Examining Interactions Between User Characteristics and Explanation Modalities on Inducing Complementarity

Tory Farmer
toryfarmer@wustl.edu
Washington University in St. Louis
St. Louis, Missouri, USA

Chien-Ju Ho
chienju.ho@wustl.edu
Washington University in St. Louis
St. Louis, Missouri, USA

Abstract

Achieving complementary performance in human–AI collaboration, where the combined efforts of humans and AI outperform either working alone, remains a significant challenge. Providing explanations for AI assistance is often considered a potential strategy to reduce human over-reliance on AI and enhance decision-making. However, empirical studies have shown mixed results regarding the impact of AI explanations on performance improvement. In this work, we extend this investigation by exploring an additional dimension: whether user characteristics influence the effectiveness of AI explanations in achieving complementarity. Using a geography-guessing task as the experimental setting, we find that user characteristics, such as openness and experience, interact with explanation modality in inducing complementarity. Our results suggest that tailoring explanations based on user characteristics could enhance complementarity and provide insights into how personalized AI explanations can improve human–AI team performance.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Explainable AI, Explanation Design, Personality, Personalization

ACM Reference Format:
Tory Farmer and Chien-Ju Ho. 2025. Examining Interactions Between User Characteristics and Explanation Modalities on Inducing Complementarity. In . ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3706599.3719998

1 Introduction

The integration of AI into decision-making processes has opened new opportunities to enhance human performance across various domains, including surgical operations [24], medical diagnoses [34], and human-robot team performance [14]. However, achieving complementarity, where the combined efforts of humans and AI exceed the capabilities of either working alone, remains a significant challenge in human-AI collaboration. Despite notable advancements in AI, there is growing recognition that simply providing AI assistance to humans does not automatically improve the joint performance of human-AI decision-making. For example, empirical evidence indicates that users often over-rely on AI assistance, which can diminish their ability to perform effectively as a team [2, 3, 32, 36].

Providing explanations alongside AI assistance has been considered a potential solution for enhancing complementarity [21]. However, empirical studies on AI explanations have yielded inconsistent results, sometimes exacerbating issues of over-reliance [1, 32]. In this work, we extend this line of research by examining whether user characteristics moderate the effectiveness of AI explanations in human-AI collaboration. To do so, our first goal is to identify a task where AI explanations contribute to complementarity. Once such a task is identified, we aim to explore ways to further improve complementarity. Motivated by research on human trust in AI [4, 15], where recent studies have shown that personalized explanations based on user traits can significantly improve trust [4, 5, 17], we investigate whether user characteristics influence the extent to which AI explanations enhance complementarity. If so, this could open opportunities for future research on personalizing explanations to strengthen human-AI collaboration.

To address the above questions, we developed a task adapted from the popular game Geoguessr[1], where participants are asked to guess which continent a given photo was taken in. We designed two types of explanations: text-based explanations, where participants receive an AI recommendation accompanied by a one-line textual explanation, and visual-based explanations, where participants receive AI recommendations with highlighted areas in the photo relevant to the recommendation. To examine the interaction between user characteristics and modalities of AI explanations, we conducted an experiment involving 400 participants. In the experiment, we first surveyed participants to collect their personal characteristics along three dimensions: openness, need for cognition, and experience with travel. Participants were then randomly assigned to one of four groups: control (no AI assistance), unexplained AI, AI with text-based explanations, and AI with visual-based explanations. We assessed how different explanation formats interact with user characteristics to affect performance. We found statistically significant interaction effects between openness and explanation modality on performance. Specifically, participants with higher/lower openness performed better with text-based/visual-based explanations. We also found an interaction effect between travel experience and explanation modality on performance. Participants with more/less travel experience performed better with visual-based/text-based explanations. These findings contribute to the growing body of work in explainable AI by suggesting that personalized explanations can indeed improve complementary performance.

[1] https://www.geoguessr.com/

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3719998

2 Related Work

Complementarity. We focus on achieving *complementarity* [15] in AI-assisted decision-making [25, 30, 33], meaning the situations when the AI-advised human decision-maker outperforms either the AI or the human individually. There has been prior work on establishing mechanisms and practices to enhance the consistency of achieving complementarity [15]. For example, environments characterized by information asymmetry, as well as those where actors differ in their ability to process and utilize specific pieces of information, have generally been more conducive to achieving complementarity [12]. There has also been research focusing on finding complementarity in AI-advised decision-making, with applications ranging from medical diagnoses [27] to predicting criminal recidivism [31]. However, merely providing humans with AI recommendations often fails to achieve complementarity and can sometimes result in worse performance [3]. This shortfall is partially due to individuals overly trusting their AI partners, often without appropriately evaluating whether the AI is offering valuable advice—a phenomenon referred to as over-reliance. Complementarity and trust are inherently connected, namely through appropriate reliance [13, 29]. To achieve complementarity, the human needs to trust the AI when it is correct, but trust themselves when the AI is wrong. This can be accomplished through users better understanding their own capabilities [22] or information asymmetry [12].

The Role of Explanation. To mitigate the negative effects of over-reliance, one approach is to provide users with additional information beyond predictions. For example, confidence meters or scores can help users determine when to trust their AI partners [13]. Stating the AI's overall accuracy can also enable users to calibrate their expectations of the AI partner [11]. An increasingly common addition to AI predictions is an explanation, broadly defined as a piece of information appended to an AI's prediction that conveys aspects of the AI's "reasoning." These explanations can take various forms, including similar data points, feature weights, or verbal descriptions, as seen with large language models (LLMs) like ChatGPT [20]. Explanations have been extensively studied for their impact on users' reception of AI recommendations, particularly in areas such as trust [19], understanding [16], and the mental effort expended [18]. The literature suggests that users actively seek information from explanations when deciding whether to adopt AI recommendations [15], and they generally prefer to receive and use as much information from AI as possible. However, explanations have also been criticized for their inconsistency in improving performance. In many cases, explanations exacerbate over-reliance, exploiting cognitive laziness or misplaced trust [1]. Much of the work in explainable AI has focused on metrics outside of complementarity, such as user trust. To address this underexplored area, recent research has proposed the concept of verifiability to provide insights into whether explanations can enhance complementarity in a given task. Specifically, the literature [7, 32] suggests that verification is the key to achieving complementarity through explanations. Explanations should enable users to verify AI outputs with greater accuracy and less effort than solving the task independently or blindly relying on the AI.

The Impact of Personalization. Prior literature highlights several benefits of personalizing explainable AI, particularly in improvements to non-complementarity metrics. Explanations tailored to user characteristics can enhance user trust [5], improve user understanding of the underlying AI mechanisms [16], and increase overall user satisfaction with the AI assistant [15]. These outcomes have been observed across various personalization methods, including adjusting the explanation's length, mode of presentation [9], tone of voice, or the material presented [23]. The literature also suggests a range of potential *axes* of personalization—metrics or characteristics along which personalization can be conducted. Factors such as Need for Cognition, Big Five Personality traits, task experience, and demographic traits have all shown significant interactions with stated trust, reliance, or perceived understanding of the AI's reasoning [4, 6, 26], as well as the ways individuals process information [8]. We suspect that these traits may also affect how users develop appropriate trust and reliance [13, 29] on AI-assisted decision-making, which could lead to complementarity. This line of literature leads us to the following research question: Interaction effects have been observed between the above axes of personalization and various user characteristics on user trust and satisfaction with the AI. Can we identify axes of personalization and user characteristics that have interaction effects on complementarity? In other words, can we leverage recently developed theories of achieving complementarity to create a personalized explainable AI that improves complementary performance?

3 Experiment Design

The experiment in this work were approved by the Institutional Review Board (IRB) at our institution and pre-registered on Open Science Framework (OSF)[2].

3.1 Research Questions

Before describing our experiment design, we first state the two research questions we aim to investigate:

- **RQ1**: Can we identify a task where we can achieve complementary performance with the use of explainable AI?
- **RQ2**: Do people with different personal characteristics perform differently with different explanations on this task?

While complementarity is a desired property for human-AI teaming, prior work often suggests that achieving complementarity is challenging for many common tasks [7]. Therefore, the goal of RQ1 is to first identify a task where complementary performance can be achieved with explanations. This enables us to address RQ2 and also provides a candidate task for future research on complementarity. In RQ2, our goal is to investigate whether individual differences influence the effectiveness of different types of explanations. If they do, this opens up the possibility of designing personalized explanations to enhance human-AI collaboration. To make our investigation more concrete, we focus on three personal characteristics: openness, relevant experience in the task, and need for cognition, and two types of explanations: visual-based and text-based (discussed in detail later). Our choice of personal characteristics and explanation

[2]https://osf.io/a6t9r/?view_only=323483088c834b02be51c4278b16dfd6

types are justified in Section 3. Specifically, we have the following three hypotheses:

- **H2A**: Participants who are more *open* perform better in AI-assisted decision making with text-based explanations than with visual-based explanations.
- **H2B**: Participants with more relevant experience in the task perform better in AI-assisted decision making with visual-based explanations than with text-based explanations.
- **H2C**: Participants with a higher *need for cognition (NFC)* perform better in AI-assisted decision making with visual-based explanations than with text-based explanations.

3.2 Experiment Task: Guessing the Continent Where a Photo Was Taken

In selecting the experiment task, we have two criteria in mind. First, we aim to design a task that ensures there exist both conditions where humans outperform AI and conditions where AI outperforms humans. To achieve this, we focus on tasks that AI typically performs better than humans and then limit the information available to AI to mimic situations where humans have additional information that AI does not have access to. Second, inspired by the work of Fok and Weld [7], which emphasizes the importance of *verifiability* in explanations for enabling complementarity, we seek tasks where AI explanations can assist humans in identifying the ground truth.

Task Description. In this work, we developed a variant of a popular geography-guessing task [10, 35]. In each round of this task, participants were presented with a photo sourced from Google Earth and asked to predict the continent where the photo was taken: North America, South America, Africa, Europe, or Asia. The data for this task was drawn randomly from selected rounds of the world map in the popular geography-guessing game "Geoguessr." This task requires minimal training and is relatively intuitive, making it suitable for administration to the general public. With respect to the two criteria above, our initial explorations demonstrate that AI achieves significantly better performance than regular human users. Additionally, we can limit the information provided to the AI (e.g., by showing the AI a partial image while providing humans with the full image). Furthermore, when the AI provides explanations (e.g., identifying the language on a sign), humans can often leverage these explanations to improve their performance. We believe that clear relative strengths of humans and AI, the ability to generate modally distinct explanations, and relatively minimal barrier to entry make this task a justifiable and well-suited option for assessing our hypotheses above. The task interfaces are shown in Figure 1 and Figure 2.

ChatGPT as AI. The AI used in our experiment is powered by ChatGPT 4o. In particular, we provide the AI a portion of the image (indicated by a red box visible to the participant, as shown in Figure 1 and Figure 2) and ask the AI to predict the continent of origin for the image. This restriction on the AI's field of view is implemented to create relative strengths for the participant and the AI. In our initial explorations, ChatGPT is generally much more accurate than average users about geography and can use visual clues more effectively. As a result, to maximize team performance, the user will have to weigh their knowledge of what ChatGPT cannot see against ChatGPT's superior analysis of what it can see. These distinct relative strengths have been shown to positively impact complementarity [12]. For many participants, the area in the image outside of the box enables relatively straightforward verification of the AI's predictions, which, as the literature suggests [7], is more likely to yield complementary performance.

Explanations. We develop two explanation types: **Text-based** explanations contain written description generated by ChatGPT in supporting its predictions (as shown in Figure 1), and **Visual-based** explanations contain blue circles that highlights the regions in the image that ChatGPT indicates that are the most relevant (as shown in Figure 2).

Procedure of Generating AI Predictions and Explanations. We use ChatGPT 4o to generate predictions and explanations. After the image of each round was selected, ChatGPT is given a randomly selected portion of the image and the following prompt: "What continent do you believe this photo was taken in? Bear in mind that it cannot be Australia / Oceania or Antarctica. Provide a very brief 1-line justification of your answer." This sentence would be provided to the participant as the "text-based" explanation. Afterwards, we provide ChatGPT with the following prompt: "List in bullet point format 1-3 features of the image you explicitly mentioned in your previous response that I can draw circles around. Provide detailed instructions for each bullet point detailing what I exactly I should circle in the image, bearing in mind that I can only draw 1 circle per bullet point." This allows us to add in circles around features of the image deemed important by ChatGPT, which define the "visual-based" explanations.

3.3 Participant Grouping

Our research examines whether participants with different personal characteristics perform differently with different explanation types.

User Characteristics. In our experiment, we measure three personal characteristics — the openness personality trait, need for cognition (NFC), and travel experience (a proxy for experience in the task). When selecting these characteristics, our goal was to identify a small set of characteristics that had interaction effects in the literature with trust in AI [4–6, 26]. More specifically, need for cognition has been associated with tending to pay more or less attention to AI-generated explanations, which in turn impacts trust [4]. Additionally, literature suggests that information that convinces experts to trust AI is different from what convinces novices [6], suggesting that task expertise should be considered. Finally, based on Nimmo et al. [26], Openness was the most closely associated with differing user trust and performance when utilizing explainable AI among Big 5 personality trait [28].

We measure the personal characteristics using 5-point Likert scales, with "Strongly Disagree" receiving a score of 1 and "Strongly Agree" given as score of 5. For openness and need for cognition (NFC), we pulled questions from the BFI-10 personality inventory, a 10-question abbreviated personality quiz commonly used in the literature [28]. The BFI-10 has been shown to be a reliable, valid, and well-used approximation of the full-length Big 5 personality quiz, which is a literature standard for measuring personality traits despite its self-reported nature. For travel experience, we could not determine a standard approach to word questions. As a result,

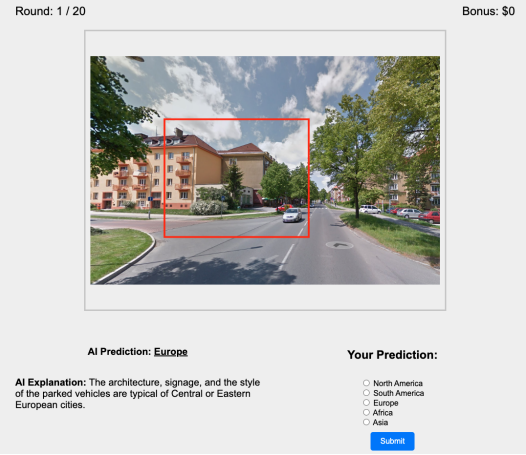

Figure 1: A sample round from the geography guessing task with text-based explanations. The red box indicates the area seen by the AI in making its prediction. In this case, the AI provides its explanation through a sentence, as textual description on the bottom left corner.

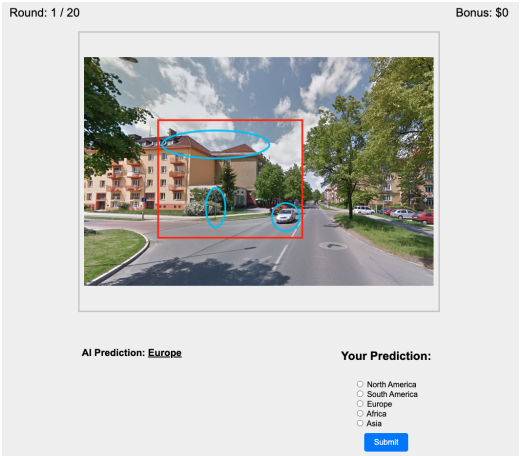

Figure 2: A sample round from the geography guessing task with visual-based explanations. The red box indicates the area seen by the AI in making its prediction. In this case, the AI provides its explanation through drawing blue circles on the relevant parts of the image.

we simply asked participants whether they agree or disagree with the following statement: "I consider myself well-traveled and/or familiar with geography outside of the United States."

Experiment Conditions. When participants arrive, they were randomly assigned to one of the four treatments:

- No AI (Control Group): Participants are not given any AI assistance when completing tasks.
- Unexplained AI: AI recommendations are offered without explanation when completing tasks.
- Text-based explanations: AI recommendations are offered with the accompany of text-based explanations.
- Visual-based explanations: AI recommendations are offered with the accompany of visual-based explanations.

3.4 Experiment Procedure

We recruited 400 participants from Prolific, restricting the study to U.S. workers, with each participant allowed to participate only once. Before conducting the main experiment, we performed a power analysis based on results from a pilot study. We determined that a sample size of $N = 98$ participants per group would be required to achieve a power of .80 at a significance criterion $\alpha = .05$ for the least statistically significant interaction effect observed in the pilot (Need for Cognition). We rounded this number up and recruited 100 participants per group. When participants arrive, they first completed the informed consent form, then were briefed on the structure of the experiment, rules, and payment structure. After the briefing, participants were given a 5-question comprehension check to ensure they understood the task. Participants who answered every question correctly were allowed to proceed; those with at least one incorrect answer were re-directed to the beginning of the briefing. Participants were provided as many attempts as desired to pass the comprehension check.

After passing the check, participants were given a short questionnaire about their Openness personality trait, travel experience, and need for cognition. Once participants completed the survey, they were given 20 rounds of the geography guessing task. To ensure consistency in task difficulty between participants, each participant was given the same 20 rounds of the task with random presented sequence. In selecting these 20 rounds, we included more challenging rounds where the AI assistant was correct only 70% of the time. Each participant was required to spend a minimum of 10 seconds per round before advancing to the next round to ensure quality responses. At the conclusion of the 20 rounds, participants were paid a flat $1.70 for completing the task, plus $0.05 per correct answer. This resulted in a maximum total payment of $2.70. Each participant is only allowed to participate the experiment once without exposing to multiple experiment conditions.

4 Experiment Results

4.1 Analysis Method

For RQ1, we first determine the average performance of the control group on the task, then compare said performance to the AI's (which is 70%). The maximum of the two performance values is the threshold for complementarity. We then compare the average performance of participants in each treatment group to the threshold using 1-sample, 2-sided t-tests. We say that our explanations achieve complementarity if at least one of the t-tests returns a statistically significant result ($p < 0.05$) after corrections for multiple comparisons. To assess RQ2, for each participant characteristic (openness, need for cognition, travel experience), we conduct a multiple linear regression using explanation type, the relevant participant characteristic, an interaction term, and a constant. We then accept or reject H2A, H2B, and H2C by assessing whether or not the interaction term in each corresponding regression is statistically significant at the .05 level.

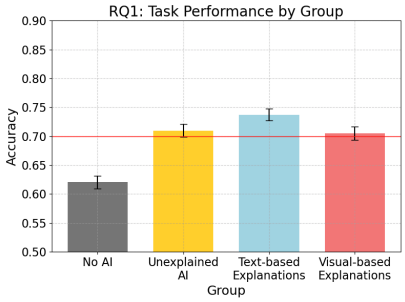

Figure 3: Results by explanation format group without differentiating based on user characteristics. The group receiving text-based explanations significantly outperformed the other groups and definitively achieved complementarity.

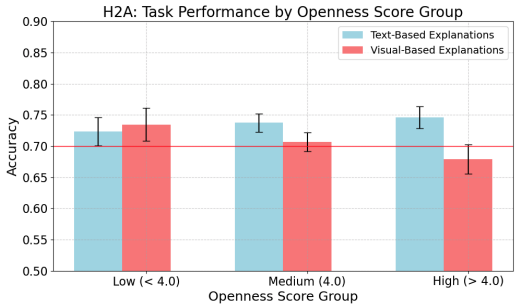

Figure 4: Performance by openness level and explanation format. Individuals with higher openness scores performed significantly better using text-based explanations compared to visual-based ones.

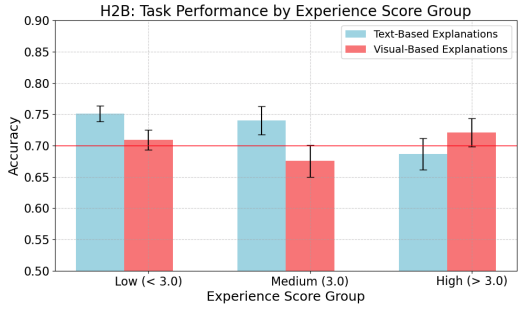

Figure 5: Performance by experience level and explanation format. Individuals with more travel experience were able to significantly better utilize visual-based explanations than their less-traveled counterparts.

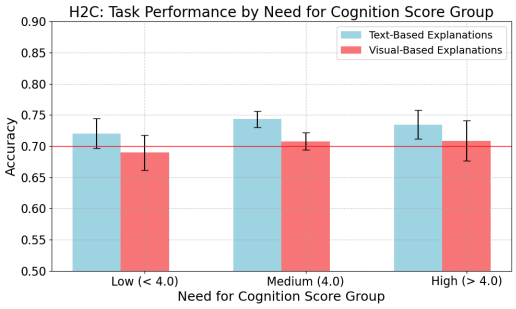

Figure 6: Performance by Need for Cognition (NFC) level and explanation format. We found no significant interaction between NFC and performance using the two explanation types.

4.2 RQ1: Identifying a Task that Enables Complementarity

The average performance of each group is shown in Figure 3. Overall, we see that participants without any AI assistance achieve an average accuracy of 62.1%, while the AI achieves a performance of 70.0%. As a result, the baseline for complementarity is 70.0%, indicated by a red line in the figure. Among the three groups receiving AI assistance, participants receiving unexplained AI assistance achieved 71.0% accuracy, participants given visual-based explanations achieved 70.5% accuracy, and participants given text-based explanations achieved 73.7% accuracy. After correcting for multiple comparisons, there was no statistically significant improvement over the complementarity baseline for the visual-based explanations ($p = 0.6802$), but the improvement for text-based explanations is statistically significant ($p < .0005$). This indicates that at least one explanation format achieves complementarity.

4.3 RQ2: Examining Interactions between User Characteristics and Explanation Modalities

We now describe the results for the three hypotheses. Again, we conduct our analysis by examining the statistical significance of the interaction term in the associated multiple linear regressions.

To more effectively visualize our data, we group participants into low, medium, and high openness based on score tercile and plot their performance with text-based or visual-based explanations.

H2A: Participants who are more *open* are more effective with text-based explanations than with visual-based explanations. As evidenced in Figure 4, participants with high openness perform better with text-based explanations, whereas participants with low openness perform slightly better with visual-based explanations. This observation is confirmed by a linear regression analysis, which reveals a significant interaction effect between openness and explanation type on performance ($p = .031$).

H2B: Participants with more relevant experience in the task are more effective with visual-based explanations than with text-based explanations. As visualized in Figure 5, well-traveled individuals tended to perform relatively better than their peers using visual-based explanations. However, our test does not reveal statistical significant effect for the interaction terms between experience and explanation types ($p = 0.065$).

Exploratory Analysis: Upon investigation, we found that individuals with high experience exhibited substantially different behavior than those with low experience. This is due, in part, to relatively few people (roughly 23% of the overall population) reporting that they are "well-traveled" (score of 4 or 5) If we choose a threshold and

treat the experience as a boolean variable: 1 if a person agreed with the notion that they were well traveled (score of 4 or 5), 0 if they disagreed (scores 1 - 3), we find a statistically significant interaction effect ($p = 0.005$). However, we find very little variance between scores 1 through 3 and between scores 4 and 5, both in our main and pilot study. While this transformation was *not* pre-registered, we believe these results suggest that some sort of interaction effect between experience and explanation modality on may well exist for a sufficiently normalized population.

H2C: Participants with a higher *need for cognition (NFC)* are more effective with visual-based explanations than with text-based explanations. As shown in Figure 6, we found that no significant interaction between NFC score and explanation type on complementary performance. As one would expect, the interaction term in our multiple linear regression was insignificant ($p = 0.307$). We conclude that previous effects seen concerning need for cognition in our pilot study are likely attributable to random chance.

5 Conclusion and Discussion

In this work, we first identify a task (geography guessing task) in which complementary performance can be achieved through explanations (**RQ1**). This task enables us to study RQ2 and also serves a candidate task for future studies on complementarity. Next, we demonstrate that participants with different personal characteristics respond differently to different types of explanations (**RQ2**). Specifically, participants with higher openness perform better with text-based explanations, while those with lower openness perform better with visual-based explanations. Similarly, participants with more travel experience perform better with visual-based explanations, whereas those with less travel experience perform better with text-based explanations. Our findings highlight the potential to enhance human-AI team performance by providing personalized explanations that tailored towards participants with different personal characteristics.

Limitations. Our results show that participants with different personal characteristics might benefit from different types of explanations in AI-assisted decision making and demonstrate the potential for personalizing explainable AI to improve complementarity. However, given the scale of the study, there are a few limitations. First, the structure of our experiment limited the number of personal characteristics we could effectively analyze, leaving many other traits in the literature worth exploring. Examples include additional Big Five personality traits, demographic features, and disabilities. Second, our study is relatively limited in scope across a few axes. We only explored two types of explanations: text-based and visual-based explanations. Our study is also limited to the geography-guessing task, making it unclear whether our findings can be generalized to other tasks. That said, our geography-guessing task was selected under two criteria: AI possesses superior task-solving skills but humans possess more information, and AI explanations provide verifiability [7]. We conjecture our findings may be more likely to generalize to other tasks with the same characteristics. Moreover, our participant pool was limited to U.S. residents recruited through Prolific, which may affect the generalizability of our findings. Third, we utilized ChatGPT to generate the AI predictions and explanations in our study. Although we believe that advancements in LLMs can further strengthen our findings, our current results are based on a tool that is rapidly evolving over time. Finally, our study requires individuals to self-report their user characteristics through answering questions from the Big Five Inventory and about their own travel experience. These self-reported user characteristics introduce additional biases in our user grouping (e.g., individuals in the group with more travel experience may be those who tend to overestimate their own experience) and our result interpretations.

Future Work. Following the above discussion, future work could explore additional user characteristics beyond the three analyzed by our work. The primary goal of our work was to demonstrate the capability to personalize explanations according to *some* user traits to increase complementarity, but more thorough work about what *specific user traits* produce said interaction effects would be clearly beneficial. Additionally, questions remain about applying the results demonstrated here to other "axes" of personalization or other tasks. Our work only considered a single axis of personalizing explanations: changing the modality of the explanations from text-based to visual-based, from more verbal to visual in nature. This single axis was sufficient to demonstrate the interaction effects of interest and thus the viability of personalized explainable AI's ability to improve complentarity. However, additional axes of personalization, such as tone, depth, or even content, of the AI explanations should also be explored. Finally, examining the generalizability of our findings to tasks beyond the geography guessing task would be important and interesting future work.

References

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.

[2] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. doi:10.1145/3449287

[3] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[4] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial intelligence* 298 (2021), 103503.

[5] Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. 2022. A survey on personality-aware recommendation systems. *Artificial Intelligence Review* (2022), 1–46.

[6] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[7] Raymond Fok and Daniel S Weld. 2023. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* (2023).

[8] Howard Gardner and Thomas Hatch. 1989. Educational implications of the theory of multiple intelligences. *Educational researcher* 18, 8 (1989), 4–10.

[9] Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1103–1116.

[10] Lukas Haas, Michal Skreta, Silas Alberti, and Chelsea Finn. 2024. Pigeon: Predicting image geolocations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12893–12902.

[11] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[12] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2024. Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. arXiv:2404.00029 [cs.HC] https://arxiv.org/abs/2404.00029

[13] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[14] Muhammad Iftikhar, Muhammad Saqib, Muhammad Zareen, and Hassan Mumtaz. 2024. Artificial intelligence: revolutionizing robotic surgery. *Annals of Medicine and Surgery* 86, 9 (2024), 5401–5409.

[15] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, 1–17. doi:10.1145/3544548.3581001

[16] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC*. doi:10.1109/VLHCC.2013.6645235

[17] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300717

[18] Vivian Lai, Yiming Zhang, Chacha Chen, Q Vera Liao, and Chenhao Tan. 2023. Selective explanations: Leveraging human input to align explainable ai. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–35.

[19] Bryan Lavender, Sami Abuhaimed, and Sandip Sen. 2024. Effects of Explanation Types on User Satisfaction and Performance in Human-agent Teams. *International Journal on Artificial Intelligence Tools* 33, 03 (2024), 2460004. doi:10.1142/S0218213024600042 arXiv:https://doi.org/10.1142/S0218213024600042

[20] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746* (2022).

[21] Rhema Linder, Sina Mohseni, Fan Yang, Shiva K. Pentyala, Eric D. Ragan, and Xia Ben Hu. 2021. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters* 2, 4 (2021), e49. doi:10.1002/ail2.49 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.49

[22] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 840, 20 pages. doi:10.1145/3613904.3642671

[23] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[24] Andrea Moglia, Konstantinos Georgiou, Evangelos Georgiou, Richard M. Satava, and Alfred Cuschieri. 2021. A systematic review on artificial intelligence in robot-assisted surgery. *International Journal of Surgery* 95 (2021), 106151. doi:10.1016/j.ijsu.2021.106151

[25] Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin. 2023. How does value similarity affect human reliance in AI-assisted ethical decision making?. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 49–57.

[26] Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. 2024. User Characteristics in Explainable AI: The Rabbit Hole of Personalization?. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[27] S.M. Atikur Rahman, Sifat Ibtisum, Ehsan Bazgir, and Tumpa Barai. 2023. The Significance of Machine Learning in Clinical Disease Diagnosis: A Review. *International Journal of Computer Applications* 185, 36 (Oct. 2023), 10–17. doi:10.5120/ijca2023923147

[28] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212. doi:10.1016/j.jrp.2006.02.001

[29] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 410–422. doi:10.1145/3581641.3584066

[30] Mark Steyvers and Aakriti Kumar. 2024. Three challenges for AI-assisted decision-making. *Perspectives on Psychological Science* 19, 5 (2024), 722–734.

[31] Guido Vittorio Travaini, Federico Pacchioni, Silvia Bellumore, Marta Bosia, and Francesco De Micco. 2022. Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction. *International Journal of Environmental Research and Public Health* 19, 17 (2022). doi:10.3390/ijerph191710594

[32] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[33] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.

[34] Bo Wang, Shuo Jin, Qingsen Yan, Haibo Xu, Chuan Luo, Lai Wei, Wei Zhao, Xuexue Hou, Wenshuo Ma, Zhengqing Xu, et al. 2021. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. *Applied soft computing* 98 (2021), 106897.

[35] Tobias Weyand, Ilya Kostrikov, and James Philbin. 2016. Planet-photo geolocation with convolutional neural networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 37–55.

[36] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.