Nuanced Effect of Human Trust in AI on Human-AI Collaboration Performance

Danushka Bandara
dbandara@fairfield.edu
Fairfield University
Fairfield, Connecticut, USA

Robert Dillon
robert.dillon@student.fairfield.edu
Fairfield University
Fairfield, Connecticut, USA

Noor Khattak
noor.khattak@student.fairfield.edu
Fairfield University
Fairfield, Connecticut, USA

ABSTRACT

In this study, we investigate the impact of trust on collaborative performance in human-AI teams, focusing on the task of image authenticity judgment. Utilizing the CASIA dataset, we conducted a human subject experiment (n=18), assessing their trust levels in AI systems and their collaborative performance in image tampering detection tasks. Participants underwent trials where they evaluated images from the dataset while considering AI output decisions, varying in accuracy. We observed that trust significantly influences collaborative performance. When participants entrusted decision-making to AI predominantly, higher trust levels correlated with superior collaborative performance, surpassing expected performance by 247%. Conversely, in scenarios where participants relied more on human judgment, lower trust in AI led to heightened collaborative performance, exceeding expectations by 406.7%. Notably, when decision-making involved a combination of human and AI inputs, lower trust in AI was associated with a notable performance increase of 467%. These findings underscore the nuanced relationship between trust and performance in human-AI collaboration.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

Human-AI interaction, trust, performance

ACM Reference Format:
Danushka Bandara, Robert Dillon, and Noor Khattak. 2024. Nuanced Effect of Human Trust in AI on Human-AI Collaboration Performance. In *Book of Extended Abstracts of the ACM Collective Intelligence Conference, 2024, June 27–28, 2024, Boston, MA.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Human trust in AI is a critical factor in determining the success of human-AI collaboration. It is important to understand when to trust or distrust AI to effectively leverage human expertise, particularly in situations where AI performance may be suboptimal. Additionally, human trust in AI is often overlooked, with a predominant focus on the machine's trustworthiness and performance. Recently, several objective methods of measuring human trust has been developed [1]. Making it a usable metric in Human Computer Interaction. Therefore it is important to understand the role that trust plays in decision making and performance.

We study the human AI system performance on the task of judging the authenticity of an image. Here we chose an existing doctored and non doctored image dataset named CASIA [2]. See an overview of the task presented in Figure 1, wherein we present the subjects with an image and obtain their response regarding whether they accept or reject the AI decision. The AI decision for the image was presented as: 'The AI system determined the above image is authentic/non-authentic. The 'AI' we used was simulated in a 'Wizard of Oz' method. Meaning the subjects were not aware of it being simulated.

The key findings of this study can be summarized as follows:

- Human AI collaboration does not always increase the performance of detecting image tampering. Especially when the AI performance is low.
- Human trust in AI improves the performance of the collaborative system only if the human mainly relies on the AI for decision-making.

2 METHODOLOGY

2.1 Images dataset

The CASIA dataset [2], short for the Chinese Academy of Sciences Institute of Automation (CASIA) dataset, is a collection of images commonly used in research related to computer vision, image processing, and machine learning. This dataset comprises a diverse range of images, including both manipulated (or "doctored") and authentic (non-doctored) images. These images are sourced from various sources and encompass a wide array of scenes, objects, and subjects. The CASIA dataset is often used in studies focusing on image manipulation detection, forensic analysis, and related fields.

2.2 Human subject experiment

In this human subject experiment, 18 participants aged 18-30 from a university in the northeastern United States underwent a series of image judgment tasks. Before the experiment, participants completed a pre-experiment questionnaire adapted from the Propensity to Trust Inventory [3] to assess their baseline trust in computers. Seated in front of a computer, participants were presented with consecutive images from the CASIA dataset and tasked with judging their authenticity, with 10 seconds of rest between trials. Using a "Wizard of Oz" methodology, participants were exposed to various simulated AI responses to the images, showcasing a range of AI behaviors and accuracy levels. Throughout the experiment, a trust measurement survey was administered to gauge subjective trust

CI '24, June 26–29, 2024, Boston, MA

levels in the AI system. Data collected included participants' image judgments, corresponding AI responses, and survey responses regarding trust. Analysis of the collected data aims to uncover the interplay between participants' baseline trust, observed AI responses, and subjective trust levels, shedding light on human-computer trust dynamics in AI-assisted tasks.

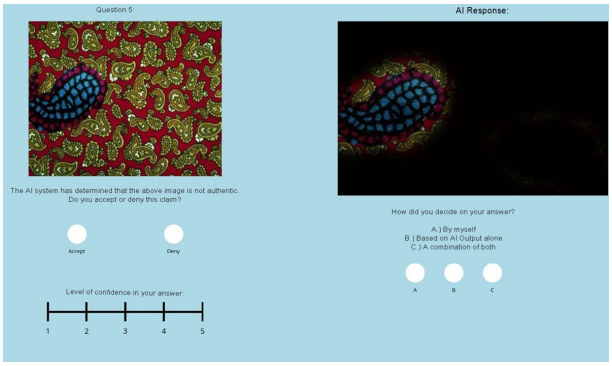

Figure 1: A testbed question which constitutes a single trial in our experiment.

3 DATA COLLECTION

The images were presented to the subjects using the Psychopy [4] platform. The subjects provided self-report responses after every image task to indicate the following,

- their acceptance/rejection of AI response
- their confidence in the acceptance/rejection decision
- whether the decision was made mainly by themselves/mainly using the AI response or a combination of both
- whether they trusted the AI during the current image task.

The subject responses were saved to a CSV file by the test bed for further analysis. The saved data also included whether the images presented were doctored/not doctored and whether the AI response presented to the participant was correct/incorrect.

3.1 Data analysis

We hypothesized that the system accuracy would be higher than the AI accuracy. However, as seen from the figure 2, this is not the case for every subject. Next, we wanted to delve more into the nuances of the performance differences. Particularly, we were interested in how the self-reported trust level of the participants affected the collaboration performance. Taking the group as a whole, we found that the average collaboration performance (accuracy) was 60% when the users reported mistrust of the AI, and 67% when they reported trust of the AI (p=0.087). Since this effect was not very significant, we explored the trust effect in more detail as described below.

Participants in this study did not make their decisions purely using the AI response, they could also use their own judgment or a combination of AI and their judgment to come to a decision. This was indicated in the self-report surveys administered during each trial. The next analysis is based on this aspect. We wanted to test how trust affected the performance when we stratified the

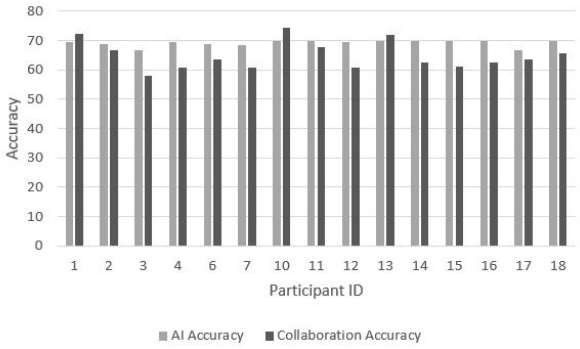

Figure 2: Variation of AI Accuracy and Collaboration Accuracy. The AI's accuracy is the number of correct responses provided by the AI. Collaboration accuracy is the number of correct responses provided by the system.

subject responses by the decision-making process of the subject. i.e. whether they made the decision mainly by themselves, mainly using AI, or a combination of both. We calculated the performance as the number of images that were correctly identified. The accuracy results for when the decision was made mainly by the participant are shown in table 1 and table 2. As seen in the tables, the collaboration accuracy depends on the AI accuracy, therefore doing a direct comparison between the collaboration performance between trust and mistrust conditions is not possible. To overcome this, we developed a model to obtain the expected collaboration accuracy.

Table 1: AI accuracy and collaboration accuracy for the mistrust condition when decision made by participant

Subject_ID	AI accuracy	Collaboration Accuracy
1	66.67	100
2	71.43	57.14
3	62.5	100
4	69.23	46.15
6	55	55
7	60	55
10	76.92	53.85
11	68.18	68.18
12	63.64	72.73
13	71.43	62.86
14	60.71	57.14
15	66.67	33.33
16	71.43	38.10
17	56.52	47.83
18	53.33	46.67

We found that the collaboration accuracy is linearly related to the AI accuracy (see figure 3). Using this linear model, we calculated the expected collaboration accuracy for each of the AI accuracy values. Then we checked how the average true collaborative accuracy compares to the expected collaborative accuracy. The results of this analysis are shown in figure 4.

Table 2: AI accuracy and collaboration accuracy for the trust condition when decision made by participant

Subject_ID	AI accuracy	Collaboration Accuracy
1	70.00	70.00
2	60.00	60.00
3	67.86	57.14
4	65.00	60.00
6	60.71	60.71
7	72.73	63.64
10	60.00	76.67
11	70.59	64.71
12	62.50	56.25
13	72.73	78.79
14	53.85	53.85
15	65.22	65.22
16	75.00	75.00
17	64.71	47.06
18	55.00	40.00

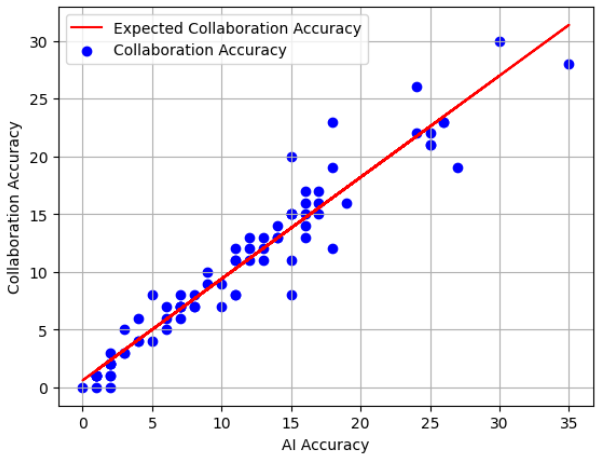

Figure 3: Linear regression model of expected collaboration accuracy based on AI accuracy.

4 DISCUSSION

We found that the average collaborative performance was higher than expected performance when they trusted the AI (247% higher) when the decision was made mainly using AI. However, when they made the decision mainly by themselves, the average performance was higher than expected (406.7% higher) when they did not trust the AI. In the case that the decision was made with a combination of AI and themselves, the performance was higher than expected (467% higher) when they did not trust the AI. This indicates that trust does have an effect on overall human-AI collaborative performance. When the trust level is lower, the performance is higher in cases where the humans make the decision themselves or with AI assistance. However, when they make decisions based mainly on AI, the higher trust levels result in higher performance. Our results show that there is a nuanced difference when it comes to the relationship

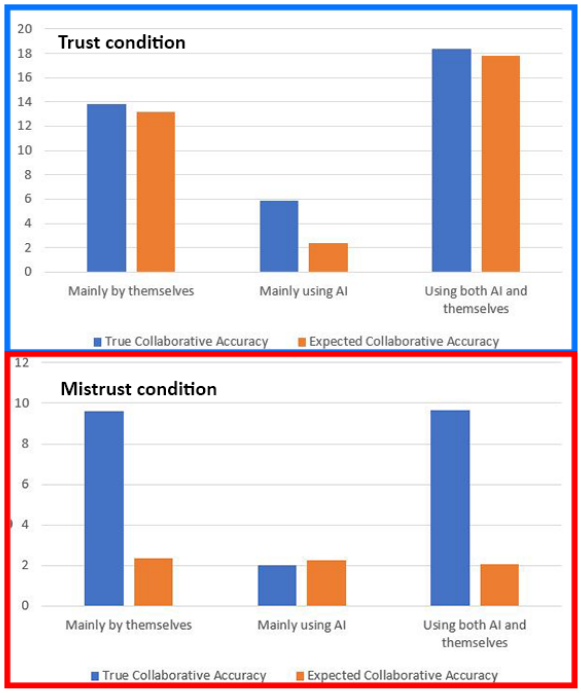

Figure 4: Comparison of true collaboration performance vs expected collaboration performance for the trust(and mistrust conditions

between trust and performance in human-AI teams. As decisions are made with more AI participation, higher trust tends to increase performance. When decisions are made with less AI participation, lower trust tends to increase performance. Of course our results are grounded in the context of image tampering as our experiment was limited to this task. Follow up studies could further explore the generalizability of our results. In essence, our findings underscore the significance of trust dynamics in shaping the performance of human-AI collaborative teams. The results indicate that the level of trust individuals place in AI systems significantly influences collaborative outcomes. Higher levels of trust in AI tend to elevate performance, particularly in scenarios where decision-making is predominantly AI-driven. Conversely, lower levels of trust in AI correlate with heightened performance when decision-making relies more on human judgment. Our study highlights the nuanced and nature of the relationship between trust and performance in human-AI collaboration, emphasizing the importance of understanding and managing trust dynamics to optimize collaborative outcomes.

REFERENCES

[1] D. Bandara and S. Sau. 2023. Are Scrutiny and Mistrust Related? An Eye-Tracking Study. In *International Conference on Human-Computer Interaction*. Springer Nature Switzerland, Cham, 539–545.

[2] Jing Dong, Wei Wang, and Tieniu Tan. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE. https://doi.org/10.1109/chinasip.2013.6625374

[3] Mark L. Frazier, Paul D. Johnson, and Stav Fainshmidt. 2013. Development and Validation of a Propensity to Trust Scale. *Journal of Trust Research* 3, 2 (2013), 76–97.

[4] Jonathan W. Peirce. 2007. PsychoPy—Psychophysics Software in Python. *Journal of Neuroscience Methods* 162, 1–2 (2007), 8–13.