HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230121

Building Appropriate Trust in AI: The Significance of Integrity-Centered Explanations

Siddharth MEHROTRA [1], Carolina CENTEIO JORGE, Catholijn M. JONKER and Myrthe L. TIELMAN
Delft University of Technology
ORCiD ID: Siddharth Mehrotra https://orcid.org/0000-0002-2067-3451, Carolina CENTEIO Jorge https://orcid.org/0000-0002-6937-5359, Catholijn M. Jonker https://orcid.org/0000-0003-4780-7461, Myrthe L. Tielman https://orcid.org/0000-0002-7826-5821

Abstract. Establishing an appropriate level of trust between people and AI systems is crucial to avoid the misuse, disuse, or abuse of AI. Understanding how AI systems can generate appropriate levels of trust among users is necessary to achieve this goal. This study focuses on the impact of displaying integrity, which is one of the factors that influence trust. The study analyzes how different integrity-based explanations provided by an AI agent affect a human's appropriate level of trust in the agent. To explore this, we conducted a between-subject user study involving 160 participants who collaborated with an AI agent to estimate calories on a food plate, with the AI agent expressing its integrity in different ways through explanations. The preliminary results demonstrate that an AI agent that explicitly acknowledges honesty in its decision making process elicit higher subjective trust than those that are transparent about their decision-making process or fair about biases. These findings can aid in designing agent-based AI systems that foster appropriate trust from humans.

Keywords. Integrity, Appropriate Trust, Trust, Explanations, AI Agents, HCI

1. Introduction

The field of Explainable AI (XAI) is expanding quickly and focuses on exploring how Artificial Intelligence (AI) can provide explanations for its internal mechanisms. These explanations are designed to offer transparency into the AI's decision-making process and internal models [1]. Studies have demonstrated that such explanations can aid users in comprehending how the AI system operates [2,3], and there are ongoing efforts to ensure that AI is appropriately trusted with the help of explanations [4,5,6].

Typically, explanations are provided to convey information about a system's ability to engender appropriate trust. However, existing literature on human trust suggests that trust is not solely based on beliefs about ability. For instance, Mayer et al. identified the

[1]Corresponding Author: Siddharth Mehrotra, s.mehrotra@tudelft.nl

three pillars of trust: Ability, Benevolence, and Integrity and their framework has been widely adopted for modeling trust [7,8,9]. Although prior work has mainly focused on ability as the primary factor for creating explanations, we argue that integrity principles should also be incorporated into explanations. To address this gap, we propose a method for generating integrity-based explanations and investigate how these explanations affect human trust in an AI agent. Our research question is: How integrity-based explanations impact user's appropriateness of trust in an AI agent?

2. Creating Integrity Based Explanations

Integrity-related principles such as fairness, transparency, and honesty have been suggested as crucial for building appropriate trust in human-human interactions [10,11]. To express these principles through explanations, we focus on these three key aspects of integrity: (1) honesty regarding the system's capabilities and confidence, (2) transparency about the decision-making process, and (3) fairness in terms of sharing potential risks such as biases. To achieve this design of explanations, three researchers collaborated to generate sentences that formed explanations expressing these principles. We adopted the approach of situation vignettes from Strackand & Gennerich [12] to create text-based explanations. Our designed explanations are presented in Appendix A.

3. Method and Preliminary Results

In our experiment, participants were asked to estimate the calories of different food dishes based on an image of the food. At the *first* step, participants were shown an image of a food dish. They were asked to select an option out of four options for estimating the calories of the food dish. At the *second* step, an AI assistant guessed the correct answer from the same options as step one. At the *third* step, participant selected their final decision by choosing between themselves or the AI assistant. Finally, at the *fourth* step, participants rated their trust in the system based on the work by Yang et al. [13].

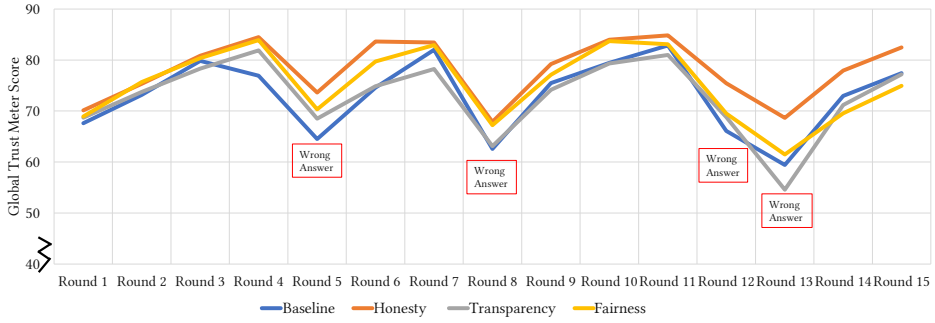

Figure 1. Illustration of mean responses for changes in Global Trust Meter over 15 rounds. The red coloured boxes represents when the AI agent provided a wrong answer *i.e.* round 5, 8, 12 and 13.

Our preliminary results indicate that trust in the AI agent dropped whenever the AI agent provided a wrong answer. We recorded an average drop of ˜15 points in trust score when a wrong answer was preceded by a right answer by the AI agent. This drop was ˜35 points when there were two wrong answers in a row. Furthermore, the honesty explanations were rated higher across all rounds and this difference populated further when AI agent provided two wrong answers simultaneously.

References

[1] Wang X, Yin M. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. ACM Transactions on Interactive Intelligent Systems (TiiS). 2022.

[2] Cai CJ, Jongejan J, Holbrook J. The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th international conference on intelligent user interfaces; 2019. p. 258-62.

[3] Páez A. The pragmatic turn in explainable artificial intelligence (XAI). Minds and Machines. 2019;29(3):441-59.

[4] Zhang Q, Lee ML, Carter S. You Complete Me: Human-AI Teams and Complementary Expertise. In: CHI Conference on Human Factors in Computing Systems; 2022. p. 1-28.

[5] Liu H, Lai V, Tan C. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. Proceedings of the ACM on Human-Computer Interaction. 2021;5(CSCW2):1-45.

[6] Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1-16.

[7] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human factors. 2004;46(1):50-80.

[8] Hoffman RR. A taxonomy of emergent trusting in the human–machine relationship. Cognitive Systems Engineering. 2017:137-64.

[9] Wagner AR, Robinette P, Howard A. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. ACM Transactions on Interactive Intelligent Systems (TiiS). 2018;8(4):1-24.

[10] Lacey J, Howden M, Cvitanovic C, Colvin RM. Understanding and managing trust at the climate science–policy interface. Nature Climate Change. 2018 Jan;8(1):22-8. Available from: https://doi.org/10.1038/s41558-017-0010-z.

[11] Mehrotra S, Jonker CM, Tielman ML. More similar values, more trust?-the effect of value similarity on trust in human-agent interaction. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society; 2021. p. 777-83.

[12] Strack M, Gennerich C. Personal and Situational Values Predict Ethical Reasoning. Europe's Journal of Psychology. 2011;7(3):419-42.

[13] Yang F, Huang Z, Scholtz J, Arendt DL. How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces; 2020. p. 189-201.

A. Appendix A

Baseline (Average length = 55 words, SD = 6 words) *The list shows the ingredients that I have accurately identified, along with their corresponding confidence scores. This information is sourced from the UNESCO food nutrition website. After totaling the calorie count, the overall sum is 738 calories. As a result, I would select the option labeled 750 since it is closest to the calculated value for the identified ingredients.*

Honesty (Average length = 125 words, SD = 23 words) *I believe that honesty is crucial, so I must confess that I am not entirely certain about identifying the total calories present on this plate. My uncertainty primarily stems from the food item marked with a dark white circle since I have limited training data for identifying this particular item. However, I have accurately identified the other items on the plate, and they are listed in the table. The information used for identification is sourced from the UNESCO food nutrition website. After adding up the calories, the total count is 750 calories, which is relatively close to 738 calories. Hence, I would select the option labeled 750 with a confidence level of 62.5%. This confidence level indicates that I am moderately certain about my response.*

Transparency (Average length = 128 words, SD = 19 words)*I have determined that the correct answer to this question is 750 calories. However, I believe in transparency, so I will explain how I arrived at this conclusion. I examined my training data from the UNESCO Food & Nutrition website and found a dish*

that closely resembled the one in the given plate in terms of calorie count. Although the dish I found was a curry, I cannot confirm its specific type. The corresponding visualization is displayed alongside the identified ingredients. Based on my training data and the dish search results, I estimate the total calorie count to be 738 calories with a confidence level of 62.5%, which is similar to the best match example. Nonetheless, I have selected 750 calories as the answer to this question.

Fairness about risk (Average length = 130 words, SD = 25 words) *As an AI language model, I strive to be fair and unbiased in all my responses. Therefore, I feel it is important to explain how I have addressed any potential biases in my answer to this question. In this instance, I cannot confidently identify the total calorie count of the plate since the food item encircled in a dark white circle poses some challenges. There is no clear consensus among human annotators regarding this item, as it has been variously classified as bay leaf, fish, meat, chicken, or beef. Moreover, while the image is labeled as an Indian Madras curry, the annotators for its ingredients only come from the western population, none of whom have a profession tag of chef. I have correctly identified the other items on the plate, which gives an estimated calorie count of 738. However, due to the uncertainties surrounding the encircled item, I have combined all the existing knowledge and selected the option labeled 750 with an overall confidence level of 62.5%.*