Explain it Like I'm Me: User Characteristics in XAI Decision-Making

Malik Khadar*
malik@umn.edu
University of Minnesota
Minneapolis, USA

Amoligha Timma*
timma062@umn.edu
University of Minnesota
Minneapolis, USA

Luka Ludden
ludde017@umn.edu
University of Minnesota
Minneapolis, USA

Harmanpreet Kaur
harmank@umn.edu
University of Minnesota
Minneapolis, USA

Abstract

Explainable Artificial Intelligence (XAI) provides tools to make the behavior of AI models more interpretable, but these tools see misuse in practice. Findings are mixed on whether XAI should be personalized via characteristics such as demographics, personality, and prior experience to address the misuse. We holistically studied the effect of these characteristics on XAI use in an experimental setting (N=149) in a manner orthogonal to recent prior work, engaging with the limitations outlined in that work. While the linear effects of separate categories of characteristics yielded similarly scant results, our exploratory and qualitative analyses revealed rich insights leading us to question the limited measurement approaches conventional to this line of research. As such, we present this work as a first step toward the more holistic, rigorous measurement of user characteristics as they relate to XAI, outlining how future work may extend even further beyond the limitations that have thus far diffused the community's collective research effort.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

Keywords

explainable AI, personalization, user characteristics, decision-making

ACM Reference Format:
Malik Khadar, Amoligha Timma, Luka Ludden, and Harmanpreet Kaur. 2025. Explain it Like I'm Me: User Characteristics in XAI Decision-Making. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732116

*Both authors contributed equally to this paper.

FAccT '25, June 23–26, 2025, Athens, Greece

ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732116

1 Introduction

Explainable AI (XAI) approaches are essential as we deploy AI and ML models in sensitive domains such as criminal justice [2], healthcare [52], and finance [20]. Many models are inherently uninterpretable "black boxes" [23] that fail to meet the expectations of transparency and accountability demanded in critical decision-making contexts, an issue that is exacerbated by the tendency for such models to perpetuate historical biases [2, 17]. In light of this issue, XAI tools provide explanations of how black box models arrive at the outputs they produce. SHapley Additive exPlanations (SHAP) [40] and Local Interpretable Model-Agnostic Explanations (LIME) [57] have emerged as the most popular of these tools, both being model-agnostic and explaining how input features contribute to model outputs.

However, a problematic gap persists between the intended and actual use of XAI tools. Prior work has consistently discovered instances of over-trust, inappropriate reliance, and a general lack of calibration when XAI tools are tested in practice, across multiple domains, contexts, and stakeholders [6, 9, 33]. Even data scientists and ML practitioners, i.e., people with ML experience, have exhibited misuse of these tools and a failure to understand them [34]. So long as this gap exists, XAI will be unable to accomplish its goal of promoting effective collaboration between people and AI.

Understanding this gap between intended and actual XAI use is perhaps the most critical problem to address for this research area. There have been two threads of research that seek to do so: 1) technical and design work grounded in theories from related disciplines (e.g., cognitive [39], social [45], and organizational [32] sciences) which are, for example, responsible for the introduction of counterfactuals in XAI [75]; and 2) experimental work seeking to investigate various human factors that affect human-AI decision-making and XAI tool use in practice. The latter work has classified user characteristics that may impact the use of XAI tools and decision-making outcomes based on cognitive (e.g., prior experience, literacy) and social (e.g., demographic, personality) factors.

Instead of adding clarity, experimental work on XAI use in practice has found mixed results on whether, which, and how user characteristics have an impact. For instance, studies have found evidence supporting [56] and refuting [68] the impact of gender on XAI interaction. Some studies found that neither domain knowledge nor knowledge about AI impacted performance in XAI decision-making tasks [33, 34, 37], while other work claims otherwise [16].

While results have indicated the significant impact of user characteristics on XAI interaction in pedagogical [13] and entertainment [43] domains, it is unclear whether these results translate from learning-oriented and satisfaction-oriented XAI use into the critical decision-making contexts that increasingly see AI integration.

Therefore, we experimentally studied user characteristics in a holistic way. Recent prior work that took a similar approach did not find any significant relationships between user characteristics and engagement with an XAI tool, warning of the potential "rabbit hole of personalization" [49]. These results were unexpected given the large history of prior work on personalization and real-world applications that rely on it (e.g., recommender systems, algorithmic curation). As such, we sought to verify their findings in a related decision-making setting. We were able to engage with the limitations outlined in their work [49] while designing our study to further test the efficacy of their findings.

We conducted an experiment where we measured demographic, personality, and AI experience characteristics of 149 participants before having them engage in an AI-assisted decision-making task. The key differences in methodology between our work and that which we build upon [49] are that we took the opportunity to gather more precise measures using validated scales and further explored the nature of our participants' reliance on our AI system by having them complete the same task both with and without its assistance.

As with prior work, the linear modeling of user characteristics failed to provide clarity to this research area, but the same was not true for our exploratory and qualitative analyses. Our non-linear models yielded intuitive results such as conscientiousness predicting changes in confidence and an interaction effect between neuroticism and hours of ML/XAI experience predicting changes in how participants trust the AI system. Our qualitative analysis revealed the key role that intuition plays in guiding XAI tool use, where this intuition is derived from a variety of sources including self-confidence, prior experience, and openness to new technology. In this work, we extend beyond the norms of measurement for this line of research and correspondingly extend beyond the limitations that prevent us from capturing the complexity of user characteristics, concluding with strategies for the research community to pursue holistic and rigorous measuring practices.

2 Related Work

2.1 Interpretability, Explanations, and AI Reliance

Many machine learning models are not inherently interpretable. In the ML context, interpretability refers to a model's "ability to explain or to present in understandable terms to a human" [14]. The ethical usage of such models is determined in part by the ability for their behavior to be understood [72], with this need being exacerbated in critical decision-making contexts such as health care [52] and criminal justice [2]. While some models are designed to be inherently interpretable *glass-boxes* (e.g. Generalized Additive Models[24], simple point systems [31], and decision trees[54]), there are many models so complicated that they require external explanations for users to understand their behavior. Those in the latter class are called *black-box* models [23]. An ongoing debate in the field of XAI considers whether black box models should be used in critical settings at all [60].

Explainability tools serve to increase understanding of black-box models. One popular post-hoc explainability tool is SHapley Additive exPlanations (SHAP), which leverages Shapley Values from cooperative game theory to explain how input feature values correspond to model outputs at both a local and global level [40]. Another commonly used post-hoc approach, Local Interpretable Model-Agnostic Explanations (LIME), relies on input perturbations to explain model behavior [57]. Most explanation approaches provide local explanations showing how a specific model prediction was made, and global explanations showing the model's behavior across all datapoints (e.g., as global feature importances and dependence plots). Both SHAP and LIME are model agnostic, i.e., they can generate explanations for any model. For a comprehensive overview of interpretability and explainability approaches, see reviews by Gilpin et al. [19], Arrieta et al. [3] Liao and Varshney [38], and Dwivedi et al. [15].

Explanations are often misused in practice and result in over-reliance on models. Prior work has found that end users and even data scientists rely on the presence of explanations to *justify* AI outputs rather than *scrutinize* them [16, 34], an issue exacerbated by information overload [36]. Prior work asserts that we have "inmates running the asylum," where explanations are designed by and for AI researchers without accommodating end users [46]. In an attempt to promote appropriate reliance on explanations, researchers have taken inspiration from human-to-human explanation techniques, drawing from theories in the social sciences to explore how explanations should be delivered and what they should convey [32, 39, 42, 45]. Another thread of research seeking to mitigate this issue investigates how user characteristics factor into XAI use, laying the groundwork for personalized XAI tailored to user needs (see below).

2.2 XAI Personalization

Prior work has studied user characteristics such as age, gender, prior experience with ML and XAI, AI literacy, and personality, and their role in people's understanding and processing of AI explanations. We categorize these user characteristics and present an overview of their (often contradictory) findings.

2.2.1 Demographics. In the personalization of XAI, the presentation and design of explanations are tailored to the user, often affecting the resulting explainability [3, 13]. Age is typically collected during research studies, and it has been found that older people, though not elderly, are more inclined to appropriately review and comprehend XAI information [29, 49]. Similarly, Reeder et al. [56] find that gender has a significant effect on user comprehension and preference for explanations, as an interaction effect with prior experience in AI/ML; though results from [68] contradict this. Gender has also been shown to be significant in the context of interacting with AI and recommender systems more generally [35].

2.2.2 Experience. The two main dimensions of experience in XAI studies are prior experience in AI/ML and domain knowledge. Here, again, we find prior work with contradictory results. Experimental work by [16, 56, 59, 73] suggests that prior experience in AI/ML can improve appropriate reliance on model and explanation outcomes;

similar studies presented in [6, 33, 34, 37] suggest otherwise. Similar contradictions arise regarding domain knowledge, where some work found it had a positive impact on XAI outcomes [11, 47] while others found counterintuitively null results [37]. Prior experience is considered significant in the realm of personalization and recommender systems, where those with prior experience possess the domain knowledge necessary to complete tasks efficiently [41]. Furthermore, different expertise levels have been shown to influence comprehension of visual, textual, and hybrid explanations [47, 66].

2.2.3 Personality. An emerging user characteristic studied in the context of personalized XAI is personality. Personality has been known to affect perception and trust in AI systems [10, 59, 70]. Furthermore, changes in the presentation of AI outputs can increase perception and trust values [27, 62], with perception and trust being correlated with specific personality traits of the user [5, 13, 74]. The Big Five personality traits—extraversion, neuroticism, conscientiousness, agreeableness, and openness—shape user preferences and are essential for personalization [74]. Specifically, users' preferences for explanations may vary based on their individual differences, such as openness and neuroticism. For example, individuals who are more open to new experiences may find innovative AI explanations easier to trust, while those with higher levels of neuroticism may experience greater uncertainty in trusting AI systems [5]. However, there has been limited research on how human personality itself impacts how XAI is utilized beyond measuring trust in a decision-making setting [13, 49]. This limited research is contradictory, and is most closely related to our motivation and work.

2.2.4 Combined Models of User Characteristics. Stein et al. [64] introduced the ATTARI-12 questionnaire, a validated tool to measure general attitudes toward AI. This work highlights significant relationships between personality traits and attitudes toward AI. For instance, individuals with higher agreeableness and younger age demonstrated more positive attitudes toward AI, while those with a greater susceptibility to conspiracy beliefs showed more negative attitudes. This study provides evidence of the importance of personality in shaping perceptions of AI, though it focuses on general attitudes rather than task-specific interactions with AI explanations.

On the other hand, Nimmo et al. [49] present an experiment showing user characteristics like personality, demographic features, and prior experience do not impact XAI use or outcomes. In this study, users were given twenty minutes to classify as many hate speech comments as possible with the help of an AI agent. Prior to the task, the user's personality was calculated using the Ten-Item Personality Inventory (TIPI), a shorter version of the Big Five Inventory (BFI) [30], the original 44 item questionnaire. They also rely on subjective ratings of prior experience and trust, though no validated scales are used for these. Nimmo et al. [49] carefully consider these limitations of measurement already known in prior work (e.g., [21, 58]), but ultimately suggest that user characteristics might be a "rabbit hole of personalization."

We build on this work by: (1) using validated scales to measure all facets of user characteristics, with supplemental use of subjective ratings; and (2) measuring outcome metrics in a human-only task baseline before having people re-do the task with AI, allowing us to measure the change in behavior when using AI and directly attribute it to user characteristics.

3 Methods

We conducted a two-part study to holistically measure and model the impact of user characteristics on XAI outcomes. First, participants completed an intake survey answering questions about user characteristics—demographics, prior experience, and personality—on validated scales [25, 30]. Participants who answered the ML literacy objective questionnaire in this intake with at least 50% accuracy were then invited to complete the second survey, a classification task. For this main task, they predicted whether 8 data points from the Adult Income census dataset [7] made more than $50k annually. Participants first completed the task unassisted and then repeated it with the help of an ML model and XAI tool. Our setup and measures are described in detail below; Figure 1 provides an overview. Our study design was approved by the University of Minnesota's Institutional Review Board (IRB).

3.1 Dataset and Datapoint Selection

We used the Adult Income dataset [7] for the main study task, a classification dataset based on 1994 census data used to predict whether a person makes over $50k a year. The dataset has 48,842 instances and 14 features including age, work class, education, marital status, occupation, race, gender, capital gain, capital loss, hours per week, native country, relationship, fnlgwt, and education num. Due to duplicate information and un-interpretable columns, relationship, fnlgwt, and education num were dropped from the dataset (this is consistent with prior work that has used this dataset [26, 34]).

We selected 8 datapoints from this dataset for our study. These specific data points were selected for two reasons. First, they were near the decision boundary, i.e., points that were inherently ambiguous for the ML model. This helped us capture people's reliance on the ML model without being confounded by its accuracy. Second, of the many decision boundary datapoints, these 8 had meaningful XAI explanation narratives (e.g., in one case, the sole deciding factor was being in a single-parent household). We anticipated that these narratives would be more engaging for task participation. In terms of model accuracy on these datapoints, half of the predictions (4/8) were accurate. We selected datapoints to have a balanced representation of prediction classifications and accuracy, selecting pairs of false positives, false negatives, true positives, and true negatives.

3.2 ML Model and Explanation Approach

We trained a LightGBM classifier to predict whether an individual makes over $50k a year. For modeling, categorical features were preprocessed through one-hot encoding, and a binary log-loss function was used to optimize the model at each iteration. We selected LightGBM for its efficiency, ability to handle categorical data, and robust performance on structured datasets. The model reached an accuracy of ~88% on both the training and test sets, leaving room for participants to question its classifications.

We chose SHapley Additive exPlanations (SHAP) [40], a widely used post-hoc method for interpreting ML outputs, to explain our model. The use of SHAP allows for direct comparison to the abundance of prior work that utilizes it [4, 33, 48, 55], and its visuals are common across most XAI tools (e.g., bar plots for global and local explanations, scatter plots to represent dependence), increasing the

generalizability of our findings. We used the Python implementation of SHAP to generate our explanations. For each of our 8 datapoints, a local SHAP waterfall plot was created to visualize the predicted f(x) value and how different features contributed to the classification of income as either less than or greater than \$50k. These visualizations provided a breakdown of positive and negative contributions of the features to the predicted outcome, allowing participants to identify key factors influencing each prediction. An example of the SHAP waterfall plot can be found in Appendix D.5.

3.3 Study Task

The study consisted of two phases: an intake survey for capturing user characteristics and a main study survey for the experiment, both detailed below. Participants completed these phases individually since the intake survey measured our inclusion criteria. Both surveys were designed using Qualtrics and administered on Prolific.

3.3.1 Intake Survey. After providing consent, participants provided three sections of intake information related to user characteristics. The first section gathered demographic information including gender, age, and education. Additionally, questions were asked regarding: the participant's occupation and whether it included ML tasks; their proficiency in ML and interpretability tool use; how much time (estimated) they spent on ML and interpretability tasks. The second section assessed the participants' personality traits using the Big Five Inventory comprised of 44 items [30]. The BFI scale remains the most widely used taxonomy of human personality for several decades: it is known for capturing these traits using a reasonable number of questions [64] and sees use in prior work on user characteristics and XAI [13, 49]. The final section evaluated people's expertise in AI/ML using Hornberger et al. [25]'s literacy questionnaire. This questionnaire provides an objective accuracy-based measure of literacy based on 31 multiple-choice questions. To reduce task burden on intake survey participants, we modified the questionnaire and included 10 of the 31 questions under the categories of: building blocks of an ML pipeline (e.g., types of data and models), data science concepts (e.g., preprocessing, overfitting), and ethical issues (e.g., bias, representation). Completion of the intake survey took approximately 10 minutes, and participants were compensated with \$2 on Prolific. Additional details on the intake survey materials are available in Appendix C.

3.3.2 Main Study. Intake participants who met the inclusion criteria—being 18 years or older, residing in the U.S., and achieving a minimum score of 50% on the ML literacy scale—were invited to the main study. Figure 1 presents an overview of all study components. This study took approximately 20 minutes to complete, and participants were compensated with \$5 upon completion.

First, the study task was introduced as follows: "You will predict whether individuals earn less than or greater than \$50,000 based on information provided about their demographics and work history." To facilitate familiarity with the task, participants completed a practice task involving four example datapoints. When a participant made their prediction about income for these datapoints, the correct answer was automatically displayed. This setup was intended to help familiarize them with the dataset features (always presented as a table), and the survey interface and format of the main task.

Following the practice round, participants proceeded to the main task, where they were presented with 8 data points from the Adult Income dataset. For each data point, they predicted whether the individual earned less than or greater than \$50k based solely on a table of feature information, i.e., no ML assistance was provided at this point. They also noted their self-confidence about each classification on a scale of 0–6 (not at all to extremely confident). We did not have participants make classifications with ML assistance during this task for two reasons: first, legislation such as the EU AI Act [71] increasingly requires explanations for black-box models involved in decision-making, making an ML-without-explanations condition less relevant; secondly, we wished to capture participants' baseline decision-making behavior influenced by nothing but their user characteristics—displaying the ML predictions at this stage would have interfered with that behavior.

Next, we asked participants to classify the same datapoints with ML assistance and SHAP explanations. Before they did so, they completed Jian et al. [28]'s trust questionnaire after reading a tutorial on the AI system to establish a baseline of their trust in the system, where the questionnaire broadly defined AI and equally balanced positive and negative statements. For these ML-assisted classifications, we provided the model's prediction alongside a SHAP waterfall plot that explained the factors influencing the prediction. For each datapoint, participants selected a classification, rated their self-confidence on the same scale as before, and also rated their confidence in the ML prediction being accurate (scale 0–6). After completing the ML-assisted classification, participants answered the trust questionnaire again, allowing us to observe any change in trust after concrete experience with the system used in our study. Finally, participants explained their decision-making for both unassisted and ML-assisted classification tasks via open-text responses. Details on the main study materials are available in Appendix D.

3.4 Participants and Data

Our surveys were administered as Prolific studies. We collected responses from 252 participants for our intake survey. Of these, 149 met the inclusion criteria—being 18 years or older, residing in the U.S., and achieving a minimum score of 50% on the ML literacy test—and completed the main study. Participant ages ranged from 18 to 66 (median=33). There were 69 male participants, 77 female participants, and 3 non-binary participants. Approximately 69% of participants had completed some form of college education (Associate's degree and beyond) and 31% held college credit or attended vocational school. Their job roles ranged from IT professionals and project managers to nurse practitioners and more.

3.5 Variables of Interest

Our goal was to model XAI outcomes based on user characteristics. We describe our independent variables collected via the intake survey and dependent variables measured during the main task.

3.5.1 Independent Variables.

(1) **Demographics.** Age, gender, and education level.
(2) **Experience.** Included four aspects of experience: (a) subjective ratings of the extent to which ML is a part of participants' job roles, (b) subjective ratings of both ML and interpretability knowledge, (c) estimates of hours spent on

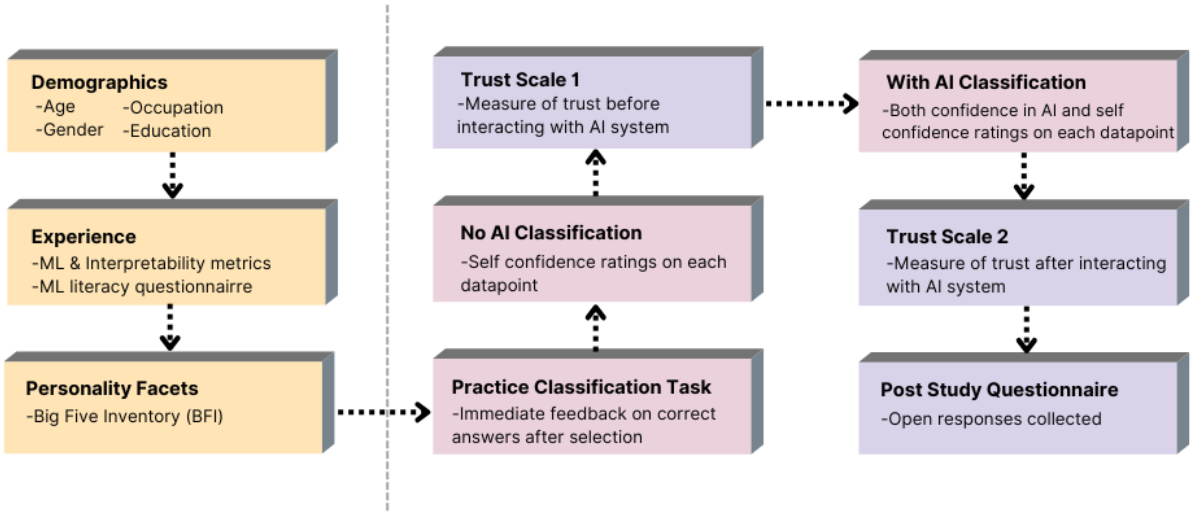


Figure 1: Overview of the study task flow. Orange boxes are associated with the intake survey. The pink and purple boxes signify the main study: pink boxes represent classification tasks and purple boxes represent points of subjective data collection.

both ML and interpretability tasks, and (d) accuracy on a modified 10-question objective ML literacy scale [25]. Given conceptual and scale similarity in the two questions asked under (b), we averaged the questions into one value representing prior experience with ML and interpretability. We verified their internal consistency using Cronbach's alpha before averaging them; the alpha was > 0.7, which is the accepted norm for merging values [33, 50]. The ML and interpretability questions under (c) underwent a similar process, resulting in an hours estimate rating representing both ML and interpretability.

(3) **Personality.** Scores for the five personality dimensions [30]—extraversion, agreeableness, conscientiousness, neuroticism, and openness—scaled to a range of 0–1. Extraversion reflects sociability and energy; agreeableness represents tendencies toward cooperation and trust; conscientiousness indicates organization and self-discipline; neuroticism encompasses emotional instability and sensitivity to stress; and openness describes curiosity and receptivity to new experiences.

3.5.2 *Dependent Variables.*

(1) **Prediction Changes:** Participants were tasked with classifying eight datapoints as either income $\geq \$50k$ or $< \$50k$ without AI assistance. They then classified the same eight datapoints (without being made aware they were the same set of datapoints) with AI assistance and explanation plots. This metric represents the number of instances (0–8) where participants altered their initial predictions after reviewing AI classifications, calculated as a percentage.

(2) **Average Change in Confidence:** For each prediction, participants indicated their level of confidence in their answer on a scale of 0–6. By comparing their first confidence rating of a given datapoint to their second, this value indicates how their level of confidence changed with the inclusion of XAI.

(3) **Change in Trust:** Participants completed a trust questionnaire both before and after interacting with the ML model and XAI outputs for the task. This questionnaire used a validated trust scale by Jian et al. [28]. The change in trust was calculated as the difference between the post-ML interaction and pre-interaction trust scores.

(4) **Inappropriate Reliance:** To measure participants' reliance on the ML and XAI outputs, we calculated the number of instances (0–8) where participants over- or under-relied on the AI during the classification task. Over-reliance occurred when participants changed an initially correct classification into an incorrect one to be consistent with the AI. Under-reliance was when participants did not change an initially incorrect classification despite the AI output being correct.

(5) **Average Confidence in AI:** For each AI-assisted prediction, participants indicated their level of confidence in the ML and XAI outputs on a scale of 0–6.

3.6 Analysis Methods

We conducted correlation analysis and statistical modeling for our variables of interest. We initially modeled our dependent variables using individual models for each of our independent variables: demographics, experience, and personality. More complex modeling using hierarchical approaches and non-linear models was also performed. We present these model decisions in the Results section to capture our process in picking models based on iterative results.

Our open-text data from the post-study questionnaire captured participants' reasoning behind classifications, both with and without AI assistance. We analyzed this data using Braun and Clarke [8]'s inductive thematic analysis approach. Two authors open and axial coded all the data, and all authors analyzed these to identify final themes. To quantify interesting themes from participants' reasoning processes, we also converted select axial codes into binary variables and computed correlations with other independent variables using Point-Biserial tests.

4 Results

4.1 Descriptive Statistics and Correlations

On average, participants in our sample leaned towards being moderately extroverted (μ=0.59, σ=0.16; scale=0–1) and neurotic (μ=0.55, σ=0.17); and more agreeable (μ=0.70, σ=0.13), conscientious (μ=0.75, σ=0.15) and open (μ=0.77, σ=0.14). In terms of prior experience, the statistics were as follows: extent of ML in role (μ=2.68, σ=1.89; scale=1–7), subjective rating of experience in ML and interpretability (μ=2.57, σ=1.38; scale=1–7), hours-based estimate of experience in ML and interpretability (μ=2.02, σ=1.14; scale=1–7), score on objective ML literacy scale (μ=6.46, σ=1.63; scale=0–10).

Table 1 presents correlation coefficients and significant results from running Spearman's rank correlation analyses for our predictors. Among the personality facets, neuroticism was significantly negatively correlated with other personality variables. Interestingly, agreeableness and conscientiousness were highly positively correlated, though they normally represent different perspectives on a spectrum of skepticism in individuals. The experience variables were significantly correlated.

Given several significant correlations, we also computed variance inflation factors (VIFs) for all predictors to check for multicollinearity, ensuring all retained variables were within acceptable thresholds of <5 (min=1.2, max=4.3).

The descriptive statistics for our dependent variables are: average prediction changed (μ=0.34, σ=0.21; scale=0–1), inappropriate reliance (μ=0.22, σ=0.14; scale=0–1), average change in confidence (μ=-0.01, σ=0.69; scale=-6–6), average confidence in AI (μ=3.48, σ=0.92; scale=0–6), and change in trust (μ=3.01, σ=8.66; scale=-72–72, with 72 being the maximum possible score on the trust scale).

4.2 Linear Models for User Characteristics and XAI Outcomes

User characteristics have primarily been modeled using linear approaches before, and we followed the same anticipated relationships. As such, we first fit linear models (LMs) for individual user characteristics and each of our dependent variables (DVs). However, LMs assume that residuals are normally distributed, which we verified was not the case for some of our data by examining the Q-Q plots of the residuals for each DV. For the dependent variables Average Change in Confidence, Change in Trust, and Average Confidence in AI, the Q-Q plots indicated that the residuals were approximately normal; i.e., LMs were appropriate for these.

We next tested Generalized Linear Models (GLMs) for our non-normal DVs (Average Predictions Changed and Inappropriate Reliance), since GLMs relax the normality assumption of residuals and allow for a broader range of response variable distributions. We modeled our non-normal DVs as Gamma distributions with a log link function. This distribution provided the best fit for the majority of our DVs, as it is suitable for right-skewed, non-negative, continuous outcomes where the variance increases with the mean (homoscedasticity), and we avoid issues regarding learning effects that would be present with count-based model families.

We also assessed the presence of influential outliers by calculating Cook's distance for each observation. Observations with Cook's distance greater than 1 are typically considered candidates for removal. No such observations were identified in our data, so all data points were retained for analysis.

By using LMs for normally distributed residuals and Gamma GLMs for right-skewed distributions, and by confirming no undue influence from outliers, we ensured that each DV was modeled with the most suitable framework, allowing us to accurately capture the relationships between predictors and outcomes.

4.2.1 Individual Model Results. Modeling the relationships using the process above for one set of user characteristics—demographics, experience, and personality—at a time did not yield significant results for any of our predictors. It is worth noting two caveats here. 1) A few intercepts were significant for our experience-characteristics model: confidence in AI (F(4,144)=2.99, p<0.05) with an estimated coefficient of 2.68 (SE=0.31), prediction changed, and inappropriate reliance, although the latter two accounted for minimal reduction in deviance for their respective models. These significant intercepts suggest that everyone, regardless of prior experience predictors, tended to have positive values for confidence in AI. 2) Gender showed a marginally significant effect on change in trust (F(4,144)=2.36, p=0.05). The estimated coefficient was 3.25 (SE=1.45).

4.2.2 Combined Model Results. While our sets of predictors showed minimal predictive power on their own, additional significances became evident once different sets of predictors were combined in a hierarchical fashion. We assessed demographic+personality and demographic+experience models, in addition to a full model that included predictors from all user characteristics. Summaries of those full models can be found in Appendix A. This exploration demonstrated that accuracy on the objective ML literacy scale significantly impacted confidence in the AI (F(8, 140)=2.36, p=0.021) with a coefficient of 0.1 (SE=0.04). Additionally, several of these models indicated significant results associated with the gender of non-binary datapoints for confidence in AI, but we believe these to be an artifact of class imbalance given the small amount of non-binary datapoints. Finally, the intercept became significant for the change in trust model when demographic and experience predictors were both included (F(8,140)=2.497, p<0.05) with a coefficient of -9.43 (SD=4.06), suggesting a general reduction in trust after participants interacted with the AI.

4.3 Exploratory Analysis

Thus far, our modeling approach has followed the same assumptions as prior work with regards to user characteristics as predictors for XAI outcomes. This was effectively our validation check for the results of these previous studies. However, we were interested in exploring more nuanced modeling approaches for this type of data given that prior work has contradictory results regarding user characteristics' impact on user behavior, both in XAI and other fields (see Section 2.2). Therefore, while our modeling results so far confirmed some prior findings (e.g., [49]), in doing so, they refuted other findings (e.g., [37, 64]). This made us question our assumptions about the relationships being considered, and we decided to further explore the data using non-linear modeling via Generalized Additive Models (GAMs)[1].

[1]Using version 1.22-5 of the gam package in R.

Table 1: Spearman correlation matrix illustrating the relationships between personality traits, demographic factors, and experience factors along with asterisks depicting significant relationships. Significance levels: ***p<0.001, **p<0.01, and *p<0.05. Hours Exp. represents hours spent on ML and interpretability tools in the past. Subj. Exp. represents self-rating of ML and interpretability tool use in the past. ML in role represents self-reported rating of ML use in daily job.

	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism	Age	Edu. Level	AI Scale	Hours Exp.	Subj. Exp.
Openness										
Conscientiousness	0.161									
Extraversion	**0.323***	**0.224***								
Agreeableness	0.118	**0.438*****	**0.265****							
Neuroticism	-0.084	**-0.484*****	**-0.394*****	**-0.513*****						
Age	0.098	**0.193***	0.101	0.160	-0.137					
Edu. Level	-0.127	**0.163***	0.101	-0.085	-0.070	0.159				
AI Scale	0.158	0.019	-0.041	-0.024	-0.018	-0.065	0.012			
Hours Exp.	**0.244****	0.138	0.129	-0.023	**-0.162***	**-0.193***	0.106	**0.229****		
Subj. Exp.	0.083	0.054	0.074	-0.042	-0.160	**-0.261****	0.025	**0.287*****	**0.827*****	
ML Role	0.095	0.022	0.122	-0.010	-0.131	**-0.218****	0.063	**0.200***	**0.755*****	**0.776*****

While GLMs are effective for modeling relationships between predictors and outcomes, they rely on the assumption that these relationships are strictly linear. In contrast, GAMs retain the linearity assumption for the overall model structure but allow for greater flexibility by relaxing the linear relationship assumption for individual predictors.

GAMs achieve this flexibility by incorporating smooth functions for predictors, which allow for the model to estimate the relationship between a predictor and the dependent variable in a nonparametric manner. Instead of fitting a single line, GAMs use these smooth functions to capture complex, non-linear trends in the data. This allows GAMs to adapt to varying data patterns while avoiding over-fitting, as the degree of smoothing can be controlled during model fitting.

With the capacity to model non-linear smooth effects for specific predictors, GAMs provide a robust framework for uncovering intricate relationships that might be missed by GLMs. This dual capability ensures that both linear and non-linear relationships are effectively captured, making GAMs a complementary approach to GLMs in our modeling process, especially when attempting to reconcile the contradictory prior work in this space.

4.3.1 Individual GAM Results. Individual GAMs for each of our user characteristics predicting each dependent variable resulted in a few significant outcomes: (1) gender significantly predicted change in trust (edf=1.00, p<0.01)[2], where the increase in trust for female participants was slightly diminished in comparison to male participants. This result aligns with a large-scale survey finding that men are more trusting of AI than women [51]. (2) accuracy on the objective AI scale significantly predicted the tendency to inappropriately rely on the AI predictions (edf=1.00, p<0.05), with an intuitive negative correlation between the variables. This finding supports claims in prior work that intuition can be used to override incorrect AI outputs in decision-making contexts [12]—this intuition would naturally be supported by a familiarity with underlying AI concepts. Finally, (3) conscientiousness significantly predicted average change in confidence (edf=3.34, p<0.01), with this non-linear relationship taking the form of a sharp rise followed by a dip and another small increase. This finding is largely consistent with prior work on visualization-assisted decision-making, where participants with average levels of conscientiousness were the most confident [1]. We similarly hypothesize that high-conscientiousness participants may have been too cautious to report high confidence while low-conscientiousness participants may have been openly unconfident in their less careful approach to the study. Conscientiousness is a well-known predictor for user behavior that makes people feel confident [53, 61, 64], and yet we had not observed this otherwise well-known effect in XAI studies thus far. These findings demonstrate the potential for non-linear modeling to capture expected results from qualitative work on XAI and user characteristics, and produce intuitive results suggested by non-XAI literature.

[2]Effective degrees of freedom (edf) indicates the complexity of smooth terms in GAMs, with higher values reflecting more flexible, non-linear relationships.

4.3.2 Combined GAM Results. Our final set of models combined all user characteristic predictors and their possible interaction effects in a single GAM model for each dependent variable. Several of these models revealed significant, smooth though non-linear, relationships. Compared to the individual GAMs, the combined GAMs had an additional significant result related to average confidence in AI, and numerous interaction effects involving personality facets became significant. Table 2 summarizes the significant results and full model results are included in Appendix B.

With these combined GAM results, we finally see the full scope of prior work on user characteristics reflected in user behavior in practice. For example, it makes intuitive sense that change in trust before vs. after directly working with an ML model and XAI outputs is impacted by some level of prior experience, and that neuroticism can make this better or worse depending on whether the neuroticism is skepticism about AI in general or about one's own familiarity with the data domain. Similarly, conscientiousness and openness are known to be the relevant personality facets for establishing notions of confidence, with one being appropriate confidence and the other having the potential for over-confidence [64]. These exploratory results are initial evidence for the relevance of prior work on user characteristics for ML and XAI contexts, with the caveat that the relationships identified in this prior work might not have a linear mapping. We elaborate on this using our qualitative data and discussion section.

Table 2: Summary of Significant GAM results from the combined GAM model with interaction effects. Note: Effective degrees of freedom (edf) indicate the complexity of smooth terms in GAMs, with higher values reflecting more flexible, non-linear relationships. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. Non-significant predictors are excluded for brevity.

Dependent Variable	Predictor	edf	p-value
Change in Trust	Gender	1.009	0.003 **
	Education Level	1.000	0.028 *
	Neuroticism × Hours of Experience	3.976	0.003 **
	Subjective Rating of Experience	4.540	0.010 **
	ML in Role	1.000	0.006 **
	Adjusted R^2	0.260	
	Deviance Explained	39.9%	
Average Change in Confidence	Conscientiousness	3.300	0.007 **
	Gender	1.000	0.007 **
	Openness × Hours of Experience	7.624	0.015 *
	Adjusted R^2	0.229	
	Deviance Explained	36.4%	
Average Confidence in AI	Openness × Hours of Experience	4.400	0.047 *
	ML in Role	1.000	0.020 *
	Adjusted R^2	0.215	
	Deviance Explained	34.6%	

4.4 Qualitative Results

We describe results derived from inductive thematic analysis of open-text responses explaining decision-making processes with and without the ML and XAI outputs. For major themes, we converted our axial codes into binary variables to calculate Point-Biserial correlations with our user characteristics and establish quantitative relationships that can support future hypothesis generation.

Without ML and XAI assistance, people applied prior heuristics about the relationship between income and demographics to make predictions. This is in line with the extensive work in cognitive science describing the bounded nature of human rationality, wherein people anchor themselves to some pieces of information and make decisions based on them [63, 69]. For our participants, the most frequently used anchors were feature values for education, hours worked per week, and occupation. In contrast to a systematic analysis of features, participants cited a reliance on their "personal experience" (P44) and "intuition" (P61) about the values of these few features for the datapoints under consideration.

With ML and XAI assistance, participant reasoning had a higher variance of factors. Most participants engaged in a "comparison of opinions" (P13) between the ML and XAI output and their own judgment, however they were evenly split between people who consulted the outputs before forming their own decision or vice versa. Confidence played a pivotal role in these interactions. Participants felt more confident when their decision matched the model outputs, while mismatches often led to reduced confidence. We found this to be of particular interest given that these were decision-boundary datapoints, and the model being used effectively had random performance on the selected datapoints.

This suggested an inherent difference between people based on their prior experience with AI and ML vs. those with more openness to newer technology. On one hand, there were participants like P67 who "fully trust the AI prediction because it was trained on a lot more data than I have ever seen, so I thought it would be a lot better of a gauge than my own opinions" and P31 who felt under-confident in their decisions because "the attempts I made during the training round were pretty unreliably bad, so I deferred to a system designed to do that."

On the other hand, participants with more prior experience or those who referenced details about the XAI outputs (e.g., SHAP values) took a more balanced approach, with some being skeptical and others remaining open to the value of AI-assisted decision making. For example, P14 noted that they "looked at the AI's result, and then did my own analysis of it based on their sector, age, and education. I took the AI's result with a grain of salt." While P147 was similarly critical about evaluating the ML and XAI outputs, they were open to accepting future outputs once they had established that the model outputs made sense: "I used the AI's prediction and the SHAP explanation plot extensively to make decisions about the individual's income. This additional information allowed me to refine my understanding and make more accurate predictions [over time]."

To quantify some of these relationships, we coded participant responses about leaning towards AI's opinion vs. their own and specific mentions of SHAP in their reasoning as two binary variables. Point-Biserial correlations showed significant positive relationships between the experience variables and these two binary variables.[3] This suggests that people with more prior experience tended to follow the AI outputs and SHAP explanations closely, and incorporated them as a part of their own reasoning. However, given that the ML model used here had little predictive power on these decision-boundary datapoints, it is unclear whether this helped them on performance metrics. What is also unclear is whether those who did not mention ML or XAI in their reasoning

[3]Correlations between leaning towards AI's opinions on predictions and experience variables: Hours-Based Experience ($r = 0.31$, $p = 0.0003$), Subjective Rating of Experience ($r = 0.33$, $p = 0.0001$), and ML Experience in Role ($r = 0.28$, $p = 0.0011$)
Correlations between mentions of specific XAI outputs (e.g., SHAP values) and experience variables: Hours-Based Experience ($r = 0.21$, $p = 0.0097$), Subjective Rating of Experience ($r = 0.18$, $p = 0.0279$), and ML Experience in Role ($r = 0.21$, $p = 0.0112$).

completely disregarded them or simply forgot to mention them in post-study responses. We hypothesize that there is a complex interaction effect at play between prior experience and openness to technology acceptance, but more in-depth qualitative methods are needed to concretely unpack that.

Overall, our qualitative analyses suggest that intuition is a key aspect of ML and XAI use. There are several sources for this intuition: inherent self-confidence, prior experience in ML and XAI, prior domain knowledge, openness to new technology. These sources also do not have linear relationships with reasoning; different people under different settings rely on the same factors in different ways. We find this to be further evidence that user characteristics—especially for this newer type of decision-making assisted by AI—are in flux both for an individual (over time) and across individuals. Therefore, we discuss the (in)feasibility of modeling them as established relationships in the following section.

5 Discussion

We conducted this research to establish an integrated understanding of user characteristics relevant to XAI use for decision-making. Contradictory results from prior studies motivated our work, with some claiming the significance of certain user characteristics and others refuting such claims. Our primary analysis did not solve this dilemma, but we discovered new insights once we challenged our approach via exploratory analysis. In an effort to provide consistency and direction to this line of research, we discuss improvements to the measurement of human-centered facets and describe more holistic approaches to collecting measurements for personalization.

5.1 Measurement of User Characteristics

We consider the lack of consensus in this line of research a measurement issue rather than a modeling one. Below, we articulate measurement issues that diffuse the community's efforts as we grapple with the complexities of user characteristics.

Disparate task settings and explanation purposes. We critically identify the need to formalize what must remain methodologically consistent between XAI studies to produce comparable results. Prior work on user characteristics and XAI yielded positive results when explanations passively educated users about system behaviors. This includes explanations about why certain recommendations appear [43, 44], or why content is displayed in a specific manner [13]. On the other hand, our work and others [37, 49] that use explanations to support decision-making have produced negative results. This discrepancy forms a pattern: XAI engagement with consequential explanations differs significantly from more supplementary explanation needs. Thus, the community should study XAI which serves different purposes separately and consistently, respecting this distinction when comparing results.

Subjective and objective metrics. Research on survey methodology consistently highlights the importance of question-phrasing [18, 22]. Yet, many XAI studies (including us, for some parts) use arbitrarily-defined sentences to capture subjective ratings, which are claimed to measure certain concepts and characteristics. This is a fundamentally flawed approach when it comes to generalizability of survey data and instruments [22]. While it is initially useful to have these types of subjective ratings, ultimately, we either need validated scales or complementary objective metrics to ensure data validity. We prioritized capturing user characteristics using either validated scales or objective metrics, and used subjective ratings like these only as supplemental measures. Significant future work is needed towards both efforts, perhaps beginning with comparing the value of existing validated scales, similar to the work of Stein et al. [64].

Lack of benchmarking and replication. The lack of benchmarking hampers the consistency and comparability of results in this research area. Benchmarks serve as common points of comparison and support the concerted, organized efforts of research communities. A systematic literature review of personalized XAI studies would benefit the design of benchmarks by retrospectively providing structure to the various methodologies, consistencies, and contradictions that have arisen in this line of research.

Overall, when we think of the measurement crisis in AI, ML, and XAI, technical facets of the field come to mind: identifying fairness metrics, representing bias and harms, or finding faithful approximations of model behavior. Much research attention has been given to these challenges, and rightfully so. However, the measurement challenges identified above instead pertain to fundamental human-centered aspects of AI, ML, and XAI use *in practice*. As a field, we need equal prioritization of both kinds of measurement problems if we ultimately intend for these technologies to be useful for people.

5.2 Holistic Measurements for Personalization

In addition to more rigorous measurement standards, we identify approaches to measurement that capture more holistic views of participants and their engagement with XAI. These approaches extend beyond the conventional methods that capture snapshots of XAI engagement. Such snapshots represent one-off engagements with proxy tasks that are severely limited both temporally and by the minimal models used to represent participants. In contrast, we propose approaches to measurement that could help establish a robust understanding of individuals and their use of XAI.

Accounting for novelty. General perceptions of ML applications are evolving and contentious [67]. These perceptions are likely to continue shifting, and will be aggravated when experiments allude to distressing applications of XAI outside of the lab such as content moderation or tasks like ours that highlight social inequalities. Given that technology acceptance and diffusion are fundamentally intertwined with user characteristics [65], research in this area should be more intentional about those mediators as general sentiment on ML shifts. Having emphasized the importance of a shifting relationship to ML on the grand scale, we do the same for the individual scale in the following point.

Capturing longitudinal use. A longitudinal study on XAI use would bring much value to this line of research, particularly given the significant effect of experiential factors we observed. Experiential factors were significant more often than personality and demographic factors, both in our quantitative and qualitative findings. As such, an individual's engagement with XAI will shift as they earn ML experience or as the technology in use becomes less novel to them. Thus, to ensure that XAI is appropriately employed in critical decision-making contexts where its use is intended to become routine, we must go beyond studying isolated proxy tasks and instead study the patterns that emerge from sustained use.

Recsys-inspired measures. Finally, we take inspiration from the study of recommender systems to propose an alternative, far more holistic, measurement approach. Rather than limit studies to arbitrarily select a handful of stable characteristics for evaluation, we can instead analyze log data to develop behavioral insights on XAI use. Just as recommender system work logs information regarding interactions, previously viewed content, and mood, there is an opportunity to log a similar set of information when individuals engage with explanations. Such logs have the benefit of capturing factors that are less stable than those typically studied in this line of research, thus accounting for how decision-making processes shift day to day and even decision to decision. Producing a dataset of such logs would allow the community to identify complex, unanticipated associations with higher ecological validity.

Overall, our results indicated that conventional approaches to studying user characteristics in relation to XAI are severely limited. Many key insights emerged once we incorporated qualitative analysis and extended beyond linear models, both practices that are not the norm for this topic. From this experience, we encourage the community to continue extending beyond the measuring practices that research in this area conforms to; we can discover richer insights once we take more holistic approaches to measuring XAI use.

6 Limitations

We acknowledge the following limitations that may affect the broader applicability of our findings. First, our use of decision boundary datapoints muddles the signal from our inappropriate reliance metric. We used these datapoints to avoid any concrete impact of the ML model's accuracy and to be able to attribute any distinctions in results between unassisted vs. assisted classification to user characteristics alone. However, the difficulty of these datapoints might have resulted in random classifications. Follow-up work could utilize datapoints with a wide range of difficulty for humans while still having low AI confidence (the AI outputs and explanations could even be fabricated to make this so).

Another limitation of our methodology is that we assume participants are not aware that they re-classify the same set of 8 datapoints. It would be valuable if this was the case so that participants would engage with the AI-assisted classifications as deeply as they did with the non-AI-assisted classifications—if participants knew they were re-classifying datapoints, they may just replicate their answers without deliberation. We never told participants that they were re-classifying the same set of datapoints, but that did not necessarily prevent them from coming to that realization. To avoid this risk, future work may consider drawing from pools of matched datapoints where each datapoint in the non-AI-assisted classifications is matched with a similar datapoint in the AI-assisted classifications. We did not use that approach as it may introduce confounding factors in the comparison between decision-making with and without AI, with this comparison being the main focus of our study; we encourage future work to consider the tradeoffs between re-using datapoints and selecting matched datapoints as they relate to the research questions being explored.

Finally, we must acknowledge that our study only observed one explainability tool, SHAP, in one proxy task without any stakes. Just as we outlined patterns corresponding to XAI that serve passive versus actionable purposes, similar patterns may persist for contexts that are inconsequential versus critical. While it is more difficult to study the critical applications of XAI, these are the very instances of the technology that motivate our research. This difficulty also informed our choice of experience-related user characteristics: recruiting participants with a wide range of domain expertise in a critical decision-making context is challenging. As such, similar to prior work [49], we focused on technical expertise and picked a dataset where domain expertise was not required. We anticipate value in both qualitative and quantitative work that deeply explores relationships between end users and the XAI tools they use to regularly make consequential decisions.

7 Conclusion

In this work, we examined the relationship between user characteristics and XAI use. We used validated metrics to collect personality, ML experience, and demographic information from participants before having them classify datapoints predicting people's income—first without XAI assistance, and then with it. While our primary quantitative analysis yielded results echoing prior work on the minimal impact of user characteristics, our qualitative and exploratory statistical analyses brought additional insights that lead us to fundamentally question how XAI personalization is measured. We step away from the conventional measuring practices limiting the possibilities for this research area, demonstrating the value in building more holistic models of end-users and their behavior. We further outline holistic and rigorous measuring practices for the community to pursue, believing they have the potential to unravel the complexities of personalized XAI.

Acknowledgments

We would like to thank our reviewers for their helpful comments. We are grateful to all of the faculty and students in the GroupLens research lab for their feedback and support. We also want to thank those who participated in our study.

References

[1] Tomás Alves, Tiago Delgado, Joana Henriques-Calado, Daniel Gonçalves, and Sandra Gama. 2023. Exploring the role of conscientiousness on visualization-supported decision-making. *Comput. Graph.* 111, C (April 2023), 47–62. https://doi.org/10.1016/j.cag.2023.01.010

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Andrés Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[4] Jackie Ayoub, X Jessie Yang, and Feng Zhou. 2021. Modeling dispositional and initial learned trust in automated vehicles with predictability and explainability. *Transportation Research Part F: Traffic Psychology and Behaviour* 77 (2021), 102–116. https://doi.org/10.1016/j.trf.2020.12.015

[5] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2022. A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human-Computer Interaction* 40, 5 (2022), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826

[6] 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. https://doi.org/10.1145/3411764.3445717

[7] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[9] 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. https://doi.org/10.1145/3449287

[10] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3517471

[11] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[12] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (Oct. 2023), 32 pages. https://doi.org/10.1145/3610219

[13] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artif. Intell.* 298, C (Sept. 2021), 23 pages. https://doi.org/10.1016/j.artint.2021.103503

[14] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning* (2017). https://api.semanticscholar.org/CorpusID:11319376

[15] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33. https://doi.org/10.1145/3561048

[16] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 316, 32 pages. https://doi.org/10.1145/3613904.3642474

[17] Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 160–171. https://proceedings.mlr.press/v81/ensign18a.html

[18] Lior Gideon. 2012. *The Art of Question Phrasing*. Springer New York, New York, NY, 91–107. https://doi.org/10.1007/978-1-4614-3876-2_7

[19] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89. https://doi.org/10.1109/dsaa.2018.00018

[20] John W. Goodell, Satish Kumar, Weng Marc Lim, and Debidutta Pattnaik. 2021. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance* 32 (2021), 100577. https://doi.org/10.1016/j.jbef.2021.100577

[21] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (December 2003), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

[22] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.

[23] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5, Article 93 (Aug. 2018), 42 pages. https://doi.org/10.1145/3236009

[24] Trevor J Hastie. 2017. Generalized additive models. In *Statistical models in S*. Routledge, 249–307.

[25] Marie Hornberger, Arne Bewersdorff, and Claudia Nerdel. 2023. What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence* 5 (2023), 100165. https://doi.org/10.1016/j.caeai.2023.100165

[26] Md Aminul Islam, Anindya Nag, Nilanjana Roy, Arpita Rani Dey, SM Firoz Ahmed Fahim, and Arjan Ghosh. 2023. An Investigation into the Prediction of Annual Income Levels Through the Utilization of Demographic Features Employing the Modified UCI Adult Dataset. In *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. 1080–1086. https://doi.org/10.1109/ICCCIS60361.2023.10425394

[27] Jun Li Jeung and Janet Yi-Ching Huang. 2023. Correct Me If I Am Wrong: Exploring How AI Outputs Affect User Perception and Trust. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (Minneapolis, MN, USA) *(CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 323–327. https://doi.org/10.1145/3584931.3606997

[28] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury and. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[29] Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies* 165 (2022), 102839. https://doi.org/10.1016/j.ijhcs.2022.102839

[30] Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology* (1991). https://doi.org/10.1037/t07550-000

[31] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein. 2017. Simple Rules for Complex Decisions. Available at SSRN: https://ssrn.com/abstract=2919024 or http://dx.doi.org/10.2139/ssrn.2919024.

[32] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 702–714. https://doi.org/10.1145/3531146.3533135

[33] Harmanpreet Kaur, Matthew R. Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 77 (April 2024), 34 pages. https://doi.org/10.1145/3637354

[34] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[35] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (Oct 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[36] Isabel Lage, Emily Chen, Jiaming He, Madhumita Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 59–67. https://doi.org/10.1609/hcomp.v7i1.5280

[37] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Comput. Hum. Behav.* 139, C (Feb. 2023), 18 pages. https://doi.org/10.1016/j.chb.2022.107539

[38] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021).

[39] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in Cognitive Sciences* 10, 10 (2006), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

[40] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[41] Jiaxin Mao, Yiqun Liu, Noriko Kando, Min Zhang, and Shaoping Ma. 2018. How Does Domain Expertise Affect Users' Search Interaction and Outcome in Exploratory Search? *ACM Transactions on Information Systems* 36, 4 (July 2018), 42:1–42:30. https://doi.org/10.1145/3223045

[42] David Alvarez Melis, Harmanpreet Kaur, Hal Daumé III, Hanna Wallach, and Jennifer Wortman Vaughan. 2021. From human explanation to model interpretability: A framework based on weight of evidence. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 9. 35–47. https://doi.org/10.1609/hcomp.v9i1.18938

[43] 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 397–407. https://doi.org/10.1145/3301275.3302313

[44] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3340631.3394844

[45] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.

2018.07.007

[46] Tim Miller, Piers D. L. Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *ArXiv* abs/1712.00547 (2017). https://api.semanticscholar.org/CorpusID:28681432

[47] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (April 2024), 39 pages. https://doi.org/10.1145/3641022

[48] Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-Based Explanation Methods: A Review for NLP Interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (Eds.). International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 4593–4603. https://aclanthology.org/2022.coling-1.406/

[49] 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 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 317, 13 pages. https://doi.org/10.1145/3613904.3642352

[50] Jum C. Nunnally. 1978. *An Overview of Psychological Measurement.* Springer US, Boston, MA, 97–146. https://doi.org/10.1007/978-1-4684-2490-4_4

[51] Nessrine Omrani, Giorgia Rivieccio, Ugo Fiore, Francesco Schiavone, and Sergio Garcia Agreda. 2022. To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change* 181 (2022), 121763. https://doi.org/10.1016/j.techfore.2022.121763

[52] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the impact of explanations on advice-taking: a user study for AI-based clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[53] Briony D. Pulford and Harjit Sohal. 2006. The influence of personality on HE students' confidence in their academic abilities. *Personality and Individual Differences* 41, 8 (2006), 1409–1419. https://doi.org/10.1016/j.paid.2006.05.010

[54] J. R. Quinlan. 1986. Induction of Decision Trees. *Mach. Learn.* 1, 1 (March 1986), 81–106. https://doi.org/10.1023/A:1022643204877

[55] Antonio Rago, Bence Palfi, Purin Sukpanichnant, Hannibal Nabli, Kavyesh Vivek, Olga Kostopoulou, and Francesca Toni. 2024. Exploring the Effect of Explanation Content and Format on User Comprehension and Trust. *arXiv preprint arXiv:2408.17401* (2024).

[56] Samuel Reeder, Joshua Jensen, and Robert Ball. 2023. Evaluating Explainable AI (XAI) in Terms of User Gender and Educational Background. In *Artificial Intelligence in HCI: 4th International Conference, AI-HCI 2023, Held as Part of the 25th HCI International Conference, HCII 2023, Copenhagen, Denmark, July 23–28, 2023, Proceedings, Part I* (Copenhagen, Denmark). Springer-Verlag, Berlin, Heidelberg, 286–304. https://doi.org/10.1007/978-3-031-35891-3_18

[57] 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* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[58] Estrella Romero, Paula Villar, J. Antonio Gómez-Fraguela, and Laura López-Romero. 2012. Measuring personality traits with ultra-short scales: A study of the Ten Item Personality Inventory (TIPI) in a Spanish sample. *Personality and Individual Differences* 53, 3 (2012), 289–293. https://doi.org/10.1016/j.paid.2012.03.035

[59] Yuxuan Rong, Thomas Leemann, Thi Thu Trang Nguyen, Lukas Fiedler, Pengfei Qian, Vaibhav V Unhelkar, and Eirini Kasneci. 2023. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). https://doi.org/10.1109/TPAMI.2023.3284738

[60] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[61] Norsaremah Salleh, Emilia Mendes, John Grundy, and Giles St. J Burch. 2010. An empirical study of the effects of conscientiousness in pair programming using the five-factor personality model. In *Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering - Volume 1* (Cape Town, South Africa) *(ICSE '10)*. Association for Computing Machinery, New York, NY, USA, 577–586. https://doi.org/10.1145/1806799.1806883

[62] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1616–1628. https://doi.org/10.1145/3531146.3533218

[63] Herbert A Simon. 1997. *Models of Bounded Rationality: Empirically Grounded Economic Reason.* The MIT Press.

[64] Jan-Philipp Stein, Tanja Messingschlager, Timo Gnambs, Fabian Hutmacher, and Markus Appel. 2024. Attitudes towards AI: measurement and associations with personality. *Scientific Reports* 14, 1, 2909. https://doi.org/10.1038/s41598-024-53335-2

[65] Gunnvald Svendsen, Jan-Are Johnsen, Live Almås-Sørensen, and Joar Vittersø. 2013. Personality and technology acceptance: the influence of personality factors on the core constructs of the Technology Acceptance Model. *Behaviour & Information Technology* 32, 4 (2013), 323–334. https://doi.org/10.1080/0144929X.2011.553740

[66] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)* (2021), 109–119. https://doi.org/10.1145/3397481.3450662

[67] H Holden Thorp. 2023. ChatGPT is fun, but not an author. , 313–313 pages.

[68] Thi Ngoc Trang Tran, Alexander Felfernig, Viet Man Le, Thi Minh Ngoc Chau, and Thu Giang Mai. 2023. User Needs for Explanations of Recommendations: In-depth Analyses of the Role of Item Domain and Personal Characteristics. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 54–65. https://doi.org/10.1145/3565472.3592950

[69] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. http://www.jstor.org/stable/1738360

[70] Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in Human-AI Interaction: Scoping Out Models, Measures, and Methods. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 254, 7 pages. https://doi.org/10.1145/3491101.3519772

[71] E Union. 2021. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. *COM/2021/206final* (2021).

[72] Heidi Vainio-Pekka, Mamia Ori-Otse Agbese, Marianna Jantunen, Ville Vakkuri, Tommi Mikkonen, Rebekah Rousi, and Pekka Abrahamsson. 2023. The Role of Explainable AI in the Research Field of AI Ethics. *ACM Trans. Interact. Intell. Syst.* 13, 4, Article 26 (Dec. 2023), 39 pages. https://doi.org/10.1145/3599974

[73] Henrique Vasconcelos, Michael Jörke, Margaret 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. https://doi.org/10.1145/3579633

[74] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Quay Au, B. Bischl, Markus Bühner, and Heinrich Hussmann. 2019. Opportunities and Challenges of Utilizing Personality Traits for Personalization in HCI. https://api.semanticscholar.org/CorpusID:204377339

[75] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law and Technology* 31, 2 (2018), 841–887.

A Combined GLM Tables

Average Prediction Changed		
Predictor	**Estimate**	**p-value**
Intercept	2.771	0.184
Age	-0.0068	0.626
Gendermale	0.185	0.553
Gendernon-binary	2.404	0.158
Education level	-0.049	0.571
AI Scale	-0.076	0.411
Hrs.Exp	0.101	0.666
Subj.Exp	-0.132	0.516
ML in Role	-0.095	0.409
Extraversion	-0.244	0.818
Agreeableness	1.284	0.353
Conscientiousness	1.173	0.335
Neuroticism	0.424	0.722
Openness	-0.609	0.622
AIC	-0.55025	

Inappropriate Reliance		
Predictor	**Estimate**	**p-value**
Intercept	4.225	0.217
Age	-0.004	0.856
Gendermale	0.366	0.479
Gendernon-binary	-1.114	0.504
Education level	-0.182	0.208
AI Scale	0.027	0.864
Hrs.Exp	-0.268	0.496
Subj.Exp	0.207	0.551
ML in Role	-0.195	0.305
Extraversion	-0.478	0.797
Agreeableness	-3.537	0.155
Conscientiousness	3.550	0.0771 .
Neuroticism	2.072	0.296
Openness	0.766	0.704
AIC	-129.82	

Average Change in Confidence		
Predictor	**Estimate**	**p-value**
Intercept	-0.736	0.370
Age	-0.001	0.853
Gendermale	0.190	0.132
Gendernon-binary	0.132	0.767
Education level	0.029	0.389
AI Scale	-0.061	0.119
Hrs.Exp	-0.157	0.118
Subj.Exp	0.122	0.164
ML in Role	0.041	0.399
Extraversion	-0.289	0.482
Agreeableness	0.603	0.274
Conscientiousness	0.778	0.103
Neuroticism	0.278	0.539
Openness	-0.222	0.642
Adj. R-squared	0.01133	
p-value	0.3391	

Average Confidence in AI		
Predictor	**Estimate**	**p-value**
Intercept	2.103	0.0502 .
Age	0.008	0.286
Gendermale	0.088	0.589
Gendernon-binary	-0.914	0.115
Education level	-0.016	0.719
AI Scale	0.105	0.039 *
Hrs.Exp	-0.145	0.264
Subj.Exp	0.101	0.370
ML in Role	0.101	0.112
Extraversion	0.573	0.285
Agreeableness	0.596	0.405
Conscientiousness	-0.351	0.569
Neuroticism	-0.254	0.666
Openness	-0.132	0.832
Adj. R-squared	0.05948	
p-value	0.06336	

Change in Trust		
Predictor	**Estimate**	**p-value**
Intercept	-2.846	0.775
Age	0.125	0.062 .
Gendermale	2.044	0.182
Gendernon-binary	-9.451	0.081 .
Education level	0.301	0.468
AI Scale	0.643	0.174
Hrs.Exp	0.488	0.688
Subj.Exp	-0.283	0.789
ML in Role	1.013	0.089 .
Extraversion	2.746	0.582
Agreeableness	2.188	0.743
Conscientiousness	-8.374	0.147
Neuroticism	-0.381	0.945
Openness	-6.012	0.301
Adj. R-squared	0.06616	
p-value	0.04788	

B Combined GAM Tables

Change in Trust		
Predictor	**edf**	**p-value**
Extraversion	1.000	0.851
Agreeableness	3.358	0.221
Conscientiousness	1.000	0.102
Neuroticism	1.000	0.960
Openness	1.000	0.373
Age	1.000	0.051 .
Gender	1.009	0.003 **
Education Level	1.000	0.028 *
Hours-Based Experience	1.000	0.956
Openness × Hours-Based Experience	1.550	0.079 .
Neuroticism × Hours-Based Experience	3.976	0.003 **
Subjective Rating of Experience	4.540	0.010 **
AI SCALE	1.000	0.064 .
ML in role	5.314	0.0059 **
Adjusted R^2	0.260	

Inappropriate Reliance		
Predictor	**edf**	**p-value**
Extraversion	1.559	0.389
Agreeableness	1.000	0.033 *
Conscientiousness	1.000	0.041 *
Neuroticism	1.000	0.284
Openness	1.000	0.710
Age	1.000	0.732
Gender	1.000	0.384
Education Level	1.000	0.271
Hours-Based Experience	1.000	0.472
Openness × Hours-Based Experience	4.47e-05	0.145
Neuroticism × Hours-Based Experience	9.86e-04	0.133
Subjective Rating of Experience	1.000	0.540
AI SCALE	1.000	0.895
ML in Role	1.000	0.219
Adjusted R^2	-0.007	

Average Prediction Changed		
Predictor	**edf**	**p-value**
Extraversion	1.000	0.969
Agreeableness	1.000	0.357
Conscientiousness	1.000	0.317
Neuroticism	1.000	0.722
Openness	2.387	0.393
Age	1.000	0.632
Gender	1.507	0.433
Education Level	1.000	0.445
Hours-Based Experience	1.000	0.666
Openness × Hours-Based Experience	8.71e-06	0.705
Neuroticism × Hours-Based Experience	1.83e-06	0.351
Subjective Rating of Experience	1.157	0.655
AI SCALE	1.000	0.411
ML in role	1.000	0.455
Adjusted R^2	0.0022	

Average Confidence in AI		
Predictor	**edf**	**p-value**
Extraversion	1.000	0.279
Agreeableness	1.000	0.324
Conscientiousness	2.926	0.064 .
Neuroticism	1.301	0.814
Openness	1.000	0.803
Age	3.497	0.285
Gender	1.811	0.144
Education Level	1.000	0.721
Hours-Based Experience	1.000	0.295
Openness × Hours-Based Experience	4.400	0.047 *
Neuroticism × Hours-Based Experience	7.82e-10	0.829
Subjective Rating of Experience	1.000	0.322
AI SCALE	3.698	0.061 .
ML in Role	1.000	0.020 *
Adjusted R^2	0.215	

Average Change in Confidence		
Predictor	**edf**	**p-value**
Extraversion	1.000	0.614
Agreeableness	1.000	0.259
Conscientiousness	3.300	0.007 **
Neuroticism	1.000	0.271
Openness	1.000	0.087 .
Age	1.000	0.857
Gender	1.000	0.007 **
Education Level	1.000	0.606
Hours-Based Experience	1.000	0.446
Openness × Hours-Based Experience	7.624	0.015 *
Neuroticism × Hours-Based Experience	2.926	0.854
Subjective Rating of Experience	1.000	0.539
AI SCALE	1.000	0.067 .
ML in role	1.000	0.057 .
Adjusted R^2	0.229	

C Intake Survey

C.1 Demographics and Background

(1) Gender (open-text)
(2) Age (open-text)
(3) Education Level
- Some high school, no diploma
- High school graduate, diploma or equivalent (e.g., GED)
- Some college credit, no degree
- Tread / technical / vocational training
- Associate degree
- Bachelor's degree
- Master's degree
- Professional degree
- Doctoral degree

(4) Your major for your education degree, e.g., Computer Science, Data Science, Information (open-text)
(5) Occupation (open-text)
(6) How long have you been in your current job or student role (please enter time in months)? (open-text)
(7) Are you currently a resident of the United States? (binary)
(8) To what extent is Machine Learning a part of your daily job or student role? (scale 1–7)
(9) How would you rate your Machine Learning knowledge? (scale 1–7)
(10) How many hours (estimate) have you spent on Machine Learning-related tasks (e.g., data preprocessing, model building)? (scale borrowed from [34])
- I have never done a Machine Learning task
- Less than 10 hours
- 10-20 hours
- 20-50 hours
- 50-100 hours
- More than 100 hours

(11) How familiar are you with interpretability tools for Machine Learning (e.g., LIME, SHAP, GAMs)? (scale of 1-7)
(12) How many hours (estimate) have you spent using interpretability tools for Machine Learning?
- I have never used an interpretability tool before
- Less than 10 hours
- 10-20 hours
- 20-50 hours
- 50-100 hours
- More than 100 hours

C.2 Personality

Participants completed the Big 5 Inventory, a validated scale used to measure personality developed by John et al. [30].

C.3 ML Literacy Assessment

For brevity, we have omitted the multiple choice options from these questions. The options can be found in the research paper about this assessment by Hornberger et al. [25]

(1) What is a key criterion for the quality of a model in machine learning?
(2) What should be considered in machine learning when dividing the data into training and test data?
(3) What is the black box problem?
(4) You are testing a machine learning model that is supposed to classify photos of animals. You notice that the model is better at recognizing cats than dogs. What could be the reason for this?
(5) What are knowledge representations in the field of AI?
(6) How does supervised learning differ from unsupervised learning?
(7) Rank the process steps in supervised learning into the correct order by dragging and dropping.
(8) Which ethical principles should be considered when developing AI?
(9) What are central risks in using AI for predictive policing?
(10) Which legal challenges do AI applications entail?

D Main Survey

D.1 Unassisted Task Overview

Participants were introduced to the unassisted classification task and given an overview of the Adult Income dataset [7].

D.2 Practice Task

For brevity, we include only one example of the four practice tasks. Once participants clicked on their classification decision, a pop-up appeared with the correct answer and feedback.

Feature	Value
Age	47
Work class	Self-emp-inc (self employed incorporated)
Education	HS-grad
Marital status	Never married
Occupation	Other-service
Race	Amer-Indian-Eskimo
Sex	Male
Capital Gain	0
Capital Loss	0
Hours per Week	56
Native Country	Scotland

(1) According to you, what is the predicted income for the person represented in this data point?
- <= 50k
- > 50k

D.3 Unassisted Task

For brevity, we include only one example of the eight unassisted data point classifications.

Feature	Value
Marital Status	Married-civ-spouse (married to a civilian spouse)
Age	39
Hours per Week	24
Occupation	Tech-support
Education	Assoc-acdm
Capital Gain	0
Capital Loss	0
Race	White
Sex	Male

(1) According to you, what is the predicted income for the person represented in this data point?
- <= 50k
- > 50k

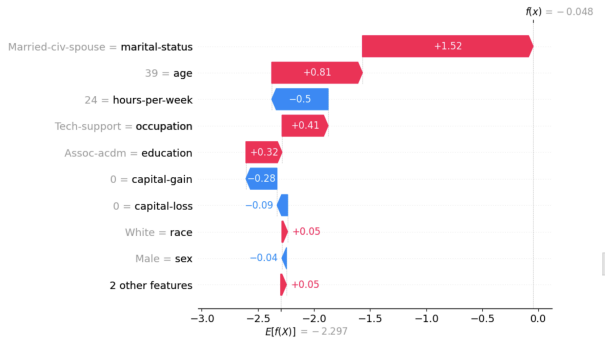
(2) How confident are you about your prediction on their income above?
- Scale 0 to 6

D.4 AI-Assisted Task Overview

Participants were told that they would classify 8 datapoints with the assistance of an AI system. We provided instructions on how to interpret the AI outputs. Here, these outputs were SHAP waterfall plots and the instructions were based on consolidated information from SHAP tutorials. Before the task, participants also completed a validated trust questionnaire developed by Jian et al. [28].

D.5 AI-Assisted Task

For brevity, we include only one example of the eight AI-assisted datapoint classifications.
The datapoint was classified as <= 50k by the AI. Here is the SHAP explanation plot for this prediction. Use it to answer the questions below.



(1) Having seen the AI's prediction, what do you think is the predicted income for the person represented in this data point?
- <= 50k
- > 50k

(2) How confident are you about your prediction on their income above?
- Scale 0 to 6

(3) To what extent do you think that the AI made the right prediction for this data point?
- Scale 0 to 6

D.6 Post-Study Questionnaire

(1) Complete the trust questionnaire developed by Jian et al. [28] again, after using the AI system.
(2) Describe your decision making process when you did not have the AI's prediction help.
(3) Describe your decision making process when you had the AI's prediction help.
(4) (Optional) How was your overall experience with the study? Any comments or concerns?