

Thesis Summary

For the last few years, the research community has observed great leaps in Artificial Intelligence (AI) research, largely due to improvements in Natural-Language Processing (NLP), which allowed for successful research in large language models (LLM). The LLM's (e.g., GPT, Claude, Gemini, etc.) have shown promising results in a wide variety of applications, ranging from the creativity sphere to healthcare and beyond. With these changes, human-AI collaboration has become a popular focal point. Human-AI collaboration (HAIC) is characterised by the joint efforts of humans and AI to achieve a common objective. Despite recent advancements, incorporating AI into tasks that typically require humans to perform them is far from trivial. The collaboration has raised concerns about trust, reliability, and bias, amongst other issues.

In this thesis, we investigate how different skill levels of people perform in a visual problem-solving task with an intermediary AI feedback assistant with regards to accuracy and confidence. To achieve this, we built a website from scratch, which presented twelve instances of “Raven’s Progressive Matrices” tests and recorded various metrics throughout the process. First, the test instance must be solved without any assistance, where an answer out of eight options is selected and a self-reported confidence level in the answer. Next, AI feedback is given, which is purposefully rigged to be incorrect in some cases, as well as its confidence level. The participant can then decide if the feedback changes their answer or confidence in any way and proceed to answer yet again. We ran the within-subjects study with $N = 25$ participants and found that despite the AI feedback being wrong occasionally, it elevates the lower and mid level skilled participants (judged by initial decision accuracy). Furthermore, we found that many participants preferred the assistant as an inspiration for patterns, suggesting that simply conveying the patterns or rulesets found without imposing an answer might be beneficial to the collaboration. Another finding was that a higher amount of trust was correlated with initial accuracy, suggesting that lower-skilled participants are more trusting of the AI. Likely for a similar reason, higher trust is correlated with a higher answer switch rate, plausibly due to increased doubt in lower skilled participants. Finally, as a future work, we suggest trying to replicate the experiment in a different domain, such as creativity.

Trust and Accuracy in AI-Assisted Visual Problem Solving

Valbjörn Jón Valbjörnsson

Aalborg University

Aalborg, Denmark

vvalbj22@student.aau.dk

ABSTRACT

Human-AI collaboration has been on the rise in recent years due to Artificial Intelligence (AI) advancements. They have shown great promise in assisting humans with various tasks but pose challenges that are non-trivial to overcome. In this paper, we investigate human-AI collaborative problem-solving in a visual task, where the AI acts as a decision support system, giving feedback mid-way through the task. We manipulate the AI assistant to purposefully be incorrect and change the level of confidence in the message it conveys. Through a within-subject study ($N = 25$), we investigate how the AI feedback impacts participants of varying skill level in terms of performance, among other metrics. We found that people with lower skill levels benefit greatly in performance, achieving greater results than without feedback, and that people prefer to use the assistant as an inspiration for patterns. Given that, we conclude that an assistant might benefit from serving as a support type, giving ideas for patterns that the human collaborators might be missing.

CCS CONCEPTS

• Human-centered computing → Empirical studies in HCI.

KEYWORDS

AI decision-support systems

ACM Reference Format:

Valbjörn Jón Valbjörnsson. 2024. Trust and Accuracy in AI-Assisted Visual Problem Solving. In *Proceedings of . ACM*, New York, NY, USA, 9 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

For the last few years, the research community has observed great leaps in Artificial Intelligence (AI) research, largely due to improvements in Natural-Language Processing (NLP), which allowed for successful research in large language models (LLM). The LLM's (e.g. GPT, Claude, Gemini, etc.) have shown promising results in a wide variety of applications, ranging from the creativity sphere[9, 31], to healthcare[36] and beyond. With these changes, human-AI collaboration has become a popular focal point. Human-AI collaboration (HAIC) is characterised by the joint efforts of humans and AI to achieve a common objective[33].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00
<https://doi.org/XXXXXXX.XXXXXXX>

Despite recent advancements, incorporating AI into tasks that typically require humans to perform them is far from trivial. The collaboration has raised concerns about trust[26, 30], reliability[8, 12], and bias[25], among other issues. In particular, it remains unclear how AI cooperation—such as feedback—affects human decision-making in different contexts. For instance, tasks can have a wide range of importance, and different collaborators may have different needs (e.g., a novice might need more details). One such context is how AI affects individuals of varying skill levels on the same task, specifically how AI feedback would affect performance (accuracy, speed, or other task-specific metrics) and confidence.

In this paper, we examined a number of these effects with the visual task known as "Raven's Progressive Matrices" (RPM) [29], which is widely used to assess fluid intelligence and is frequently linked with IQ tests. To guide our research, we devised a few research questions of interest:

- RQ1. How does intermediary AI feedback affect the accuracy and success of collaborators of different skill levels?
- RQ2. How does AI confidence affect human confidence?
- RQ3. What role does a human's trust in AI have in decision-making?

We ran a within-subject experiment on a digitalised RPM version[34], where participants solved 12 RPM puzzles with intermediary AI feedback being correct or incorrect and confident or unconfident in a randomised manner.

We find that, despite the AI feedback being rigged to be incorrect, it elevated lower- and middle-skilled participants consistently. The participants also seemed to enjoy the feedback as an inspiration to see patterns they had not seen before. We suggest that it may be enough for a decision-support agent to reveal patterns and rule sets instead of imposing an answer. Confidence changes were present the most when participants were in doubt, suggesting that the feedback did not matter when they were certain. Furthermore, lower skills correlate with higher trust, and higher trust correlates with answer changes.

2 RELATED WORK

2.1 AI in Decision Support Systems

Artificial intelligence (AI) has been on a sharp rise in Decision Support Systems (DSS) in many fields over the past few years, such as healthcare[5, 15, 26], education[19, 21], business[3, 17] and many other areas. With the growing usage of AI support, researchers have identified common issues in human-AI collaboration. Namely, under-reliance[10, 27], when people show a direct or indirect reluctance to use AI suggestions, and conversely, over-reliance[5, 22], which generally manifests itself when people rely too much on the AI support or are unable to determine how much they should trust it. The research community has recently made an effort to learn more

about the factors that may or may not influence people’s reliance on AI feedback and suggestions. Lu et al. found that when second opinions are presented with a suggestion, regardless if the second opinion is from a peer or AI-generated, decision-makers showed reduced over-reliance and increased under-reliance[23]. Which can be favorable in cases where users are susceptible to over-reliance. However, humans can exhibit the Dunning-Kruger Effect (DKE) (i.e., overestimating their own abilities) and then tend to rely less on AI systems, which can hinder their performance. He and colleagues found that the DKE can be mitigated by introducing a tutorial task, which would act as feedback before the actual task[13].

3 METHOD

We conducted a within-subjects study to find out how an AI that offered intermediary feedback affected people with different skill levels. In our study, we utilised the well-known and established psychometric test "Raven’s Progressive Matrices"[29], where participants needed to identify a missing figure in a 3x3 matrix. See Section 3.1 for further clarification. We designed an AI that would provide a one-shot interaction with the user once the user locked in an answer. The participants experienced all four AI conditions, which were split evenly and randomly among the tasks.

3.1 Task

A widely studied psychometric task named "Raven’s Progressive Matrices"[29] was chosen for several qualities that it possesses for the experiment. The matrices come in all variations of difficulty and can be easy to understand, which makes it a convenient task to administer. Specifically, we used Set II from the Advanced Progressive Matrices variation, which contains 36 tasks of increasing difficulty. Carpenter et al. found a great variance in error rates among people of diverse backgrounds[6], which serves the experiment well by introducing different skill levels. Administering 36 tasks is still time-consuming (40–60 minutes), and thus, several studies have developed means to shorten the amount of time it takes. Hamel and colleagues suggested that a 20-minute timed version is an adequate predictor for scores acquired with no time limit[11]. Two other studies took a different approach and shortened it to 12 tasks, each using their own methodologies to identify the most impactful matrices in the test[4, 35]. Due to worries about encouraging participants to disregard AI feedback, we opted for the shortened task, specifically the tasks listed in Bors et al[4].

3.2 Experimental Setup

3.2.1 Task Design. For the study design, we introduced a two-phased decision-making process to evaluate the impact of the AI feedback. The phases were identical, with the difference being that the first phase was done before AI feedback and the second one after AI feedback. The decisions made before AI feedback will henceforth be referred to as **HD1** (human-decision 1), and the decisions made after AI feedback will be called **FD1** (final-decision 1). In each phase, there are four different steps. In the first step, participants look at the RPM puzzle. Then, they evaluate eight different options, pick one of them, and indicate how much confidence they have in the answer. Afterwards, AI feedback was given for the participants to read and evaluate. The interface can be viewed in Figure 5.

3.2.2 AI Assistant Design. As shown in Table 1, we designed four different conditions for the AI assistant. Each of them varies in correctness and the level of confidence it expresses.

	Correct Feedback	Incorrect Feedback
High Confidence	Correct-Confident	Incorrect-Confident
Low Confidence	Correct-Unconfident	Incorrect-Unconfident

Table 1: AI Assistant Feedback Conditions

The responses from the AI were predetermined messages where each condition had one answer for each task. We implemented an artificial response delay of 2–5 seconds, as some research has suggested that slowing down response times may reduce algorithm aversion[20, 27]. To generate the responses, we used GPT-4 by feeding it the image of the task along with the option it was supposed to advocate for.

3.3 Experimental Procedure

Participants were given a link to our web interface and supplied with a participant code in order to start the task. Upon entering, the participants received text instructions and image illustrations to aid them in the upcoming task. Once the instructions were read, the task began by showing the first of twelve Advanced Progressive Matrix (APM) puzzles on the left-hand side of the interface. The right-hand side contained three elements: an area to display the AI response, a selection of the eight possible solutions to the matrix, and a slider to indicate the confidence in the selected answer. When an answer and confidence have been locked in, the participants receive an analysis from an AI as a one-shot interaction and are given the opportunity to reconsider their answer and confidence. Following the completion of the twelve tasks, participants were required to complete a demographic survey and the TXAI[14] AI trust scale. Finally, participants were asked three open-ended qualitative questions regarding their experiences with the task and AI interaction.

3.4 Data collection

We collected a variety of data, including a set of diverse data from task performance and self-reported measures from surveys. Additionally, we asked three open-ended qualitative questions regarding the participants experience with the AI interaction throughout the tasks.

3.4.1 Task Performance. Records of various metrics throughout the task were stored. First of all, we split the task decisions into two parts (HD1 and FD1). For the HD1 and FD1 decisions, we stored: an answer, the participant’s confidence level, time used (speed) in milliseconds, and whether the answer was correct. Furthermore, we recorded whether the answer or confidence changed between the decisions.

3.4.2 Self-reported Measures. The TXAI survey was administered to the participants. The TXAI by Hoffman et al.[14] is an eight-question questionnaire featuring a five-point Likert scale that ranges from 1 ("I strongly disagree") to 5 ("I strongly agree"). As per the

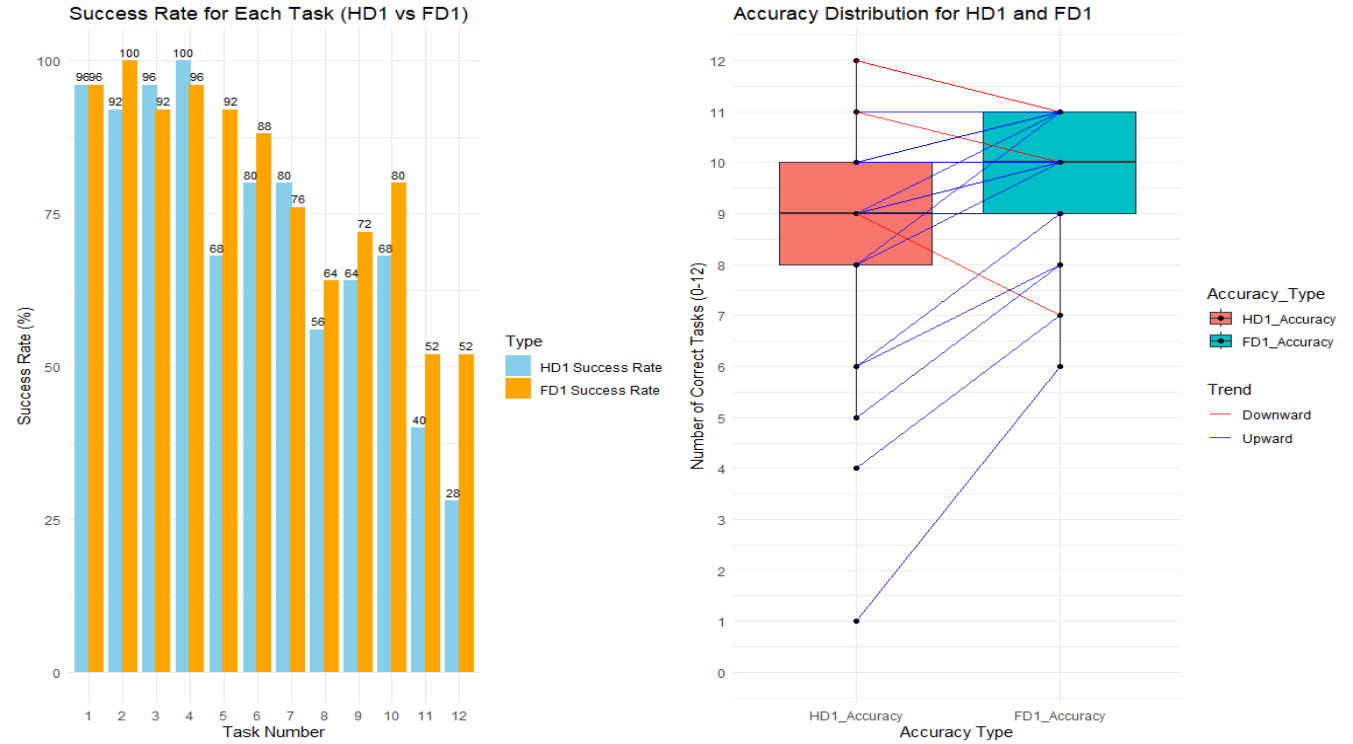


Figure 1: (a - left) Success rate across tasks by HD1/FD1 | (b - right) Trend between accuracy of initial decision versus final decision

findings from Perrig et al., we did not include item six as their findings indicate that the scale works better without it[28]. Aside from removing question six, the scale was administered in its original form.

3.4.3 Open-ended Questions. As the final step, we collected responses to three qualitative questions. The questions were designed by focusing on how the participants perceived the support from the AI, both positively and negatively.

- How did the AI assist you in conducting the task?
- How did the AI hinder you in conducting the task?
- How and why did the AI recommendations impact your confidence?

3.5 Technical Implementation

In the creation of the experiment interface, we utilised several different components to make a smooth experience for the participants.

3.5.1 Technicalities. Our tech stack in the project ended up consisting of Vue-3, Node Express, and PostgreSQL. Vue-3 was responsible for the client-side rendering, providing data to the client from the server, as well as sending the experiment results back to the Node Express server. The server acted as an intermediary in a classic client-server-data structure.

3.5.2 User Interface. We elected for a simple design that made clear distinctions between the different elements of the task. The

main task figure took up around half of the screen on the left-hand side, whereas the AI responses, figure options, and confidence bar shared space on the right-hand side (Figure 5 in the appendix).

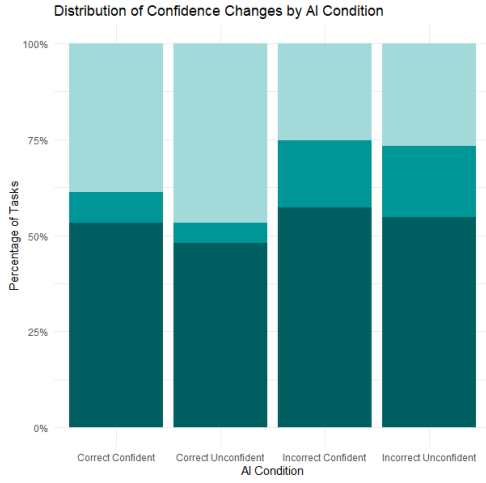
4 RESULTS

In this section, we present our findings from the study, where we recruited 25 participants (17 men and 8 women). Each participant solved the same set of 12 RPM tasks, resulting in data from 300 instances. Out of the 300 instances, the AI conditions are a quarter each overall, totaling 75 instances per condition.

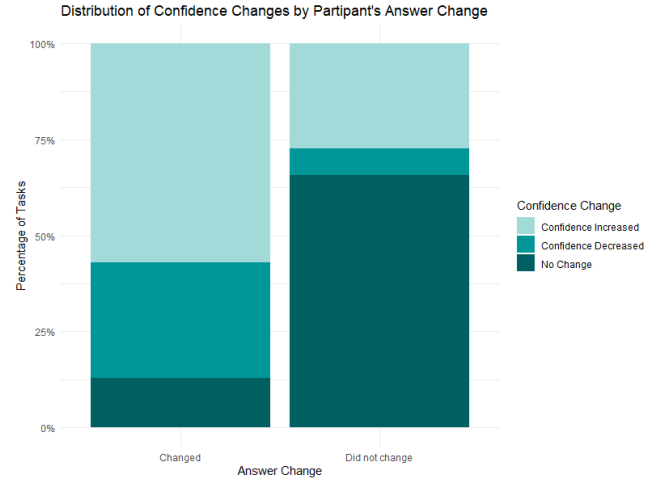
4.1 Impact of AI Assistance On Success Rate

Throughout the study, participants completed a total of 300 tasks combined. We observed a mean of 72.3% success rate ($SD = 22.9$), and for FD1, we found 80% success rate on average ($SD = 17.0$). The overall distribution of the success rate per task can be seen in Figure 1 (a). By using a paired t-test, we have identified a significant difference between the success rate of HD1 and FD1 in FD1's favor ($t = -2.7331$, $p = 0.01947$). This test does not adjust for conditions, i.e., the data includes when the AI is incorrect as well

In a similar fashion, we investigated the overall accuracy of the participant's HD1 and FD1. With an accuracy of 9.6 ($SD = 1.44$), FD1 outperformed HD1 with an average accuracy score of 8.68 ($SD = 2.54$). Calculating the statistical difference between the two accuracy's with a paired-test yielded a significant difference ($t =$

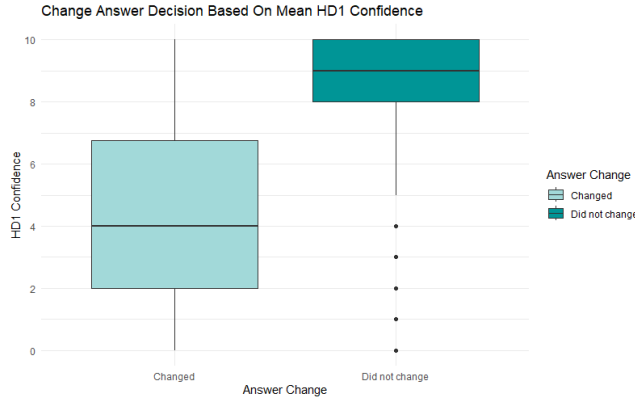


(a) Distribution of confidence changes by AI condition



(b) Distribution of confidence changes by answer changes

Figure 2: Overview of confidence changes



(c) Answer change decision based on HD1 confidence

-4.03, $p = 0.00019$). This shows the upward trend for most less skilled (i.e., with less HD1 accuracy) participants (RQ1). (see Figure 1 - b)

4.2 Changes In Answer & Confidence Post-AI Feedback

4.2.1 Distribution of Changes. Participants changed both the answer and confidence on average 2.44 times ($SD = 2.60$). When only counting the actual answer changes, participant changed their answer an average of 2.8 times ($SD = 2.61$). The chi-square test found that there was no statistically significant correlation among AI conditions and changes made by participants ($\chi^2(9) = 0.00$, $p = 1.0$).

4.2.2 Changes In Confidence And Decisions. The change in confidence is shown in more detail in Figure 2 (a). We split confidence decisions into three parts: increased, decreased, and no change. The conditions differ mostly between the correct and incorrect modes. When the AI is correct, the confidence increases in 42.7% of the instances and only decreases in 6.7% of the instances. Conversely,

the incorrect condition had 26% of instances where the confidence increased and 18% where it decreased (RQ2).

On Figure 2 (b), we show results from looking at the distribution of confidence changes between when participants changed and did not change their answer. It is clear that there are substantial differences when the answer is changed. 57.14% of the time, the confidence increased. The confidence decreased in 30% of the instances and did not change at all in 12.85%. Contrarily, when there was no change in the answer, 65.65% of the instances had no change in confidence at all.

To better understand when changes are being made, we checked for the initial confidence (HD1 Confidence) when a decision was made to either change the answer or keep it. Figure 2 (c) shows that in instances where the answer did not change, the initial confidence was far greater ($M = 8.27$, $SD = 2.43$). Conversely, when participants changed their answer, their initial confidence was significantly lower ($M = 4.41$, $SD = 2.98$).

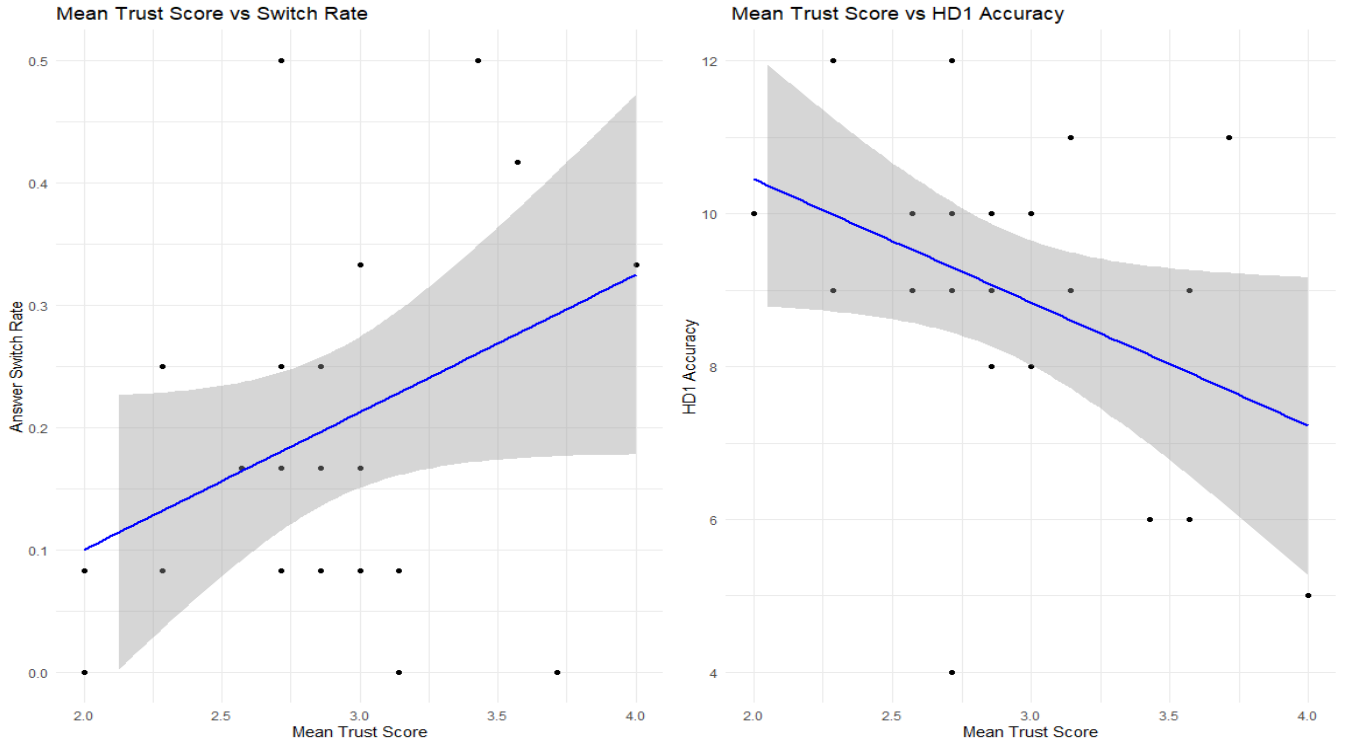


Figure 3: (a - left) Mean Trust Score paired up against switch rate | (b - right) Mean Trust Score paired up against HD1 accuracy

4.3 Impact of Trust

We measured the participants trust in AI via the self-reported TXAI[14] questionnaire; the average trust score was $M = 2.95$ ($SD = 0.98$).

4.3.1 Switch rate. Figure 3 (a) depicts the relationship between our participants mean trust score and how often they switched to the AI suggestion. Furthermore, we calculated the Pearson correlation coefficient (PCC), and the mean trust score and switch rate were found to be positively correlated ($df(23)$, $p = 0.004$) (RQ3).

4.3.2 Accuracy. From the mean trust score versus switch rate (see Figure 3 - b), we observed a trend where a higher switch rate was associated with a higher mean trust score. Given that, we are interested in comparing the mean trust score with HD1 accuracy. The PCC test found a positive correlation (i.e., higher HD1 accuracy is correlated with a higher mean trust score) ($df(23)$, $p = 0.003$) (RQ1, RQ3).

4.4 HD1-AI Alignment

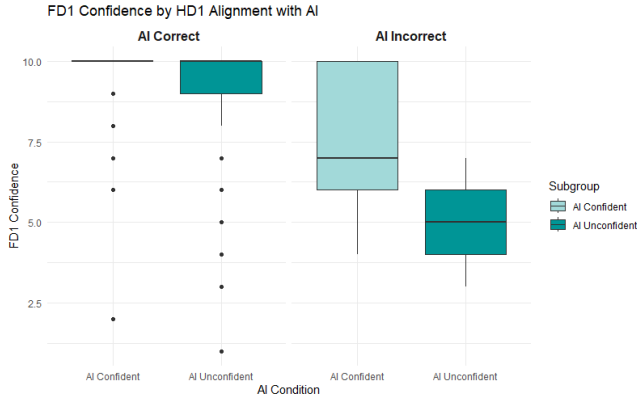
When you compare "Correct-Confident" ($t = 3.31$, $p = 0.005$) and "Correct-Unconfident" ($t = 3.6$, $p = 0.002$) to "Incorrect-Unconfident," you can see that the confidence in FD1 when HD1 (the first answer) aligns with the AI suggestion is significant. Furthermore, there is also a significant difference between "Correct-Confident" and "Incorrect-Confident" ($t = 2.59$, $p = 0.043$). No significant differences were found within each correct type (i.e., "Correct-Confident" versus "Correct-Unconfident").

One of our interests was seeing how users behaved in terms of time (task speed) when the AI feedback was or was not aligned with their expectation (HD1 answer). We divided all of the instances into two parts: When HD1 is aligned with AI and when HD1 is not aligned with AI recommendation. Figure 4 (b) depicts the results: when HD1 is aligned with AI, the mean time taken is 15.2 seconds ($SD = 9.8$ seconds), and when they are not aligned, the mean time taken is 28.4 seconds ($SD = 20.3$ seconds). We found significant differences in FD1 speed between the alignment groups ($p < 0.001$). (RQ3)

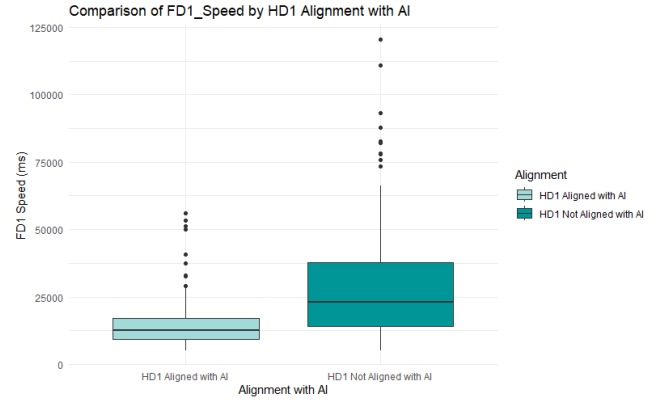
4.5 Qualitative Analysis

As the final step in the procedure, the participants were asked to provide answers to three open-ended questions, aiming to understand how the participants perceived the AI assistance and how the recommendations might have influenced their confidence or decisions. We identified a few common angles and concerns that the participants expressed.

4.5.1 Inspire Me With Patterns. An interesting finding through the qualitative questions was that the participants emphasised that the primary assistance of the AI feedback was identifying patterns they had missed: "I think the AI most of the time just strengthened my answer; in some of the tasks, it gave me a different perspective for the puzzle that also made sense." (P8). Several other participants expressed similar views, e.g., "It helped by providing angles I didn't notice." (P20) and "If I was completely unsure of the answer, the tool would help by describing the patterns to look for. This was very helpful."



(a) FD1 confidence (i.e. final confidence) when HD1 and AI agree



(b) FD1 speed (i.e. time taken when HD1 and AI agree)

Figure 4: Overview of HD1 alignment

(P9). It suggests that some participants preferred to use the AI feedback as inspiration to find the right answer rather than directly looking for the answer. Similarly, literature on explainable AI (XAI) supports this by emphasising that human-AI collaboration is more effective with explanations than working as a black box[1, 2]. As such, we suggest that perhaps imposing an answer is not necessary, it might be more beneficial for the collaboration to give an analysis of the patterns and rule sets that the AI finds, which could inspire human collaborators to make more informed decisions.

4.5.2 Losing Trust. Some participants expressed loss of trust in the AI after experiencing responses they thought were incorrect: "At the start of the test the AI was consistently wrong, to the point where if it produced the same answer as me, I'd double check my answer." (P9). Another participant echoed this sentiment: "I expect the tool to do well and be trained on such a task, so I expect it to be correct. During the first few answers, I decided that the AI recommendations are unreliable." (P14). Losing trust in AI due to errors is consistent with various researches on AI trust[30, 32].

4.5.3 Impact When In Doubt. Consistent with the trend of the quantitative confidence results (e.g., shown by Figure X+3), the impact the AI feedback had on confidence was mostly when the participants were in doubt. "It did not impact my confidence when I was very confident, but when I was not confident, the AI impacted it a little." (P30).

5 DISCUSSION

In this study, we investigated how one-shot interaction with AI assistance impacts people's answers and confidence under various conditions in a visual task. One of our primary results is that the AI assistance delivered an overall upward trend (i.e., final accuracy was higher than initial accuracy), especially for lower-skilled participants, for the overall accuracy score of the test (RQ1). For some of the more skilled participants, there was a downward trend (i.e., initial accuracy was higher than final accuracy). A plausible explanation for this might be that the AI feedback confused them, and they deferred to the AI recommendation, whereas the upward trend deferred due to a lack of confidence. Another possible angle

is through social comparisons. Michinov and colleagues found that comparing your ideas to partner's you perceive as superior in the task has benefits for your performance[24].

The success rate of the specific tasks started out relatively high, with a 90–100% success rate for the initial accuracy but predictably descending for the more complex tasks. Interestingly, there are cases where the HD1 success rate outperforms the FD1 success rate (e.g., task 3), likely due to the incorrect AI being more likely to confuse than for the correct AI to be of benefit in a relatively easy task.

The distribution of the confidence changes (see Figure 2 - a) turned out to be somewhat similar across the conditions, with the group types being similar to each other, only differing significantly on correct/incorrect. This is yet another indicator that the differences between confident and unconfident are not great enough to matter. However, as expected, we observed a higher rate of decreased confidence in the incorrect conditions and a lower amount of increased confidence.

When we look at the distribution of the confidence changes when the participant decided to switch answers, there is quite a contrast between the confidence changes in each case. When the participants changed, we saw that, more than half of the time, the confidence increased. In fact, a very small percentage of answer switches had no change. This is consistent with our qualitative analysis in Section 4.5.3, as participants were more likely to switch, mentioning that their confidence had mostly an effect when in doubt. Furthermore, this is also consistent with findings from existing literature[7].

Lower trust in AI correlated with higher initial accuracy, perhaps slightly surprisingly, though it can be explained similarly to experts having lower trust in AI[16, 37]. Higher proficiency participants may have higher confidence in themselves, and therefore be less confident in the AI as soon as it makes a single mistake.

If the initial answer (HD1) aligned with the AI suggestion (see Figure 4 - a), we observed contrasting confidence differences between the correct and incorrect groups, but less so between the confident and unconfident sub-groups. We observe a far more diversified confidence when the AI answer is incorrect, suggesting that the participants were in fact cast in doubt by the incorrect AI.

Under the same circumstances, we also saw changes in time-taking patterns when there was an alignment. After getting their answers verified by the AI, people might feel less inclined to double-check. All responses are about 40–60 words at most; for an average adult to read (around 240 WPM on average), it would take around 10–14 seconds, which seems consistent with the time taken on HD1 and AI alignment.

5.1 Limitations and Future Work

We acknowledge a number of limitations to our work that need to be considered when interpreting our findings. First, our analysis was done by looking at a single task, however, problem-solving comes in many shapes and forms. Second, at his current time, GPT-4 does not possess the capabilities to solve RPM's reliably, much less the more complex ones. Due to that, the prompt given to the model ranged from a simple instruction to a detailed hint, depending on the complexity of the problem. This may not accurately reflect the future prospects of GPT or similar models. Third, it might have been useful to screen participants for their visual prowess to ensure a wider spread of skill levels.

Trying to communicate confidence through the tone of the message proved difficult. Future work could try to improve that method by using recommendations from Kim et al, where they studied different versions of uncertain responses and found some of them to be effective[18]. Additionally, it could entail switching away from trying to convey confidence through the text and rather trying to give AI confidence scores in its solution. As of now, a generative AI such as GPT-4 would likely not be sufficient by itself; it would need to be assisted by a model designed to complete these kinds of visual puzzles. Additionally, it could be interesting to conduct studies to find comparable upward and downward patterns in skills other than problem-solving, like creativity.[24].

6 CONCLUSION

In this paper, we present the results of a within-subjects study that evaluates the effects of intermediary AI feedback on problem-solving complex visual tasks. We investigated the impact by manipulating the AI feedback to give correct or incorrect answers, as well as confident or unconfident answers. We found that (1) the AI feedback was beneficial to lower and middle-skilled participants, providing an upward trend when compared to their initial decision (HD1) to the final decision (FD1). (2) Participants with lower skills correlated with higher trust, and higher trust correlated with switching answers. (3) Wrong answers by the AI decrease self-confidence in the answer, and participants lost faith in the AI's reliability when encountering them. Based on our findings, we argue that AI feedback can still be beneficial for people of lower skill, even with some inaccuracies. Finally, we further suggest that it may be enough for a decision-support agent to reveal patterns and rule sets instead of imposing an answer.

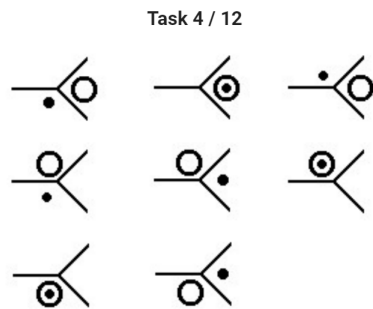
ACKNOWLEDGMENTS


I would like to thank my advisor, Niels for his patience and help over the past two semesters. I'd also like to thank Rune and Sander for using their own time for my benefit; you guys are all awesome!

REFERENCES

- [1] Anastasia Angelopoulou, Epaminondas Kapetanios, David Harris Smith, Volker Steuber, Bencie Woll, and Frauke Zeller. 2022. Editorial: Explanation in human-AI systems. *Frontiers in Artificial Intelligence* 5 (2022). <https://doi.org/10.3389/frai.2022.1048568>
- [2] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579. <https://doi.org/10.3390/make4020026>
- [3] S K R Anumandla. 2018. AI-enabled Decision Support Systems and Reciprocal Symmetry: Empowering Managers for Better Business Outcomes. *International Journal of Reciprocal Symmetry and Theoretical Physics* 5 (2018), 33–41.
- [4] Douglas A. Bors and Tonya L. Stokes. 1998. Raven's Advanced Progressive Matrices: Norms for First-Year University Students and the Development of a Short Form. *Educational and Psychological Measurement* 58, 3 (1998), 382–398. <https://doi.org/10.1177/0013164498058003002> arXiv:<https://doi.org/10.1177/0013164498058003002>
- [5] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 41 (apr 2024), 32 pages. <https://doi.org/10.1145/3637318>
- [6] Patricia A Carpenter, Marcel A Just, and Peter Shell. 1990. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychol. Rev.* 97, 3 (1990), 404–431.
- [7] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (oct 2023), 32 pages. <https://doi.org/10.1145/3610219>
- [8] David "davidad" Dalrymple, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, Christian Szegedy, Ben Goldhaber, Nora Ammann, Alessandro Abate, Joe Halpern, Clark Barrett, Ding Zhao, Tan Zhi-Xuan, Jeannette Wing, and Joshua Tenenbaum. 2024. Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. arXiv:2405.06624
- [9] Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Proceedings of the 14th Conference on Creativity and Cognition* (Venice, Italy) (CC '22). Association for Computing Machinery, New York, NY, USA, 623–627. <https://doi.org/10.1145/3527927.3535197>
- [10] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 1 (Feb. 2015), 114–126.
- [11] Ronald Hamel and Verena D. Schmittmann. 2006. The 20-Minute Version as a Predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement* 66, 6 (2006), 1039–1046. <https://doi.org/10.1177/0013164406288169> arXiv:<https://doi.org/10.1177/0013164406288169>
- [12] Yugo Hayashi and Kosuke Wakabayashi. 2017. Can AI become Reliable Source to Support Human Decision Making in a Court Scene?. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17 Companion). Association for Computing Machinery, New York, NY, USA, 195–198. <https://doi.org/10.1145/3022198.3026338>
- [13] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. <https://doi.org/10.1145/3544548.3581025>
- [14] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023). <https://doi.org/10.3389/fcomp.2023.1096257>
- [15] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans. Comput. Healthcare* 1, 1, Article 6 (mar 2020), 20 pages. <https://doi.org/10.1145/3344258>
- [16] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. (2020).
- [17] S Kaggwa, T F Eleogu, F Okonkwo, O A Farayola, P U Uwaoma, and A Akinoso. 2024. AI in decision making: transforming business strategies. *International Journal of Research and Scientific Innovation* 10, 12 (2024), 423–444.
- [18] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (, Rio de Janeiro, Brazil,) (FAccT '24). Association for Computing Machinery, New York, NY, USA, 822–835. <https://doi.org/10.1145/3630106.3658941>









- [19] Sandra Leaton Gray. 2020. Artificial intelligence in schools: Towards a democratic future. *Lond. Rev. Educ.* 18, 2 (2020).
- [20] Anastasia Lebedeva, Jaroslaw Kornowicz, Olesja Lammert, and Jörg Papenkordt. 2023. The Role of Response Time for Algorithm Aversion in Fast and Slow Thinking Tasks. In *Artificial Intelligence in HCI*, Helmut Degen and Stavroula Ntoa (Eds.). Springer Nature Switzerland, Cham, 131–149.
- [21] Xiaoshuang Liu, Mohammad Faisal, and Abdullah Alharbi. 2022. A decision support system for assessing the role of the 5G network and AI in situational teaching research in higher education. *Soft Computing* 26, 20 (01 Oct 2022), 10741–10752. <https://doi.org/10.1007/s00500-022-06957-5>
- [22] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- [23] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does More Advice Help? The Effects of Second Opinions in AI-Assisted Decision Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 217 (apr 2024), 31 pages. <https://doi.org/10.1145/3653708>
- [24] Nicolas Michinov, Eric Jamet, Natacha Métayer, and Benjamin Le Hénaff. 2015. The eyes of creativity: Impact of social comparison and individual creativity on performance and attention to others' ideas during electronic brainstorming. *Computers in Human Behavior* 42 (2015), 57–67. <https://doi.org/10.1016/j.chb.2014.04.037> Digital Creativity: New Frontier for Research and Practice.
- [25] Lama H. Nazer, Razan Zatarah, Shai Waldrip, Janny Xue Chen Ke, Mira Moukheiber, Ashish K. Khanna, Rachel S. Hicklen, Lama Moukheiber, Dana Moukheiber, Haobo Ma, and Piyush Mathur. 2023. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health* 2, 6 (06 2023), 1–14. <https://doi.org/10.1371/journal.pdig.0000278>
- [26] Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. 2023. Co-design of Human-centered, Explainable AI for Clinical Decision Support. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 21 (dec 2023), 35 pages. <https://doi.org/10.1145/3587271>
- [27] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (nov 2019), 15 pages. <https://doi.org/10.1145/3359204>
- [28] Sebastian A. C. Perrig, Nicolas Scharowski, and Florian Brühlmann. 2023. Trust Issues with Trust Scales: Examining the Psychometric Quality of Trust Measures in the Context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (, Hamburg, Germany,) (CHI EA '23). Association for Computing Machinery, New York, NY, USA, Article 297, 7 pages. <https://doi.org/10.1145/3544549.3585808>
- [29] John C Raven and JH Court. 1938. *Raven's progressive matrices*. Western Psychological Services Los Angeles, CA.
- [30] Beau G. Schelble, Jeremy Lopez, Claire Textor, Rui Zhang, Nathan J. McNeese, Richard Pak, and Guo Freeman. 2024. Towards Ethical AI: Empirically Investigating Dimensions of AI Ethics, Trust Repair, and Performance in Human-AI Teaming. *Human Factors* 66, 4 (2024), 1037–1055. <https://doi.org/10.1177/0018720822116952> arXiv:<https://doi.org/10.1177/0018720822116952> PMID: 35938319.
- [31] Oliver Schmitt and Daniel Buschek. 2021. CharacterChat: Supporting the Creation of Fictional Characters through Conversation and Progressive Manifestation with a Chatbot. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) (C&C '21). Association for Computing Machinery, New York, NY, USA, Article 10, 10 pages. <https://doi.org/10.1145/3450741.3465253>
- [32] Jeff C. Stanley and Stephen L. Dorton. 2023. Exploring Trust With the AI Incident Database. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67, 1 (2023), 489–494. <https://doi.org/10.1177/21695067231198084> arXiv:<https://doi.org/10.1177/21695067231198084>
- [33] Timo Sturm, Jin Gerlach, Luisa Pumplun, Neda Mesbah, Felix Peters, Christoph Tauchert, Ning Nan, and Peter Buxmann. 2021. Coordinating Human and Machine Learning for Effective Organizational Learning. 45 (09 2021), 1581–1602. <https://doi.org/10.25300/MISQ/2021/16543>
- [34] John Eustis Williams and David M. McCord. 2006. Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. *Computers in Human Behavior* 22, 5 (2006), 791–800. <https://doi.org/10.1016/j.chb.2004.03.005>
- [35] JR. Winfred Arthur and David V. Day. 1994. Development of a Short form for the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement* 54, 2 (1994), 394–403. <https://doi.org/10.1177/0013164494054002013> arXiv:<https://doi.org/10.1177/0013164494054002013>
- [36] Simona Wójcik, Anna Rulkiewicz, Piotr Pruszczyk, Wojciech Lisik, Marcin Poboży, and Justyna Domienik-Karłowicz. 2023. Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era. *Cardiology Journal* 30, 6 (2023), 1018 – 1025. <https://doi.org/{}>
- [37] Lingrui Xu, Zachary A. Pardos, and Anirudh Pai. 2023. Convincing the Expert: Reducing Algorithm Aversion in Administrative Higher Education Decision-making. In *Proceedings of the Tenth ACM Conference on Learning @ Scale* (Copenhagen, Denmark) (L@S '23). Association for Computing Machinery, New York, NY, USA,






I'm not entirely sure, but option 1 might fit if we consider that the black dot rotates anti-clockwise, and the empty circle only shifts when moving to a new row. Its final position on the stem seems plausible, assuming this pattern holds consistently.

? You can now reconsider or confirm your **original answer**. Please provide both your answer and the confidence you have in this answer.

1 	2 	3 	4 
5 	6 	7 	8 

How confident are you in your answer?

Not confident at all  Fully confident

ANSWER

Figure 5: Web interface

A WEB INTERFACE