an extensively researched method of estimating human trust. The difference between run-time trust and ad-hoc trust is not responsible for a correlation value which lies this close to being uncorrelated. The human trust modelling is, therefore, most likely flawed. The seemingly sound intuition of calculating run-time trust through suggestions ignored proved too simple.

The correlations in table 1 also provide insights into different variables influencing trust. Both collaboration fluency and explanation satisfaction are moderately correlated with subjective trust. This most likely means that the questionnaire performs correctly in measuring different factors. The number of suggestions ignored is used in the objective trust calculation; this clarifies the high correlation between them. Notably, the number of suggestions ignored is not correlated to the subjective trust measurement, further consolidating the claim that modelling trust on suggestions ignored is too simple. The most straightforward measurement of agent performance is its score and completeness; being uncorrelated to both signifies a simple agent performance to human trust modelling would have also been insufficient.

Going over the different assets the research question wanted to tackle signifies a broader problem. It was hypothesised that the trust agent would improve the follow-

ing metrics: subjective trust, objective trust, suggestions ignored, collaboration fluency, explanation satisfaction, score and completeness. However, not even one metric proved a statistically significant improvement after ruling out objective trust as a reliable metric subjective trust is the most important measurement. The trust agent failed to improve human trust with only a mean difference of 0.248 and a p-value of 0.339. The only two variables which came close are explanation satisfaction (p-value: 0.124) and goal (p-value: 0.064). Both performed slightly better in the trust agent.

It also is not the case of the baseline agent outperforming the trust agent according to this experiment; all measurements are simply statistically insignificant. Essentially the baseline agent and trust agent produce the same results on the measured variables. The hypothesis is therefore rejected.

6.1 Reflection

A rejected hypothesis begs the question of why the experimental setup has either failed or is inconclusive. The variables measured and statistical analysis are sound. It could be argued that some of the objective metrics are inadequate, but the subjective metrics are tried and tested. Any differences between the two agents would have been signified in the questionnaire. Another reason is the disparity of data. Ideally, four more trust agent participants should have been added to comply to the minimum standard of 15. However due to unforeseen personal circumstances they were left out. Gathering more data, however, is done to prevent significant outliers. The fact that the 11 data entries already performed this similar to the baseline denotes a more significant issue.

In my opinion, results are caused by either a flawed assumption or information overload. A case can be made for weaknesses in the trust agent's design. The results indicate no statistically significant difference between baseline and trust, implying that the trust agent failed to differentiate its behaviour. This could have been caused by the fact that both agents present the same explanation types, just in a different order. Whilst the trust agent assumes that low trust requires more explanations and vice-versa, the baseline agent combines the explanation types randomly. Analysing the results proves that the trust assumption is flawed and that the experimental setup has failed. Seeing no difference between agents indicates that the assumption performs the same as the random approach. Therefore further research into the tailoring of explanations should dive deeper into what is most crucial for aligning human trust.

Another possibility for the results is based on the task causing information overload. Feedback received by most participants after the experiment was conducted stated that they did not notice the suggestions had changed over time. This problem occurred to both baseline and trust agents. When asked to describe their thought process on a prompted suggestion, many described only looking at the possible actions and deciding with intuition. Factors such as time pressure and the learning curve were given as feedback as to why this is happening, possibly indicating an information overload. This approach to decision-making skips the essential tailored explanations necessary to this research question causing the hypothesis to be inconclusive. It could explain why the results indicate no difference between baseline and trust agent. The metrics' suggestions ignored' and 'explanation satisfaction' indicates the trust agent behaved the same as the baseline agent, but in practice, it did change its behaviour. However, these arguments are improperly collected and cannot be fully backed up with the existing metrics. Therefore, more research is required into the participants' usage of the suggestions in the current experimental setup.

7 Conclusion

In this work, we have designed a user-aware agent capable of tailoring its explanations to human trust. The effects of tailoring were studied through implementing a HAT environment in software called MATRX and installing multiple quantitative metrics. The provided baseline agent was modified to model human trust and tailor explanations to it. The experiments were conducted on 26 participants and analysed through comparison t-tests and correlation tests. However, the results indicate no statistically significant difference between the two agents, causing the hypothesis to be rejected. There are two possible causes of the rejection, leading to either a failed or inconclusive experimental setup. First, the assumption on which tailoring is based could be flawed, resulting in an agent performing as skillfully as its baseline. This would have caused the experimental setup to fail; further research into tailoring explanations to human trust should be required. Second, feedback received after the experiments indicated an information overload which caused the participants to skip the essential tailored explanations. This would have caused the experimental setup to be inconclusive and requires more research into the usage of suggestions in the current setup.

A Chat Interface

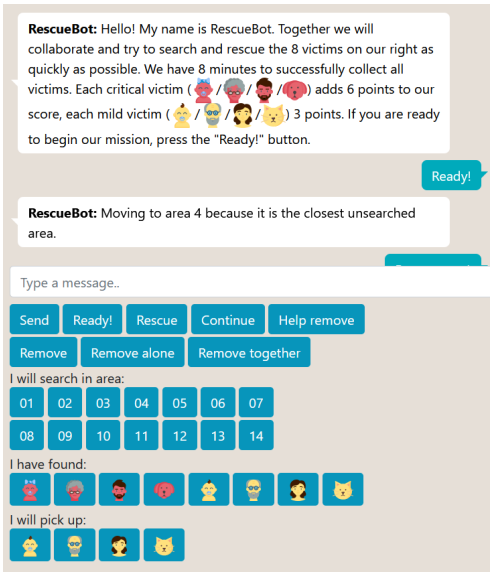

Figure 3: Chat interface with response buttons

B Trust Phases

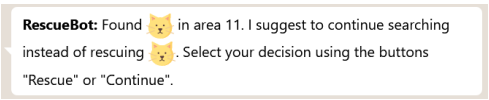

Figure 4: Agent message in a high trust phase

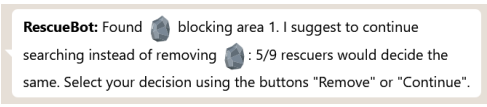

Figure 5: Agent message in a normal trust phase

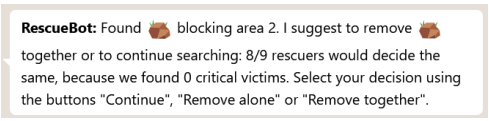

Figure 6: Agent message in a low trust phase

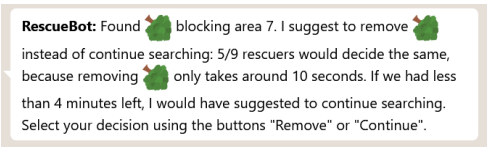

Figure 7: Agent message in a very low trust phase

References

[1] G. Klien, D. Woods, J. Bradshaw, R. Hoffman, and P. J. Feltovich, "Ten challenges for making automation a "team player" in joint human-agent activity," *Intelligent Systems, IEEE*, vol. 19, pp. 91 – 95, 5 2004.

[2] J. van Diggelen, J. S. Barnhoorn, M. M. M. Peeters, W. van Staal, M. L. Stolk, B. van der Vecht, J. van der Waa, and J. M. Schraagen, "Pluggable social artificial intelligence for enabling human-agent teaming," 9 2019.

[3] J. van Diggelen, M. Neerincx, M. Peeters, and J. M. Schraagen, "Developing effective and resilient human-agent teamwork using team design patterns," *IEEE Intelligent Systems*, vol. 34, pp. 15–24, 3 2019.

[4] M. Johnson and A. Vera, "No ai is an island: The case for teaming intelligence," *AI Magazine*, vol. 40, pp. 16–28, 3 2019.

[5] S. T. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," 2019.

[6] C. C. Jorge, S. Mehrotra, M. Tielman, and C. Jonker, "Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams," 4 2021.

[7] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, pp. 230–253, 6 1997.

[8] M. Lewis, K. Sycara, and P. Walker, "The role of trust in human-robot interaction," 2018.

[9] M. Johnson, J. M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. V. Riemsdijk, and M. Sierhuis, "Coactive design: Designing support for interdependence in joint activity," *Journal of Human-Robot Interaction*, vol. 3, p. 43, 3 2014.

[10] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," 5 2019.

[11] R. S. Verhagen, M. A. Neerincx, and M. L. Tielman, "A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable," 2021.

[12] P. Kaniarasu, A. Steinfeld, M. Desai, and H. Yanco, "Robot confidence and trust alignment," pp. 155–156, 2013.

[13] M. W. Floyd, M. Drinkwater, and D. W. Aha, "Learning trustworthy behaviors using an inverse trust metric," 1 2016.

[14] Y. Guo and X. J. Yang, "Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach," *International Journal of Social Robotics*, vol. 13, pp. 1899–1909, 12 2021.

[15] C. Neubauer, G. Gremillion, B. S. Perelman, C. L. Fleur, J. S. Metcalfe, and K. E. Schaefer, "Analysis of facial expressions explain affective state and trust-based decisions during interaction with autonomy," 2020.

[16] I. B. Ajenaghughrure, S. C. Sousa, I. J. Kosunen, and D. Lamas, "Predictive model to assess user trust: A psycho-physiological approach," pp. 1–10, Association for Computing Machinery, 11 2019.

[17] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[18] L. Hendrickx, C. Vlek, and H. Oppewal, "Relative importance of scenario information and frequency information in the judgment of risk," *Acta Psychologica*, vol. 72, pp. 41–63, 1 1989.

[19] TNO, "Matrx software."

[20] G. Hoffman, "Evaluating fluency in human-robot collaboration," *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS*, vol. 49, 2019.

[21] R. Hoffman, S. Mueller, G. Klein, and J. Litman, "Measuring trust in the xai context," *Michigan Tech Publications*, vol. PsyArXiv Preprints, 11 2021.

[22] 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.

[23] R. S. Verhagen and M. Vogel, "Project repository." https://github.com/mijnnaamismarin/TUD-Research-Project-2022, 2022. Accessed: 16-06-2022.