User Profiling in Human-AI Design: An Empirical Case Study of Anchoring Bias, Individual Differences, and AI Attitudes

Mahsan Nourani[1], **Amal Hashky**[2], **Eric D. Ragan**[2]

[1]The Roux Institute at Northeastern University
[2]University of Florida
m.nourani@northeastern.edu, ahashky@ufl.edu, eragan@ufl.edu

Abstract

People form perceptions and interpretations of AI through external sources prior to their interaction with new technology. For example, shared anecdotes and media stories influence prior beliefs that may or may not accurately represent the true nature of AI systems. We hypothesize people's prior perceptions and beliefs will affect human-AI interactions and usage behaviors when using new applications. This paper presents a user experiment to explore the interplay between user's pre-existing beliefs about AI technology, individual differences, and previously established sources of cognitive bias from first impressions with an interactive AI application. We employed questionnaire measures as features to categorize users into profiles based on their prior beliefs and attitudes about technology. In addition, participants were assigned to one of two controlled conditions designed to evoke either positive or negative first impressions during an AI-assisted judgment task using an interactive application. The experiment and results provide empirical evidence that profiling users by surveying them on their prior beliefs and differences can be a beneficial approach for bias (and/or unanticipated usage) mitigation instead of seeking one-size-fits-all solutions.

Supplemental Material: https://osf.io/qmv6k/

1 Introduction

Due to the rapid advancements in Artificial Intelligence and Machine Learning (AI/ML) and their deployment across various contexts and domains, many laypeople have begun to incorporate AI tools into their daily tasks. Laypeople, despite limited or no technical knowledge of AI/ML, may already possess views and beliefs on this technology. These shaped influences stem from various media sources, such as social, news, literature, and entertainment (Nader et al. 2022). While the formed perceptions and prior beliefs may not accurately reflect the true nature of an AI system, it can potentially influence the way users engage and interact with it. On the other hand, during initial interactions with an AI-supported tool, users form new, evolving "beliefs" of the system. For instance, users may form positive or negative first impressions of an AI system based on when they are exposed to model errors (Nourani, King, and Ragan 2020; Nourani et al. 2021; Tolmeijer et al. 2021). These formed impressions influence user mental model formations, trust, and other human factors. However, there is limited understanding of how prior beliefs about AI and long-term interpersonal and experiential differences influence users' perception formation and other behaviors during interactions. This is particularly challenging because these beliefs are often difficult to capture and vary widely in type and level of abstraction. Furthermore, there are no known approaches to collectively capture and encompass this diversity.

In this paper, we investigated the influences of users' long-term past differences and AI beliefs on their formed impressions of an AI system during usage, as well as the interplay between these factors. We hypothesize that these differences may influence usage behaviors more prominently than designers typically anticipate. To address this question, we need to: i incorporate methods to capture and measure a wide range of prior beliefs and differences, and ii examine a method that collectively encompasses these captured behaviors into similarity-based profiles, allowing us to compare usage behaviors based on these profiles.

We designed a between-subjects user study to control the order of users' exposure to model errors. Previous studies have used similar designs to demonstrate that the timing of errors can lead to positive or negative anchoring bias, with each anchoring group showing varied behaviors with the same model (Nourani, King, and Ragan 2020; Tolmeijer et al. 2021; Nourani et al. 2021). We extend these studies by accounting for prior user beliefs and interpersonal differences, studying how these factors affect user anchoring and other usage behaviors, despite their controlled anchor group. To capture these differences, (i.e., i), we propose using standardized behavioural questionnaires, as they are both easy to implement and allow for capturing multiple differences simultaneously. Moreover, these questionnaires are often designed with predefined scores that categorize subjects into respective groups. In the user study, the participants completed a selection of five standardized questionnaires to capture personality traits, experiences with, and perspectives of the AI technology, *before*[1] engaging in a human-AI collaborative task (Figure 3). These questionnaires cumulatively produced 13 metrics based on their stan-

[1]To ensure new impressions from AI usage would not affect their responses.

dardized scores. We used these metrics as features to train a clustering model to profile the users into similar groups based on their responses to all the questionnaires (ii). Using this metric, we investigated the statistical differences and interactions between *user profiles* and controlled *anchoring group*, and their influences on user behaviours in human-AI collaborative decision-making. The study provides strong evidence that behavioral and background questionnaires can be effectively used for user profiling, with individuals in each cluster sharing similar beliefs and traits. Additionally, our results indicate that people of different profiles may exhibit distinct anchoring behaviors, even under controlled conditions. We maintain that profiling users based on their past beliefs of AI and interpersonal differences allow designers to better navigate and understand anchoring behaviours, and potentially mitigate or limit these and other harmful or unexpected usage behaviours. This paper establishes foundational evidence that user profiling based on past experiences and beliefs can be effective, marking an initial step toward personalization for bias mitigation.

2 Related Work

In alignment with our work, researchers have examined the role of interpersonal differences, prior experiences, and beliefs toward AI in human-AI collaboration and interactions. Among others, personality traits and AI or domain experience are the topics most explored (Kouki et al. 2019; Ehsan et al. 2021; Millecamp et al. 2019; Lindvall, Lundström, and Löwgren 2021). For instance, Cai et al. (2022) found that three types of personal characteristics (*personality traits*, *domain knowledge*, and *trust propensity*) can significantly influence *how* users develop trust in conversational recommender systems. Similarly, Nourani et al. (2020) investigated the effects of domain experience on first impression formations and trust. They demonstrated that only individuals with domain experience exhibit first impressions, and these impressions are reflected in the changes of their trust over time unlike the novices.

In the context of content moderation, Molina and Sundar (2022) discovered that fear of AI, combined with other personal differences such as trust in humans, power usage, and political ideology, can predict whether users will exhibit positive or negative heuristics. Among these studies, the latter is perhaps the closest to our work, as it specifically explores the interplay of past beliefs and personality traits with biased behaviors.

Personalization and user profiling are not novel concepts, and numerous papers have explored adapting outcomes based on individuals or groups with shared similarities. Proposing user adaptation is not a unique contribution of our paper. Instead, the novelty in our work lies in the criteria used for user profiling. Prior work in the AI and HCI community has focused on personalizing the content based on what the models learn about a user over time and through interactions (Jahanbakhsh et al. 2023; Siirtola and Röning 2019; Schelenz et al. 2023). For example, Jahanbakhsh et al. (Jahanbakhsh et al. 2023) designed a personalized AI system that feeds the user input back to the model as training data to predict their assessment of future content. Unlike prior work, the purpose of our study is to establish a better understanding of using personal differences, attitudes, and beliefs toward AI as means for user profiling; specially, as an approach for bias or harmful behaviour mitigations. A study closely related to ours is from the field of Human-Robot Interaction (HRI): the work by Huang et al. (2021) employs The Big Five personality traits questionnaires to validate the usability of asynchronous questionnaires (as opposed to embedding them in the conversation) in extracting relevant behavioral information to build user profiles based on those traits. They observed that the correlation between the two is stronger among younger participants than their adult counterparts. However, even in HRI, behavioral questionnaires were not utilized for user profiling but rather to comprehend users' backgrounds before and after interaction (Ahmad, Mubin, and Orlando 2017). Our work concentrates on exploring *various* types of prior beliefs and differences among users, as opposed to *only* focusing on one factor (domain knowledge); including user AI literacy, personality traits, and beliefs about AI. Furthermore, we explore utilizing these factors for user profiling to understand how individuals with varying profiles exhibit different anchoring behaviors, even in the presence of other anchoring factors. In this sense, our contribution is novel and provides further insights into improving AI systems by considering user differences and past experiences.

3 User Experiment

3.1 Research Goals and Hypotheses

Individuals construct their own interpretations of AI technology through firsthand experiences and implicit learning (Reber 1989). When collaborating with an AI system, their prior beliefs, misconceptions, and attitudes can shape their interactions and decision-making processes, often leading to harmful behaviours. Such misconceptions are more pronounced among laypeople, who lack a robust foundational knowledge of AI to calibrate their judgments and use as a safety net (Nazaretsky, Cukurova, and Alexandron 2022). Our goal is to explore how user differences influence human-AI collaborations, assess the potential of grouping users based on similar traits and attitudes, and evaluate whether profiling users based on similarities in interpersonal differences can accurately predict usage and anchoring behaviors. Acknowledging that user differences and prior beliefs stem from various sources, we consider differences of three main sources as a proof-of-concept to investigate our research aims. These sources are 1) implicit knowledge of AI (non-expert knowledge), 2) individual characteristics (personality traits), and 3) attitudes toward AI (i.e., anxiety, fear, acceptance, and positive/negative attitudes).

In particular, this user study is motivated by the following research questions:

- How do individual differences and past experiences affect usage behaviours with intelligent systems?
  - **RQ1**: Which type of preexisting beliefs and experiences exerts a stronger influence on human-AI col-

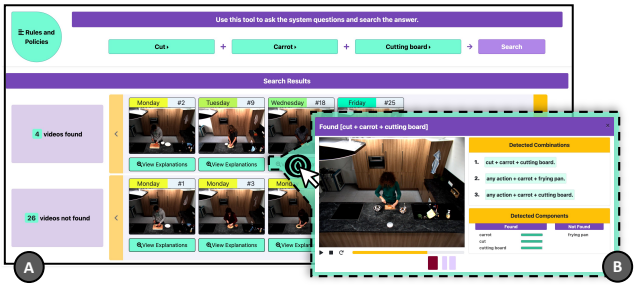

Figure 1: The XAI application used in the study. (A) shows the query-building tool, which allowed users to search activities within videos. By clicking "Rules and Policies" at the top left corner, a sliding panel appeared that included a list of kitchen policies. The search results are grouped based on if the video was matched with the query. Each video is represented with a thumbnail preview, a unique number, and its corresponding weekday. Upon selecting *View Explanations*, a modal window appeared, featuring a video player and the explanations for the model's prediction for that query (B).

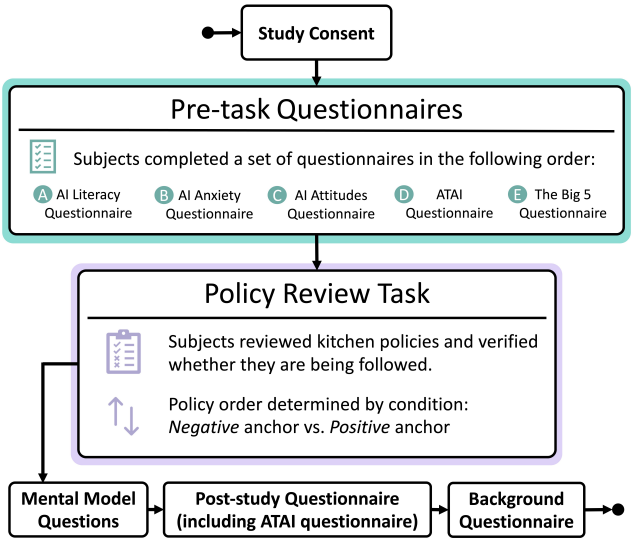

Figure 2: An overview of the study procedure.

laboration: *implicit knowledge*, *individual characteristics*, or *attitudes and feelings towards AI*?

- **RQ2**: How can we categorize people into distinct groups by profiling based on these measured differences?
- **RQ3**: How do user's first impressions and usage behaviours change based on their assigned profiles?

We will select a variety of standardized questionnaires to measure these constructs. **RQ1** is designed to examine the questionnaire measures and their correlations. **RQ2** examines an approach to use questionnaire metrics/measures in identifying and building user profiles. Finally, **RQ3** allows us to investigate how these user profiles canpredict outcomes of human-AI collaboration, as well as the beliefs and biases they develop during the process. We will use anchoring bias as an example of these biases, as it is a behavioral effect that is known to develop differently depending on a user's initial experiences with an AI system (Nourani, King, and Ragan 2020; Tolmeijer et al. 2021; Nourani et al. 2021). This profiling approach shows promise as a strategy for mitigating bias in the future, as we hypothesize that individuals from different profiles may form distinct first impressions, regardless of how they are anchored in the user study.

3.2 Experimental Design

Study Task For this study, we aimed to identify a task that involves elements of human-AI collaboration. We sought a real AI system that is sufficiently complex and open-ended to enable measuring users' decision-making performance and their mental models. We used an XAI system that we previously designed and developed al. (Nourani et al. 2021) for an open-ended and exploratory human-AI collaboration task[2] (Figure 1). For simplicity, we will refer this prior work as *Nourani et al. (2021)* for the rest of this paper. In summary, this system is designed for video activity recognition in a kitchen setup. In each video, a person is seen performing various activities related to cooking. The interface allows users to query the XAI using combinations of {*action, object, location*}. Along with the tool, we replicated the open-ended task from Nourani et al. (2021), as their task selection aligned with ours. The study task revolved around video review requiring participants to assess whether kitchen employees were adhering to a set of rules and policies. The hypothetical policies were designed to necessitate participants to formulate and test multiple queries in order to validate each one. This human-AI collaborative task is beneficial in two ways: (1) it allowed individuals to observe the model perform enough to build mental models of the XAI system, and (2) it allowed us to control *when* participants would be exposed to the *model weaknesses* and *strengths*, to control anchoring bias. Nourani et al. (2021) achieved the latter by designing four policies that expose participants to *XAI weaknesses* and four policies that expose them to the *XAI strengths*. Each policy was created as a true/false statement, and participants had to verify their correctness by responding *yes* or *no*.

Conditions To address our research hypotheses and questions, we used a between-subjects experimental design by controlling the order of the policies in the list. Because people tend to start at the top of the list and work their way down, controlling the ordering can anchor people's first impressions of the algorithm (Nourani et al. 2021). Regardless of condition, all participants observed *the same set of policies*. However, participants in one ordering group first encountered policies that expose model weaknesses at the top of the policy list (*negative* anchor), while the other ordering moved these "problem" policies to the bottom of the list (*positive* anchor).

[2]The open-source software from our previous work (Nourani et al. 2021) was cloned from: https://github.com/MahsanNourani/Pineapple.

Questionnaire	Author(s)	Scale	Item #	Measure Code Name(s)
A Non-Experts' AI Literacy	Laupichler (2023)	11-point	38	NXAI
Description: Captures people's ability to critically evaluate AI technology competencies to use them in daily lives.				
B AI Anxiety	Li (2020)	7-point	20	AIANX
Description: Captures people's anxiety toward AI technology, with questions covering multiple sub-categories.				
C AI Attitudes	Schepman (2020)	5-point	20	ATT_{Pos}, ATT_{Neg}
Description: Captures people's positive and negative attitudes toward AI systems.				
D Attitudes Toward AI	Sindermann (2021)	11-point	5	$ATAI_{Fear}$, $ATAI_{Acc}$
Description: Captures people's broad fear and acceptance of AI technology.				
E Big-5 Personality Traits	Lang (2011)	7-point	15	5 measures – See description
Description: Describes five dimensions of personality; extraversion, agreeableness, conscientiousness, neuroticism, and openness.				

Figure 3: Questionnaires employed in the user study. Higher values across all measures indicate increased effects.

Our anchoring bias conditioning is done similarly to the experiment in (Nourani et al. 2021). However, unlike their study, ours did not manipulate the presence or absence of explanations. Instead, all participants had access to full explanations throughout the task.

User Study Measures We collected participants' responses to the main policy task, along with comprehensive interaction logs. Participants also completed a post-study task, rating their perception of accuracy for nine model activity components on a scale of 0-100% and giving a binary confidence rating (High vs. Low) for their responses. These measures, designed by Nourani et al. (2021), assessed user mental models of XAI strengths and weaknesses. The question set included four components reflecting model strengths (high detection accuracy) and five components indicating model weaknesses (low detection accuracy).

At the end of the study, participants were asked to provide an overall estimation of the model's accuracy in percentage, followed by a short background questionnaire to collect their demographics. To measure participants' prior perceptions of AI technology and their personality traits, we included various questionnaires prior to the main task. These questionnaires are described in the following section.

Pre-Task Questionnaires To explore and capture the impact of interpersonal differences and prior beliefs of human-AI collaboration, we utilized standardized behavioral questionnaires. Administering the questionnaires before the main AI-assisted task replicates the scenario of a new user interacting with an AI system, allowing us to profile them for better customization. Our goal was to categorize participants based on their questionnaire responses, utilizing the questionnaire "scores" as features for clustering them into groups with shared patterns of responses. Questionnaires, commonly used for measuring individual and cognitive differences, serve as suitable data collection tools for this purpose. We selected six different instruments that allow us to capture people's individual differences and experiences towards AI. The questionnaires are included in the supplemental material[3], and Figure 3 provides an overview of them.

Procedure We conducted the study online and through a web-based interface. Several attention checks were included for quality control with the crowd-sourced data; additional information is available in the Supplemental Material. The study was approved by the Institute's Review Board (IRB), and took approximately 30–40 minutes to complete. Figure 2 provides an overview of the study procedure.

After consenting to participate in the study, participants completed the questionnaires introduced in Section 3.2, that were chosen to measure their pre-existing perceptions and attitudes towards AI as well as their personality traits. These questions were administered prior to the main task because we aimed to minimize: (1) the effects of survey fatigue by limiting the number of questions per page as they progressed through the last questionnaire; (2) the availability bias that might be caused by seeing the negative statements about AI right before using the XAI system for the main task; and (3) the potential biases caused by seeing negative or positive statements as the first questionnaire. For these criteria, we chose the order as seen in Figure 2.

After completing a short tutorial, the participants completed the main policy review task (in one of the two study conditions) where they were asked to verify 9 kitchen policies. Note that the order of these policies were altered based on the assigned condition (as described in Section 3.2), while the 5th policy was the same policy and was designed to serve as attention check. The participants concluded the study by answering background and mental model questionnaires. To constrain overall study duration, we aimed to keep these questions lightweight and require low mental demand. However, to find out if using a real AI system can shift participants' attitudes toward AI, they were asked to fill out the ATAI questionnaire ($N = 5$) again.

Participants Participants were recruited through University of Florida's crowdsourcing platform (received up to 1% extra credit) and via word-of-mouth (no compensation). To

[3]Supplemental material is available on https://osf.io/qmv6k/.

address **RQ2** and increase diversity when categorizing people into distinct groups, we continuously recruited participants through the crowdsourcing systems from June to December 2023 without limiting the number of participants, resulting in a total of 471 study completions. They consisted of 161 females, 295 males, and 8 non binary people (while 7 preferred not to disclose their gender), with a majority ($> 97\%$) reporting their age between 18 to 34 years old. The participants were randomly assigned to a condition. There were 227 in the *negative* anchor condition and 244 in the *positive* anchor condition. Additionally, 72.3% of the participants stated they have not taken any AI/ML related courses.

4 Results

4.1 Data Processing

We conducted a thorough data review process to ensure the integrity and quality of the dataset used for the final analysis, resulting in the exclusion of incomplete/corrupt data points *or* failed to pass any of the attention checks. This resulted in excluding 50 and 45 total from the *negative* anchor and the *positive* anchor conditions, respectively, and remaining data from 376 participants was used for statistical analysis (*negative* anchor: 177, *positive* anchor: 199).

Prior to analyzing the results, we made adjustments to some of the measures to ensure they were aligned with the research goals. As the study's online format may have led to attention lapses among participants, we aimed to minimize these lapses to ensure consistent data analysis by defining "time per trial" as the duration between answering two policies. Using Tukey's fences outlier detection method (Seo 2006), we replaced trial times beyond the $\pm 1.5 \times IQR$ with average time in the condition (excluding the outlier), and then calculated average trial time per participants over all 9 trials. Additionally, we evaluated the influence of AI predictions and explanations on task performance by creating an approximate user agreement measure based on policy responses. This measure considered how often participants correctly assess policies showcasing model strengths or incorrectly assess those exposing weaknesses. The fifth policy, an attention check, was excluded. While this measure approximates participant performance that may differ due to various factors, it offers a reasonable assessment of alignment between user decisions and model advice.

Following our prior work (Nourani et al. 2021), we categorized participants' mental models into two factors to assess perceptions of model strengths and weaknesses separately. From 9 post-study tasks, 5 reflected model strengths, and 4 reflected weaknesses. We calculated estimation errors by comparing reported accuracies to ground truths, with positive values indicating overestimation and negative values indicating underestimation. For each of these categories, we calculated the average percentage of error and participants' self-reported confidence in their mental models. In the final step, we calculated the questionnaires' scores per their scoring guidelines, yielding accurate measures for intended constructs: five for the Big Five questionnaire and one or two measures for the rest. This process involved reversing or grouping questions before calculating each participant's average measure value, ensuring higher values reflect greater levels of the construct. Figure 3 summarizes these measures and their assigned code names for further analysis.

4.2 Questionnaire Analysis

As described in Section 4.1, each pre-study questionnaire produced multiple aggregated measures. In total, 11 items were extracted per participant for further analysis. Additionally, 2 extra measures from the post-study background questionnaire, self-reported AI/ML and Computer Science familiarity, were included.

Correlations Among Questionnaire Measures We computed Pearson correlation coefficients to assess the linear pairwise relationship between these 13 measures. Our analysis reveals that several measures are significantly correlated with one another, as seen in Table 1 in the Supplemental Materials. For instance, the test revealed significant positive correlations of AI acceptance with ML familiarity, non-expert AI literacy, positive AI attitude, and *extroversion*. Simultaneously, AI acceptance has significant negative correlations with AI anxiety, negative AI attitude, neuroticism, and fear of AI. It is crucial to note that the correlations tested and reported here *do not imply causation*. For example, while we can observe a similar trend of responses across measures that gauge negative feelings and attitudes towards AI, we can not conclude that AI anxiety leads to negative attitudes towards AI and vice versa. The correlation analyses was performed merely to draw a better picture of the questionnaires and measures employed for this study.

4.3 Linear Regression Model of Questionnaires for Outcomes

To assess the significant influences of questionnaire measures on usage behaviors in the main study task, we employed separate multivariate multiple regression models for each usage behavior measure, utilizing the questionnaire measures as features (**RQ1**). Despite correlations among measures (as discussed in Section 4.2), we retained all measures because each questionnaire item represents distinct constructs. To confirm this decision, we conducted statistical tests, including a multivariate analysis of variance (*MANOVA*). More details are available in Supplemental Material. Figure 4 shows coefficients for each dependent variable regarding each outcome. The results presented here offer compelling evidence that various factors such as user backgrounds, preexisting beliefs, attitudes, and experiences with AI, as well as individual differences, can have a simultaneous impact on human-AI collaborations. For instance, AI acceptance affects task accuracy, perceived mental models, and user confidence. Extroversion and conscientiousness traits significantly affect participants' perceived mental model and confidence ($p < 0.05$). According to our results, **altogether, *personality traits* have a greater influence on various human-AI collaboration outcomes** (**RQ1**). These results do not consider the experimental treatment; the upcoming sections address the simultaneous influences of questionnaire measures and anchoring conditions.

	$ATAI_{Acc}$	$ATAI_{Fear}$	Neurotic	Extrovert	Open	Agreeable	Conscien.	ATT_{Pos}	ATT_{Neg}	AIANX	NXAI	mlFam	csFam
Agreement	0.18	-0.68	0.12	-0.13	0.74	0.94	-0.28	-0.30	0.22	1.07	**-1.59**	-0.25	0.94
Time	-0.56	0.08	-2.14	-2.33	0.80	1.21	**2.73**	-0.21	0.55	-1.76	0.11	-0.55	0.34
Error	-0.82	0.90	0.61	0.00	1.49	-0.93	1.03	-0.41	0.49	-1.99	-0.35	1.46	**-2.95***
EstAcc	**1.60**	0.07	-0.06	0.03	**1.57**	-0.30	1.16	**-1.11**	0.16	-0.07	-0.09	0.45	0.49
MM_{Stren}	**0.04***	0.00	-0.01	-0.02*	0.01	0.02*	0.02*	0.00	0.03*	-0.02	0.02*	-0.02	0.01
MM_{Weak}	0.02	0.01	-0.01	-0.02	0.01	0.02	0.02	0.02	0.00	0.01	**0.03***	-0.02	0.02
$MMConf_{Stren}$	**4.23***	-0.36	**-4.20***	-2.80	3.25*	1.47	3.90*	-1.69	2.37	1.68	-0.17	0.26	-1.90
$MMConf_{Weak}$	3.62	-1.84	-3.63*	-4.17*	0.80	3.29*	**4.66***	0.68	4.10	0.62	-0.07	-1.31	-0.31

Figure 4: The Multivariate Linear Regression coefficients illustrate correlations between measures (columns) and outcomes (rows), indicating its direction and strength. Larger absolute values imply stronger correlations, but significance differences may vary. Asterics (*) indicate significant coefficients ($p < 0.05$) and bolded items show the strongest coefficient. The colors *Green* (*B&W: lighter color*) shows positive and *Purple* (*B&W: darker color*) show negative regressions.

Clustering and User profiling A primary goal of this study (**RQ2**) was to examine using questionnaires on individual traits and attitudes towards AI to cluster users based on similarities. The profiling approach aims to enable researchers to predict potential biases and harmful behaviors, allowing for customized interactions based on assigned groups. To achieve user profiling, we employed a bottom-up, exploratory approach treating each of the 13 questionnaire measures as a feature in an unsupervised clustering problem. The K-means algorithm (Lloyd 1982; Forgy 1965) was used for participant grouping. Crucial considerations included determining the number of clusters (K) for our sample size and ensuring data scaling across all measures. Despite some measures being correlated, we chose *not* to remove any for clustering, as each questionnaire captured distinct constructs contributing differently to the final clustering.

To determine an appropriate number of clusters (K), we adopted a comprehensive approach using statistical methods and data-specific considerations. Balancing meaningful variability with adequate cluster size in our 376-point dataset was crucial. We used the elbow method and silhouette plots to determine the optimal number of clusters. While the elbow method lacked a clear "elbow" point, silhouette analysis suggested (k=2) as the optimal number, resulting in 194 and 182 data points in clusters 1 and 2, respectively. Further clustering details are available in the Supplemental Material.

To interpret cluster results, we assigned meaningful labels representing most data points within each cluster. Using clustering as the independent factor with two levels, we conducted a one-way ANOVA for each feature to identify significant contributors to each cluster (Figure 5). Based on our findings, only *agreeableness* trait does not contribute to clustering. These labels are later used for a more meaningful discussion of our results in the paper:

1. *AI Skeptics* (profile 1): Individuals in this group display strong fear, anxiety, and heightened negative attitudes towards AI, along with high neuroticism and conscientiousness traits.
2. *AI Optimists* (profile 2): This group displays the highest positivity in AI attitudes AI, being more accepting and less fearful, anxious, and negative. Individuals in this cluster are more open, extroverted, and literate in AI.

4.4 Clusters vs. Anchoring in Outcome Prediction

We consider *user profiles* as a secondary independent variable to assess their interactions and influences in relation to the study condition. Our goal was to examine the impact of user profiles (created based on individual differences) on users' decision behaviors, particularly as they develop new behaviors such as anchoring biases while using the system (**RQ3**). A 2-way independent ANOVA tested differences across the 2×2 comparison, with a Tukey-HSD post-hoc test facilitating pairwise comparison across seven dependent variables wne the interaction effect was significant.

User Task Performance The analysis of the main task showed that compared to their counterparts, participants from the *positive* anchor condition were significantly faster ($F(1, 372) = 5.91$, $p < 0.05$) but made significantly more errors ($F(1, 372) = 5.827$, $p < 0.05$). These results align with our previous work (Nourani et al. 2021) showing users get anchored based on how soon they observe errors. This suggests we successfully replicated the study and findings. The test did not detect interaction effects or effects of user profiles.

User Perceptions and Mental Models Our analysis did not demonstrate significant differences on mental model measures based on either of the conditions, meaning that neither the user profiles nor the policy order independently impacted the either of the mental model measures.

However, a significant interaction effect was observed between user profiles and the policy order on mental models of system strengths ($F(1, 372) = 4.35, p < 0.05$). Pairwise comparison revealed marginally significant differences between people from the two clusters, only for participants from *positive* anchor condition ($p = 0.06$). Of those who observed model strengths early-on (*positive* anchor), those who identify with *profile 1* (AI Skeptics) tended to underestimate the model's accuracy more than those from *Profile 2* (AI Optimists). This marginal significance suggests a differential effect based on the order in which information is pre-

$ATAI_{Acc}$	$ATAI_{Fear}$	Neurotic	Extrovert	Open	Agreeable	Conscien.	ATT_{Pos}	ATT_{Neg}	AIANX	NXAI

Figure 5: A one-way ANOVA assessed significant differences among features between participants from Profiled 1 and 2. *Yellow* (*B&W: lighter color*) cells indicate Profile 2 > Profile 1 for each given feature (column), *Lavender* (*B&W: darker color*) cells show Profile 1 > Profile 2, and no cell background means no sig. differences observed. E.g., those in Profile 2 were sig. more extroverted than their counterparts.

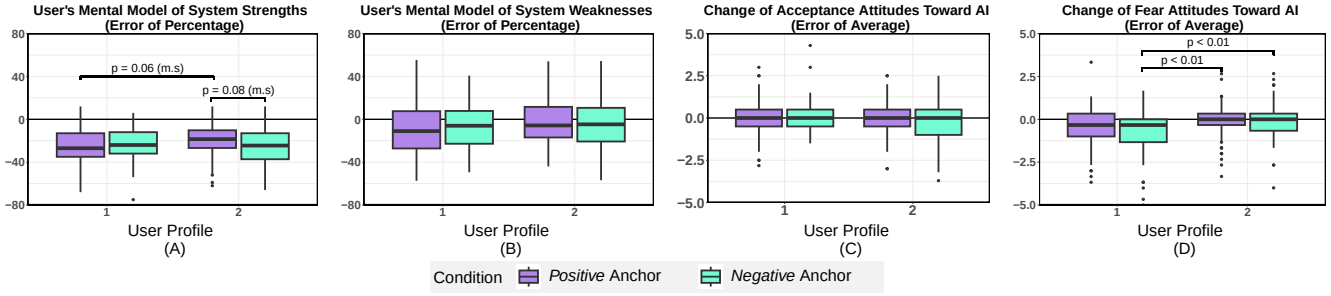

Figure 6: **(A, B)** The distributions of the mental model of weaknesses and strengths measures based on the two-way analysis of anchoring bias and user's background profile. A post hoc analysis revealed marginal significance (m.s) between some of the conditions as shown in the left figure. **(C, D)** The distribution of changes in acceptance and fear attitudes toward AI, by comparing scores pre/post-study. Fear significantly decreased *after* using the AI system for participants from profile 1.

sented, but further investigation is needed to confirm those findings with a statistical significance. Figure 6 (left) shows the data distribution across the different conditions.

The tests also revealed a suggestive main effect based on *user profile* on participants' perceptions of the system weaknesses, with $F(1, 372) = 3.24$, $p = 0.07 (m.s.)$. Comparing *Profile 1* and *Profile 2* distributions in Figure 6 (left), both exhibit extensive ranges, indicating variability in users' estimations of system weaknesses. Despite a slightly higher median for Profile 1, overlapping interquartile ranges suggest a similar spread in the central 50% of data for both groups. This implies that participants in both profiles tend to overestimate the system's accuracy. With a p-value below 0.1, not statistically significant conventionally, we interpret this as a suggestive trend. Further research is needed to explore if this trend holds under different conditions.

Change of Attitudes Toward AI To assess changes in participants' attitudes towards AI following the study, we re-administered the ATAI questionnaire to both control groups, allowing participants to opt-out if desired. In each group, 22 participants chose not to participate. Missing values were then filled using mean imputation, and attitude changes were calculated by subtracting post-study scores from pre-study scores. Positive values indicated an increase in acceptance or fear, while negative values indicated a decrease.

Our results show a significant main effect based on user profile, $F(1, 372) = 14.86$, $p < 0.001$, where participants in *Profile 1* significantly decreased their fear of AI. This indicates the so-called AI Skeptics' fear of AI lowered significantly *after* using the AI system and participating in the collaborative task. The distribution of the responses shows that *AI Optimists* remained indifferent on average regarding changes in fear, with their initial fear toward AI being relatively low on average ($M = 4.15$ out of 11, $SD = 1.21$).

5 Discussion

This experiment aimed to use questionnaires to identify constructs based on long-term influences and individual differences, and to assess their impact on users' behaviors when collaborating with AI systems. By categorizing users into distinct profiles with unsupervised methods, we conducted an in-depth analysis of how these profiles affect users' anchoring behaviors and understanding of the system.

We built multivariate linear regression models based on the study's dependent variables and the measures extracted from the pre-study questionnaires (See Figure 3). **Our results showed that multiple types of user differences and attitudes toward AI can simultaneously affect the outcomes of human-AI collaborations.** For example, individuals with higher levels of AI acceptance, agreeableness, and AI literacy tend to have more accurate mental models of AI's strengths. Conversely, higher levels of ML familiarity and extroversion are associated with less accurate mental models of strengths. This counter-intuitive finding suggests that the ML familiarity question used may not be as effective in measuring lay users' AI literacy compared to the NXAI questionnaire with multiple questions. Caution is needed in interpreting these results as the observed relationships do not imply causation, requiring further research to explore underlying mechanisms. The negative correlation between extroversion and mental model formation warrants further investigation in future research to understand its underlying reasons. Another notable observation is that users heavily depend on their AI literacy to construct improved mental models of an AI system, indicating that higher literacy results in better mental models. **This emphasizes the need for AI education using simple and novice-friendly language to**

enable individuals to form more realistic perceptions of intelligent systems.

Our research also focused on building user profiles based on personal differences and the differences in behaviours based on these profiles. After careful considerations and thorough analyses, we ended up with two user profiles labeled as *AI Skeptics* (profile 1), and *AI Optimists* (profile 2). We then compared these profiles to the controlled anchoring conditions. Our analysis confirms the anchoring bias affecting decision-making behaviours, which is consistent with the previous work (Nourani et al. 2021). For mental models of strengths, our results find a strong interaction effect between the anchoring bias and user profile. Post-hoc tests indicate that **even when users are prompted to form positive first impressions of the algorithm, they are unlikely to develop accurate mental models unless they belong to *profile 2*[4].** Moreover, participants from *profile 2* form better mental models when they are positively anchored. This observation has several implications. First, it provides empirical evidence that user differences and prior experiences can be effective user profiling methods and are important factors in user behaviors. Second, it presents an interesting case where user profiling can be used for personalization. Previous work has found that users with positive first impressions build more accurate mental models (Nourani et al. 2021). However, we see here that positive anchors are not enough for those from *profile 1*[5], and designers might need other personalized techniques to improve their mental models of the system.

Another notable finding is that **those in *profile 1* experience a significant decrease in fear of AI *after interacting* with it**. This surprising observation suggests that their initial fears may have been rooted in a lack of experience with AI systems and a lower level of AI literacy compared to their counterparts (See Figure 5). It also implies that novices can adjust their expectations and misconceptions of AI when they have the opportunity to interact with an AI model. However, further investigations are needed to gain a deeper understanding of why this phenomenon occurs.

Our profiling approach focuses on grouping individuals for anchoring bias mitigation. In other words, if system designers can identify the right questionnaires and methods to measure user differences before they engage with the system, they may be able to personalize initial experiences based on user groups to mitigate the negative anchoring effects, as demonstrated in previous research. **Overall, our results provide supporting evidence that combinations of user differences do indeed affect their anchoring behaviors regardless of observed model performance.** We hope the lessons learned from this study will underscore the importance of considering user backgrounds and serve as a starting point for designing and personalizing systems to better meet user needs.

Designers and practitioners can devise domain-specific strategies to understand and address user differences influenced by preexisting beliefs, prior experiences, and personality traits. These strategies should prioritize: (1) identifying interpersonal differences, (2) selecting appropriate user profiling methods based on these factors, (3) predicting interaction patterns of new users by assigning them to relevant profiles, and (4) personalizing the system based on predicted behaviors. An iterative, user-centered process can tailor the experience for prospective users.

6 Limitations and Future Work

A practical challenge for adopting questionnaire items is covering a sufficient range of scales while limiting total scope of questions. While we found smaller versions for some, like the Big Five, we opted not to modify others to preserve their integrity. This resulted in a total of 98 questions before the study, and some post-study. Despite the median study duration being 35 minutes, indicating a reasonable time commitment, some online participants might have found it tiring, which is a limitation of this profiling approach. Future research may explore questionnaires designed to efficiently elicit various factors with a minimized set of questions, or alternative approaches to collect this information over an extended period.

Identifying an optimal unsupervised clustering approach was challenging. Our goal was to demonstrate a proof-of-concept user study for utilizing pre-usage questionnaires to elicit long-term past influences for user profiling, rather than proposing a specific algorithmic technique. While we used K-Means among several options, we adhered to its assumptions and requirements and employed evaluation methods to assess cluster separation. Future work can benefit from exploring unsupervised algorithms for user profiling based on long-term past influences, considering the specific domain, task, and type of questionnaires involved.

Additionally, we emphasize that the tasks and questionnaires were used for investigative purposes. While standardized questionnaires helped encompass a variety of background factors, including AI attitudes and personality traits, we recognize they may be limited in capturing complex constructs and unintentionally lead to challenges related to ethics and privacy. For instance, incorporating some questionnaires may lead to stereotypical user profiles or obscure the root causes of observed behaviors. To minimize this risk, we removed our assigned profile labels from the Discussion section. However, our true contribution lies beyond these examples, and we firmly believe that this approach can be applied to different questionnaires, tasks, and contexts. We encourage others building on our findings to embrace and leverage individual differences in their designs, specially when combating biases and misconceptions. It is important to use due diligence in selecting questionnaires or other data collection methods that **celebrate and account for user diversity. This is the core message of our paper.** It is also important to note that our findings might reveal different trends with a substantially larger and more diverse sample size. Future researchers can extend this work by actively recruiting people from different backgrounds, demographics, and degrees/occupations.

[4]We labeled users of this profile as *AI Optimist*. We refrained from using this label here to avoid misinterpretations of these findings in future work. This is discussed more in the limitation section.

[5]We labeled users of this profile as *AI Skeptics*.

Acknowledgments

The authors would like to thank the anonymous reviewers and the study participants for contributing their time. This paper was supported in part by NSF award 1900767 and by DARPA under HR00112390063.

Ethics and Privacy Statement

As researchers, we believe it is crucial for future research and practitioners to be mindful of the ethical and privacy implications associated with our user profiling technique. Profiling based on users' backgrounds and interpersonal differences may carry significant risks and should be approached with caution. We urge those applying our methods to prioritize ethical conduct by obtaining clear and informed consent from users before collecting or utilizing their data. Data privacy should be ensured by anonymizing information when training the profiling model, processing it anonymously when categorizing new participants, and using it solely for profiling. Additionally, if profiles are iteratively refined, user data must be securely stored to prevent unauthorized access. Practitioners should remain vigilant about the potential misuse of user data and strive to protect privacy throughout all stages of implementation.

References

Ahmad, M. I.; Mubin, O.; and Orlando, J. 2017. A systematic review of adaptivity in human-robot interaction. *Multimodal Technologies and Interaction*, 1(3): 14.

Cai, W.; Jin, Y.; and Chen, L. 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*, 1–14.

Ehsan, U.; Passi, S.; Liao, Q. V.; Chan, L.; Lee, I.; Muller, M.; Riedl, M. O.; et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509*.

Forgy, E. W. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21: 768–769.

Huang, T.-R.; Liu, Y.-W.; Hsu, S.-M.; Goh, J. O.; Chang, Y.-L.; Yeh, S.-L.; and Fu, L.-C. 2021. Asynchronously embedding psychological test questions into human–robot conversations for user profiling. *International Journal of Social Robotics*, 13: 1359–1368.

Jahanbakhsh, F.; Katsis, Y.; Wang, D.; Popa, L.; and Muller, M. 2023. Exploring the Use of Personalized AI for Identifying Misinformation on Social Media. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–27.

Kouki, P.; Schaffer, J.; Pujara, J.; O'Donovan, J.; and Getoor, L. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 379–390.

Lang, F. R.; John, D.; Lüdtke, O.; Schupp, J.; and Wagner, G. G. 2011. Short assessment of the Big Five: Robust across survey methods except telephone interviewing. *Behavior research methods*, 43: 548–567.

Laupichler, M. C.; Aster, A.; and Raupach, T. 2023. Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence*, 4: 100126.

Li, J.; and Huang, J.-S. 2020. Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. *Technology in Society*, 63: 101410.

Lindvall, M.; Lundström, C.; and Löwgren, J. 2021. Rapid assisted visual search: Supporting digital pathologists with imperfect AI. In *26th International Conference on Intelligent User Interfaces*, 504–513.

Lloyd, S. 1982. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2): 129–137.

Millecamp, M.; Htun, N. N.; Conati, C.; and Verbert, K. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 397–407.

Molina, M. D.; and Sundar, S. S. 2022. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media & Society*, 14614448221103534.

Nader, K.; Toprac, P.; Scott, S.; and Baker, S. 2022. Public understanding of artificial intelligence through entertainment media. *AI & society*, 1–14.

Nazaretsky, T.; Cukurova, M.; and Alexandron, G. 2022. An instrument for measuring teachers' trust in AI-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference*, 56–66.

Nourani, M.; King, J.; and Ragan, E. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 112–121.

Nourani, M.; Roy, C.; Block, J. E.; Honeycutt, D. R.; Rahman, T.; Ragan, E.; and Gogate, V. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *26th International Conference on Intelligent User Interfaces*, 340–350.

Reber, A. S. 1989. Implicit learning and tacit knowledge. *Journal of experimental psychology: General*, 118(3): 219.

Schelenz, L.; Segal, A.; Adelio, O.; and Gal, K. 2023. Transparency-Check: An Instrument for the Study and Design of Transparency in AI-based Personalization Systems. *ACM Journal on Responsible Computing*.

Schepman, A.; and Rodway, P. 2020. Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in human behavior reports*, 1: 100014.

Seo, S. 2006. *A review and comparison of methods for detecting outliers in univariate data sets*. Ph.D. thesis, University of Pittsburgh.

Siirtola, P.; and Röning, J. 2019. Incremental learning to personalize human activity recognition models: the importance of human AI collaboration. *Sensors*, 19(23): 5151.

Sindermann, C.; Sha, P.; Zhou, M.; Wernicke, J.; Schmitt, H. S.; Li, M.; Sariyska, R.; Stavrou, M.; Becker, B.; and

Montag, C. 2021. Assessing the attitude towards artificial intelligence: Introduction of a short measure in German, Chinese, and English Language. *KI-Künstliche intelligenz*, 35: 109–118.

Tolmeijer, S.; Gadiraju, U.; Ghantasala, R.; Gupta, A.; and Bernstein, A. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, 77–87.