How Do AI Explanations Affect Human-AI Trust?

Lam Bui, Marco Pezzola, and Danushka Bandara(✉)

Fairfield University, Fairfield, CT 06825, USA
dbandara@fairfield.edu

Abstract. We conducted a study to explore the effect of different types of Artificial intelligence (AI) explanations on the human trust in AI systems. An in person user study was conducted (n = 7) and the trust condition was induced by varying the accuracy of AI response. The human trust was measured by surveys administered after each trial. The participants physiological data was also collected during the experiment. Our results show that image based explanations induced higher level of arousal in the participants, and the participants preferred image based explanations. We also found that the AI response accuracy had an effect on the user acceptance of AI's decision in the following trial. We also found that Photoplethysmography results had statistically significant correlation with the level of trust. The implications of this study are that AI performance and type of explanations both have an effect on the level of user trust in AI. Also this work could be extended to develop an objective measure of trust using physiological data.

Keywords: Human AI interaction · Trust · Physiological measures

1 Introduction

Most of human life today is integrated into Machine Learning (ML) and Artificial Intelligence (AI) systems, with more complex and accurate algorithms being developed each day. So much of our lives are influenced through AI, whether it be our car's GPS giving us the fastest route home, smart homes adjusting our thermostat, or even the advertisements we see online relating to a product we just searched on the internet. In many past studies, researchers have studied human interaction with ML/AI. This paper will discuss our own research between humans and ML, we study this relationship using physiological sensors and self report surveys.

Human trust in machine learning refers to the confidence and reliance that individuals place in the ability of ML systems to make accurate predictions and decisions. Building trust in ML systems is important for their successful deployment and adoption in various industries and applications. According to literature, there are several aspects that affect human trust in ML systems.

- Explainability or the ability to understand the internal logic behind the decision [1, 2].
- Performance: People are more likely to trust a ML system when it demonstrates high accuracy and low error rates [3].

H. Degen and S. Ntoa (Eds.): HCII 2023, LNAI 14050, pp. 175–183, 2023.
https://doi.org/10.1007/978-3-031-35891-3_12

- Control and autonomy: People are more likely to trust a system when they have control over it and can override its decisions [4].

Yang, Huang, Scholtz, and Arendt [5] found that subjects demonstrated higher trust when both understandability and layout of the explanation is provided, with 95% bootstrap confidence intervals; where image-based explanations outperformed rose chart-based explanations since image-based explanations "increase appropriate trust, decrease overtrust and undertrust, improve self-confidence, and show more usability". In another study, Yin, Vaughan, and Wallach [3] demonstrated that high user trust comes most commonly when models observe or state a higher accuracy percentage than the user and decrease their trust otherwise. This demonstrates that subjects are biased to believe in a machine's prediction so long as it exceeds their own.

1.1 Physiological Manifestations of Trust

Heart Rate Variability
Trusting environments or individuals may be associated with increased heart rate variability, which is a measure of the variation in the time interval between heartbeats. Higher heart rates would imply a higher level of trust due to the correlation with higher levels of expectations [6, 7].

Cortisol Levels
Trusting environments or individuals may be associated with lower cortisol levels, which is a hormone associated with stress. In a study conducted by Ditzen, Schaer, Gabriel, Bodenmann, Ehlert, and Heinrichs [8], both men and women would have reduced cortisol levels while in a state of positivity.

Skin Conductance
Trusting environments or individuals may be associated with changes in skin conductance, which is a measure of electrical conductance of the skin. Showing trust in ML that concludes with a negative or uncomfortable result should increase conductance. Such an example is with self driving vehicles deciding on routes to take [9]. Researchers found that Skin Conductance would be higher prior to highly rewarding and risky decisions; meaning "if the participant had not registered the fact that the decision was risk, skin conductance did not increase."

Brain Activity
Trusting environments or individuals may be associated with changes in brain activity, specifically in regions related to social cognition, emotion regulation, and decision-making [10].

2 Study Design

2.1 Measurement of Trust

The aim of this study is to explore the concept of trust in ML systems and investigate the physiological, performance and explainability factors that contribute to building trust between humans and ML. We will discuss the implications of our findings for the design, development and deployment of ML systems. In our study, subjects were presented with a task of classifying authentic and non authentic (doctored) images. The trust and distrust conditions were induced by a simulated AI system providing correct or incorrect suggestions.

- If the participant accepted the AI decision, we consider that the participant trusted the AI.
- If the participant rejected the AI decision, we consider that the participant distrusted the AI.

We analyzed the data through different statistical techniques to show the relationship between different variables in the experiment and trust.

2.2 Experimental Design

We conducted an experiment where seven subjects participated in a ‘Human AI Interaction’ experiment. The participants were approximately 20 years of age and were from a university in the northeast United states. Participants completed a pre-experiment questionnaire to gauge their base level of trust in computers. We adopted the disposition to trust inventory [11] and adjusted the questions to reflect human-computer trust. When participants arrived for the experiment, they were seated in front of the computer and presented various images consecutively with 10 s of rest between each trial. Their job was to judge whether the images were authentic or ‘doctored’ (manipulated in any way). We used the CASIA [12] dataset as the source for the images (Fig. 1).

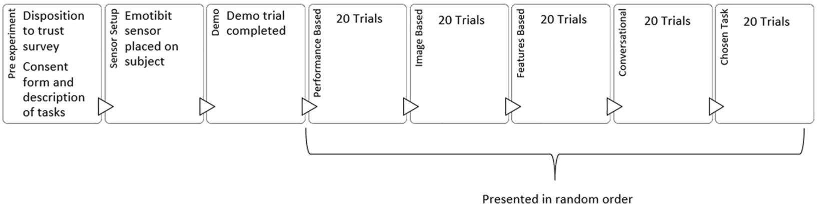


Fig. 1. Experimental design

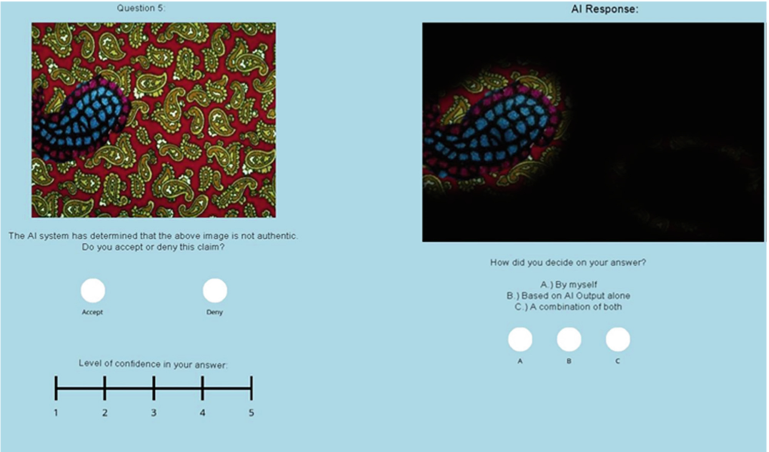
In addition to the images, participants saw the AI decision of whether the image was authentic or not authentic. We used a ‘wizard of oz’ methodology for this study (i.e., we simulated the AI response) to show the participants a variety of AI responses. We also administered a trust measurement survey during the experiment to gauge how trust varies subjectively [13].

2.3 Types of Explanations

We used four types of explanations in this experiment.

1. Image-based explanations: In case of non-authentic images from the CASIA dataset, the image was occluded and the part of the image relevant to the decision was uncovered. For authentic images, the whole image was shown.

2. Performance based explanations: Had a textual description of the machine learning model performance and the performance metrics such as Area Under the Curve (AUC), Precision, Recall, Accuracy.
3. Feature-based explanations: The features (Color and noise patterns, dense field copy move, JPEG dimples, self consistency splice, splicebuster) were chosen to display the contribution of different features to the AI decision.
4. Conversational explanations: A representation of the AI agent was shown with a short sentence about the confidence level of the AI in the answer (Figs. 2 and 3).



Question 5:

The AI system has determined that the above image is not authentic. Do you accept or deny this claim?

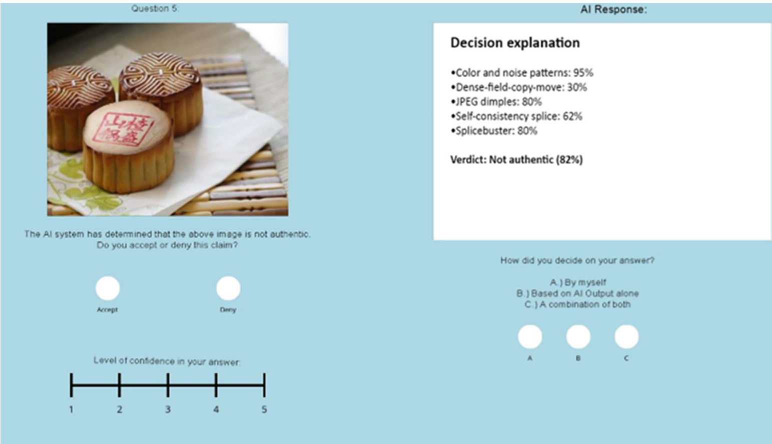
Level of confidence in your answer:

AI Response:

How did you decide on your answer?

A.) By myself
B.) Based on AI Output alone
C.) A combination of both

Fig. 2. An example trial with an image based explanation for a non-authentic image.



Question 5:

The AI system has determined that the above image is not authentic. Do you accept or deny this claim?

Level of confidence in your answer:

AI Response:

Decision explanation

•Color and noise patterns: 95%
•Dense-field-copy-move: 30%
•JPEG dimples: 80%
•Self-consistency splice: 62%
•Splicebuster: 80%

Verdict: Not authentic (82%)

How did you decide on your answer?

A.) By myself
B.) Based on AI Output alone
C.) A combination of both

Fig. 3. An example trial with a feature based explanation.

2.4 Collected Data

Throughout our study, we collected physiological data from participants utilizing emotibit, an open-source biosensor that measures the following data,

- Electrodermal activity
  - EDA-Electrodermal Activity
  - EDL-Electrodermal Level
  - Skin Conductance Response (SCR) Amplitude
  - Skin Conductance Response (SCR) Rise Time
  - Skin Conductance Response (SCR) Frequency
- Heart Rate
  - Heart Rate
  - Heart Inter-beat Interval
- Photoplethysmography (PPG)
  - PPG-Infrared
  - PPG-Red
  - PPG-Green
- Body movement (Using accelerometer, gyroscope and magnetometer)
  - Accelerometer (X, Y, Z)
  - Gyroscope (X, Y, Z)
  - Magnetometer (X, Y, Z)
- Temperature
  - Temperature via Medical-grade Thermopile

After each question/trial, the participants were presented with several questions relating to their trust perspective. The questions include,

- Do you accept or Deny the AI decision?
- How confident are you about your answer?
- How did you arrive at your answer? (By myself, Based on AI output alone, A combination of both)

2.5 Research Questions

RQ1: Effect of the AI explanation types (image based, performance based, feature based, conversational) on trust
RQ2: How does the AI's previous performance affect trust?
RQ3: What are the physiological measures most correlated with participant trust level?

3 Results

3.1 RQ1

See Figs. 4 and 5.

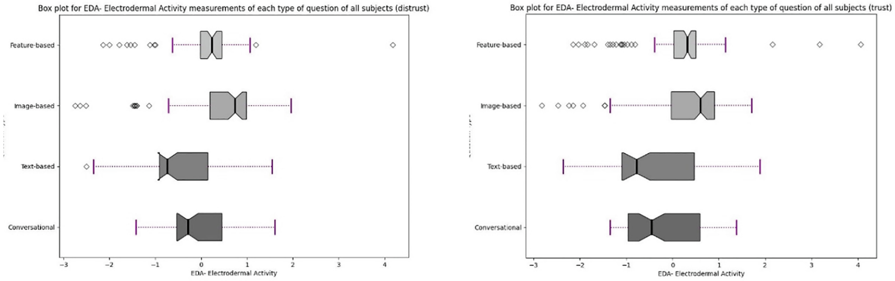


Fig. 4. EDA values for the distrust (left) vs trust (right) conditions do not show a noticeable difference. However, we do see a noticeable difference between the types of explanations. The image based explanations induced the maximum EDA levels in both conditions. Feature based explanations induced the next highest EDA levels.

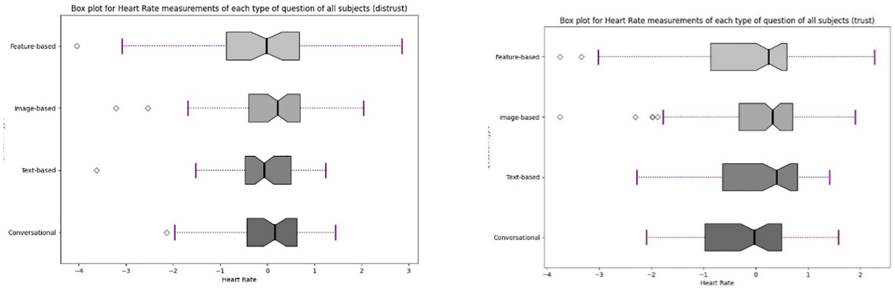


Fig. 5. Text based (performance based) explanations induced a higher heart rate in the trust condition (right) than the distrust condition (left).

3.2 RQ2

See Fig. 6.

3.3 RQ3

The physiological measures having highest correlation with trust level are, PPG green ($p < 0.005$), PPG infrared ($p < 0.05$) and PPG red ($p < 0.1$).

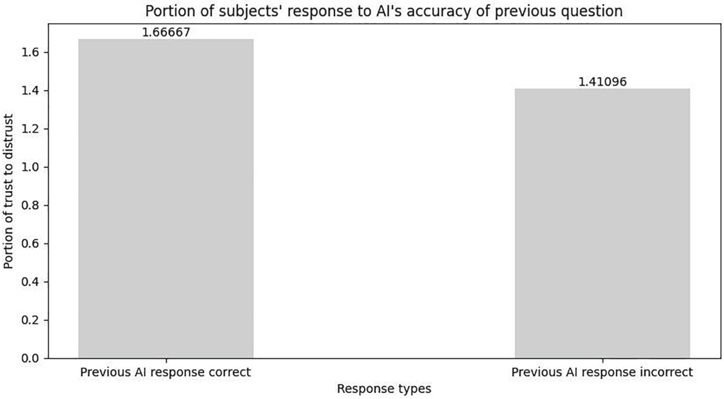


Fig. 6. Considering the ratio of trust to distrust, we can see that the participants had a higher ratio of trust if the AI provided the correct suggestion in the previous question vs if AI had provided an incorrect suggestion in the previous question.

4 Discussion

The results show that the various types of explanations do induce different physiological responses among the participants. Specifically, image based explanations induced higher EDA levels which is correlated with arousal. This could be due to the fact that images elicit a visceral reaction compared to the other methods of explanations.

Surprisingly, we did not notice a large difference between the heart rate variability for the trust and distrust conditions. It could possibly be due to the fact that our stimulus was not enough duration to create a difference in stress response in the subjects.

We did find that the AI's previous performance has an influence of the trust level. If the AI had given the correct answer to a question, the participants were more likely to trust the AI in the following question.

We also found that the Photoplethysmography data show statistically significant variation with the trust levels. Since PPG is an indicator of cardiovascular and respiratory activity, this could mean that it is capable of capturing the physiological response to trust vs distrust conditions.

Our findings indicate that image based explanations seem to have a larger effect on the participants, creating a higher level of arousal (Electrodermal activity). We also found that given a choice, participants preferred the image based explanations. Out of the seven subjects, all but one subject chose the image based explanation over the others. These results show that AI designers should pay attention to the type of explanations provided by AI as it has implications for trust in the AI and for human acceptance of AI decisions. Although we have a limited dataset, we can still see that trust and distrust conditions can be observed in the physiological data. Further studies could explore this further using machine learning models.

5 Conclusion

We designed and conducted a study to measure the effect of different types of AI explanations on trust. We used subjective surveys as well as objective physiological measures to develop a comprehensive view of human AI trust. We conclude that the types of AI explanations as well as past performance of the AI influence human trust in the AI system. Future work could explore the physiological correlates of trust with a larger dataset.

Acknowledgements. We would like to acknowledge the contribution to the trust testbed development by Robert Dillon and Noor Khattak. We would also like to thank Hien Tran and Miguel Cuahuizo for their contributions to the data analysis phase of the project.

References

1. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inform. **113**, 103655 (2021)
2. Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F., Kerren, A.: The state of the art in enhancing trust in machine learning models with the use of visualizations. Comput. Graph. Forum **39**(3), 713–756 (2020)
3. Yin, M., Wortman Vaughan, J., Wallach, H.: Understanding the effect of accuracy on trust in machine learning models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12, May 2019
4. Israelsen, B.W., Ahmed, N.R.: "Dave...I can assure you ...that it's going to be all right ..." a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. ACM Comput. Surv. **51**, 6, Article no. 113 (2019), 37 p. https://doi.org/10.1145/3267338
5. Yang, F., Huang, Z., Scholtz, J., Arendt, D.L.: How do visual explanations foster end users' appropriate trust in machine learning? In: Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 189–201, March 2020
6. Mitkidis, P., McGraw, J.J., Roepstorff, A., Wallot, S.: Building trust: Heart rate synchrony and arousal during joint action increased by public goods game. Physiol. Behav. **149**, 101–106 (2015)
7. Merrill, N., Cheshire, C.: Trust your heart: assessing cooperation and trust with biosignals in computer-mediated interactions. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 2–12, February 2017
8. Ditzen, B., Schaer, M., Gabriel, B., Bodenmann, G., Ehlert, U., Heinrichs, M.: Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict. Biol. Psychiatry **65**(9), 728–731 (2009)
9. Morris, D.M., Erno, J.M., Pilcher, J.J.: Electrodermal response and automation trust during simulated self-driving car use. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 61, no. 1, pp. 1759–1762. SAGE Publications, Los Angeles, September 2017
10. Zak, P.J.: The neuroscience of trust. Harv. Bus. Rev. **95**(1), 84–90 (2017)
11. McKnight, D.H., Choudhury, V., Kacmar, C.: Developing and validating trust measures for e-commerce: an integrative typology. Inf. Syst. Res. **13**(3), 334–359 (2002). Modified to cater to trust in AI instead of trust in people. Disposition to trust inventory

12. Pham, N.T., Lee, J.W., Kwon, G.R., Park, C.S.: Hybrid image-retrieval method for image-splicing validation. Symmetry **11**(1), 83 (2019)
13. Madsen, M., Gregor, S.D.: Measuring human-computer trust (2000)