Who Needs What Explanation? How User Traits Affect Explanation Effectiveness in AI-Assisted Decision-Making

Torrence S 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

Can personalized AI explanations improve human-AI team performance? Motivated by research on individual differences in cognitive science, we examine whether user characteristics influence the effectiveness of AI explanations in AI-assisted decision making. We study this question through preregistered experiments in two tasks. In a sentiment-analysis task, we find that individual differences in user characteristics shape how users respond to explanations, but these differences do not lead to human-AI complementarity, where the joint performance of humans and AI exceeds that of either alone. Motivated by this limitation, we design a new geography-guessing task in which humans and AI possess complementary strengths. In this setting, we again observe that user characteristics interact with explanation types, and now these effects also contribute to complementarity. These results suggest that tailoring explanations to individual users can improve performance and provide valuable insights into how personalization may enhance human-AI collaboration.

CCS Concepts

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

Keywords

Explainable AI, Explanation Design, Personality, Personalization, Complementarity

ACM Reference Format:
Torrence S Farmer and Chien-Ju Ho. 2025. Who Needs What Explanation? How User Traits Affect Explanation Effectiveness in AI-Assisted Decision-Making. In *Proceedings of ACM Conference on Intelligent User Interfaces (IUI '26).* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3742413.3789089

1 Introduction

AI-assisted decision making, which leverages AI assistance to enhance human decision making, has opened new opportunities across various domains [22, 35, 37], from medical diagnosis [41] to criminal justice [54]. However, 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 [5, 6, 56, 59]. A commonly proposed remedy is to accompany AI recommendations with explanations, additional information that conveys aspects of the AI's reasoning, to help users decide when to follow or override AI advice [25, 32, 34]. Ideally, such explanations should enable users to detect AI errors and adjust their reliance accordingly. However, empirical findings remain mixed, with studies showing that explanations can even intensify over-reliance on AI [4, 56]. To better understand these mixed outcomes, recent research has examined properties of explanations that may support more effective human-AI collaboration. For example, explanations that are verifiable, those that allow users to check the AI's recommendations against information available to them, have been argued to encourage more appropriate reliance [12].

In this work, we focus on a complementary yet underexplored dimension of explanations for AI-assisted decision making: the users themselves. We investigate how individual traits, such as personality characteristics and prior experience, shape how people utilize AI explanations. Our focus on individual differences draws on insights from several disciplines showing that personal characteristics influence how people process and respond to information. In education, aptitude–treatment interaction theory [44, 47] posits that the effectiveness of instructional methods depends on learners' prior experience and skill: step-by-step guidance may benefit novices but hinder those with greater expertise. In cognitive science, research shows that personality traits—particularly the Big Five [42]—affect how people interpret system outputs [38], solve problems [10, 52], and acquire new skills [21]. Traits often studied alongside the Big Five, such as need for cognition (NFC), have also been linked to how deeply users engage with detailed information from AI-driven systems, such as music recommendations [34]. Collectively, these findings suggest that individual characteristics shape how users process and respond to explanations—a perspective that may be crucial for understanding when and for whom AI assistance improves decision performance. While some studies have begun to examine the role of user traits in explainable AI [8, 34, 38], few have explored how these traits interact with explanation design to influence human-AI outcomes.

To address this gap, in this work, we investigate whether individual differences shape how people respond to AI explanations and whether these differences influence the potential for humans and AI to reach *complementarity*, i.e., the joint performance of humans and AI exceeds that of either alone. To examine these questions, we conducted two preregistered experiments across two tasks.

In the first experiment, we examined a standard sentiment analysis task [1, 40, 53]. We recruited 400 participants from Prolific

IUI '26, Paphos, Cyprus

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/3742413.3789089

to judge the sentiment of movie reviews, deciding whether each review's associated rating was positive or negative. Participants were assigned to one of four conditions: no AI assistance, AI assistance without explanations, AI assistance with sparse explanations, or AI assistance with dense explanations. In addition to completing the task, participants filled out a survey measuring their personality traits, including need for cognition, openness, and prior experience with movie reviews. Our results revealed a significant interaction effect between need for cognition and explanation length on task performance and provided evidence for interaction between experience and explanation length on performance. However, overall, participants performed worse with AI assistance than without it, regardless of whether explanations were provided. The results suggest that while users characteristics do influence how people respond explanations, there is limited potential for personalization to enhance complementarity in this setting.

In our second experiment, we developed a geography-guessing task adapted from the popular game Geoguessr[1] , in which participants guessed the continent where a photo was taken. We designed the task so that humans and the AI possessed different strengths, mimicking a common scenario in which the AI is generally more accurate while humans have access to additional private information [3, 17, 20]. Participants were randomly assigned to one of four conditions: no AI assistance, AI assistance without explanations, AI assistance with text-based explanations, or AI assistance with visual explanations. In the text-based condition, participants received an AI recommendation accompanied by a one-sentence rationale; in the visual condition, they received an AI recommendation with highlighted regions in the photo relevant to the decision. We again recruited 400 participants from Prolific and found a significant interaction effect between openness and explanation modality on performance, as well as evidence for an interaction between travel experience and modality on performance. Participants who received AI assistance with explanations showed improved overall performance and achieved complementarity, and these benefits were more pronounced when the explanation format aligned with their personality traits. These findings suggest that the influence of user traits on how individuals incorporate AI explanations generalizes across different tasks and explanation types. They also highlight the potential of leveraging individual differences to better support effective human-AI collaboration.

2 Related Work

To situate our work within the broader literature, we review three strands of research that inform our study. First, we examine prior work investigating the impacts of AI explanation design and personalization. Next, we review research on individual differences, focusing on identifying user traits that might shape information processing and interaction with AI systems. Finally, we connect these perspectives to the literature on human-AI complementarity, which seeks to understand when and how human-AI collaborative decision making can achieve performance exceeding that of either partner alone.

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

The Impact of AI Explanations. A common approach to improving AI-assisted decision making is to provide users with additional information beyond the AI predictions. For example, confidence scores can help users determine when to trust AI recommendations [19], and stating the AI's overall accuracy can help users calibrate their expectations [16]. An increasingly common addition to AI outputs is an explanation, information describing why the AI made a particular decision. Explanations can take various forms, including similar data points, feature weights, or verbal rationales, as seen in large language models (LLMs) such as ChatGPT [31]. Explanations have been mostly studied for their effects on users' perceptions and behavior, particularly in terms of trust [30], understanding [27], and mental effort [29]. The literature suggests that users actively seek information from explanations when deciding whether to adopt AI recommendations [14, 25, 34], and generally prefer to receive and use as much information from the AI as possible.

Recent work further explores personalizing explanations to better match user characteristics. Personalized explanations have been shown to enhance user trust [10, 28], improve understanding of the AI's reasoning [27], and increase satisfaction with AI assistance [25, 28]. These benefits have been demonstrated through adjustments to explanation properties such as amount of content [27], modality [14], level of detail [27, 32], tone, or content [33]. Outside the domain of AI-assisted decision making, personalization has improved outcomes across various fields. For example, research in aptitude–treatment interaction (ATI) [44, 47] shows that aligning instruction with individual aptitude improves learning outcomes, and personality-based recommender systems have been found to increase user satisfaction and engagement [10, 34]. In robotics, some work has been done on trying to personalize explanations to improve performance [46], but this work personalized based on previously observed interactions with the single participant it's working with instead of readily-measurable personality traits prior to any experiment run.

Overall, studies on personalized explanations overwhelmingly focus on perceptual or attitudinal outcomes rather than performance. In contrast, our work examines how tailoring explanations to directly-measurable user characteristics affects decision performance and whether such personalization can promote human-AI complementarity in AI-advised decision-making.

What User Traits Might be Relevant. Prior work has identified a range of user traits that shape how individuals process information and interact with AI systems. Among the most commonly studied are the Big Five personality traits [10, 42], need for cognition (NFC) [8, 34], and task experience [44, 47]. These traits have been shown to interact with users' trust, reliance, and perceived understanding of AI explanations [8, 11, 38]. Because they capture how individuals process and engage with information [13], they may also influence how users calibrate their reliance on AI assistance [19, 45], ultimately shaping team performance.

Cognitive theories further illuminate how these traits might interact with explanation design. Cognitive Fit Theory [51, 57] posits that performance improves when the representation of information aligns with both the task and the user's cognitive style. For example, in a geography-guessing task, spatial cues in visual explanations may better suit highly experienced users who can draw on visual

memory from prior travel [8, 11]. Conversely, the symbolic and conceptual nature of text-based explanations may align better with individuals high in openness, who are more receptive to abstract reasoning and less reliant on concrete cues [11, 24]. Cognitive Load Theory [49, 58] provides an additional perspective, suggesting that overly detailed explanations can impair performance by increasing extraneous cognitive load. In a sentiment-analysis task, for instance, dense explanations that highlight numerous low-weight words may distract users and reduce efficiency. Such effects are particularly detrimental for users with high experience, who may suffer from the expertise reversal effect [24, 50], and for those with high need for cognition, who may overanalyze uninformative details.

Building on this literature, we focus on three characteristics, openness (from the Big Five), need for cognition, and experience, as they represent distinct yet complementary dimensions of cognitive style, motivational orientation, and domain expertise. Together, these dimensions capture key sources of individual variation that are most likely to moderate how users interpret, evaluate, and benefit from AI explanations.

Towards Human-AI Complementarity. One central goal in AI-assisted decision making is to achieve complementarity—where human-AI teams outperform either the human or AI alone [4, 6, 19, 33]. However, empirical findings remain mixed: users often over-rely on AI assistance, which can diminish their ability to perform effectively as a team [5, 6, 56, 59]. When explanations are provided, they also often fail to improve performance and, in some cases, even exacerbate over-reliance on AI by reinforcing misplaced trust or cognitive laziness [4, 12, 14, 38]. Recent theoretical work identifies the conditions under which explanations can promote complementarity, emphasizing the importance of private human information [15] and the role of verifiability—the extent to which explanations allow users to check AI outputs against available information [12, 56]. Building on these insights, we design our geography-guessing task to establish complementary strengths between humans and AI and to examine how explanation modality and user characteristics jointly influence team performance.

In summary, our work bridges these three threads of research by examining how user traits interact with explanation design to shape complementarity in AI-assisted decision making. Through a sentiment-analysis task and a novel geography-guessing task, we show that traits such as openness, need for cognition, and experience influence how users engage with AI explanations and when such personalization can improve human-AI team performance.

3 Experiment 1: Sentiment Analysis Task

The goal of this work is to examine whether user characteristics influence the effectiveness of AI explanations in AI-assisted decision making. In the first experiment, we use a standard sentiment-analysis tasks [53, 55]. While this task is well studied in AI-assisted decision making, there is little prior work examining whether different users respond differently to explanations in this setting. This makes it an ideal first task for our experiment. In this task, users are presented with IMDB movie review and asked to predict whether the reviewer's rating is high (6 or more out of 10) or low (4 or less out of 10), while reviews with a score of 5 are excluded for ambiguity. Participants are randomly assigned to different conditions that vary in whether they receive AI assistance and, if so, in the type of explanation accompanying the AI's recommendation. We also measure user traits through a post-task survey. We then examine whether there are interaction effects between user traits and AI explanations. More specifically, we ask the following core research question.

- **RQ1**: Do individuals with different characteristics perform differently under different explanation conditions?

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

3.1 Experiment Design

We describe our experimental design below, including the explanation types, user traits, and corresponding hypotheses.

3.1.1 Explanation Types. To vary the explanations provided to users and examine the effects of personalization, we followed prior work [32] and manipulated the level of explanation detail. We used LIME [43] to generate feature-based explanations for the AI's sentiment predictions, highlighting words that contributed most to each prediction. Accordingly, the task includes two explanation types: sparse and dense. Sparse explanations display the top three words in LIME's output, while dense explanations include all words with non-zero influence (up to a maximum of ten).

Participants were randomly assigned to one of four conditions:

- **No AI (Control Group)**: Participants received no AI assistance.
- **Unexplained AI**: Participants received AI recommendations without explanations.
- **Sparse Explanation**: Participants received AI recommendations with sparse explanations.
- **Dense Explanation**: Participants received AI recommendations with dense explanations.

3.1.2 User Characteristics. Following prior literature, we focus on the Big Five personality traits [42], need for cognition (NFC) [8], and domain experience. Based on the results of a preliminary pilot, our main experiment concentrated on openness (from the Big Five), need for cognition, and experience. Openness reflects creativity and an individual's willingness to consider ideas from external sources, including AI systems. Need for Cognition captures how much individuals seek out activities that involve substantial cognitive effort, such as engaging in more thorough analysis of presented information. To measure experience, we used the number of movies viewed as a proxy for relevant domain expertise.

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. The questionnaire for measuring user traits is included in Appendix E. Specifically, for openness, we drew questions from the BFI-10 personality inventory [42], a 10-item abbreviated personality measure commonly used in the literature. The BFI-10 has been shown to be a reliable and valid approximation of the full-length Big Five personality inventory [23], a standard instrument for measuring personality traits despite its self-reported nature.

[2]The pre-registration can be found at: https://osf.io/4j5cd/?view_only=55e3c35e986b483caf8421c4dad8b1b3

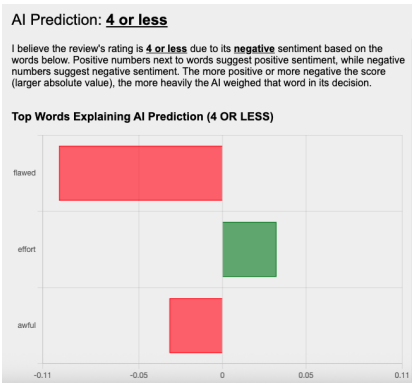


Figure 1: A sample explanation in the sentiment analysis task. This is a sparse explanation containing 3 words.

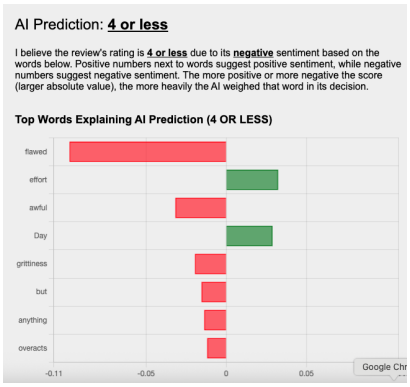


Figure 2: A sample explanation in the sentiment analysis task. This is a dense explanation containing 8 words.

We use the BFI-10 to balance measurement accuracy with survey efficiency. For need for cognition (NFC), we adapted questions from the Need for Cognition Scale [7], a widely used and well-studied questionnaire. For experience, there is no standard set of questions. Accordingly, we asked participants to indicate their agreement with the following statements: "I watch a substantial amount of movies," and "I have read movie reviews before and am generally knowledgeable about what may cause a movie to receive a high or low score."

3.1.3 Hypothesis. Our hypotheses are grounded in established psychological theories, as discussed in the related work section. Specifically, prior studies suggest that individuals high in openness tend to exhibit greater intellectual curiosity and a stronger appreciation for complex, abstract information [10, 38, 42]. Accordingly, we expect that more open users will perform relatively better with denser, more detailed explanations. Theories of expertise reversal [24, 50] and cognitive load [49] suggest that low-knowledge or low-experience individuals benefit disproportionately from well-guided instruction, whereas overly verbose information can increase cognitive load and hinder high-experience users. Thus, we expect individuals with higher experience to perform better with sparser explanations that omit less informative words. Finally, because individuals with higher need for cognition are more willing to expend mental effort processing additional information [58], we anticipate that they may overanalyze verbose explanations and therefore perform worse when provided with denser ones. These theoretical insights inform the hypotheses presented below.

- **H1A**: Participants who are more *open* will perform better with more verbose (denser) explanations.
- **H1B**: Participants with a higher *need for cognition (NFC)* will perform better with less verbose (sparser) explanations.
- **H1C**: Participants with greater *experience* will perform better with less verbose (sparser) explanations.

3.2 Experiment Procedure

3.2.1 Task Implementations. We draw movie reviews from the IMDB dataset [53, 55]. Reviews with scores of 6 or higher (out of 10) are treated as positive, and those with scores of 4 or lower as negative. Reviews with a score of 5 are discarded to avoid ambiguity, and positive and negative reviews are balanced in the dataset before being drawn. For the implementation of the AI model and explanations, we adopt a standard setup. Specifically, we use BERT [9] to generate sentiment predictions and LIME [43] to produce explanations. For each review, LIME identifies the words most relevant to BERT's prediction and visualizes them in a bar chart indicating their relative importance. During the task phase, both humans and the AI are provided only with the review text and need to predict whether the associated rating is positive or negative.

3.2.2 Recruitment. We recruited 400 participants from Prolific, restricting the study to U.S. workers. Before conducting the 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.

3.2.3 Experiment Procedure. Participants 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. After passing the check, participants were given a short questionnaire about their openness personality trait, need for cognition, and movie-watching experience. Once participants completed the survey, they were given 20 rounds of the task. We included more challenging rounds where the AI assistant was correct only roughly two-thirds of the time. This was done in line with the literature [5] to better observe and analyze challenging scenarios where the user actually has to think through the AI's advice instead of having the answer be too obvious.

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.

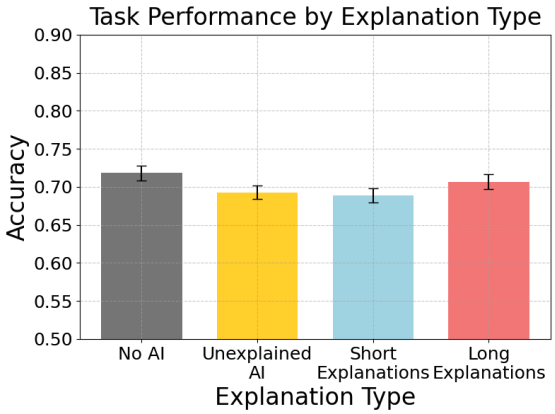


Figure 3: Results by explanation density for the sentiment analysis task without differentiating based on user characteristics.

3.3 Experiment Results

We first examine the overall performance across the four experimental conditions. As shown in Figure 3, at the group level, user performance is at a comparable level across all treatments.

In the following analysis, we focus on our main research question (RQ1) and examine whether user characteristics moderate the effectiveness of AI explanations on task performance. Specifically, we focus our analysis on the two groups that received explanations to test for the interaction effects of interest. To assess RQ1, for each participant characteristic (openness, need for cognition, and travel experience), we perform a multiple linear regression that includes explanation type, the relevant participant characteristic, their interaction term, and a constant. We then evaluate H1A through H1C by testing whether the interaction term in each regression is statistically significant at the 0.05 level using standard one-sided t-tests.

We now describe the results for the hypotheses. Tables containing abbreviated results of our regressions are presented, with full readouts available in the appendix. To more effectively visualize our data, we also group participants into low, medium, and high "score groups" based on score tercile of their characteristics compared to the overall population and plot their performance with dense or sparse explanations. This grouping is only for visualization purposes and has no impact on the underlying regression analysis, the results of which are also reported below. The accuracy needed for complementarity is indicated by a red line in Figure 4.

3.3.1 H1A: Participants who are more open *perform better with with more verbose (denser) explanations in the sentiment analysis task.* As shown in Figure 4a, openness appears to have little interaction with explanation density in affecting performance. Overall, users perform slightly better with denser explanations regardless of their openness scores. This observation is supported by the linear regression results in Table 1, which show no significant interaction effect between openness and explanation density on performance ($p = .694$).

3.3.2 H1B: Participants with a higher need for cognition (NFC) *perform relatively better with sparser explanations.*

Table 1: H1A: Openness and AI Explanation Modalities' Effect on Accuracy

Predictor	Coeff.	Std. Error	T-statistic	p-value
Intercept	0.6652	0.049	13.557	–
Open. Score	0.0030	0.006	0.490	–
Exp. Type	0.0497	0.064	0.782	–
Interaction	-0.0040	0.008	-0.509	**0.694**

Looking at the results in Figure 4b, we find that participants with low need for cognition perform quite significantly better when given denser explanations. However, as need for cognition increases, the effects of the two explanation densities begin to diverge. Medium-NFC participants only perform 1.8 percentage points better with denser explanations compared to sparser ones, and this effect flips to a gap of 4 percentage point for high NFC users in favor of sparser explanations. This observation is confirmed by the regression results summarized in Table 2, where there is a statistical significant interaction between need for cognition and explanation ($p = 0.017$).

Table 2: H1B: Need for Cognition and AI Explanation Modalities' Effect on Accuracy

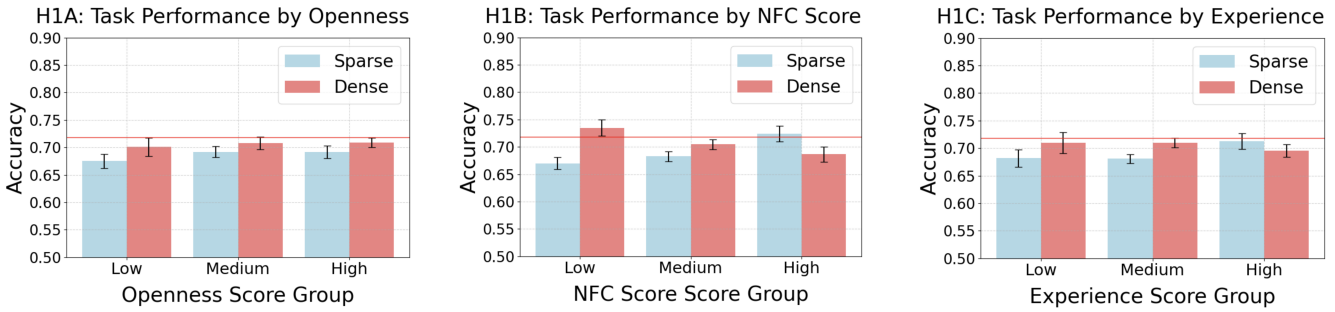
Predictor	Coeff.	Std. Error	T-statistic	p-value
Intercept	0.6308	0.041	15.306	–
NFC Score	0.0151	0.010	1.446	–
Exp. Type	0.1407	0.059	2.391	–
Interaction	-0.0316	0.015	-2.141	**0.017**

3.3.3 H1C: Participants with more relevant experience *perform relatively better with sparser explanations.* As visualized in Figure 4c, individuals who have watched more movies tended to perform relatively better when given sparser explanations compared to their peers. However, our test shown in Table 3 does not reveal a statistically significant effect for the interaction terms between experience and explanation types ($p = 0.103$).

Table 3: H1C: Experience and AI Explanation Modalities' Effect on Accuracy

Predictor	Coeff.	Std. Error	T-statistic	p-value
Intercept	0.6255	0.057	11.012	–
Exp. Score	0.0080	0.007	1.131	–
Exp. Type	0.1098	0.073	1.498	–
Interaction	-0.0116	0.009	-1.272	**0.103**

To explain this discrepancy, we conducted an exploratory analysis and found that participants with greater movie experience exhibited substantially different behavior from those with less experience, with a sudden shift towards superior performance using sparser explanations. Among individuals with an average or lower number of movies watched, however, performance did not differ meaningfully between the dense and sparse explanation conditions. Consequently,



(a) We found no significant interaction between openness and performance using the two explanation densities.

(b) Individuals with higher NFC scores performed better using sparser explanations versus denser ones.

(c) Individuals with more movie experience performed better with sparser explanations versus denser ones.

Figure 4: Performance by user trait and explanation modality in the sentiment analysis task. The error bars represent standard errors. The red line represents the threshold for complementarity, corresponding to the higher of the human or AI performance when each acts alone.

we conducted an exploratory analysis treating experience as a binary variable. When grouping participants into two categories—high-experience (top tercile) and not-high-experience—we observed a statistically significant interaction effect ($p = 0.018$). This transformation was *not* pre-registered and is therefore considered *exploratory*. Nonetheless, we believe these findings offer suggestive evidence that users with higher levels of expertise may benefit more from personalized explanations, an effect that warrants further investigation in future work.

3.4 Discussion

Overall, our results indicate statistically significant interactions between need for cognition (NFC) and explanation density on performance. We also find suggestive evidence that experience may interact with explanation density in the exploratory analysis. These results make sense in light of the prior psychological literature; Sweller [49] suggests that additional information with limited predictive value can be detrimental for users with high NFC, as such users may overanalyze less informative words. Similarly, the psychological literature suggests that high-experience users may be disproportionately hindered by longer explanations containing less relevant information, consistent with the expertise reversal effect [24, 50], which also can partially explain our user experience results.

However, a main motivation of this study was to examine whether personalization could enhance complementarity. When comparing the overall performance of the AI to the overall performance of the users, the bar for complementarity in this task, indicated by the red line in Figure 4, was 71.8%. Because not all users performed equally well on the task with no AI assistance, we also considered complementarity for each subset of users bucketed by openness, NFC, and experience. We found that only one subset of users, those with particularly low need for cognition using long explanations, achieved complementarity. Extended results and visualizations for performances across all four subgroups in this experiment can be found in the appendix. This outcome is perhaps not surprising given prior findings that explanations in sentiment analysis tasks rarely lead to complementarity [4, 12]. More broadly, many common tasks in explainable AI have struggled to consistently demonstrate complementary performance [12], motivating research into the conditions under which complementarity is more likely to emerge.

4 Experiment 2: Geography Guessing Task

There are two primary motivations for this experiment. First, we aim to examine whether our finding—that user characteristics influence how individuals utilize explanations—generalizes to other tasks and explanation formats. Second, we seek to design a task in which human-AI complementarity is more likely to occur. To this end, the literature [15] suggests that human users having private information is often a main driver of complementarity. Humans and AI can use their asymmetric information [18] to assist each other improving their joint decision-making, improving performance. Some literature [12, 56] also suggests that verifiability—the ability of users to confirm the AI's reasoning using available information—is key to fostering appropriate reliance. Explanations should help users verify AI outputs with greater accuracy and less effort than solving the task independently or blindly following the AI.

Motivated by these insights, we developed the geography-guessing task designed to promote complementarity while supporting multiple explanation modalities that could yield benefits from personalization. In this task, users are presented with locations drawn from Google Street View and asked to identify the continent from which each image originates. Some users are assisted by an explainable AI system that provides predictions and explanations. To foster conditions conducive to complementarity, we restrict the information available to the AI by showing it only a partial image while providing users with the full image. This setup mimics the common scenario that the AI is more accurate than humans when provided the same information, but humans might possess additional private information. This setup promotes distinct and complementary strengths between the two. The task interfaces are shown in Figure 5 and Figure 6. We have the following two research questions.

- **RQ1**: Do individuals with different characteristics perform differently under different explanation conditions?

- **RQ2**: Do the AI explanations enable human-AI teams to achieve complementarity in the geography-guessing task?

The experiment and analysis for experiment 2 are also preregistered on the Open Science Framework (OSF).[3]

4.1 Experiment Design

To address the research questions, we describe our experimental design below, including the explanation types, user traits, and corresponding hypotheses.

4.1.1 Explanation Types. The geography-guessing task includes two explanation modalities to explore whether differences in explanation modality can affect performance in AI-advised decision-making. Specifically, we design two explanation modalities: **Text-based** explanations that contain written description generated by the AI in supporting its predictions (Figure 5), and **Visual-based** explanations that contain blue circles that highlight the regions in the image that the AI indicates are the most relevant (Figure 6). We select these two modalities because textual rationales [43] and visual saliency overlays [36] are commonly bundled with contemporary classification systems. Because they occupy complementary verbal and spatial channels, dual-coding [39] and cognitive-fit theories [57] suggest that each modality may differentially benefit differing users depending on their individual characteristics.

Participants were randomly assigned to one of four conditions:

- **No AI (Control Group)**: Participants received no AI assistance.
- **Unexplained AI**: Participants received AI recommendations without explanations.
- **Text-Based Explanation**: Participants received AI recommendations with text-based explanations.
- **Visual-Based Explanation**: Participants received AI recommendations with visual-based explanations.

4.1.2 Fostering Complementarity in Task Design. This design for the geography guessing task has several desired properties that foster the conditions for complementarity. First, by limiting the information available to the AI, we create complementary strengths between humans and the AI [12]. This also reflects many real-world scenarios in which the AI has stronger predictive power given shared information, while humans possess additional private knowledge. [11, 14] Second, our explanations are often verifiable: [12, 56] users can leverage information from parts of the image outside the AI-accessible region to check whether the AI's explanation aligns with reality. Lastly, the task itself is intuitive and has a low barrier to entry. We believe that the ability to establish relative strengths for humans and AI, the suitability for generating explanations in distinct modalities, and the low barrier to entry make this task well suited for testing our hypotheses and for serving as a promising candidate in future research on human-AI collaboration.

4.1.3 User Characteristics and Hypothesis. We examine the same user characteristics as in the sentiment-analysis task: openness, need for cognition (NFC), and experience. To measure experience, we use travel experience as a proxy for relevant expertise.

[3]The pre-registration for the geography-guessing task can be found at: https://osf.io/a6t9r?view_only=323483088c834b02be51c4278b16dfd6

We look to align our hypotheses with the psychological literature. Specifically, we hypothesize that more open users will perform better with the more extensive text-based explanations, as the literature suggests that openness is characterized by higher levels of intellectual curiosity and a greater appreciation for complex information [10, 38, 42]. Expertise reversal [24, 50] and cognitive load theories [49] suggest that individuals suggest that low-knowledge or low-experience individuals benefit disproportionally from well-guided instruction, which suggests increased performance using the more hand-holding text-based explanations. However, it's not as clear from the literature what affect NFC will have on the geography-guessing task, as users with higher NFC scores are more willing to spend mental energy analyzing and utilizing additional information [58], and it's not obvious which explanation modality requires more mental energy to fully utilize. Building off of this literature, we develop the hypotheses below.

- **H1A**: Participants who are more *open* will perform relatively better with text-based explanations.
- **H1B**: Participants with a higher *need for cognition (NFC)* will perform relatively better with visual-based explanations.
- **H1C**: Participants with more relevant *experience* will perform relatively better with visual-based explanations.
- **H2**: Overall, participants given explanations will achieve complementarity, outperforming those without AI assistance or with unexplained AI.

4.2 Experiment Procedure

4.2.1 Task Implementations. The geography-guessing task is based on the game Geoguessr, where participants guess the location of a photo pulled from Google Streetview. Note that Australia was excluded from this exercise due to extreme levels of confusion in a pilot study, and Antartica was excluded due to lack of coverage. In our task, users are given a still image from Streetview and are asked to guess what continent the photo was taken in. Their AI partner is able to see part of the photo, shown using a red box (as in Figure 5 and 6) and can use this photo to make a prediction of its own. Some users will also receive explanations as part of the AI's output. We use ChatGPT 4o to generate both the prediction and explanation. In particular, we provide ChatGPT with the image in the red box and prompt it to predict the continent of origin for the image. We then create a separate chat for each round using the same image to generate an explanation. To align the information provided by the two explanation types, we prompt ChatGPT to have a one-to-one mapping between the circled image element in the visual-based explanation and the corresponding item mentioned in the text-based explanation. The prompts we used to generate AI predictions and explanations are included in the appendix.

4.2.2 Recruitment and Experiment Procedure. The recruitment and procedure mirror Experiment 1. We recruited 400 participants from Prolific, restricting the study to U.S. workers, with 100 participants assigned to each of the four groups. Participants first completed the informed consent form, took a 5-question comprehension check, then answered short questionnaire about their openness personality trait, need for cognition, and travel experience.

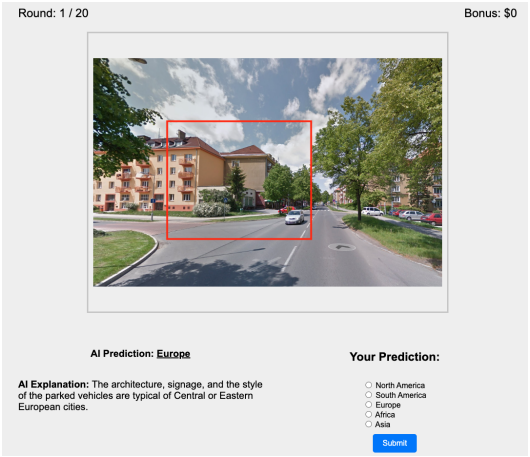


Figure 5: A sample round from the geography guessing task with text-based explanations. The red box indicates the area seen by the AI in making its prediction. In this case, the AI provides its explanation through a sentence, as textual description on the bottom left corner.

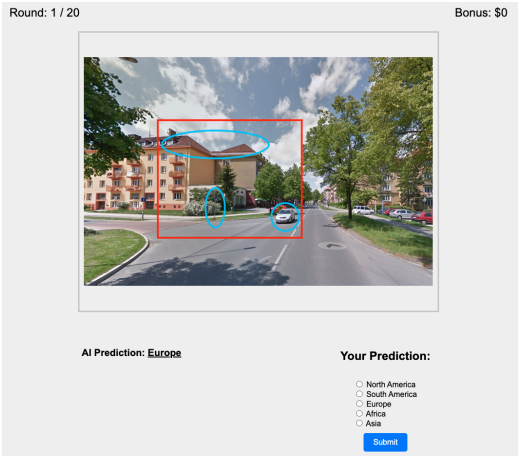


Figure 6: A sample round from the geography guessing task with visual-based explanations. The red box indicates the area seen by the AI in making its prediction. In this case, the AI provides its explanation through drawing blue circles on the relevant parts of the image.

The setup of the questionnaire was identical to the setup in the sentiment-analysis task, but with the experience questions replaced with questions about experience with geography and geography-guessing games. Once participants completed the survey, they were given 20 rounds of the task before taking a short ending questionnaire and getting paid out. Additionally, as in Experiment 1, we gave users harder rounds than average in line with the literature [5] to better observe and analyze challenging scenarios where the user has to think through the AI's advice instead of having the answer be too obvious. In this experiment, the AI was 70% accurate. The payment structure of the geography-guessing experiment is also identical to sentiment analysis: $1.70 for completion plus $0.05 per correct answer.

4.3 Experiment Results

We examine the overall performance across the four experimental conditions. As shown in Figure 7, We observed that users without AI assistance significantly underperformed the other groups, and that among users given AI assistance, those who received text-based explanations performed the best as a whole (see more discussion in H2 below). We now focus on running statistical analyses (multiple linear regressions) to answer our research questions of interest, beginning with RQ1.

To address RQ1, we conduct a multiple linear regression for each participant characteristic that includes explanation type, characteristic score, an interaction term, and a constant. We then evaluate our hypotheses by determining whether the interaction term in each corresponding regression is statistically significant at the 0.05 level using standard one-sided t-tests.

4.3.1 H1A: Participants who are more open *perform relatively better with text-based explanations in the geography guessing task.* We focus the remainder of our analysis on the two groups receiving explanations to analyze the interaction effects of

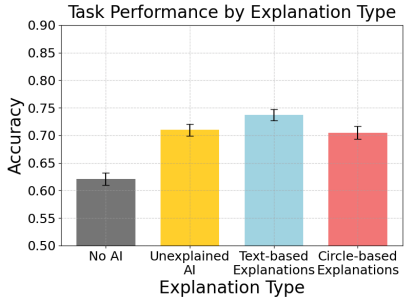
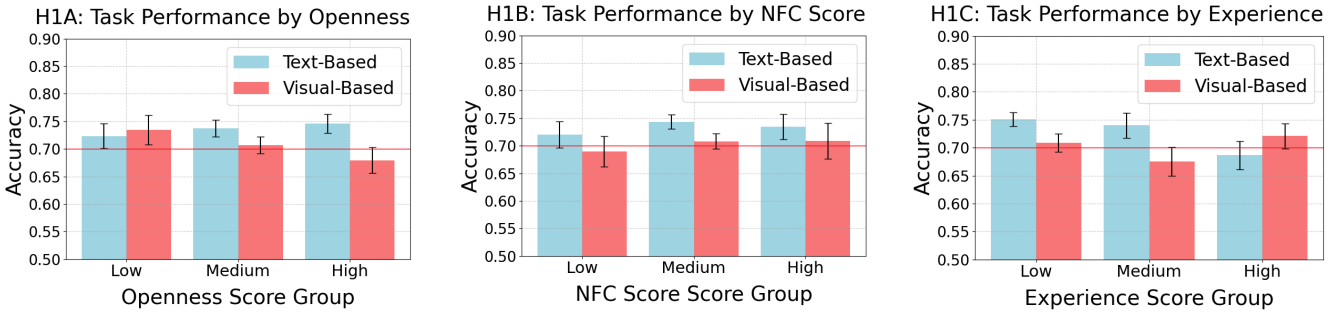


Figure 7: Results by explanation format for the geography-guessing task without differentiating based on user characteristics. Participants receiving text-based explanations significantly outperformed the other participants and definitively achieved complementarity.

interest. Looking to Figure 8a, 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 as show in Table 4, which reveals a significant interaction effect between openness and explanation type on performance ($p = .0313$).

Table 4: H1A: Openness and AI Explanation Modalities' Effect on Accuracy

Predictor	Coeff.	Std. Error	T-statistic	p-value
Intercept	0.6674	0.048	13.881	–
Open. Score	0.0182	0.012	1.489	–
Exp. Type	0.1002	0.073	1.374	–
Interaction	**-0.0346**	0.019	**-1.863**	**0.0313**



(a) Individuals with higher openness scores performed better using text-based explanations.

(b) No significant interaction between NFC and performance using the two explanation types.

(c) Individuals with more travel experience performed better with visual-based explanations.

Figure 8: Performance by user trait and explanation modality in the geography-guessing task. The error bars represent standard errors. The red line represents the threshold for complementarity, corresponding to the higher of the human or AI performance when each acts alone.

4.3.2 H1B: Participants with a higher need for cognition (NFC) *perform relatively better with visual-based explanations in the geography guessing task.* As shown in Figure 8b, we found no significant interaction between NFC score and explanation type on performance, as the interaction term in our multiple linear regression shown in Table 5 is not statistically significant (p = 0.4892). As a result, this hypothesis is not supported.

Table 5: H1B: Need for Cognition and AI Explanation Modalities' Effect on Accuracy

Predictor	Coeff.	Std. Error	T-statistic	p-value
Intercept	0.7005	0.054	13.054	–
NFC Score	0.0093	0.013	0.701	–
Exp. Type	-0.0341	0.081	-0.419	–
Interaction	**0.0006**	0.020	**0.027**	**0.4892**

4.3.3 H1C: Participants with more relevant experience *perform relatively better with visual-based explanations in the geography guessing task.* As visualized in Figure 8c, well-traveled individuals tended to perform relatively better than their peers using visual-based explanations. However, like in the sentiment-analysis task, our test shown in Table 6 does not reveal a statistically significant effect for the interaction terms between experience and explanation types (p = 0.0693).

Table 6: H1C: Experience and AI Explanation Modalities' Effect on Accuracy

Predictor	Coeff.	Std. Error	T-statistic	p-value
Intercept	0.7884	0.026	30.416	–
Exp. Score	-0.0219	0.010	-2.143	–
Exp. Type	-0.0810	0.037	-2.175	–
Interaction	**0.0208**	0.014	**1.481**	**0.0693**

To understand the discrepancy, we conducted an *exploratory* analysis, similar to Experiment 1, by dividing users into high-experience (top tercile) and not-high-experience groups, treating experience as a binary variable. We found a statistically significant interaction effect in this regression (p = 0.0128).

4.3.4 H2: Overall, individuals given explanations in the geography guessing task achieve complementarity. Returning to experiment-group-level performance, the average performance of participants given each explanation type is shown in Figure 7. Among the three sets of participants receiving AI assistance, those receiving unexplained AI assistance achieved 71.0% accuracy, those given visual-based explanations achieved 70.5% accuracy, and those given text-based explanations achieved 73.7% accuracy. The AI alone achieved 70% accuracy, significantly outperforming users with no AI assistance, meaning that the baseline for complementarity was 70%. Overall, users given explanations achieved 72.1% accuracy, which is statistically significantly higher than the baseline (p = .0017). We suspect that this complementarity stems from our task design, which promotes relative strengths between users and their AI companions, allowing users to leverage their own expertise when interpreting the AI's advice. However, an alternative explanation for this performance gain is that participants may have learned more about the AI over time through few-shot learning, possibly by memorizing patterns in the AI's behavior. While we did not identify patterns in the AI's behavior that participants could plausibly learn from, future studies could more directly investigate potential learning effects when interpreting results and designing tasks.

4.4 Discussion

In our experiments, we observed that there exist interaction effects between the Openness personality trait and explanation modality on task performance. We also found suggestive evidence that experience might interact with explanation modality as well. These results are sensible with respect to the existing literature, as cognitive fit theory [57] suggests that text-based explanations may better match the cognitive style of users with high openness, who are more receptive to abstract reasoning and don't need elements of a round to be

circled in red to be used effectively. In terms of experience, we see little deviation in responses among individuals without high travel experience, then a sudden shift for very experienced individuals. This resulted in the interaction term for experience being not statistically significant when treated as a raw score, but becoming statistically significant once thresholded. We saw no significant interaction effect from need for cognition (NFC).

We want to note that both experiments show similar patterns when examining the effects of experience on explanation effectiveness and performance. In both cases, the interaction term between explanation type and experience is not statistically significant when experience is treated as a continuous variable, but it becomes significant once thresholded. We suspect that this may be due to the use of a Likert scale to measure experience. If a certain level of experience is necessary to elicit different reactions to different explanations, it may be more appropriate to ask users directly whether they possess that level of experience (using an explicit threshold) rather than relying on a Likert scale. This phenomenon should be examined and validated further in future work.

We also identify that this task achieves complementarity, with the users who received an AI-generated explanation outperforming the unassisted users or the AI alone. In particular, at the population level, the improvement for text-based explanations compared with the complementarity threshold was statistically significant ($p < .0005$). There were also subgroups (such as high-openness individuals) where visual-based explanations achieved complementarity at a .05 significance level. More importantly, every single subgroup of users based on Openness, NFC, and Experience, had at least one explanation type where complementarity (average performance of 70%) was achieved, as shown in Figures 8a,8b, and 8c. This indicates that unlike in the sentiment analysis task and in the literature to this point, we can personalize explanations to improve complementarity for a broad population of users.

5 General Discussion and Future Work

Recap and Interpretation. In this work, we examine how user characteristics interact with explanation design to shape performance in AI-assisted decision-making. Through preregistered experiments involving two tasks, a standard sentiment-analysis task and a geography-guessing task designed with complementarity in mind, we demonstrate that user traits such as openness, need for cognition, and experience influence how users utilize explanations in AI-assisted decision-making. Specifically, we find statistically significant interactions between need for cognition (NFC) and explanation length on performance in the sentiment-analysis task, as well as between Openness and explanation modality on performance in the geography-guessing task. We also find suggestive evidence that experience may significantly interact with explanation features across both tasks once thresholded. These cross-task observations suggest that experience might moderate the effectiveness of explanations and that alternative methods of measuring experience should be explored in future work.

In the geography-guessing task, we further observe that users who received explanations overall achieved complementarity, and that each subgroup (by NFC, openness, and experience) achieved complementarity for at least one explanation modality. This concurrent demonstration of the benefits of personalization and complementarity distinguishes our work from prior literature and highlights the potential of personalized explainable AI to enhance human–AI team performance by tailoring explanations to user traits.

Individual Differences and Personalized Explanations. A growing body of research demonstrates that personalized explanations tailored to individual differences can influence a range of non-performance outcomes. For example, personalization has been shown to improve learning rates in educational systems [8] and enhance user satisfaction [10, 34] and trust [28] in recommender systems. Other work has identified performance-related benefits of personalization in specific domains such as robotics [46] and medical diagnosis [48], although their personalization is learned through repeated interactions over multiple rounds of the same task, not based on independently measurable personality traits. Many of these studies also consider user traits similar to ours—such as the Big Five personality dimensions, need for cognition, and prior experience. In this context, our work complements the existing literature by showing that task performance in explainable AI–assisted decision-making can also be systematically shaped by these commonly studied user characteristics. At the same time, other studies have reported that tailoring explanations to user traits (including the Big Five) does not meaningfully improve users' understanding of the underlying AI system [38]. While such findings may appear contradictory, direct comparisons are difficult due to the heterogeneity of tasks, user characteristics, and evaluation metrics used across studies. In this regard, our focus on task performance offers a unifying and interpretable metric that may help bridge findings across the personalization literature. Performance-based evaluation not only provides practical insight into how to maximize the benefits of AI assistance, but also facilitates more consistent cross-study comparisons.

Task performance has not often been the central focus in studies of personalized explanations, in part because achieving complementarity, where human-AI teams outperform either alone, has proven challenging [12, 56]. We suspect that two design choices in our study contributed to achieving complementarity in the geography-guessing task: (1) designing the task to feature clear relative strengths between humans and the AI, and (2) creating explanation modalities that present information in ways potentially better suited to different user types and cognitive styles. These design elements distinguish our work from prior studies on personalized explainable AI by demonstrating how individual differences and explanation design can jointly promote complementary performance in decision-making settings. Of course, not every task naturally lends itself to distinct human-AI strengths. Nonetheless, future work could generalize our approach by tailoring explanations to optimize each user's ability to verify AI outputs [12], thereby increasing the likelihood of complementarity across a broader range of tasks and domains. Taking a more performance-centric approach to evaluating personalized explanations may ultimately enable a more rigorous and systematic understanding of when and why personalization enhances human-AI collaboration.

The Role of Stakes. The stakes in our experiment were relatively low compared to those in many real-world decision-making domains, such as medicine or criminal justice. Participants could earn up to $1.70 in base pay and an additional $1.00 in performance-based

bonuses. This payment structure is common in studies of AI-assisted decision making [4], but it raises questions about how our findings might generalize to contexts where users bear greater responsibility or face higher consequences for their decisions. To discourage minimal engagement, we required participants to pass a comprehension test with perfect accuracy and imposed minimum time thresholds per round and per survey question. Nonetheless, given the small per-round rewards (approximately five cents), we cannot be certain that participants invested the same level of cognitive effort as individuals would in higher-stakes settings.

That said, many real-world applications of explainable AI remain inherently low-stakes, where personalization could meaningfully improve user experience without major ethical or social risks. For example, personalized explanations could enhance everyday tasks such as music recommendation [34] or adaptive learning systems [8]. While errors in these domains rarely carry serious consequences, incremental improvements across repeated interactions can accumulate into substantial benefits in efficiency and user satisfaction. By contrast, in high-stakes settings, personalized AI explanations designed to optimize performance may introduce additional ethical challenges—such as concerns about fairness, transparency, or differential access to information—that warrant careful consideration before deployment.

Measuring User Traits. Our study relied on self-reported user characteristics collected through standardized questionnaires assessing personality-related traits and questions assessing relevant experience. Self-report measures are widely used in both psychology and human–computer interaction research, but they can introduce biases in participant grouping. For instance, participants who report higher levels of travel experience may overestimate their actual exposure, potentially influencing interpretation of our findings. Alternative approaches exist for inferring user traits more implicitly [2, 26], though these typically require substantial time and resources. In contrast, short self-report questionnaires enable efficient, large-scale data collection from a diverse participant pool, such as those recruited via Prolific. While self-reports are imperfect proxies for the underlying traits of interest, our results indicate that they nonetheless provide sufficiently strong signals to detect the interaction effects under investigation. Moreover, their practicality and scalability make them appealing for real-world applications that seek to personalize AI systems based on user characteristics.

Other Limitations. Our results show that participants with different personal characteristics may benefit from different types of explanations. However, as is common in human-subject research, our findings are shaped by the specific design choices of our experiments. Below, we discuss additional limitations and considerations for generalizing our results.

First, while our experiments spanned two different tasks in distinct domains, further research is needed to test the robustness of these findings across a broader range of settings. In particular, the literature would benefit from studies that simultaneously examine multiple domains and user subgroups to establish when and how complementarity and interaction effects reliably emerge. The geography-guessing task was deliberately chosen based on two criteria: (1) the AI system holds superior task-solving abilities due to its extensive knowledge base, while (2) human participants possess additional perceptual information that the AI lacks, and the AI's explanations offer verifiability [12]. We conjecture that our findings are most likely to generalize to tasks with similar properties and that these characteristics could serve as useful guidelines for designing future studies of complementarity. To further validate the generalizability of our results, future studies could explore tasks in which humans and AI perform roughly equally well without assistance, or in which humans outperform their AI counterparts. Future work could also examine the impact of overlap between average participants' task knowledge and the AI's, as we suspect this overlap is substantial in sentiment analysis but much smaller in geography guessing.

Second, our experimental design necessarily constrained the range of personal characteristics we could examine, leaving other potentially relevant traits unexplored—such as additional Big Five dimensions, demographic factors, and cognitive or physical disabilities. Furthermore, for institutional review board compliance, our participant pool was limited to U.S. residents recruited via Prolific, which may limit generalizability. Expanding future studies to include a wider variety of explanation types, cultural contexts, and populations would strengthen the external validity of this line of research. Future research could also analyze personalizing different features of the explanations beyond density and modality, such as structure or uncertainty cues.

Finally, our study used ChatGPT to generate AI predictions and explanations in the geography-guessing task. While this approach is increasingly common, we acknowledge that our findings are based on a model that continues to evolve rapidly, and human–AI interaction dynamics may shift accordingly. Likewise, although LIME remains a foundational tool for generating localized explanations in sentiment-analysis tasks, newer methods and variants may yield different outcomes. Replicating our study with evolving models and explanation techniques would help clarify which effects persist across technological advances.

Future Work. Building on the above discussion, future work should examine the generalizability of our findings to additional tasks with varying levels of stakes, such as medical diagnosis or criminal recidivism prediction. Another promising direction is to manipulate the verifiability of explanations to better understand their role in supporting performance and moderating interaction effects. For example, future studies could employ tasks such as maze solving [56], restrict the human's field of view to match that of the AI, or limit the information accessible to the AI to systematically vary the balance of knowledge between the two. Further research could also expand the range of user characteristics examined. Our primary goal was to demonstrate that personalized explanations can influence performance, and secondarily, that such personalization may enhance complementarity. A deeper understanding of which specific traits most strongly drive these effects would help guide the design of more effective personalization strategies. Finally, future work could explore personalization along additional axes beyond those studied here. We focused on explanation modality and explanation length, given their prominence in the literature, but other dimensions—such as tone, depth, or informational content—may also meaningfully affect how users engage with AI explanations. Understanding how these factors interact with individual traits will be essential for developing truly adaptive and human-centered explainable AI systems.

6 GenAI Usage Disclosure

We did not use generative AI to write the content of this paper. Generative AI tools were used to assist with coding the websites used to host the experiments, to generate the AI explanations for the geography-guessing task as described in the paper, and to provide proofreading and minor language refinement during manuscript preparation.

References

[1] Nehal Mohamed Ali, Marwa Mostafa Abd El Hamid, and Aliaa Youssif. 2019. Sentiment analysis for movies reviews dataset using deep learning models. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol* 9 (2019).

[2] Danny Azucar, Davide Marengo, and Michele Settanni. 2018. Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and individual differences* 124 (2018), 150–159.

[3] Maya Balakrishnan, Kris Johnson Ferreira, and Jordan Tong. 2025. Human-Algorithm Collaboration with Private Information: Naïve Advice-Weighting Behavior and Mitigation. *Management Science* (2025).

[4] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. doi:10.1145/3411764.3445717

[5] 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

[6] 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.

[7] John T Cacioppo and Richard E Petty. 1982. The need for cognition. *Journal of personality and social psychology* 42, 1 (1982), 116.

[8] 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.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[10] 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.

[11] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? Evaluating achine learning interpretations in cooperative play.. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. doi:10.1145/3301275.3302265

[12] Raymond Fok and Daniel S Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* (2024).

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

[14] 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.

[15] Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. 2025. The Value of Information in Human-AI Decision-making. *arXiv preprint arXiv:2502.06152* (2025).

[16] 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.

[17] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2025. Complementarity in human-AI collaboration: Concept, sources, and evidence. *European Journal of Information Systems* (2025), 1–24.

[18] 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

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

[20] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–20.

[21] John Wesley Hostetter, Cristina Conati, Xi Yang, Mark Abdelshiheed, Tiffany Barnes, and Min Chi. 2023. XAI to increase the effectiveness of an Intelligent Pedagogical Agent. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*. 1–9.

[22] 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.

[23] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology* (1991).

[24] Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational psychology review* 19, 4 (2007), 509–539.

[25] 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

[26] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.

[27] 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

[28] 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

[29] 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.

[30] 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

[31] 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).

[32] 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.

[33] 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.

[34] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th international conference on intelligent user interfaces*. 397–407.

[35] 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

[36] T Nathan Mundhenk, Barry Y Chen, and Gerald Friedland. 2019. Efficient saliency maps for explainable AI. *arXiv preprint arXiv:1911.11293* (2019).

[37] 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.

[38] 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.

[39] Allan Paivio, James M Clark, et al. 2006. Dual coding theory and education. *Pathways to literacy achievement for high poverty children* 1 (2006), 149–210.

[40] Saeed Mian Qaisar. 2020. Sentiment analysis of IMDb movie reviews using long short-term memory. In *2020 2nd International Conference on Computer and*

Information Sciences (ICCIS). IEEE, 1–4.

[41] 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

[42] 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

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[44] Gavriel Salomon. 1972. Heuristic models for the generation of aptitude-treatment interaction hypotheses. *Review of Educational Research* 42, 3 (1972), 327–343.

[45] 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

[46] Andrew Silva, Pradyumna Tambwekar, Mariah Schrum, and Matthew Gombolay. 2024. Towards balancing preference and performance through adaptive personalized explainability. In *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*. 658–668.

[47] Richard Snow. 1989. Aptitude-Treatment Interaction as a Framework for Research on Individual Differences in Learning. In *Learning and Individual Differences*, Philip Ackerman, Robert J. Sternberg, and Robert Glaser (Eds.). W.H. Freeman, New York, 13–59.

[48] Harishankar V Subramanian, Casey Canfield, and Daniel B Shank. 2024. Designing explainable AI to improve human-AI team performance: a medical stakeholder-driven scoping review. *Artificial Intelligence in Medicine* 149 (2024), 102780.

[49] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation*. Vol. 55. Elsevier, 37–76.

[50] John Sweller, Paul Ayres, and Slava Kalyuga. 2011. The expertise reversal effect. In *Cognitive load theory*. Springer, 155–170.

[51] Jay M Teets, David P Tegarden, and Roberta S Russell. 2010. Using cognitive fit theory to evaluate the effectiveness of information visualizations: An example using quality assurance data. *IEEE transactions on visualization and computer graphics* 16, 5 (2010), 841–853.

[52] Xian Teng. 2024. *Discoverability and interpretability of spurious associations in data-driven decisions*. Ph. D. Dissertation. University of Pittsburgh.

[53] Kamil Topal and Gultekin Ozsoyoglu. 2016. Movie review analysis: Emotion analysis of IMDb movie reviews. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 1170–1176.

[54] 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

[55] Sandesh Tripathi, Ritu Mehrotra, Vidushi Bansal, and Shweta Upadhyay. 2020. Analyzing sentiment using IMDb dataset. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 30–33.

[56] 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.

[57] Iris Vessey. 1991. Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision sciences* 22, 2 (1991), 219–240.

[58] Josephine Zerna, Christoph Scheffel, Corinna Kührt, and Alexander Strobel. 2023. Need for Cognition is associated with a preference for higher task load in effort discounting. *Scientific Reports* 13, 1 (2023), 19501.

[59] 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.

Technical Appendix

A Multiple Linear Regression Extended Results

Tables 7 to 14 are the full readouts for our multiple linear regressions for each user trait in both tasks.

Table 7: Multiple Linear Regression Extended Results for Openness (m3) in the sentiment analysis task.

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.665	0.049	0.568	0.762	< .001
Openness	0.003	0.006	−0.009	0.015	.624
Long Explanation	0.050	0.064	−0.076	0.175	.435
Interaction	−0.004	0.008	−0.020	0.012	.612

Note. $N = 186$. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 8: Multiple Linear Regression Extended Results for NFC (1003) in the sentiment analysis task.

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.631	0.041	0.550	0.712	< .001
Need for Cognition	0.015	0.010	−0.006	0.036	.150
Long Explanation	0.141	0.059	0.025	0.257	.018
Interaction	−0.032	0.015	−0.061	−0.002	.034

Note. $N = 186$. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 9: Multiple Linear Regression Extended Results for Experience (m1) in the sentiment analysis task.

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.626	0.057	0.513	0.738	< .001
Experience	0.008	0.007	−0.006	0.022	.260
Long Explanation	0.110	0.073	−0.035	0.255	.136
Interaction	−0.012	0.009	−0.030	0.006	.205

Note. $N = 186$. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 10: Multiple Linear Regression Extended Results for Experience (m1_high) in the sentiment analysis task with thresholding into bins of (1,2,3) and (4,5). (Exploratory.)

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.676	0.012	0.652	0.699	< .001
Experience (High)	0.037	0.020	−0.003	0.077	.068
Long Explanation	0.038	0.016	0.006	0.070	.021
Interaction	−0.060	0.028	−0.116	−0.004	.036

Note. $N = 186$. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 11: Multiple Linear Regression Extended Results for Openness in the geography-guessing task.

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.667	0.048	0.573	0.762	< .001
Openness	0.018	0.012	−0.006	0.042	.137
Treatment	0.100	0.073	−0.043	0.243	.169
Interaction	−0.035	0.019	−0.071	0.002	.063

Note. $N = 3448$. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 12: Multiple Linear Regression Extended Results for NFC in the geography-guessing task.

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.701	0.054	0.595	0.806	< .001
Need for Cognition	0.009	0.013	−0.017	0.035	.483
Treatment	−0.034	0.081	−0.193	0.125	.675
Interaction	0.001	0.020	−0.039	0.041	.978

Note. $N = 3448$. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 13: Multiple Linear Regression Extended Results for Experience in the geography-guessing task.

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.788	0.026	0.738	0.839	< .001
Experience	−0.022	0.010	−0.042	−0.002	.032
Treatment	−0.081	0.037	−0.154	−0.008	.030
Interaction	0.021	0.014	−0.007	0.048	.139

Note. N = 3448. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

Table 14: Multiple Linear Regression Extended Results for Experience in the geography-guessing task with thresholding into bins of (1,2,3) and (4,5). (Exploratory.)

			95% CI		
Effect	Estimate	*SE*	*LL*	*UL*	*p*
Intercept	0.748	0.011	0.726	0.770	< .001
Experience (Binned)	−0.062	0.027	−0.114	−0.009	.022
Treatment	−0.049	0.017	−0.083	−0.015	.005
Interaction	0.084	0.037	0.010	0.157	.026

Note. N = 3448. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit.

B Additional Experiment Details

B.1 Extended Results of Group-By-Group Performance in the Sentiment Analysis Task

Below are full results of performance for each group studied in the sentiment analysis task, including control groups. We find that low-NFC users are the only subgroup that achieves complementarity.

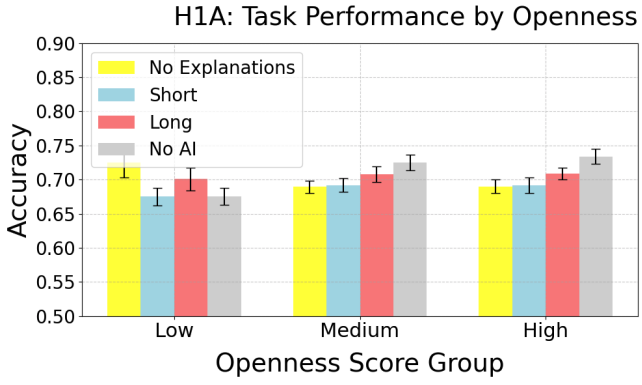


Figure 9: Results for all four experimental groups in the sentiment analysis task, broken down by Openness.

The AI's performance sits at sixty-five percent for this task.

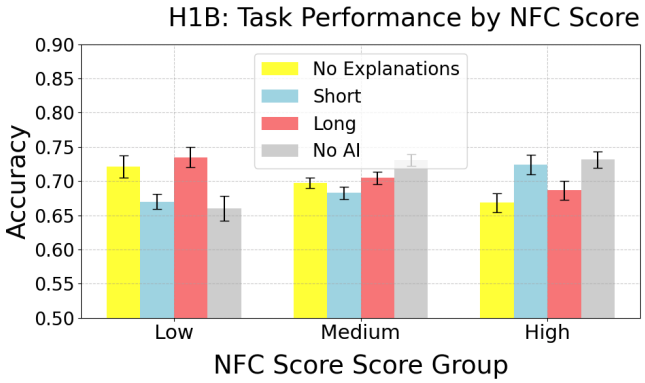


Figure 10: Results for all four experimental groups in the sentiment analysis task, broken down by NFC score.

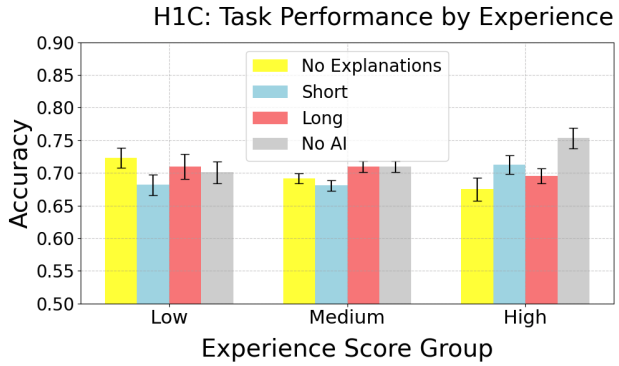


Figure 11: Results for all four experimental groups in the sentiment analysis task, broken down by Experience score.

B.2 Extended Results of Group-By-Group Performance in the Geography-Guessing Task

Below are full results of performance for each group studied in the geography-guessing task, including control groups. We find that users with no AI assistance consistently underperform every other subgroup.

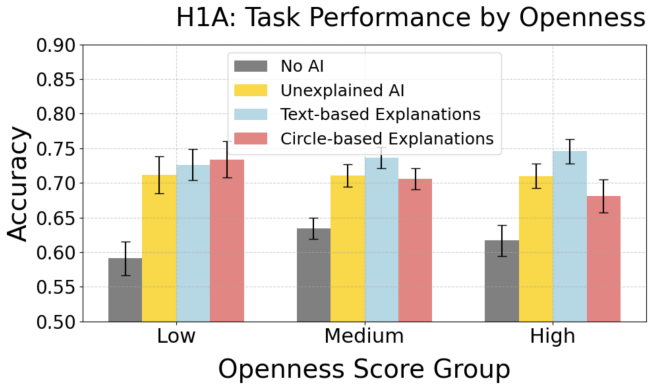


Figure 12: Results for all four experimental groups in the geography-guessing task, broken down by Openness.

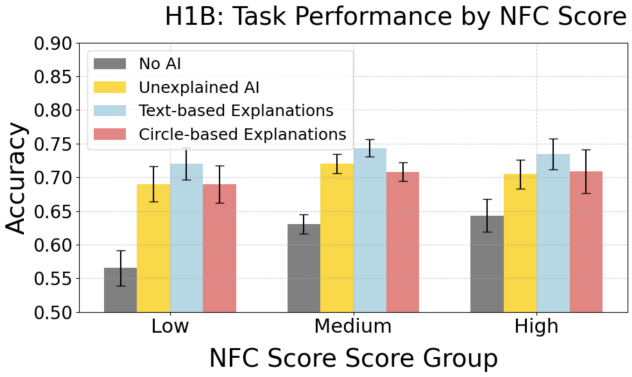


Figure 13: Results for all four experimental groups in the geography-guessing task, broken down by NFC score.

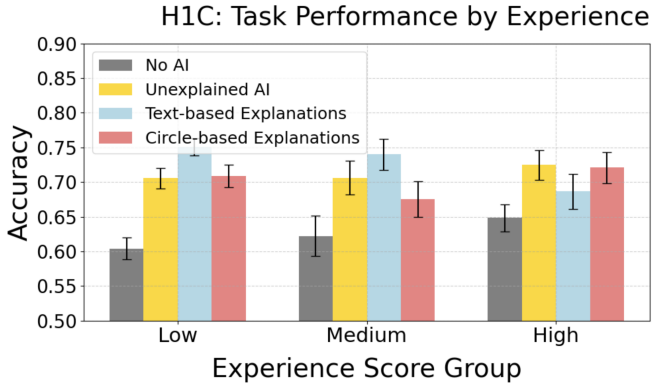


Figure 14: Results for all four experimental groups in the geography-guessing task, broken down by Experience score.

B.3 Procedure of Generating AI Predictions and Explanations for the Geography-Guessing Task

For the geography-guessing task, 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 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.

C Exploratory Analysis: Trust and Reliance in the Geography Guessing Task

Beyond the pre-registered analysis, we also examined traditional metrics of subjective trust and reliance through an exploratory analysis in the geography-guessing task to dig deeper into the origins of the complementarity we observed. For subjective trust, we asked participants at the end of the experiment to rate their trust in the AI using a 5-point Likert scale in response to the statement: "I generally trust the AI's recommendations," ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). For reliance, we measured the ratio of participant answers aligning with the AI assistance. We analyzed whether or not individuals with certain user traits provided with certain AI assistance (unexplained, text-based explanations, visual-based explanations)[4] had significantly different levels of subjective trust or reliance compared to the population as a whole.

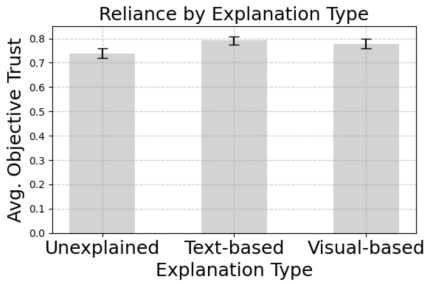


Figure 15: Participant reliance by AI explanation type.

The full set of results is included below D. Overall, most of the comparisons we made are not statistically significant after correction for multiple comparisons. Below, we describe three findings that are statistically significant after corrections First, as shown in Figure 15, we found that participants given visual-based explanations had significantly higher reliance on the AI as a whole compared to participants not given explanations at all (p = .0232 after correction for multiple comparisons). This is despite the fact that both sets of participants performed roughly equally well on the task as a whole. This highlights that complementarity and reliance, while related, capture different aspects of AI-assisted decision making.

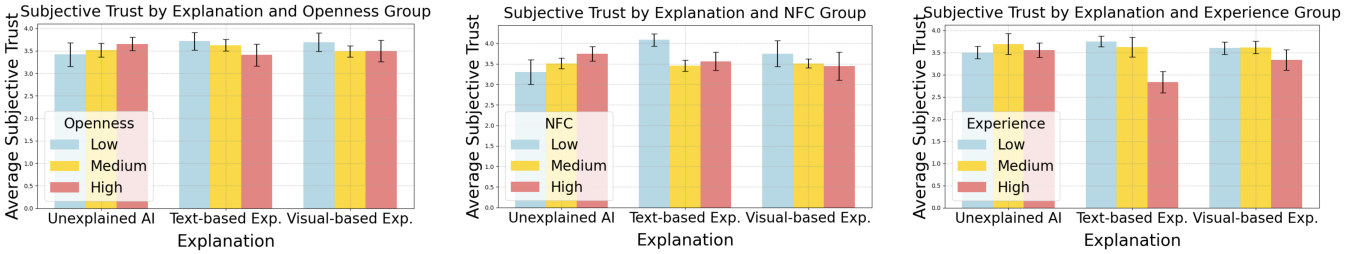
Second, as shown in Figure 16c, we found that participants with high experience reported similar levels of trust as their peers when using unexplained AI, but lower trust when using visual-based explanations—and much lower trust when using text-based explanations (p = 0.02268). This may help explain why high-experience participants performed relatively poorly with text-based explanations: their travel knowledge may have made them more aware of the information the AI was missing, leading them to trust it less, regardless of its correctness. However, further experiments are needed to more definitively identify the root causes of this phenomenon.

Third, as demonstrated in Figure 16b, participants with low need for cognition (NFC) subjectively trusted the text-based explanations at a rate that exceeded the population average (p = .0144). We conjecture that this may be due to the relative ease of use for the text-based explanations; the AI in this condition more clearly spells out exactly why it made the decisions it did, requiring less cognitive effort from the user for interpreting the explanations. Interestingly, this increased subjective trust did not manifest in higher reliance or any significantly different performance. Further experiments are necessary to more concretely identify the root cause of this phenomenon.

D Additional Figures for the Exploratory Analysis on the Geography-Guessing Task

For completeness, we include the additional figures from the above exploratory analysis. The main results that directly address our research questions are illustrated using Figures 17 to 24. We only included the p values when the result is statistically significant after corrections for multiple comparisons. Specifically, we examined our results using the metrics of subjective trust and reliance. For subjective trust, we asked participants at the end of the experiment to rate their trust in the AI using a 5-point Likert scale in response to the statement: "I generally trust the AI's recommendations," ranging from "Strongly Disagree" (1) to "Strongly Agree" (5). For reliance, we measured the ratio of participant answers aligning with the AI assistance. We analyzed whether individuals with specific user traits, when provided with different forms of AI assistance (unexplained, text-based explanations, or visual-based explanations), exhibited significantly different levels of subjective trust or reliance compared to the overall population. Specifically, we conducted two-sided t-tests for each explanation format and user characteristic pairing, comparing their trust and reliance scores to the overall population averages, and applied a Bonferroni correction for multiple comparisons.

[4]Given AI is not mentioned in the control, we did not measure the subjective trust and reliance on AI in the control.



(a) Participant subjective trust by AI explanation format and participant openness groups.

(b) Participant subjective trust by AI explanation format and participant NFC groups.

(c) Participant subjective trust by AI explanation format and participant travel experience groups.

Figure 16: Subjective trust by user trait and explanation modality for the geography-guessing task.

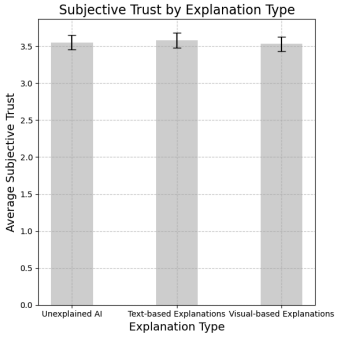


Figure 17: Subjective trust in the AI by explanation type. No group has statistically higher or lower subjective trust than the overall average trust.

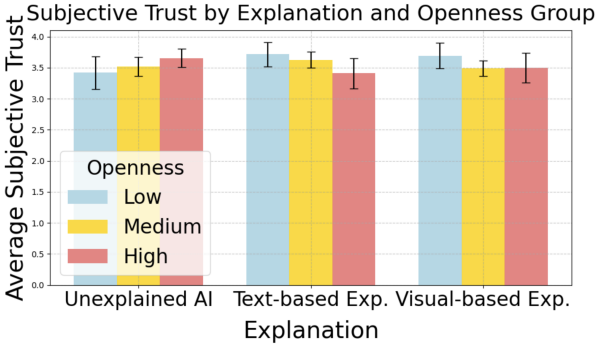


Figure 18: Subjective trust in the AI by explanation type and Openness level. No group has statistically higher or lower subjective trust than the overall average.

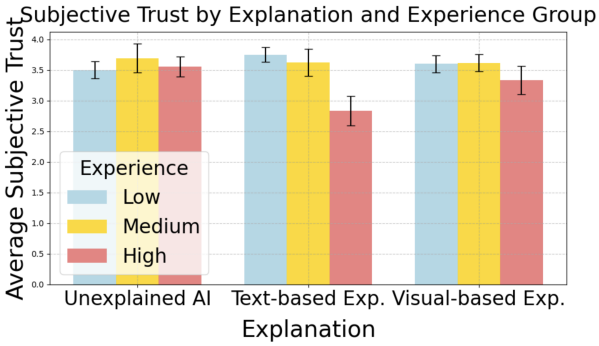


Figure 19: Subjective trust in the AI by explanation type and Experience level. High-experience individuals receiving text-based explanations subjectively trusted AI more when explanations were provided, particularly text-based explanations (p = 0.02268)

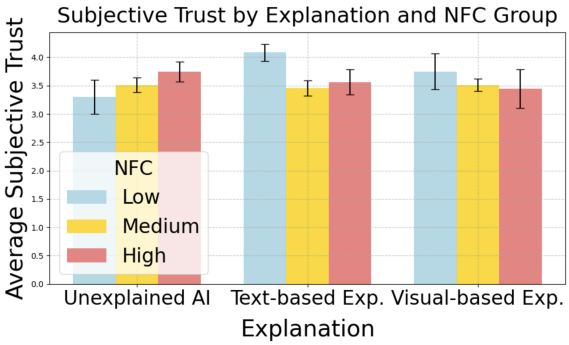


Figure 20: Subjective trust in the AI by explanation type and NFC level. Low-NFC individuals receiving text-based explanations subjectively trusted AI more when explanations were provided, particularly text-based explanations (p = .0144)

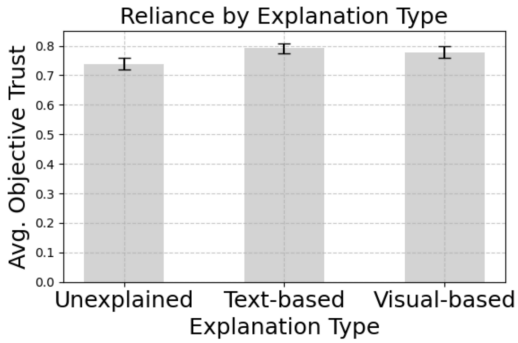


Figure 21: Reliance in the AI by explanation type. Individuals given unexplained AI recommendations were less likely overall to rely on the AI than those given explanations.

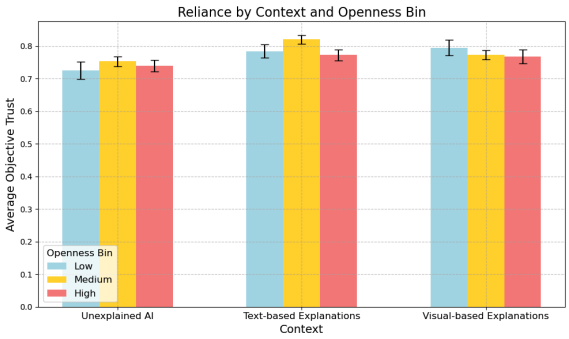


Figure 22: Reliance in the AI by explanation type and Openness level.

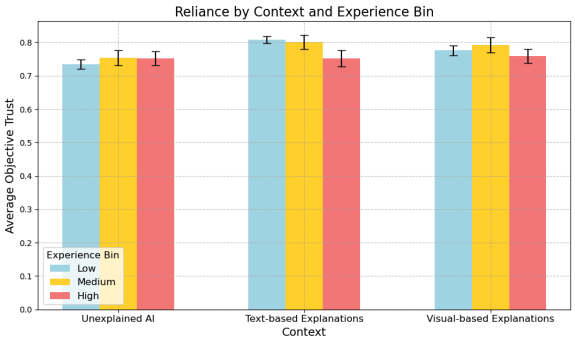


Figure 23: Reliance in the AI by explanation type and Experience level.

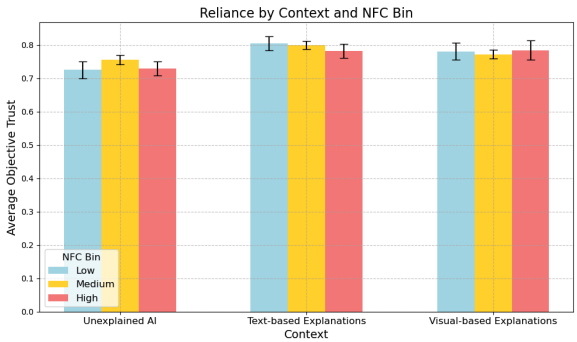


Figure 24: Reliance in the AI by explanation type and NFC level.

E Survey Design for Measuring User Traits

Below are the questions for the surveys administered to participants of the Sentiment Analysis and Geography-guessing tasks. Each survey features 2 questions about relevant task experience, 1 question about Need for Cognition, and 2 questions about the Openness personality trait pulled from the BFI-10 [42], as specified in the linked pre-registrations. Users were required to spend a minimum of 5 seconds per question and had to submit responses to all 5 questions to advance in the experiment. Survey questions for the sentiment analysis task:

(1) I have read movie reviews before and am generally knowledgeable about what may cause a movie to get a high or low score.
(2) I watch a substantial amount of movies.
(3) I enjoy solving challenging puzzles or complex problems.
(4) I have few artistic interests.
(5) I have an active imagination.

Survey questions for the geography guessing task:

(1) I have extensive experience with geography-guessing games (ex. Geoguessr).
(2) I consider myself well-traveled and/or familiar with geography outside the USA.
(3) I enjoy solving challenging puzzles or complex problems.
(4) I have few artistic interests.
(5) I have an active imagination.