Emerging Technologies and Future of Work, Vol. 117, 2023, 125–132
https://doi.org/10.54941/ahfe1004412

Exploring Trust and Performance in Human-Automation Interaction: Novel Perspectives on Incorrect Reassurances From Imperfect Automation

Jin Yong Kim[1], Szu-Tung Chen[1], Corey Lester[2], and X. Jessie Yang[1]

[1]Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48105, USA
[2]College of Pharmacy, University of Michigan, Ann Arbor, MI 48105, USA

ABSTRACT

Previous studies examining humans' trust towards automation have primarily focused on when both automation and humans receive the same input information. In this experiment, we explored human's trust towards an automated decision aid using the information provided by humans as the input. This unique scenario highlighted a new challenge wherein automation's incorrect verification of humans' wrong actions can lead to humans receiving (incorrect) reassurance that their action was right, even when it was not. The results indicated that incorrect reassurance leads to lower performance and greater trust decrement. We also observed outcome bias, where the incorrect automation recommendations were penalized less when the final performance was less harmed.

Keywords: Trust in automation, Automated decision aids, Mental rotation task, Outcome bias, Human-automation interaction, Human-AI-interaction, Human-computer interaction

INTRODUCTION

Consider the following hypothetical scenario:

Sarah, a highly skilled pharmacist, is responsible for filling the medication bottles for each prescription order. Recently, the pharmacy Sarah works for has introduced an AI computer vision system that can scan the filled bottle and identify the specific medication inside. The AI system is introduced as another layer of verification before medication dispensing. Today, Sarah received a prescription order for one patient, Noah, who needs to take medication **X**.

Sarah had a lapse when filling the bottle and incorrectly filled it with medication **Z.** The filled bottle is then scanned by the AI system, which unfortunately makes an error in recognizing the filled medication. Now two cases could occur:

Case A: The AI system scans the filled medication and identifies it as medication **Y.** As medication **Y** is not the ordered medication, the AI system signals a red flag. We will refer to this case as the **simply incorrect** case.

Case B: The AI system scans the filled medication and incorrectly identifies it as medication **X** and signals a green light. We will refer to this case as the **incorrect reassurance** case.

The hypothetical scenario presents novel characteristics that are not covered in existing research paradigms examining trust in and dependence on automation. To the best of our knowledge, previous studies have primarily investigated scenarios where the automated decision aid makes a prediction based on raw information, presents the prediction to humans, and humans subsequently make the final decision. For example, an automated combat identification (CID) aid can scan the environment, identify a threat, and make recommendations to soldiers (Du, Huang, & Yang 2020; Neyedli, Hollands, & Jamieson, 2011). In this paradigm, the human and the automation have access to the same raw materials based on which their decisions can be made.

In contrast, in the hypothetical pharmacy scenario (Lester et al., 2021), the input to the AI system is human-processed data (i.e., Sarah filled the bottle, and the AI scanned the bottle). This difference leads to unique opportunities for human errors. For example, in Case B, Sarah incorrectly filled with medication Z but thought she filled with X. The AI system's incorrect identification of the filled medication Z as X serves as an incorrect reassurance to Sarah.

The primary objective of the study is to examine the difference between Case A (simply incorrect) and Case B (incorrect reassurance), both involving an initial wrong human choice. However, in Case B, the AI's incorrect decision will provide (incorrect) reassurance to the human that may mislead them to believe that their initial decision was correct. For example, Sarah could trust the automation prediction and dispense medication Z. The potential outcome may be most damaging because this action could potentially harm Noah's health. If Sarah questions the automation prediction and checks with her eyes, she may replace the bottle with medication X. Therefore, we hypothesize:

Compared to the simply incorrect case, incorrect reassurance case will result in lower performance and larger trust decrement.

METHODS

Participants: This research complied with the American Psychological Association code of ethics and was approved by the Institutional Review Board at the University of Michigan. Thirty-five university undergraduate and graduate students (average age = 22.31 years, SD = 3.14) participated in the experiment. Participants were required to have a normal or corrected-to-normal vision and received a compensation of ten dollars base rate with a bonus of up to ten dollars based on their performance.

Experimental Task: Prior to starting the experiment, participants completed a consent form, and a demographics survey (i.e., age and gender). Subsequently, they were provided with a video that explained the experimental task.

In the experiment, participants were given the mental rotation test, which was derived from spatial visualization tests (Shepard and Metzler, 1971; Vandenberg and Kuse, 1978). Python Tkinter package was utilized for task development.

During each trial, participants were shown a reference image alongside five options, consisting of one correct alternative and four distractors. The correct alternative shared the same structure as the reference image but appeared in a rotated position. Two distractors were randomly selected from rotated mirrored images of the reference image and the other two distractors were randomly chosen from rotated images of other reference images.

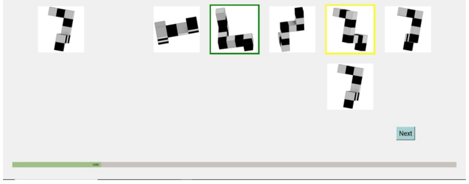
Initially, participants were instructed to select, as accurately as possible within 15 seconds, the image that represented the same 3D object as the one depicted in the reference image. After making the selection, they were required to click the "Next" button to proceed. If no selection was made within the given time limit, the text displayed "You did not make a selection within the time limit, -4pt".

Following the initial choice, participants rated their confidence in their initial choice using a visual analog scale. The scale ranged from "Not confident at all" on the leftmost point to "Absolutely confident" on the rightmost point.

Next, participants were presented with the AI system's recognition. The AI system attempted to identify the 3D object depicted in the participant's chosen image. The AI system correctly displayed the reference image of the correctly recognized 3D object 70% of the time, but incorrectly recognized the selected image 20% of the time. Additionally, in 10% of the trials, the AI system recognized the participant's initial choice as the reference image for the trial. If the participant's initial choice was right, it resulted in a correct prediction. On the other hand, if the participant's initial choice was wrong, incorrect reassurance case (Case B) appeared. Therefore, the automation reliability ranged from 70-80%, based on the participant's initial choice.

After viewing the AI's recognition, participants were given the option to either confirm their initial answer by clicking the "I was right" button or reject their initial answer by clicking the "I was wrong" button within 10 seconds. If no button was pressed within the time limit, the text on the final choice page displayed "You did not make a selection within the time limit, -2pt".

Subsequently, participants were presented with their performance feedback (see Figure 1). Points were allocated based on both the initial and the final choices. Participants received 2 points for the right answers and lost 2 points for the wrong answers. Additionally, participants lost points if they failed to respond within the time limit. Each trial resulted in either a gain of 4 points, a loss of 4 points, or no change in points for the participants.



Investigation of human trust change toward an automated decision aid in success and failure scenarios

The right answer is highlighted in Green
Your Initial Choice is Incorrect, -2 pt
You Incorrectly Commited to your answer, -2 pt

Current Score:
0

Figure 1: Participants were presented with the feedback page. Their initial answer is highlighted in yellow, machine prediction is shown below their initial answer, and the correct answer is highlighted in green. This is an example of the incorrect reassurance.

At last, participants were asked to rate their trust in the AI system after each trial using a visual analog scale. The trust scale ranged from “I don’t trust the decision aid at all” on the leftmost anchor to “I absolutely trust the decision aid” on the rightmost anchor.

The above steps were repeated 60 times throughout the experiment, presented in random order. The flowchart of the experimental task is illustrated in Figure 2.

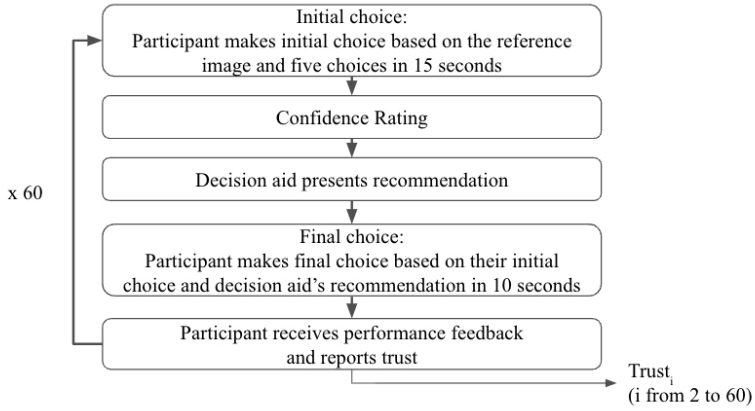

Initial choice:
Participant makes initial choice based on the reference image and five choices in 15 seconds
Confidence Rating
Decision aid presents recommendation
Final choice:
Participant makes final choice based on their initial choice and decision aid’s recommendation in 10 seconds
Participant receives performance feedback and reports trust
x 60
Trust$_i$
(i from 2 to 60)

Figure 2: Experimental procedure.

The experiment employed a within-subjects design. The independent variable was the patterns of performance. Based on the participant’s initial answer choice and the decision aid’s recognition, five patterns were identified (see Table 1).

The presence of each pattern or outcome was participant-dependent as the participant’s initial and final responses could not be manipulated.

Table 1. Patterns of performance.

Initial Answer Choice	Decision Aid Prediction	Pattern Name
Right	Correct	
Right	Incorrect	
Wrong	Correct	
Wrong	Incorrect	Case A: Simply Incorrect
Wrong	Incorrect – Matches the reference	Case B: Incorrect Reassurance

Measures

Trust change: We used Lee and See’s definition of trust, which is defined as the attitude that an agent will help achieve an individual’s goal in a situation characterized by uncertainty and vulnerability (2004). After each trial *i* participants reported their *trust(i)* in the decision aid. We calculate trust change as:

Trust change (i) = Trust(i) – Trust(i-1), where *i*=2, 3, …, 60

Since the moment-to-moment trust is reported after each trial, only 59 trust adjustments are obtained from each participant.

Final Performance: Another dependent variable was the final performance of the experimental task. The final performance was calculated as the percentage of correct final answers for each pattern for each participant.

Confidence: After participants selected their initial answer, they rated their confidence in their selection on a visual analog scale from 0 to 100. The confidence was calculated for each pattern by taking the average rating from each of the trials for each pattern.

Reaction Time: Initial reaction time was measured in seconds from when the five answer choices appeared until when participants pressed the next button. The final reaction time was measured in seconds from when the machine prediction appeared until when participants pressed either the "I was right" or "I was wrong" buttons.

RESULTS

The number of occurrences for the 5 patterns of performance was calculated post-experiment. For this paper, we focus on the patterns discussed in the introduction. The simply incorrect pattern occurred for each of the 35 participants. However, the incorrect reassurance pattern did not occur for four participants. Because the patterns of performance was a within-subject variable, one-way repeated measures ANOVA was conducted.

Table 2. Mean and SD of important measures of the experiment.

Patterns Of Performance	Trust Change Mean (SD)	Final Performance (%) Mean (SD)	Confidence Rating Mean (SD)	Initial Reaction Time (s) Mean (SD)	Final Reaction Time (s) Mean (SD)
Case A: Simply Incorrect	-0.8 (5.0)	62.7 (29.6)	42.6 (20.1)	12.5 (1.8)	4.8 (2.2)
Case B: Incorrect Reassurance	-7.0 (11.4)	24.2 (32.2)	52.0 (21.6)	12.1 (2.9)	3.3 (2.8)

There was a significant effect of patterns of performance on trust change *(F(1,64) = 8.57, p<.01)*. The incorrect reassurance pattern showed a greater decrement in trust change compared to the simply incorrect pattern *(difference = 8.98)* (see Figure 3).

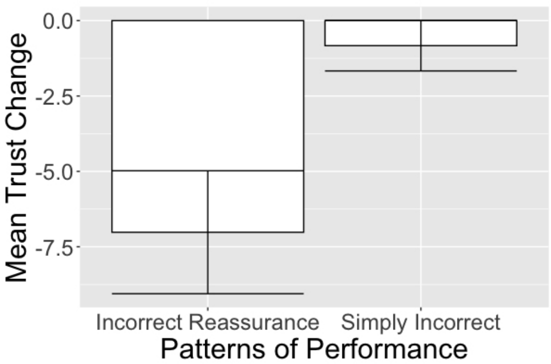


Figure 3: Comparison of trust change between simply incorrect pattern and incorrect reassurance pattern.

Patterns of performance had a significant effect on final performance *(F(1,64) = 25.63, p<.001)*. The simply incorrect pattern showed higher final performance compared to the incorrect reassurance pattern *(difference = 38.53)* (see Figure 4).

There was also a significant effect of patterns of performance on final reaction time *(F(1,64) = 5.42, p<.05)*. The incorrect reassurance pattern showed shorter reaction times compared to the simply incorrect pattern *(difference = 1.43 seconds)* (see Figure 5).

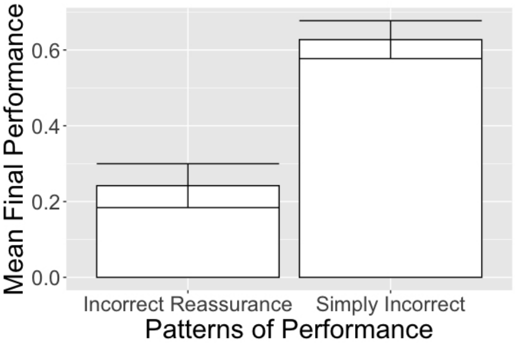


Figure 4: Comparison of final performance between simply incorrect pattern and incorrect reassurance pattern.

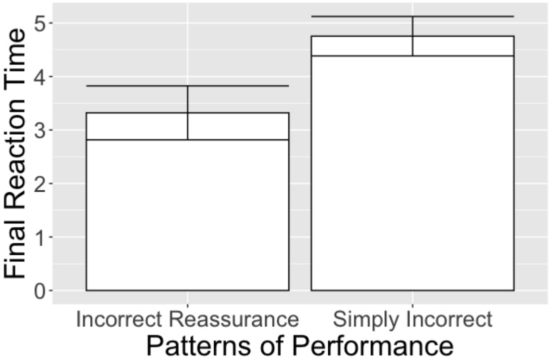


Figure 5: Comparison of final reaction time between simply incorrect pattern and incorrect reassurance pattern.

No significant effect of patterns of performance was found on confidence *(F(1,64) = 3.31, p=.074)* and initial reaction time *(F(1,64) = 0.45, p=.51)*.

DISCUSSION AND CONCLUSION

Results revealed that the incorrect reassurance patterns, when participants received incorrect reassurance from automation for wrong initial answers, resulted in a shorter final reaction time, and worse final performance than the simply incorrect patterns, when participants received incorrect automation prediction that did not match either the reference image or participants' initial answer choice. The incorrect reassurance from the AI led the participants to quickly make the wrong decision for the final choice.

Participants also had a larger trust decrement from incorrect reassurance patterns compared to simply incorrect patterns. This aligns with the previous studies of outcome bias (Baron and Hershey, 1988; Yang et al., 2021) that even though automation made incorrect predictions for both patterns, the trust decrement of incorrect reassurance pattern was greater as the final outcome was worse for this pattern compared to the simply incorrect pattern.

From the Swiss Cheese model, a damaging failure can occur when holes from multiple layers align (Reason, 2000). In the context of the incorrect reassurance pattern, the first hole in the Swiss Cheese model is the participant's wrong initial choice and the second hole is the incorrect reassurance that suggests the participant had initially made the correct choice. When the two holes lined up, it's highly likely that the participant could not recognize their error and kept committing to the error they made initially. This result confirms our hypothesis that the incorrect reassurance case (Case B) is problematic and requires further in-depth investigation.

These findings contribute to a fundamental understanding of how human trust is influenced in scenarios of automation failures when input information to the automation is processed instead of the raw information that humans have. These insights have practical implications for the design and implementation of semi-automated decision aids in domains where safety and effectiveness are critical.

ACKNOWLEDGMENT

This work was supported in part by the University of Michigan Industrial and Operations Engineering departmental fellowship and the Rackham Graduate Student Research Grant awarded to the first author, and the National Institutes of Health's National Library of Medicine grant R01LM013624.

REFERENCES

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of personality and social psychology*, *54*(4), 569.

Du, N., Huang, K. Y., & Yang, X. J. (2020). Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. *Human factors*, *62*(6), 987–1001.

Lester, C. A., Li, J., Ding, Y., Rowell, B., Yang, J. X., & Kontar, R. A. (2021). Performance evaluation of a prescription medication image classification model: an observational cohort. *NPJ Digital Medicine*, *4*(1), 118.

Neyedli, H. F., Hollands, J. G., & Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human factors*, *53*(4), 338–355.

Reason, J. (2000). Human error: models and management. *Bmj*, *320*(7237), 768–770.

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science (American Association for the Advancement of Science)*, *171*(3972). https://doi.org/10.1126/science.171.3972.701

Vandenberg, S. G., & Kuse, A. R. (1978). Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization. *Perceptual and motor skills*, *47*(2). https://doi.org/10.2466/pms.1978.47.2.599

Yang, X. J., Schemanske, C., & Searle, C. (2021). Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation. Human Factors, 00187208211034716. https://doi.org/10.1177/00187208211034716