SHADOWS OF THE PAST: THE EFFECTS OF USER'S DIFFERENCES AND PAST EXPERIENCES ON HUMAN-AI PARTNERSHIP

By

MAHSAN NOURANI

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2023

To my parents and my brother, for being the moon in my darkest nights.

ACKNOWLEDGEMENTS

I take immense pride in my Ph.D. journey and the choices that have led me to become the researcher and individual I am today[1]. Surely, the experiences of these past years will undoubtedly influence my future, providing a strong foundation for the person I aspire to be and the scientific contributions I hope to make. But let me tell you, I did not walk this path alone. The incredible support I received from the people around me made all the difference. I honestly cannot believe I would have made it without "my village", or as I would like to put it, my "star cluster"[2]. Their belief in me and encouragement have been the backbone of my success, and I am forever grateful for their unwavering support. Admittedly, I cannot possibly name every person who contributed to this journey, as my memory may fail me, and doing so could fill pages, but I will attempt to acknowledge some of the most significant individuals who made all the difference.

First, I would like to thank my advisor, Eric Ragan, for his substantial support all these years. I started my Ph.D. in his lab at Texas A&M university in 2017, and because of his great personality, his research expertise, and advising style, I took a chance and moved to University of Florida to continue pursuing my Ph.D. degree under his supervision. In hindsight, choosing this career path was the best decision I ever made, and I would make it again without hesitation. Throughout these challenging 6 years, especially during the COVID-19 pandemic, I couldn't have asked for a more supportive and exceptional mentor. Eric's jokes and positive mentality made research conversations more enjoyable and reminded me that research can be fun. I will always remember how he openly acknowledged hard work, complimented my qualities, and supported me during moments of imposter syndrome. He never placed blame when things did not go as planned and generously shared credit for mistakes and challenges, demonstrating his qualities as a great advisor. Eric, you were my guiding northern light, helping me overcome obstacles and grow

[1]When I was writing my acknowledgement section, I was listening to Hans Zimmer's original soundtrack for "Interstellar", one of my all-time favorite movies. If you would like to, I suggest listening to "Where We're Going" as you continue reading. Also relevant to this section, is a quote from that movie that is forever engraved in my head: "Love is the one thing we're capable of perceiving that transcends dimensions of time and space." — Dr. Brand, Interstellar.

[2]My community of people, akin to a star cluster held together by self-gravitation, radiated brightness and offered unwavering support during the darkest moments of my Ph.D. journey. They illuminated my path and guided me through the challenges, making my experience all the more meaningful and enriching: https://en.wikipedia.org/wiki/Star_cluster.

as a researcher and person. I'm forever grateful for your support, mentorship, and passion for research, and for treating me as an equal when the time was right.

I extend my gratitude to the members of my committee, Juan Gilbert, Jennifer Wortman Vaughan, Peter Kvam, and Vincent Bindschaedler, for their time, support, and feedback while I was planning and working on my dissertation research. Particularly, I want to extend my heartfelt appreciation to Jenn, for graciously accepting the role of a special committee member despite her demanding schedule and commitments. Along with Hal Daumé III, Solon Barocas, and Forough Poursabzi-Sangdeh, Jenn provided invaluable mentorship during and after my summer internship with the Microsoft Research FATE team, offering me a fresh perspective on conducting human-centered research. Jenn, you are my role model, and I aspire to become a remarkable researcher and mentor just like you one day. Your willingness to share your experiences and offer valuable advice has been extremely appreciated. Thank you.

I would like to thank the members of the INDIE lab and the HCC cohort at University of Florida and the members of the TEILab at Texas A&M University for providing me with a research home and continued support throughout my Ph.D. journey. I am especially grateful to Shaghayegh Esmaeili, Amal Hashky, Jeremy Block, Donald Honeycutt, Sina Mohseni, Fernando Rodríguez, Joseph Wiggins, Kimberly Ying, Memet Celepkolu, Yerika Jimenez, Lydia Pezzullo, Osazuwa Okundaye, and Angela Chan, for their collaborations on my projects, for lending me an ear during my rants about graduate school, and for the memorable game nights and gatherings that nurtured my mental well-being and sense of community. Thank you all sincerely.

A special thank you goes to my collaborators and colleagues, including Vibhav Gogate, Chiradeep Roy, and Tahrima Rahman from University of Texas in Dallas, Stephan Bruckner and Fabian Bolte from University of Bergen, Marco Cavallo from Apple, Emily Wall from Emory University, and Alireza Karduni. Particularly, a special shout out goes to Fabian, for not just being an exceptional research collaborator, but also being my partner in life, who has been patiently supporting me through hours of procrastination and self doubt, ensuring I do not starve while preparing and writing my dissertation draft, and for always lending me a hand in my time of need.

He is like that one person in the audience who proudly smiles at you and stands up to cheer for you with unmatched enthusiasm. Thank you, Fabian, for being there with me throughout the emotional roller coaster of my Ph.D. journey, especially considering you've already navigated and healed from your own. Your constant support and understanding have meant the world to me.

Last and foremost, I would like to express my appreciation and acknowledge the support given to me by my loving family, who shaped the person I am today. They believed in me and encouraged me to aim higher and dream bigger; from my childhood days of dreaming to be a NASA astronaut, when I decided to become a fashion designer, to when I made up my mind to be a computer engineer, they backed me up and never forced me to live "their" dreams instead. To my mom, Fariba, who is my utmost role model and the first woman in science I ever knew; to my dad, Hossein, who helped me fight the social norms and challenges against women and supported me through the thick and thin; and to my only brother, Ehsan, who has inspired me throughout our childhood and continues to do so every day: this one is for you; it is all you. My family not only financially and emotionally supported me my whole life, but they endured years of me studying abroad without the ability to be physically together. I love them more than words can express, and I wholeheartedly dedicate my research and future career to the three of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figures page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

SHADOWS OF THE PAST: THE EFFECTS OF USER'S DIFFERENCES AND PAST EXPERIENCES ON HUMAN-AI PARTNERSHIP

By

Mahsan Nourani

August 2023

Chair: Eric D. Ragan
Major: Computer Science

Recent advancements in Artificial Intelligence (AI) have increased public awareness and engagement with this technology, prompting individuals to assume decision-making roles that complement AI capabilities. In the pursuit of ethical AI systems that prioritize users and stakeholders, it becomes crucial to understand how people's backgrounds, particularly their personal differences and past experiences, influence their behaviors and usage patterns in their collaboration with AI. In my dissertation, I address these open questions by introducing a conceptual model of users' past experiences that describe various types of long-term and short-term past experiences as well as individual differences. This model is derived empirically and supported by a thorough semi-structured literature review, surveying body of work in the field of human-AI collaboration. The literature review complements the conceptual model by gathering and presenting empirical evidence that aligns with its framing, while providing a valuable overview of the current landscape of user-centered research in this domain and identifying gaps in knowledge. The conceptual model is a substantial piece of my dissertation, motivating further investigations of prior experiences based on short-term and long-term past effects, and how they influence one-another during periods of usage. To substantiate its applicability of, I design and conduct a series of human-subjects experiments, which serve to provide additional evidence and practical examples of how the conceptual model can be utilized to frame important research questions and advance our understanding in the field. I first present a

user study on how anchoring bias (as an example of short-term past experiences) affects user perceptions of the system, as well as the efficiency and effectiveness of the human-AI collaborations. In another study, I investigate the effects of anchoring bias when controlling people's level of domain expertise (as an element of long-term experiences) to study the interplay between long-term and short-term past experiences and how they affect user trust and its changes over time. Both studies provide compelling evidence that people's prior experiences play a significant role in shaping their usage behaviors and perceptions. However, due to the intricate nature of people and the wide array of their long-term past experiences, it becomes crucial to delve deeper into understanding the effects of these experiences on usage behaviours during their collaborative interactions with AI. To address this need and gain insights into the impact of various experiences on individual usage differences, a third user experiment is introduced. In this study, the focus is on investigating the potential of using a questionnaire to capture how people's personality traits, pre-existing expectations and knowledge of AI, and attitudes toward AI influence their usage behaviors and mental models of the AI systems. I demonstrate how utilizing questionnaires to measure people's differences and long-term past experiences prior to first-time usage can help predict differences in usage behaviours as a baseline for user profiling to mitigate anchoring bias. My dissertation research provides significant contributions to the fields of human-computer interactions and human-centered AI, serving as a valuable foundation and motivation for future work in these areas.

CHAPTER 1
INTRODUCTION

Over the last two decades, Machine Learning and Artificial Intelligence (ML/AI) approaches have fundamentally altered users' analytical reasoning and decision-making by making computer systems smarter and more intelligent. In fact, a large number of existing ML/AI systems have achieved relative autonomy in that they can decide and act on their own with minimal human intervention. While many AI-powered technology automates outcomes for people's needs, many require people as agents to monitor, analyze the outcomes, and collaboratively make decisions together with the AI. In many systems, partial autonomy is preferred over full autonomy for cases when the final decision is significant and/or the user's expertise is required to evaluate the predictions and reason analytically [203]. This motivates the need for mixed-initiative systems [72] and collaborative human-AI systems, where the goal is to exploit the merits of both human and the machine for better decision making (also known as AI-assisted decision making [203] or data-driven decision-making for visual analytics systems [4]).

The challenge to understanding human behavior with intelligent systems is that many types of human differences may lead to distinct usage behaviors, influencing the outcomes of human and AI collaborative efforts. People's unique personal experiences shapes their understanding of technology and their views of the world. This applies to our context of intelligent systems as well, where individual differences are partially rooted in their mental model of AI as a concept and the technologies/devices supported by it.

Studying these differences is crucial given the complex nature of human minds; people think and reason differently based on their unique past experiences, backgrounds, demographic differences, existing biases, and impressions of AI and technology. The implied impressions of AI may lead to divergence in interactions, perceptions, utilization, and understanding of AI-powered tools and models. Ultimately, these assorted behaviors can take place when people of various backgrounds tend to make high-stakes decisions with the aid of an AI algorithm. The body of work in the field of Human-centered AI (HCAI) consists of numerous work investigating how certain attributes originated from past experiences affect people's usage behaviors. For

instance, researchers have studied how knowledge of AI or domain affects their perceptions of the system and their trust [48, 19, 129]. Investigating the how question is critical for improving intelligent systems while keeping their users and stakeholders in mind. However, little is known about what these different attributes are in the first place. Identifying which prior experiences associate with individual differences in usage of AI systems can not only improve our search for which hows to investigate, but can also motivate finding techniques to circumvent imposed challenges or augment behaviors that are caused by people's differences. Ultimately, we can find solutions for leveraging people's differences in building tools that satisfy their needs and facilitate their usage of such AI systems.

My research is motivated by these questions to investigate users' differences (in usage behaviors) based on their past experiences. In the following section, I will formulate the research goals and questions more elaborately.

1.1 Research Questions and Goals

The primary focus of my dissertation is to comprehend the impact of user differences and past experiences on interactive human-AI systems, with the goal of improving the design and implementation of AI systems. By delving into user differences shaped by prior experiences, a comprehensive investigation of these encounters was undertaken to characterize and explore their implications on human-AI collaboration. This understanding of how people's unique differences and experiences influence their behaviors with intelligent systems is essential in constructing AI/ML systems that are fairer and uphold ethical principles. In particular, my research is motivated by the following Guiding Research Questions (GRQs):

- GRQ1: What past experiences lead to differences with current usage behaviors with intelligent systems that rely on human-AI partnerships?
- GRQ2: How do these individual past experiences affect human-AI collaborations?

- GRQ3: How does identifying differences spawned from past experiences help with predicting usage behaviors and the effectiveness of collaborative decision-making between humans and AI?

I explored these questions with studies and literature review analysis through a human-centered perspective. Each presented work in this manuscript is designed to cover some of these questions by addressing more specific research questions. In the next section, I will provide a summarized overview of my dissertation.

1.2 Summarized Overview

It is important to understand how individual differences and past experiences affect their perceptions and usage of AI-powered systems and tools. People's differences are rooted in prior experiences and circumstances, including, but not limited to: their exposures to AI, their expertise and knowledge of AI and domain, their cognitive and societal biases, and their perceptions of AI. For the purposes of this proposal, we will refer to these by people's past experiences. These prior experiences affect human-AI collaborations and the outcomes and decisions achieved through this partnership.

Before studying how these past experiences affect human-AI teaming, it is critical to understand what these past experiences are, i.e., which elements of the past can impact this relationship. While there has been prior work in the field to organize our understanding of human-AI teaming (e.g., [115, 1, 78, 68, 92]), there is no concise organization that characterizes human's prior experiences that influence people's behaviours with AI system. Such a taxonomy can serve as a road map for intelligent system engineers and designers to build AI systems responsibly. Furthermore, it helps raise their awareness and attention towards people's experiential differences and elevating those to improve the efficiency of human-AI teaming.

To achieve these objectives, I conducted a semi-structured literature review to explore the existing body of work in the field of human-AI partnership. The findings from the literature review were then analyzed and organized to develop a conceptual model that captures the diverse range of users' prior experiences and differences. We categorized the model based on the stability

versus transience of user experiences and characteristics. The model and the results of the literature review analyses are extensively presented in Chapter 3.

Following the framing of the conceptual model of users' past experiences, a first step was to verify and better understand how people' transient experiences with intelligent systems affect their perceptions of the model and current usage; specifically, when forming mental models of systems that they have not used before, as mental models play an important role on usage behaviours. By using anchoring bias (a cognitive bias that refers to people's tendency to rely heavier on the information they receive early-on) as an example of short-term past experience, I designed and conducted a user study to understand how people's mental models and confidence in the AI's outputs is affected their negative or positive impressions of the intelligent system. To investigate the interactions between short-term past experience and the development of mental models, the study investigated the interplay between system explainability and anchoring bias. This study contributes new empirical knowledge of potential effects of short-term recent experiences and provides insights into what to consider when designing explainable AI tools. This study is presented in Chapter 4.

Additionally, to better understand the interplay between transient and stable past experiences as described in the conceptual model, I designed another study to compare elements from each of these categories by controlling anchoring bias and domain expertise. The goal was to understand how present usage behaviours (such as trust) is anchored by people's first impressions of an intelligent system and how past task domain knowledge affects these behaviours and impressions. This study provides evidence to support that prior experiences of recent and long-term/stable nature each play a different role in how people understand and trust an intelligent system, highlighting the importance of investigating human-AI partnership by accounting for people's prior differences, as well as how such differences can be augmented and accounted for when designing AI systems. This study is described in more detail in Chapter 5.

Finally, Chapter 6 extends the work done by the first two studies in order to understand the effects of multiple stable and transient factors simultaneously. In the study presented in this

Chapter, I examined the relationship between people's prior perceptions and experiences of AI and their anchoring effects on their collaborations with AI systems. Recognizing that people's understanding and judgments of AI can be influenced by various sources of information, I explored how these prior perceptions and experiences impact their initial impressions of AI systems and subsequently shape their usage behaviors. To capture the diversity of individuals' prior perceptions and experiences, I incorporated a range of questionnaires to assess their stabilized differences and past experiences related to AI. Through this data collection process, I aimed to identify distinct user profiles based on their responses and investigate whether these different profiles lead to variations in usage behaviors. By analyzing the collected data and examining the correlations between user profiles and usage behaviors, this study aimed to provide insights into the influence of prior perceptions and experiences on people's interactions with AI systems.

My dissertation contributes empirical evidence to demonstrate the importance of studying AI systems from the perspectives of human users and their differences, and advocates for designing tools that elevate these prior differences to improve human-AI collaborations and partnership.

CHAPTER 2
LITERATURE REVIEW

Over the last two decades, Machine Learning and Artificial Intelligence (ML/AI) approaches have fundamentally altered users' analytical reasoning and decision-making by adding levels of automation. In many applications, partial autonomy (as opposed to full autonomy) is preferred when the final decision is significant and/or the user's expertise is required to evaluate the predictions and reason analytically [203]. This creates an opportunity for human-AI teaming, with the goal of exploiting the merits of both human and the machine to improve decision making.

With many applications and used in various domains, many AI models suffer from a lack of transparency, which hampers user's ability to make justified conclusions. Metaphorically, these models are often times referred to as black-boxes (similar to human mind [51]); they transform the input data into predictions or recommendations, without showing their inner workings or rationale for how one outcome is generated as opposed to another. This may result in user confusions and limited understanding, and might induce other (harmful) behaviours when working with black-box models. Due to these challenges, many researchers have been advocating for increased model transparency. In the recent years, there has been growing interest in building explainable artificial intelligence (XAI)—specifically explainable machine learning—systems as a solution to transparency [116, 1]. Notable examples include explainable recommendation systems (e.g., [188, 27, 202]), classification systems (e.g., [151, 86, 6, 187]), activity recognition systems (e.g., [205, 155, 110, 10]), and visual analytics tools (e.g., [69, 165]).

In this chapter, I will first present an overview of the body of work in the fields of XAI and later, delve deeper into human-centered AI/XAI research and its relevant topics to this dissertation.

2.1 Explainable AI

As machine learning and artificial intelligence algorithms are increasingly used to assist with making decisions in high-stakes and impactful scenarios, such as criminal justice systems [156, 13] and medical diagnosis [19, 22, 55], transparency and explainability of the model predictions become vital. Various researchers have explored explainability and

interpretability techniques in AI/ML models and systems. Du et al. [44] present a survey on different interpretability techniques, including post-hoc explanations (where explanations are extracted from the model from local or global perspectives, e.g., [57, 94, 88]), intrinsic explanations (where the model is self-explanatory and is built to be interpretable globally or locally, e.g., [179]), and model specific/agnostic explanations (e.g., [152, 151]). Some researchers have explored explanation by example, where a model provides examples of relevant instances from the training set for a given input instead of attempting to explicitly explain how models reason (e.g., [82, 20, 121]. For instance, Cai, Jongejan, and Holbrook [20] defined and explored two types of example-based explanations in the visual domain and investigated their effectiveness with humans: (1) normative explanations, which establish a norm/trend for the target class by showing training examples, which would help the users understand classifications, and (2) comparative explanations, that show the most similar examples (which can be of different classifications) from the training set to the input.

Researchers also describe types of focus for what is being explained by different explanations. For example, Keane and Kenny [80] argue that transparency tries to reflect how an AI system works, while post-hoc interpretability focuses on the whys in the AI system, providing justification for its outputs. In another work, Hohman et al. [68] provide an interrogative survey on a large number of works in Deep Neural Network (DNN) visualization papers and organize the literature into six categorise based on how the visualization could reveal different aspects of a DNN. These categorise include: (1) Why visualize deep learning models? (e.g., [117, 102]), (2) What data, features, and relationships can be visualized? (e.g., [61]), (3) When is visualization used in deep learning? (e.g., [145, 76]), (4) Who would use and benefit from visualization of deep learning?, (5) How to visualize data, features, and relationships? (e.g., [76]), and (6) Where has the visualization of deep learning been used? (e.g., [153, 197]. While these categories are presented for visualization of deep learning approaches, similar questions are relevant for all interpretable and explainable AI systems.

Some researchers focus on the scope and amount of details for explanations. Following Shneiderman's mantra [162], designers can provide a summarized overview of the explanations before showing too much detail to the users. However, previous research show evidence that oversimplified explanations can create challenges for users as they might decrease the chances of capturing users' attention [90]. Furthermore, Lim and Dey [100] found that more descriptive, detailed explanations cause user disagreement with the system. From these two studies, it can be concluded that balancing the appropriate amount of explanations is as vital as providing explanations itself.

Explanations can either focus on describing the model as a whole (global explainability) or justifying the model's rational on producing specific outputs (local or instance-level explainability) [1, 116]. Most of the current efforts on explainability in Human-Computer Interaction (HCI), AI, and ML communities have been focusing on instance-level explanations. This approach can be useful for when it is important to understand decisions on a single instance (e.g., to understand why a person is being denied a loan). On the other hand, global explanations are powerful to help users expose patterns in the models that can lead to identification of model biases [42], improve user mental models [131, 69], and/or debugging the model [165, 69]. Since information and data visualization approaches are often used to help users make sense of the underlying data and find patterns [45], most of the current body of work on global transparency is in form of visualizations and visual analytics tools (e.g., [69, 165]). Although it may seem binary at first, explanation scope is more of a spectrum and depends on the context, the model, and the application. For instance, Neto and Paulovich [124] present Explainer Matrix; a visualization matrix for interpretable understanding and exploration of random forests that support both global and local explanations.

2.2 Human-Centered AI/XAI

A variety of different research communities have studied the design of user-centered (X)AI systems. The goal is to design while being mindful of users and their needs. In the visualization community, researchers have focused on building and designing tools for XAI systems in an

attempt to improve understanding [165], analytical process [142, 200], debugging [170, 200], and fairness [4]. Work in the HCI community has resulted in user-centered design guidelines and frameworks for XAI systems [49, 186, 190], while others have compared and evaluated design variations for explanations to understand how they can improve user experience in an XAI system [81, 103]. Some HCI researchers focused on building empirical knowledge around explainable intelligent systems by studying user behaviors. Evaluation of mental models is one topic that is explored in the HCI community. In the context of intelligent systems, mental models refer to the user's built and formed concept of how a system works based on their interactions with the system. In 1993, Staggers and Norcio [167] argued that most of the research attempts infer the presence of mental models by comparing users' performances and observing differences in problem-solving between novice and expert users within a certain domain; however, it is hard in general to measure mental models. Some researchers focused on studying how transparency can improve user mental model of an intelligent system. Eiband et al. [49] proposed guidelines on how to improve the transparency in intelligent user interfaces by studying expert, user, and target mental models iteratively to understand what to explain to the users. In a work by Hoffman et al. [67], various metrics for measuring human factors in XAI systems have been proposed, including human-machine trust, user-machine task performance, and mental model for user evaluation in XAI systems. In their paper, they lists multiple techniques to elicit mental models in XAI systems, e.g., think-aloud problem solving, card sorting, and prediction task (where users are presented with a test case and asked to predict the results of the system given the input). For instance, Kulesza et al. [90] studied how explanation soundness and completeness affect user mental model through the think-aloud approach. Another fairly common method to measure mental models is through prediction task, where users are asked to estimate an algorithm's prediction behaviour after a period of working with the tool (e.g., [128, 8, 108]).

Numorous prior HCI work investigate the relationship between user mental models and task performance [192, 132]. Some work has shown that a proper mental model of a system can have positive effects on user performance [38, 206], while others show a contradicting effect [126, 15].

Overall, the existing body of work in HCI and psychology does provide evidence that user-machine task performance and mental model are correlated [192]. How they are correlated, however, could depend on the task, context, study population, or other external factors [191].

Helping users build a proper mental model in intelligent systems is critical for various reasons. One of the common reasons is to improve user-machine trust [53, 195, 131], as humans might find a hard time trusting what they do not understand. While transparency is known as an approach to improve user understanding and the fidelity of their formed mental models [116], previous research has shown other factors, such as user's cognitive biases, can affect the quality of these formed mental models, despite the presence of explanations [131].

2.3 User Trust

Generally speaking, trust is a complicated concept [65]. Social and psychological researchers have been studying human trust for many years. Although there is not one agreed upon definition of trust in these areas, human-human trust is commonly based on believing that the trustee will do what is expected and in a predictable manner [54, 105]. Similarly, we can define trust in automation as a user's ability to rely on and predict the results from the automated system. In 2007, Madhavan and Wiegmann [107] discussed the subtle differences and similarities of human-human and human-automation trust. Despite human teams and human and machine teams sharing similarities, one shall not assume these two are interchangeable concepts. Similar to with other humans, humans do form "relationships" with automated systems (e.g., as suggested by Bowers et al. [18, 17]). People might go to the extremes of showing politeness or reflecting gender stereotypes with automated computing systems, with some showing more attraction to computers that echo their own personalities [123]. However, Madhavan and Wiegmann maintain that a human's so-called trust in their automated teammate plays an important role in decision success.

Similar to human-human trust, once human-machine trust is lost, it is hard to reestablish it [65]. However, research has shown that humans are more forgiving towards humans than machines when their invested trust is violated [33]. Moreover, Jian et al. [74] empirically showed

that humans are more cautious to rate high distrust against their human teammates and preferred to rate their trust low or negative. This was not the case for human-machine collaborations, showing that humans are more likely to use the term distrust with automated machines rather than humans. Both of these previous work are among a few notable examples that signify human-human trust is not quite the same as human-machine trust, and highlight the importance of maintaining trust in automation.

The concept of human trust in automated systems have been studied and discussed as early as 1986 with Muir's [120] paper focusing on the importance of human trust with automated decision aids. Prior to that, other researchers have pointed out the importance of considering the trust between humans and machines, without systematically studying it [161, 160]. Muir maintains that human trust in automated systems follow three expectation categories, namely persistence of the natural world, technical competence, and fiduciary responsibility. By combining and working with Barber's taxonomy of human expectations in trust [12] and Rempel, Holmes and Zanna's taxonomy on dynamics of trust [150] (both on human-human trust), he proposed a two-dimensional framework for studying and understanding the trusting relationship between humans and automated systems).

Specifically, as argued by Muir, the third expectation—fiduciary responsibility—becomes critical when the machine's competence exceeds that of the humans', or is not known to them. In the task of making decisions, this unbalanced competence should be considered, particularly, when humans consult the machine, with its predictions and recommendations designed to either replace the person in some ways or to remedy their deficiencies.

Unable to assess the system's competence, the human must rely on their own judgment of the system intentions and purposes. Muir believes this is still challenging for the said-users because "The machine's intentions in asking for certain information or providing certain solutions may be ill-understood by the user either because (1) an explanation capability is not included in the system, or (2) the explanation given is opaque (e.g., if it is expressed in terms of the machine's cognitive system (rule structure) rather than in the terms which the human uses to think about the

problem).” This can cause people difficulty in assessing the reliability of the system and propose issues to the users when calibrating their trust. Muir’s paper provides insightful discussions on various subjects of interest for the recent human-centered AI attempts, such as human-machine trust, the need for explainability (and the importance of effective, human-meaningful explanations), and considerations for user domain expertise when discussing trust.

As also hinted by Muir [120], the transparency of a decision-support system is closely connected and contributes to user trust [66]. In the recent years, with the utilization of machine learning and artificial intelligence algorithms in decision-aid systems, more researchers and practitioners are focused on techniques to improve the lack of transparency for these so-called black-box models. With that, psychologists and human-computer interaction researchers have also been studying trust in intelligent systems, particularly, those with transparent and explainable models. For instance, Hoff and Bashir [64] provide design recommendations for improving user trust in automated systems (for example, by simplifying interface design and increasing transparency).

A variety of literature in this direction focuses on how explanation characteristics can impact user trust in the outcomes. For instance, in my previous paper, we demonstrated that users tend to trust the same model with the same outcomes significantly more when the explanations are closer to what they perceive as meaningful (Nourani et al. [128]). In this work, using a binary image classification task with heat map explanations, we compared user’s perception of model accuracy—testing two levels of high and low simulated accuracies—and comparing three explanation presentations: 1) strong and 2) weak explanations on the basis of their meaningfulness, and 3) no explanations (See examples in Fig. 2-1). In our user study, we observed that participants underestimate the model’s accuracy significantly more when they do not find the explanations meaningful with respect to the prediction. Additionally, participants with no explanations had a significantly more accurate estimation of the model than those with weak explanations. With previous research demonstrating user’s observed accuracy can directly influence user trust [195], through our user study, we found that providing users with explanations

Figure 2-1. Example explanations from my prior work [128]. We tested the effects of explanation meaningfulness on user trust and perception of accuracy. Our results demonstrated that when users perceive explanations as weak (with respect to human meaningfulness), they significantly underestimate the model accuracy and their trust in the model decreases, even when compared to when no explanations are available.

they do not find meaningful can cause a decrease or loss of trust. Similarly, Papenmier et al. [138] studied the interplay between model accuracy and explanation fidelity, and how they affect user trust in intelligent systems. Their results show that model accuracy plays a more important role on user trust than explainability. They also found that users cannot be tricked into trusting a bad classifier when the system provides high fidelity explanations.

Another examined aspect is how explanation scope can affect user trust. Much of the work on explanation scope has been focused on local explanations, to justify the single outputs as opposed to the model as a whole. Riberio, Singh, and Guestrin [151] describe two types of human-machine trust: trust a prediction vs. trust a model. They argue that in many situations involving a decision-making task, trusting a prediction becomes more important than other times. This mainly depends on how critical and impacting the effect of the decisions is. For instance, with medical diagnosis, blindly following model decisions on the basis of fate can lead to catastrophic outcomes [151]. Similarly, Adadi and Beradda [1] assert that local explanations can

provide higher sense of trustworthiness than global explanations, as they provide a sense of understanding and trust by relying on the data and explaining individual outcomes. Global explanations indeed provide abstract views of a model's inner-workings while local explanations demonstrate the model's performance in action; i.e., they are a more concrete and visible presentation of the model's decision-making. This might indeed affect users and their trust formations.

In some cases, humans tend to mistrust (or over-trust) an intelligent system, which will often times result in over-relying on the model. This can be problematic especially since these models tend to err, in which case over-relying on the system outputs and predictions can lead to wrong decisions. This phenomenon is also known as automation bias [119, 5]. As previous research suggests, many times, this can be due to the lack of user's confidence in their competencies and capabilities as opposed to the machine's [19, 169]. Other times, users who suffer from automation bias tend to believe such tools maintain knowledge and intelligence that is superior to theirs [169]. Some might be under the impression that the model bases its predictions on conditions or information not known or apparent to them. Previous research demonstrates that novice users can suffer from this problem [119, 19]. On the contrary, misplaced distrust can cause users to underestimate the system and to move towards self-reliance, eventually causing them to stop using the automated system in the future, even though the model is trustworthy. For example, in a previous work [132], we showed that when users observe system weaknesses (in form of weak explanations), they tend to disagree with the system even when it is right. This self-reliance affects novice users solving non-domain-specific tasks, as well as more experienced users solving critical and domain-specific tasks [134].

In general, similar to how people's trust in other people can change over time [65], human's trust in automation may also go through changes with usage over time. Though, this is not expected when users start distrusting the system, as that might trigger discontinued usage [129, 60]. There are a variety of researchers who have been exploring user's trust calibration in machine learning systems. Yu et al. [196] studied changes of lay user's trust in a

decision-aid system over time based on different levels of model accuracy. They found that with lower accuracies, trust tends to decrease over time. In another previous work (Honeycutt, Nourani, and Ragan [70]), we studied the effects of users' providing feedback in a human-in-the-loop intelligent system. In this between-subjects user study, we varied the trends of accuracy changes over time (i.e., increasing, decreasing, and maintaining a constant observed accuracy) and compared user trust based on whether they were asked to provide feedback or not. Our results, to our surprise, demonstrated that users tend to mistrust on the model's predictions when they were asked to provide feedback. This is against one's expectations that observing model's performance allows the user to appropriately adjust their trust. In fact, in the upcoming sections, we discuss other aspects that affect user's calibration of trust and other factors, such as their mental models of an intelligent system.

A principal challenge in studying trust with intelligent systems is the lack of appropriate, objective measures to calculate trust. Trust is subjective by nature and therefore challenging to quantify. Several methods have been suggested to measure trust in intelligent systems [115]. There are many suggested trust in automation questionnaires such as [67] that can be used to measure trust explicitly. Some of the implicit trust measurements include checking user agreement with wrong system outputs [138, 132]; repeatedly asking for trust ratings [196]; and measuring user perception of system accuracy as an indication of user trust [195, 128]. Being an spectrum, as opposed to strict, solid states, trust is currently being measured through assigning specific values. While there are no effective way to quantitatively measure trust, studying other quanitfiable measures and drawing their relationships with trust (similar to [195]) may allow us to conclude where the user's trust is situated without explicitly asking them.

2.4 Past Influences

Semi-automated intelligent systems heavily rely on human analysis and sense-making for the final decision or analysis. Human judgments and decisions come from their complex brains, structure and inner-workings of which is yet to be fully understood. Human reasoning can be influenced by various conscious and subconscious factors, some of which are rooted in prior

experiences and formed expectations of AI. Preexisting factors, such as prior knowledge, background culture, experiences, cognitive abilities, age, and expertise can influence human perceptions and cognition abilities [64, 113]. Researchers in HCI, HCAI, and CSCW communities have evaluated how factors of these natures affect people's usage behaviors and understanding of AI. For instance, Zhang et al. [201] studied how people perceive and form expectations of their AI teammates. Through an initial surveying study and a follow-up interview with invited participants, they found that while people have mixed feelings towards AI teammates, they have positive attitudes towards AI in general, and argue that these attitudes are prune to change based on relative factors, such as their willingness to team-up with AI tools. This willingness can be affected by their pre-existing attitudes towards AI and prior collaborative experiences, even those with humans.

People's biases and cognitive heuristics are one of the more studied preexisting factors. Cognitive biases are a subcategory of these biases that are caused by erroneous decision-making heuristic [182] and may influence usage and human behaviours of intelligent systems. While such biases have been studied in various areas, HCI research attempts to acknowledge their existence and impacts when humans work with computing systems. Here, we specifically focus on how cognitive biases influence machine learning and artificial intelligence tools.

Today, many known variations of cognitive biases exist [1]. In 2018, Dimara et al. [37] organized a taxonomy of 154 known cognitive biases in seven categories to make visualization designers aware when designing tools. Fig. 2-2 shows these categories at a high-level.

In the context of human-centered AI, some of the most common studied biases include automation bias [129, 186, 131], confirmation bias [125], anchoring effect [129, 181], and availability heuristic [159]. Cognitive biases have been studied in various human-centered domains of interest to this review, including visualization and visual analytics [37, 183], decision-making [182, 37], and XAI [186, 131, 129].

[1] List of Available Biases on Wikipedia–Accessed: [May 1, 2021]: https://en.wikipedia.org/w/index.php?title=List_of_cognitive_biases&oldid=791032058

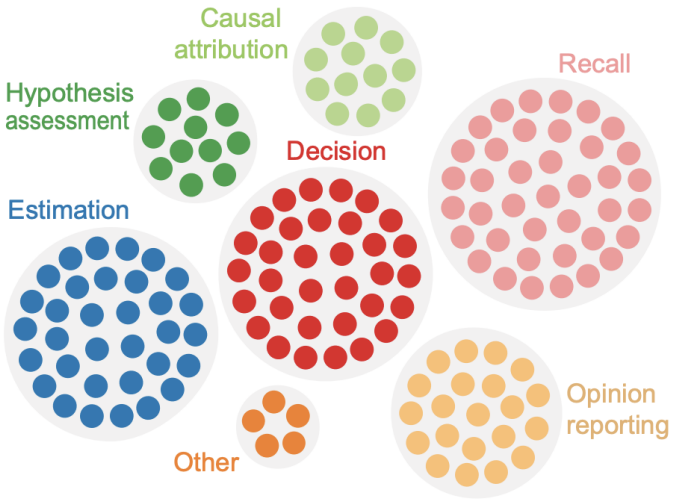


Figure 2-2. An overview of the seven cognitive bias categories, from a taxonomy that organized 154 known cognitive biases, proposed by Dimara et al.[37]. Each circle represents a cognitive bias.

Machine learning and artificial intelligence models are prone to error, as are humans. In the human-AI collaborative setting, models should rely on human strengths with reasoning and discovering patterns by referring to existing evidence or prior knowledge not available to the models. Simultaneously, humans rely on these automated systems to achieve their task. In a recent paper, Bogert, Schecter, and Watson [14] found evidence to support that, amidst decision-making, users tend to rely more on the algorithm than social influences as the task becomes more difficult. This was persistently observed regardless of the quality of decision advice the algorithm provided. However, when humans perceived low-quality advice from the algorithm, they were less forgiving of the model compared to when they received low-quality advice from wise humans.

Either over-reliance or under-reliance can affect the quality of user experience and more importantly, could cause fatal errors in decision-making processes [180, 19]. Design choices for a machine learning system can lead to either of these reliance behaviours. For instance, Bussone, Stumpf, and O'Sullivan. [19] show that while more detailed explanations increase user trust, they may lead to over-reliance; meanwhile less detailed explanations can result in self-reliance as users

might not understand model's justifications enough to rely on it for their high-stakes decisions. One common example of over-reliance is automation bias [30], a cognitive bias where users sometimes tend to default to trusting and relying on the outputs of an intelligent system when they become comfortable with the system's performance or think it "seems smart". Automation bias can cause problems such as subconsciously ignoring or missing system errors (omission), and ignoring the contradictory factors in the decision-making process and following system suggestions (commission) [141, 30].

Confirmation bias refers to the human tendency to find redundant evidence to support an existing hypothesis rather than looking for evidence for contradicting possibilities [186]. For this bias, more mitigation suggestion exist from the related work; for example, "analysis of competing hypotheses" and "evidence marshalling" [63]. In the context of medical XAI systems, Wang et al. [186] propose showing the findings first before formulating a hypothesis, as well as showing the prior probabilities of diagnoses.

Somewhat similar to confirmation bias, anchoring effect (also known as first impressions and primacy effect) describes a cognitive bias where users are anchored by the first piece of information encountered [37]. Research on first impressions has shown that human's early observations and judgements can bias and affect their behaviours towards people [198, 50], systems [127], and/or agents [144, 34]. In the machine learning community, some researchers have focused on the effects of order of training data on model's performance accuracy [24, 199]. In visual analytics, Wall et al. [183] argue that heuristics such as anchoring effect can occur during the analytical process, causing the bias to propagate to the findings or reaching false conclusions.

In their book chapter, Vaughan and Wallach [180] discuss ML intelligibility, defined as the stakeholders' ability to monitor machine learning systems enough to achieve their tasks or goals. They argue that, since humans (be it those who implement or build such systems, or those who use it or are affected by their predictions) are the center of intelligent systems, the strategy to intelligibility should start from user needs. Here, I will also focus on the group of stakeholders who are directly using the system, i.e., the people who are among the target users.

Table 2-1. The example of the categorization for stakeholders expertise based on the context and three types of knowledge manifest, from the framework by Suresh et al. [171].

	Knowledge		
Context	Formal	Instrumental	Personal
ML	The math behind model architectures, training processes, optimization	ML toolkit familiarity, off-the-shelf models	Tricks of the trade (e.g., hyperparameter values, feature engineering)
Domain	Data domain theories, like symptoms & treatments	Experience working with other related technology (e.g., medical devices, document mining tools)	Lived experience (e.g., prior memories of similar events)
Milieu	Sociocultural theories (e.g., redlining, mass incarceration)	ML system familiarity (e.g., virtual assistants, recommender systems)	Lived experience and cultural knowledge (e.g., values, attitudes)

There are many ways of studying and categorizing the differences between stakeholders. For example, one may categorize the stakeholders based on their demographic information. Level of expertise—specifically domain expertise—can be yet another factor that may cause differences. Especially if a model is used in a specific domain, possessing levels of domain expertise might affect the decision-making and system usage. Also, users can bring about domain proficiency that AI models lack; i.e., a human-AI collaboration for effective decision-making or analytical reasoning [203]. As Hoff and Bashir [64] maintain, prior expertise and knowledge of the domain can influence user reliance and trust in decision-aids. Preexisting knowledge can also impact users' understanding of and ability to detect errors and understand the uncertainties of model predictions [180, 129]. This is also observed by Merrit et al. [111], where they learned that users who are able to perform a task without aid are able to detect its failures, which indicates the importance of expertise and training in improving machine-user task performance. Such ability to find/observe errors can also affect user understanding, mental models, reliance, and trust.

It is, thus, compelling to study and account for user domain expertise in the field of human-centered AI. With domain familiarity and expertise serving as preexisting factors that may reflect on user behaviours and experiences, in HCI, researchers mainly focus on measuring the

interplay between expertise and human behaviours with intelligent systems. For example, Schaffer et al. [157] designed an experiment to test how presence of explanations, level of automation, and level of system error influences user's acceptance of advice from the intelligent system. Their results show explanations helped people with less task familiarity more, while showing explanations to users with higher task familiarity led to automation bias. Researchers in fields of psychology and social sciences have explored the notion of domain expertise to understand its impacts on human social and personal behaviours. A well-known example is the Dunning-Kruger effect [46], demonstrating that people with more expertise are less confident about their expertise (i.e., rate it as lower) while novices are overconfident with their knowledge (i.e., rate it higher). This review will only focus on the role of and concerns with domain expertise with intelligent systems.

Studying domain expertise in the artificial intelligence and machine learning context is getting more attention in the recent years, with many open questions yet to be explored in the future. One of the more studied relationships is the effects of domain expertise on trust. For example, Zhang, Liao, and Bellamy [203] studied how users' trust calibration can be affected by knowing that their domain knowledge is higher than the model's. They argue that user's ability to calibrate their trust is not enough for effective decision making, but humans need to bring their domain knowledge and expertise to replace model errors and deficiencies. Doshi-Velez and Kim [42] discuss the importance of user domain expertise on model transparency. They maintain that the amount and purpose of explanations should differ based on user's level of familiarity of the task and domain and motivations for using the system. Similarly, a recent study by Wang and Yin [189] reveals that explanation's impact on user-task performance varies based on user's level of domain expertise. Regarding the scope of explainability, Cai et al. [22] found that expert users of human-AI collaboration systems prefer general decision-making justifications from the model (i.e., global explanations) as opposed to case-specific, local explanations.

Many researchers studied the interplay between domain expertise and model transparency, with different contributions and suggestions: from design suggestions for interpretable ML/AI

systems [42, 79, 19] to explanation scope preference [22] based on domain expertise. Additionally, some researchers studied the perceptual comparison of novice and expert users on how they analyze outputs of intelligent systems [178]. These studies have focused on different domains, such as medical [19, 178, 22], data science [79], visual analytics [32], aviation [119], and criminal analysis [73, 168]. For instance, in the context of pathology, when a machine learning model fails to return images that are clinically relevant, the domain experts lose their trust and abandon the usage of system for their own expertise [83, 87, 21]. In their recent CHI paper, Suresh et al. [171] propose a framework to characterized the stakeholders of an XAI model. In the first part of their framework, they describe the knowledge of stakeholders based on their level of knowledge/expertise as well as the contexts or task that make the expertise relevant. For instance, in the context of milieu (i.e., the domain where the human-AI collaboration takes place in), user's Personal knowledge can be a reflection of past life experiences, such as cultural values or personal attitudes. Table 2-1 shows these categorise in more detail.

It is important to bear in mind that novice and expert terminologies are task and domain dependent and their definitions might vary from one system to another. For example, some researchers define novice users as students or those who have a ground-level of knowledge in the domain (e.g., [19, 119]), while many times terms as novice or lay users are referring to the general public with close to zero knowledge in the domain (e.g., [42]).

CHAPTER 3
CONCEPTUAL MODEL OF USER'S PRIOR EXPERIENCES AND DIFFERENCES IN HUMAN-AI COLLABORATION

Given the crucial role that people play in human-AI collaborations, whether as users or stakeholders, there is a growing recognition of the need to understand how individual differences can impact the outcomes of these partnerships. As individuals accumulate experiences over time and possess diverse cognitive influences and traits, their behaviors within the context of intelligent systems can vary significantly. This diversity poses both design and ethical challenges when it comes to building AI systems that effectively support and accommodate these individual differences in order to enhance the team performance and overall experience.

In this chapter, we begin by providing an overview of the potential stable and long-term versus short-term and transient past experiences, influences, and traits that can shape human-AI collaboration. We recognize the importance of these experiential factors in understanding how individuals interact with AI systems and the impact it has on their behaviors and outcomes. Building upon this perspective, we conduct a semi-structured literature review to explore the various types of past experiences and differences that have been studied within the field of human-centered AI. Our aim is to identify the connections and relationships between these factors and gain insights into the broader themes and topics that emerge from our analysis. By organizing and synthesizing the findings from our literature review, we develop and propose a conceptual model that captures the interplay between users' past experiences, individual differences, and their usage of AI systems. This model serves as a valuable resource for future researchers, providing a roadmap for understanding how people's prior states and traits are intertwined with their interactions with AI systems. Additionally, it highlights the existing challenges and gaps in our understanding of these topics, paving the way for further investigation and exploration.

3.1 Situating the Research

Individuals accumulate experiences throughout their lives, which shape their perceptions, actions, and decisions. These experiences can be transient, subject to change or replacement, or more enduring, lasting a lifetime or being resistant to change. In the realm of human-AI collaborations, individuals' past experiences with AI, their understanding of the technology, and

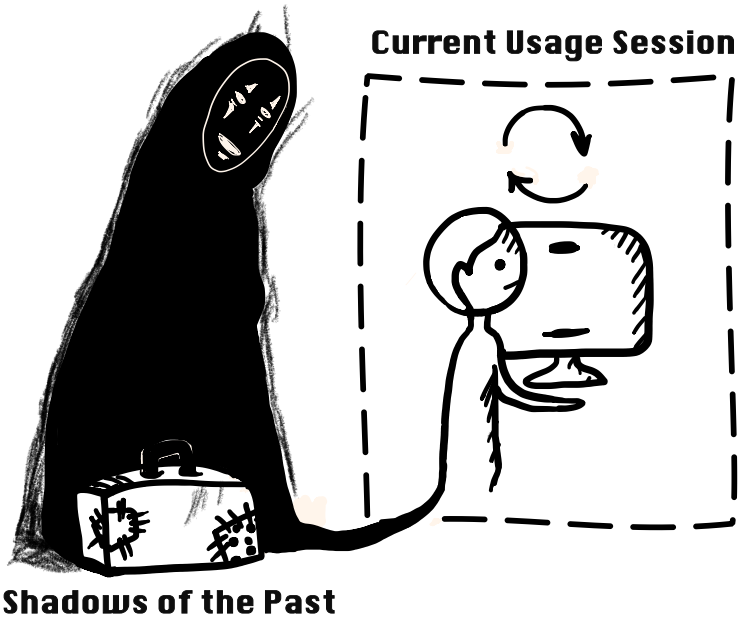


Figure 3-1. When engaging with an AI-supported tool, users bring their past experiences, beliefs, and traits, which can influence their current session. These "shadows of the past" have the potential to impact users' behaviors and experiences during the session. Over time, the accumulation of these experiences can become part of their long-term "baggage". Even after the usage session ends, the recent states and experiences contribute to the ongoing development of this baggage, shaping future usage sessions.

their personal beliefs, biases, traits, and behaviors all play a role in shaping their interactions with AI algorithms. As social beings, individuals are influenced by the societies they belong to, the people they interact with, and the collective norms and values that surround them. These communal influences can intersect with personal experiences and individual influences, further shaping and contributing to the diversity of behaviors and attitudes in human-AI collaborations.

To comprehensively examine the impacts of people's experiences and differences on human-AI collaborations, we adopt two perspectives: (a) the duration of their effects, distinguishing between lasting and stable influences versus short-term and evolving ones, and (b) the origin of these influences, differentiating between personal characteristics and communal factors. Our data labeling approach is grounded in these perspectives, allowing us to categorize and analyze the literature on human-AI collaborations based on people's experiences and differences.

3.2 Survey Methodology

We conducted a semi-structured literature review to understand the state of research in the fields of Human-Centered AI (HCAI) and human-AI collaboration that include some aspects of long-term and/or short-term past experiences and differences in their user evaluation, as influenced by the framing provided in Section 3.1. We searched for peer-reviewed manuscripts that were published in some of the top and most relevant HCAI venues (conferences and journals), as listed in Table 3-1. Our search focused on the peer-reviewed manuscripts within these venues that were published within the past five years (at the time of the research), covering 2018 through 2022. Due to the large number of publications within these years and venues, we performed advanced keyword search that matched our research goals and fit our inclusion criteria[1]. We probed these databases based on keywords that reflect the nature of human-AI collaborative systems (e.g., AI-supported decision-making, human-AI partnership, mixed-initiative AI systems); Fairness, Accountability, Transparency, and Ethics (FATE) keywords (e.g., responsible-AI, explainable AI, AI ethics; and keywords relating to user studies (e.g., user experiment, user evaluation, participants). This resulted in an initial paper set including 65 papers from IEEE Xplore and 395 papers from ACM Digital Library.

To reach our final set of papers for the analysis, we performed iterative verification against our initial inclusion criteria in order to get a first set of papers to examine. After this round, we ended up with a total of 184 papers (ACM: 138, IEEE: 30, HCMOP: 16). Our initial criteria for skimming and cleaning up the search results were:

1. The manuscript must be a full length peer-reviewed publication; and
2. There must be human-subject experiment(s) in the paper; and
3. There must be an AI/ML component in the paper.

In this round of selection, we quickly scanned the metadata and abstract information, acknowledging the possibility of including papers that may not strictly adhere to the main

[1] HCOMP was an exception to our keyword search process, as it is a smaller venue with a limited number of papers ($\tilde{80}$) within the specified date range. Consequently, for this conference, we conducted a manual search.

Table 3-1. The list of journals and conferences searched for collecting the initial list of papers for analysis. The venues listed under the Main section were our initial target list. However, we ended up including hits from other venues of IEEE due the small number of hits resulted from our search. These venues are listed under Miscellaneous.

Main	
TVCG	IEEE Transactions of Visualization & Computer Graphics
TOCHI	ACM Transactions of Computer-Human Interactions
TiiS	ACM Transactions on Interactive Intelligent Systems
CHI	ACM Conference on Human Factors in Computing Systems
IUI	ACM Conference on Intelligent User Interfaces
HCOMP	AAAI Conference on Human Computation & Crowdsourcing
VIS	IEEE Conference on Visualization & Visual Analytics
Miscellaneous / Other	
	IEEE Transactions on Affective Computing
FUZZ-IEEE	IEEE International Conference on Fuzzy Systems
ARSO	IEEE International Conference on Advanced Robotics & It's Social Impacts
ROMAN	IEEE International Conference on Robot & Human Interactive Communication

inclusion/exclusion criteria (as discussed further in the upcoming paragraph). The primary goal was to identify and eliminate obviously irrelevant papers, reducing the size of the paper set for a more thorough read-through. Nevertheless, it is important to note that some papers chosen in this round may still require further assessment during the thorough read-through phase to ensure alignment with the established inclusion/exclusion criteria.

3.2.1 Inclusion/Exclusion Criteria

The 184 papers were then used as the base dataset for the main analysis. Each paper was examined at least once and based on the inclusion and exclusion criteria. Papers that were determined to be out-of-scope for the analysis were flagged to be removed for the final analysis and report. Occasionally, the same paper appeared twice in our initial list, as a conference paper and as an extended journal version. In these cases, we removed the duplicates. If a paper was in

scope, we extracted information for the paper to be used in the final analysis. We summarized the main inclusion and exclusion criteria in Table 3-2. After iterative review and meticulous consideration of each paper based on our predefined inclusion/exclusion criteria, we have arrived at a final selection of 49 papers that align with our research goals. This final subset of papers underwent comprehensive analysis to summarize and conceptualize the trends and patterns regarding the aspects of long-term and short-term past experiences that the field of human-centered AI has prioritized thus far.

3.2.2 Data Annotation

We reviewed each of the 184 papers thoroughly and collected the following information (per paper) that was essential to our research goals and questions:

1. Metadata: Title, authors, year, and publication venue.
2. Research contribution(s): The top two paper contribution types, in the order of importance, including evaluation, system, application, framework, algorithm, and/or technique.
3. Task-specific information: Domain, target users, machine learning algorithm, and task type (e.g., decision-making, human-in-the-loop).
4. Experiment setup: Intelligent system and its characteristics, such as explainability, visualization/interface, and interactivity.
5. Experiment design: Evaluation method, study variables, and data collection methods.
6. Past-experiences and time-line specific information.
7. Research questions/goals and key findings.

3.2.3 Analysis Methodology

Once the final selection of papers was made and the data collection process concluded, we proceeded to organize them through iterative rounds of coding. This involved categorizing the papers into specific codes based on recurring patterns and shared analytical objectives. This semi-systematic approach allowed us to establish a structured framework for analyzing and synthesizing the findings from each paper, facilitating the identification of common themes and

trends within the literature. Eventually, we used the findings from this analysis to formulate and present the conceptual model of user's past experiences. In the upcoming sections, we will present the results of our analysis.

3.3 Contribution Types and Domains

Out of all the 184 papers reviewed, approximately 57.61% ($N = 106$) were deemed out of scope and did not to meet the main exclusion criteria (as seen in Table 3-2), and thus, were excluded from further analysis due to their lack of relevance to the research objectives. Most of these were HCAI and human-AI collaboration papers that failed to satisfy the third inclusion criteria, i.e., they did not explicitly study elements of long-term/short-term past experiences and differences. Furthermore, it is worth noting that approximately 13.61% ($N = 29$) of the reviewed papers were subsequently removed from consideration as they did not meet the initial inclusion criteria. These papers were identified during the more thorough read-through phase, indicating that our initial quick scanning round may have failed to detect their irrelevance. This subset included duplicate papers, as well as papers from completely irrelevant fields that were the result of false positive search hits from the ACM Digital Library and IEEE Xplore databases. These duplicates and irrelevant papers were inadvertently included during the search process, and they needed to be identified and removed to ensure the integrity and relevance of the final paper set. For the remainder of this section, we first discuss some aggregate analysis for papers that were in- and out-of-scope for our search ($N = 106 + 49 = 155$), and then we will shift our towards analyzing the data and discussing the findings from the papers within the research scope, which comprise the final set of 49 papers.

3.3.1 Contribution Analysis

We analyzed the paper set based on their top-two contribution types, and we found that only 42.58% ($N = 66$) were primarily evaluation papers. Specifically, approximately 36.13% ($N = 56$) were empirical human-centered evaluation papers with no additional contributions, while a small percentage of the papers had other secondary contributions (6.45%, $N = 10$). For instance, Agrawal and Cleland-Huang [3] study the effects of explainability on behaviours of the operators

Table 3-2. The inclusion and exclusion criteria used for the analysis for the final list of 184 papers.

Inclusion Criteria
(A) The paper must include human-subjects experiment(s), regardless of the experiment being the main contribution or not.
(B) The study / paper must include a human-AI collaboration element. We are inclusive in the types of AI, as long as humans are involved in the loop; e.g., it can be based on a real AI model or be synthesized; only the outputs of the AI can be shown (i.e., no interactivity involved); AI usage scenarios may be discussed; the AI might be explainable or black-box;$\cdots$
(C) Some element of user's past experiences or differences is explicitly studied and be included in the experiment design, either as independent variable or as measure/outcome of the study. This includes, but is not limited to, participants' comparison of backgrounds, attitudes/behaviours, biases, perceptions, and trust. It could also mean measures related to time (or affected by time), such as mental models and expectations, as well as measures that evolve in time, such as trust over time.
Exclusion Criteria
(A') The paper includes user experiment(s), but only to evaluate a system / application (for usability testing). The only exception to this rule is if the usability study follows the inclusion criteria (C).
(B') The paper includes AI/ML systems, but no human-AI partnership elements.
(C') None of the experiment design elements are related to past experiences, individual differences, or time-related components.

of multiple small Unmanned Aerial Systems (sUAS), such as drones (primary: evaluation), and provide design guidelines for designing human-in-the-loop interfaces (secondary: guidelines / framework). Moreover, the majority of the set consisted of papers with primary contributions of anything other than human-subjects evaluation ($N = 89$, 57.42%), with 11 papers not including any evaluation at all, and hence, failing the inclusion criteria (A). The remaining 78 papers (50.32%) were primarily focused on developing a system (44), technique (21), framework (8), application (4), and algorithm (1), while using human-subjects experiments to verify and/or gain further understanding of what was developed. Figure 3-2 provides a visual representation of the paper distributions based on primary and secondary contribution types via stacked bar charts.

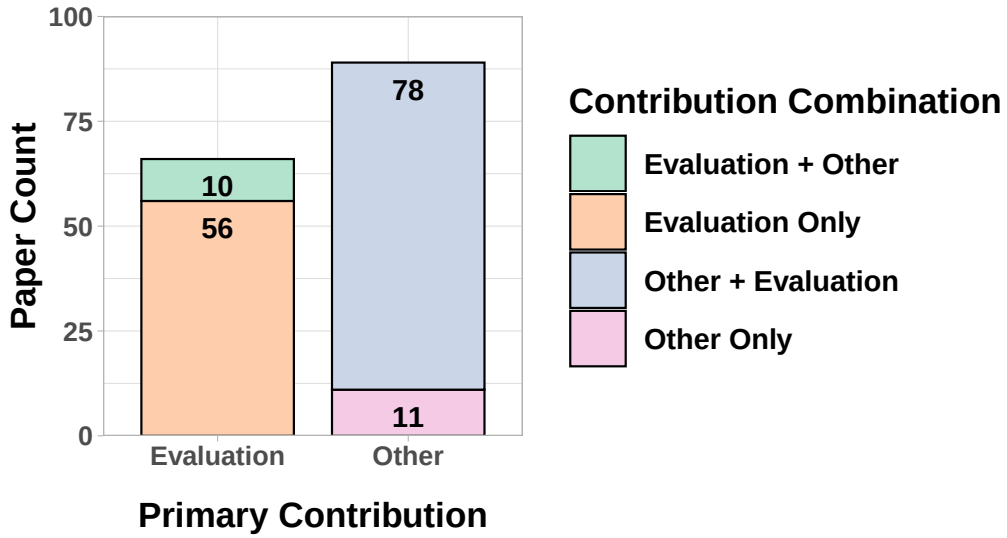


Figure 3-2. The distribution of in- and out-of-scope papers based on their primary contribution. The bar on the left shows the breakdown for papers where human-subjects evaluation was a primary contribution, while the one on the right shows that of other primary contributions, such as system, application, technique, algorithm, or framework.

Of the finalized list of 49 in-scope papers, 36 primarily contribute to the field of human-centered AI through empirical human-subjects evaluations. Additionally, 13 papers make secondary contributions of this sort. It is important to highlight that we deliberately chose not to immediately dismiss the papers with evaluation as their second contribution, as we speculated that some of them might contain thorough and promising studies that are relevant and align with our search criteria.

3.3.2 AI Approach, Domain, and Task Analysis

To synthesize the domains extracted by the papers, we incorporated a deductive coding approach based on the decision-making domain categories by Lai et al. [92]. They classify decision-making domains into 9 high-level categories, including Law & Civic, Medicine & Healthcare, Finance & Business, Education, Leisure, Professional, Artificial (made-up or artificial domains), Generic (generic tasks that can be applied to various domains), and Other. Our research focuses not only on human-AI collaborative decision-making but also on human-in-the-loop systems. As such, we kept an open mind in adjusting the codes and the baseline groups from their work. Using affinity diagramming and deductive coding, we classified papers into similar

categories. Our analysis resulted in 10 high-level categories, with Autonomous Vehicles introduced as a new category. In a few cases, the paper was categorized into more than one category. For each category, we also provide the high-level AI/ML approach used. Examples of approaches include Human-Robot Interaction (HRI), Conversational AI, Classification, Recommender System, Human-in-the-Loop (HitL), and Fairness. Table 3-3 shows the summary of this analysis.

3.3.3 Discussions on Gaps and Future Work

The analysis of the examined papers revealed a predominant focus on supervised classification and recommendation algorithms in the context of human-AI collaboration. While these algorithms have been extensively studied and applied in various domains, it is crucial to acknowledge that human-centered AI encompasses a wider range of AI/ML approaches and tasks. We acknowledge that the emphasis on specific AI techniques in the analyzed papers may have been influenced by the communities or domains from which the papers were selected. This potential selection bias may have resulted in an overrepresentation of these algorithms compared to others. To ensure a comprehensive understanding of human-centered AI, future research should strive to explore and investigate a broader spectrum of algorithmic approaches. This could involve considering papers from diverse communities and domains, as well as exploring emerging AI methodologies that have the potential to contribute to the field of human-AI collaboration

It is not entirely surprising that many papers in the field of human-centered AI focus on specific applications or systems and include studies that evaluate those specific contexts. However, it is important to recognize that the field can benefit from a stronger emphasis on human-centered empirical studies that prioritize the people involved, including users and stakeholders. By shifting the focus from system-centric perspectives to human-centric perspectives, researchers can address important socio-technical issues and prioritize the well-being, privacy, and individual differences of users and stakeholders. This approach would involve integrating insights and methodologies from fields such as psychology and social sciences to study the psychological aspects of human-AI collaborations more comprehensively.

3.4 Human-Subjects Study Design & Control

In order to gain insight into researchers' research goals and focus, we categorized their studies based on the type of experiment design they employed. By examining the independent

Table 3-3. The breakdown of tasks and AI/ML approaches from included papers.

Domain Category	Human-AI Collaborative Task	ML/AI Approaches
Law & Civic	Recidivism prediction [40, 189, 56]; Child maltreatment [26].	· Classification (3) · Fairness (1)
Medicine & Healthcare	Clinical Decision-Support [137]; Online symptom checking [177]; Cancer assessment [101].	· Classification (2) · Conversational AI (1)
Finance & Business	Loan decision and risk assessment [29, 122]; Real estate valuation [146, 28]; Car insurance pricing [16].	· Classification (5) · Fairness (1)
Professional	Offensive language detection [139]; Career selection & Hiring [184, 143]; Email management [25]; Creative writing [164].	· Classification (2) · Fairness (1) · HitL (1) · Conversational AI (1)
Autonomous Vehicles	Autonomous Driving [135]; Small Unmanned Ariel Systems [3].	· Autonomous Agents (2)
Education	Robot Tutoring [106, 148]; Math E-learning[136]; School admission decisions [2].	· HRI (2) · Classification (1) · Recommender Sys. (1)
Leisure	Interactive story-telling [204]; AI in games [31, 59, 52]; Speed dating [56]; Music [23, 89, 112], Movies [98], Video ads [35], Artistic images [41], and News [97] recommendation.	· Recommender Sys. (7) · Classification (2) · Conversational AI (2) · HRI (1)
Artificial	Defective object pipeline [11]; Diner Guru's dilemma [157]; News article writing [173].	· Classification (2) · Recommender Sys. (1)
Generic	Classification verification [128, 129, 140, 7]; AutoML and data science workflow [185, 43]; object recognition feedback [71].	· Classification (4) · HitL (3)
Other	Video activity recognition [130, 85]; Emotion detection [62, 166]; Forest coverage detection [189]; Age estimation [56]; Leaf-based tree classification [193].	· Classification (3) · Computer Vision (2) · Recommender Sys. (1) · Conversational AI (1)

variables used in the studies, we identified common trends in the research focus and aimed to identify current research directions and potential research gaps. In the following section, we present the findings of this analysis. It is important to note that a single paper may be assigned to multiple categories, reflecting the diverse nature of the research conducted in the field. Also, the analysis and results provided in this and the rest of the paper are based on the final set of 49 papers.

3.4.1 Effects of AI/XAI Presence on Human-AI Collaborations

As anticipated based on our empirical observations, a significant portion of the papers in our analysis primarily focused on evaluating the impact of incorporating transparency and explainability into AI models on people's usage behaviors and task performance. This research question is of great importance as we continue to explore the implications of explainable AI in real-world applications.

Some papers in our analysis focused solely on controlling the presence of explanations or transparency in order to examine their effects on usage behaviors. For instance, Millecamp et al. [112] studied how personal characteristics affect user perceptions and interactions when music recommendations are explained, and designed a within-subjects user study by controlling whether they observe explanations or not. Springer et al. [166] also controlled explanation presence, but in a between-subjects design, to understand people's preferences to using and their reactions to transparency when detecting text emotions. Researchers have also controlled additional variables to examine the relationship between explainability and other factors. In their study, Schaffer et al. [157] implemented a comprehensive approach by controlling multiple variables to examine the effects of explainability in the context of other factors. By manipulating variables such as the level of automation, percentage of error, and presence of explanations, they aimed to explore the complex interplay between factors like overconfidence, automation bias, system errors, and the presence of explanations in human-AI collaborations.

Additionally, some studies included no-explanation conditions as baselines for comparison to different types and characteristics of explanations (e.g., [128, 189, 62, 41]). Two studies within

our set also examined the impact of including an AI-supported automation approach on user task performance, mental models, and other behaviors [101, 157].

3.4.2 AI/XAI Model Specifics, Characteristics, and Behaviours

Some studies have controlled the type of AI/XAI model used and the behaviors it exhibits, including comparing different models or algorithms [41, 98, 193], manipulating the features and model information presented to users [7, 146], and varying levels of accuracy and error behaviors [11, 139, 25, 140, 128, 70]. Additionally, Grgić-Hlača et al. [56] examined how the similarity between AI and people's error-making behaviors influenced perceptions and decision-making. This research focus contributes to the design of user-centered AI systems and enhances our understanding of building more effective tools.

3.4.3 Types and Characteristics of Explanations

The majority of papers in our dataset focused on the exploration and understanding of various types of explanations. The primary objective was to investigate and compare alternative approaches to presenting explanations and determine the relevant information to include, taking into account the specific task and domain. Some researchers specifically focused on evaluating the effectiveness of different explanation methods. For example, Aechtner et al. [2] compared four explanation methods, including LIME and SHAP, to assess which approach was more effective in enhancing user trust and perceptions of a model designed for school admission decision-making.

Various studies have also controlled for explanation type and medium as factors of interest. Although researchers often use the term "type" to refer to these factors, it encompasses a range of aspects related to explainability. For instance, Szymanski et al. [173] compared visual and textual explanations, Gao et al. [52] compared explanations based on heuristics versus mind modeling, and Dodge et al. [40] compared four types of explanations in the context of recidivism prediction and fairness, including demographic-based explanations and sensitivity-based explanations. Additionally, many studies focused on comparing different explanations based on their quality, particularly in terms of their faithfulness to the model (e.g., [139, 136]) and/or their meaningfulness to humans (e.g., [128]).

3.4.4 AI Interactivity

A handful of papers we analyzed have controlled AI interactivity and human-in-the-loop interactions in their study design. Honeycutt et al. [70] studied the interplay between changes of accuracy and feedback interactivity when participants were tasked to provide feedback to the model when it made mistakes. Chiang and Yin [28] study how providing different types of user tutorials to improve non-expert's ML literacy affects their reliance on the model. Uniquely, they control the interactivity, scope, and presence of tutorials across different conditions in their study, which is different than most papers that control ML/AI parameters or the main task. Another example is from Wang et al. [185], who propose AutoDS, an interactive AI system to aid in data science pipelines, and compare user performance to classical data science approaches of using Jupyter notebook (less interactivity).

Several papers in our analysis have incorporated the control of AI interactivity and human-in-the-loop interactions as part of their study design. For example, Honeycutt et al. [70] examined the interplay between changes in accuracy and feedback interactivity by asking participants to provide feedback to the model when it made mistakes. Chiang and Yin [28] investigated the impact of different types of user tutorials on improving non-experts' machine learning literacy, controlling for interactivity, scope, and presence of tutorials across different conditions. Another example is the work of Wang et al. [185], who proposed AutoDS, an interactive AI system to support data science pipelines and compared user performance to traditional data science approaches using Jupyter notebook, which offers less interactivity. All of the aforementioned studies provide compelling evidence that interactivity considerations in AI systems play a significant role in shaping user behaviors and influencing task performance. These findings underscore the importance of further exploring and studying the impact of interactivity in future research.

3.4.5 AI Assistance and Intervention

Another notable trend in study design we observed was comparing and controlling the interplay between AI systems and user initiation, as well as the type and role of assistance

provided by the AI system. Many researchers have compared different modes of intervention and initiation when users collaborate with an AI system and seek assistance. This includes comparing user-initiated and on-demand interactions, system-initiated interactions, and mixed-initiative approaches [23, 148, 164]. Notably, Ramachandran et al. [148] even studied the interplay between children's help-seeking motivation behaviours and the mode of intervention in the context of child-robot interactive tutoring.

Another aspect of AI assistance that has been studied is the role that AI assumes as an assistant, as seen from work by Liao and Sundar [97]. They investigated how AI assistants should communicate the need and reasoning for the collection of sensitive or private data based on the role they play, whether as a help-provider or a help-seeker. From a different perspective, Kim et al. [85] examined level of AI assistance by controlling explanation presence (AI with and without explanations) and the timing for explanations (only first-half vs. only second-half). While this study design was particularly employed to evaluate the effectiveness of their proposed explainable video activity recognition system for improving task performance and mental models, it serves as an interesting example of studying AI assistance and the timing of providing extra information, and whether extra information and assistance is always needed or necessary. Future studies can further explore the complex relationship between users' behaviors, intervention modes, and system outcomes in various contexts.

3.4.6 Data Fairness

Within our dataset, several papers explored the manipulation or control of fairness in the outputs shown to users, particularly in relation to the underlying data. For instance, both Peng et al. [143] and Wang et al. [184] studied gender-specific stereotypes in hiring decision-making and whether debasing approaches could help mitigate these biases. In the study by Peng et al. [143], the researchers employed debiasing techniques by manipulating the representations of female profiles and the types of professions assigned to address persistent gender biases in hiring decision-making. Similarly, Wang et al. [184] compared two types of recommender systems: one

that was gender-biased and another that was gender-aware. The study aimed to examine whether debiasing AI systems can overcome people's internal stereotypes and societal biases.

3.4.7 Individual Differences and Long-Term Past Experiences

A subset of the papers analyzed in our study employed study designs that specifically controlled variables related to the goals of our analysis framing, as discussed in Section 3.1. These papers aimed to compare different levels of past experiences and/or differences to gain a better understanding of human-AI collaborations, with a central focus on the "human" aspect of the interaction. However, it is important to note that only six papers, accounting for approximately 12% of the total, incorporated such study designs.

These papers adopted two main approaches to control or manipulate study variables related to people's past experiences and differences. The first approach involved measuring and analyzing based on personality traits, such as the Big Five personality traits. For example, Kouki et al.[89], Cai et al.[23], and Dey et al. [35] examined the impact of personality traits on various aspects of human-AI collaborations. The second approach involved recruiting participants or stakeholders with different characteristics or expertise levels. Cheng et al.[26], Dominguez et al.[41], and Nourani et al. [129] specifically focused on controlling the differences among participants and their impact on human-AI collaborations.

3.4.8 Other Task-Related Controls

Finally, some researchers focused on designing controlled user studies with task-/ study-specific independent variables. For instance, Chromik et al. [29] designed a study to understand how non-expert users gain a global understanding of the model based on local explanations. Non-expert users are often considered as heuristic thinkers compared to their counterparts. In order to enforce and study the effect of type 1 and type 2 thinking systems on gaining a systematic overview of the model, they included a moderator for half of the participants.

3.4.9 Discussion of Gaps and Future Work

This analysis provides valuable insights into the landscape of human-AI collaboration and human-centered AI research. It reveals that while there is a significant emphasis on enhancing AI

algorithms and models from a user-centered perspective, there is comparatively less attention given to studying the role of people in the context of human-AI collaboration. The majority of research focuses on exploring and addressing challenges related to explainability, interactivity, and system/model design, with limited efforts directed towards understanding people as cognitive and unique contributors to the collaboration process. This highlights a potential gap in the current research landscape and underscores the importance of further investigation into the human aspects of human-AI collaboration. By placing a greater emphasis on understanding the characteristics, behaviors, and experiences of individuals involved in these collaborations, we can enhance the design and development of more effective and user-centric AI systems.

There is a notable emphasis in the literature on investigating the effectiveness of explanations in the decision-making process and human-in-the-loop scenarios. However, it is evident that the answer to whether the presence of explanations is effective is highly context-dependent. Indeed, there is a valuable opportunity to redirect research efforts towards designing more personalized explanations that cater to individuals based on their unique differences and experiences. For example, by exploring explanations that adapt to factors such as personality traits or cognitive styles, we can enhance the effectiveness and relevance of explanations in human-AI collaborations.

We strongly advocate for the research community to broaden their focus on complementary aspects of human-AI collaborations. There are numerous important questions that can be explored to improve decision-making, such as determining which initiation methods are most effective, identifying the optimal level of information detail for AI interventions, and understanding how people perceive and benefit from complementary advice or explanations based on different styles of intervention. Additionally, investigating the role of technology-mediated nudges in leveraging user biases to enhance the outcomes of human-AI collaborations is an intriguing avenue for future research. By delving into these areas, we can go beyond the traditional scope of explainability and leverage the unique differences and experiences of individuals to foster more effective and beneficial collaborations with AI.

In addition, conducting user studies and examining the effects of algorithms on different groups of participants, such as those with varying levels of expertise, can offer valuable insights for designing and selecting effective algorithms. For instance, comparing two explanation methods like SHAP and LIME may not provide a comprehensive understanding of their effectiveness, as both methods can be challenging to comprehend for certain user groups. Instead, by recruiting diverse users and stakeholders, we can contextualize our investigation of the effectiveness of these methods and gain a deeper understanding of how they perform in different user contexts. his approach allows us to account for the unique perspectives and needs of various user groups, leading to more inclusive and effective algorithmic design.

3.5 Past Experiences and Differences

In this section, we present an organized overview of the papers based on the elements of user's past experiences and differences they addressed. This organization is influenced by our described framing in Section 3.1. Each paper contributes to our understanding of human-AI collaboration by either controlling or measuring these factors, observing their outcomes, or inferring underlying latent variables. Table 3-4 presents a comprehensive summary of the themes and categories identified in each paper, providing an overview of the relevant patterns and trends in the sample literature. In this section, we will explore these categories in detail.

3.5.1 Long-Term Past: Personal

These are components of the long-term past that are specific to each individual. We divide them into two categories based on whether they are experiences and temporary states that are more prone to change over time or more enduring and consistent traits. In this section, we discuss each of these categories based on the body of work we examined.

3.5.1.1 Traits

Personal Traits (PT) are those long-term individual factors that are less prone to change and create unique personalities, leading to differences of behaviours. Based on our analysis, we describe them in three categories.

Table 3-4. Paper topics organized by timeline of state / experience and evaluation techniques.

		Long-term										Hybrid,			Current Usage											
		Personal							Communal			Interplay			Short-term						Over Time					
		Traits			Experiences																					
Citation (abbreviated)	Evaluation Method	Cognitive Biases & Priors	Societal Biases	Characteristics & Dispositions	Domain Knowledge	Tech. Expertise & Exposures	Attitudes Toward AI Tech. & Tools	Expectations & Views of AI	Knowledge & Shared Experiences	Communal Roles	Group Similarities & Stereotypes	Demographics	Automation Bias	Task Familiarity	Interactions	Cognitive processes & biases	Reliance & Aversion	Trust	Mental Models	Perceptions & Impressions	Other Attitudes (e.g., Confidence)	Trust	Mental Models	Task Performance	Perceptions	Learning Effects	Changes of Attitudes
Aechtner 2022 [2]						■												■	■								
Agrawal 2021 [3]					■								■														
Alqaraawi 2020 [7]																			■		■						
Bansal 2019 [11]																							■	■			
Bove 2022 [16]																			■								
Cai 2022 [23]				■	■													■	■								
Chen 2022 [25]																						■	■			■	■
Cheng 2021 [26]										■	■																
Chiang 2022 [28]					■	■							■	■													
Chromic 2021 [29]		■																	■								
Das 2020 [31]																								■		■	
Dey 2020 [35]				■												■											
Dodge 2019 [40]				■			■				■																
Dominguez 2019 [41]					■							■			■			■									
Drozdal 2020 [43]																		■	■								
Gao 2020 [52]																			■								
Grgić-Hlača 2022 [56]																			■					■			■
Guo 2022 [59]				■		■						■		■						■							
Heimerl 2022 [62]					■	■							■					■	■		■						
Honeycutt 2020 [71]		■																■		■					■		
Kim 2021 [85]																	■	■	■		■						
Kouki 2019 [89]				■		■									■					■							
Liao 2021 [97]				■									■					■									
Liao 2022 [98]		■	■	■							■		■					■									
Lindvall 2021 [101]					■								■			■		■									
Ma 2022 [106]															■	■			■	■							
Millecamp 2019 [112]				■	■	■								■				■		■	■				■		
Nakao 2022 [122]											■	■								■							
Nourani 2019 [128]																		■	■								
Nourani 2020 [129]					■								■		■	■						■					
Nourani 2021 [130]													■		■	■			■	■	■						
Omeiza 2021 [135]		■					■											■	■				■				
Ooge 2022 [136]											■	■					■	■									
Panigutti 2022 [137]							■						■	■			■	■									
Papenmeier 2022 [139]				■								■			■				■			■					
Papenmeier 2022 [140]																			■								
Peng 2019 [143]											■	■															
Poursabzi 2021 [146]																■	■		■		■						■
Ramachandran 2019 [148]				■								■			■											■	
Schaffer 2019 [157]					■								■		■										■		■
Singh 2022 [164]								■					■			■			■								■
Springer 2019 [166]		■														■		■	■								
Szymanski 2021 [173]						■									■	■											
Tsai 2021 [177]																		■	■		■						
Wang 2021 [185]					■	■	■				■																■
Wang 2021 [189]											■				■												
Wang 2022 [184]																			■		■	■					
Yang 2020 [193]				■	■	■									■			■				■					
Zhang 2022 [204]										■																	
# times explored (underlined: ≤ 5):		5	1	11	11	9	4	1	0	2	8	7	11	4	11	9	4	19	21	7	8	5	3	3	3	3	5
# times per category:		23			25				14				28			68						22					

Legend: Quantitative Methods: Qualitative Methods: Mixed Methods:

3.5.1.1.1. (PT1) Cognitive Biases & Priors People have deeply rooted prior beliefs, cognitive biases, and heuristics that affect their decision-making and views of the world. Most cognitive biases are long-term, unconscious, and hard to change. As such, they can influence people's collaborations with AI systems and the decision-making outcomes. When studying user's preference and trust in a movie recommender system based on the quality and type of its recommendations, Liao et al. [98] found evidence of strong self-serving biases based on the quality of recommendations; i.e., people tend to take credit for the decision when the recommender system provides good suggestions, while blaming the system when it provides bad recommendations. They also observe that such bias can be circumvented based on the type of recommendation. Other long-term cognitive biases and heuristics measured or found were illusion of Explanatory depth [29], confirmation bias [166], and the hard-easy effect [29].

3.5.1.1.2. (PT2) Societal Biases Aside from cognitive biases, people may form societal biases that are judgments or formed perceptions towards individuals, groups, and opinions within the context of a society, which may or may not be accurate. Here, we will focus on biases that operate at the personal, as opposed to the communal level. These personal biases are formed based on a person's presence and interactions in the society and might share some resemblance with some cognitive biases. Examples include group conformity, proximity bias, and associative bias. Algorithmic fairness and AI ethics research primarily center around societal biases at the communal level, particularly those related to gender, race, and other sensitive societal groupings or stereotypes [2]. These are categorized as long-term communal elements, and we will discuss them later in this section. From our analysis, we only detected one example of societal biases that match this category. In the work by Liao et al. [98], it is shown that people prefer recommendations based on collaborative-filtering over content-filtering due to bandwagon effect,

[2]It is worth noting that certain biases can overlap across multiple categories, such as cognitive and societal biases, or personal and collective societal biases, depending on the context. This presents an avenue for future research in the field of Human-Computer Interaction (HCI) and Human-Centered AI (HCAI) to explore and distinguish these categories further. With that said, we will conclude the discussion at this juncture, leaving room for future exploration and investigation.

which is people's tendency to adopt a behaviour others in their surroundings have adopted (Irrespective of whether it aligns with their own beliefs).

3.5.1.1.3. (PT3) Characteristics & Dispositions People's personal characteristics, traits, and dispositions play an important role in their behaviours. Researchers have examined a variety of personal traits and characteristics and their influences on human-AI collaborations. Among the papers we examined, one of the more commonly-observed studied elements are the Big Five personality traits [7, 35, 98], a taxonomy introduced by John and Srivastava [75]. For instance, Dey et al. [35] studied the effects of personality traits on people's tone preference for the video advertisement that are recommended to them, and found that connections between the two; e.g., highly extroverted people preferred active ads while people with high agreeableness were more inclined to alert-toned ads.

Locus of control was another on-theme characteristic. Guo et al. [59] studied people's locus of control (and other usage behaviours) in explanation-driven interactive ML systems through an interactive AI-supported game of Tic-Tac-Toe. They found that people's perception of control is significantly affected by their perception of quality of the feedback provided by the system, and that people from different backgrounds perceive their level of control differently than those from other groups. For instance, those with high school degree or lower felt less control of the system when the explanations were visual (as opposed to textual). Other types of studied characteristics and dispositions were cognitive style (e.g., holistic vs. analytic) [40], learning motivation and help-seeking attitudes [148], and propensity to trust [193, 23].

3.5.1.2 Experiences

Individuals' actions, decision-making, and behaviors are shaped by their extensive past experiences, which encompass a range of accumulated and evolving long-term prior experiences. Unlike static traits, experiences possess a more dynamic nature as they can be accumulated and altered over time. Moreover, experiences can have a profound impact on the development and manifestation of enduring traits. In this section, we describe long-term past individual

experiences in the context of human-AI collaborations based on the themes emerged from our literature review.

3.5.1.2.1. (PE1) Domain Knowledge Researchers have investigated the role of domain knowledge and familiarity in human-AI interactions, defined as level of understanding and expertise in the domain that the AI system is being utilized. To incorporate domain expertise into study design, one effective approach is to designate it as an independent variable. Researchers can achieve this by recruiting individuals with diverse levels of domain expertise in a between-subjects setup. This enables a comparison of the behavioral differences in human-AI collaborations based on varying levels of domain expertise [129]. This is specifically useful in cases where a system is designed to assist users of varied domain familiarity and expertise. However, most papers did not explicitly regulate domain expertise during participant recruitment. Instead, they assessed participants' task or domain familiarity levels using pre- and post-study questionnaires. The findings were subsequently analyzed and interpreted based on the varying levels observed within their study sample. For instance, the level of leaf familiarity was assessed in a leaf classification task [193], musical sophistication was measured in the context of music recommendation [23, 112], the number of hours spent creating artwork was considered in the domain of artistic image recommendation [41], and the level of seniority in pathology was determined to study clinical workflows.

3.5.1.2.2. (PE2) Technical Expertise & Exposures In addition to domain experience and knowledge, individuals' technical skills, including their familiarity with AI/ML and data science, as well as their prior exposure to AI technology or systems, can significantly influence their behaviors and decision-making outcomes. Similar to the assessment of domain expertise, the majority of papers in this category also employed questionnaires to measure participants' levels of familiarity with these technical aspects, either before or after the primary task, and analyzed the findings by considering people's technical and exposure differences. In the study conducted by Guo et al.[59], it was discovered that individuals with a higher level of XAI familiarity

experienced greater ease in providing and receiving feedback from the AI algorithm, ultimately leading to higher overall satisfaction. Similarly, Aechter et al.[2] explored various explanation methods, such as SHAP and LIME, and observed differences in behaviors and understanding of the explainability method based on participants' self-reported levels of AI familiarity. Notably, the findings revealed a contrast with existing literature: novice users demonstrated no particular preference for the scope of explanations, while experts exhibited a stronger preference for global explanations compared to their less experienced counterparts.

3.5.1.2.3. (PE3) Attitudes Towards AI Technology & Tools Users' behaviors during interactions with AI systems can be influenced by their underlying attitudes toward AI and technology. These attitudes, formed over the long term, can be shaped by a combination of individual traits and past experiences with AI systems. Dodge et al. [40] found that people's general attitudes toward ML trust and fairness strongly influence how they judge fairness of an algorithm, even more so than their cognitive styles (PT3). They also demonstrate it is critical to not only design explanations in ways to enable users to judge ML fairness but also to account for people's general attitudes and prior positions toward fairness. Fear of AI is another attitude explored by researchers [137, 135]. For example, Omeiza et al. [135] found that people are disinclined to trust and grant full control to autonomous vehicles due to their fear of AI caused by the track records of accidents reported in news articles and media. Some work also explored attitudes towards tools that are fused by AI/ML algorithms, e.g., attitudes toward AutoML systems [185].

3.5.1.2.4. (PE4) Expectations & Views of AI People's expectations of AI systems can be shaped by various factors, although these expectations may not always align with the actual capabilities or characteristics of AI. It is important to differentiate between expectations (PE4) and attitudes (PE3) towards AI. Attitudes primarily encompass individuals' general perspectives and emotions towards AI as a whole, while expectations can be more specific and related to particular types of AI systems or tasks performed using AI, or even specific systems. These

expectations can vary and may not always correspond to the reality or actual performance of AI systems. For example, someone may have a positive attitude towards AI in healthcare but expect a conversational AI to accurately diagnose complex medical conditions, even though the technology might not currently possess that level of capability. In our review of the papers, Singh et al. [164] stood out as the only study that investigated the influence of people's expectations on AI-supported writing tools and how those expectations, in turn, impacted interactions with the tool. They utilized a hybrid *Expectation-Process-Outcome* model, which involved gathering participants' mental models, priors/expectations, and folk theories before the primary interaction, followed by an evaluation of the outcomes post-interaction. Their findings provide compelling evidence of the existence and influence of prior expectations and assumptions on human-AI interactions. Note that how people act differently based on their expectations can be influenced by other personal elements, such as traits and biases. For example, someone's formed expectations can bring about confirmation bias, where they actively seek evidence that aligns with their preconceived expectations. Exploring expectations, their correlations with other long-term factors, and their implications constitute a crucial avenue for future research, as it holds the potential to enhance our understanding of human behavior in relation to expectations.

3.5.1.3 Discussions of Gaps and Implications

Our analysis reveals candid research examining the role of long-term personal past experiences and traits in human-AI collaborations. This indicates that researchers are actively considering these factors across different topics and domains. However, there is still much more work to be done to gain deeper insights into how these factors influence decision-making and the dynamics of human involvement with AI systems. In the papers encompassing this category, the primary emphasis revolved around assessing two key aspects: domain knowledge (PE1) and characteristics & dispositions (PT3). Concerning the former, there exist prospects for deliberate testing and recruitment of individuals based on their level of domain expertise, as opposed to measuring expertise and subsequently analyzing the observed levels within the collected sample.

This presents an opportunity for more targeted investigations and controlled evaluations of domain expertise's impact on human-AI collaborations.

On the other hand, we found minimal focus on societal biases (PT2) on a personal level and expectations & views of AI (PE4). There is more to be known about societal biases and/or cognitive biases that find meaning in the context of society. Future research should aim to explore a wide range of long-term societal and cognitive biases, as well as priors and expectations of AI, to better understand their impact on human-AI collaborations. Additionally, it is essential to investigate how collaborative systems can effectively leverage these biases to enhance decision-making processes and improve outcomes in human-in-the-loop scenarios. By gaining deeper insights into these areas, we can develop strategies and design approaches that harness these biases in a constructive manner, leading to more effective and beneficial human-AI collaborations. For example, a potential avenue of exploration lies in the design of explainable and interactive approaches that can capitalize on people's bandwagon effects. One possible idea to explore involves offering comprehensive overviews of a task or system, considering the prior decisions and usage patterns of other individuals, with the intention of influencing users to align their behavior with the choices made by others. Such approach seeks to prime users towards adopting a similar usage pattern based on bandwagon social influence. It can be beneficial as a mechanism to train users or to circumvent/prevent other types of hazardous biases.

3.5.2 Long-Term Past: Communal

It is crucial to consider the long-term aspects of human-AI interactions within societal and communal contexts. Whether focusing on AI users or stakeholders, a comprehensive examination of these sociotechnical elements is necessary for understanding the dynamics of human-AI partnerships. In this section, we delve into the emerging themes identified through our literature review, which encompass the sociotechnical factors that play a significant role in shaping these partnerships.

3.5.2.1 (C1) Demographics & Socioeconomic

Exploring demographic and socioeconomic information can provide valuable insights into understanding the sociotechnical dimensions of human-centered AI. Among the studied socioeconomic and demographic factors, researchers have studied various dimensions, including age groups categorized based on specific ranges or adolescence level [148, 59, 139, 136], gender [41, 143, 139], education level [59], as well as country and cultural differences [122]. For instance, Nakao et al. [122] studied the effects of cultural differences based on the country of origin on people's perceptions and views of fairness. Researchers have examined these factors from the perspectives of both system users and stakeholders. For example, Peng et al. [143] conducted a study that demonstrated the impact of self-reported gender on the fairness of hiring decisions made by decision-makers regarding candidates within specific gender groups. They found evidence supporting the influence of gender on decision-making outcomes.

3.5.2.2 (C2) Communal Roles

Our review revealed an intriguing research focus in two of the papers, which involved examining how people's perceptions and usage behaviors vary depending on the roles they assume within a societal context. This investigation sheds light on the ways in which individuals' societal roles influence their interactions with AI systems, offering valuable insights into the nuanced dynamics of human-AI collaborations within larger social frameworks. Cheng et al. [26] conducted a study investigating how non-expert users can solicit perceptions of fairness from a community regarding child maltreatment detection. The research focused on two key decision-making stakeholders: parents and social workers, both of whom were recruited and studied to gather insights into their fairness perceptions based on their societal roles. While they did not found evident differences between the responses between these roles, their study provides an interesting perspective on communal elements of human-AI collaboration. In the second paper by Zhang, Xu, and colleagues [204], they present an AI-supported storytelling app called *StoryBuddy*, designed for use both with and without parental involvement. The authors evaluate the system to gain insights into the interactions between children and their parents with

StoryBuddy in both modes. This research focuses on the roles of parents and children within the societal context of the family, and also highlights the under-explored area of human group and AI collaboration, adding an additional intriguing dimension to the study.

3.5.2.3 (C3) Group Similarities & Stereotypes

In our analysis, we identified a recurring communal pattern among the papers, which involved examining and/or utilizing similarities, differences, and societal biases, such as stereotypes associated with various societal groups. In this context, societal groups refer to collectives formed and bound by societal factors, including, but not limited to race, gender, culture, or occupations. Certain studies have focused on developing human-in-the-loop and decision-support systems that leverage similarities among the groups with which users identify to enhance human-AI collaborations. For instance, Ooge et al.[136] presented a historical dashboard that offered insights into how adolescents within the same age group performed in answering mathematical questions. Additionally, Liao et al.[98] studied a music recommendation approach that took into account similarities within the user's demographic group. However, more studies have focused on group differences and stereotypes associated with the groups, especially in the context of AI fairness and ethics. Nakao et al. [122] conducted a study comparing perceptions of fairness across different cultural groups using a loan-application scenario. By examining six distinct cultural dimensions, including masculinity and long-term orientation, they explored how different cultural groups perceive fairness differently. The study revealed that despite similarities in cultural dimensions within the same culture or country, different cultural groups can exhibit contrasting fairness perceptions. For instance, participants from Portugal generally perceived the loan applications as less fair compared to participants from Sweden. In the studies conducted by Wang et al.[184] and Peng et al.[143], the focus was on decision-makers' stereotypes regarding career preferences based on gender. Despite the implementation of bias mitigation techniques in both studies, persistent preferences and stereotypical biases towards specific genders for certain jobs were identified. These findings highlight the continued need for investigating and designing

interventions to address communal prejudices and stereotypes among decision-making stakeholders in AI-supported contexts.

It is worth noting that there are some connections between this category and C1, in that researchers often manipulate or study demographic variables to examine group-based communal factors. The connection between these categories does not render either of them irrelevant or redundant since they are not always mutually inclusive.

3.5.2.4 (C4) Knowledge & Shared Experiences (Proposed Category)

Before concluding this section, we would like to highlight a category that was not explicitly observed in our literature review but is nonetheless relevant and important: the influence of communal knowledge and shared experiences on people's usage of and collaboration with AI systems.

Communal knowledge and experiences refer to the collective understanding, values, traditions, beliefs, experiences, and mental models shared within a community. These shared experiences and knowledge can have a significant impact on individuals when they interact with AI systems. For instance, the acceptance or resistance towards AI systems may vary across different countries or communities based on their attitudes towards technology. Additionally, certain communities may possess specific priors or biases that influence their interactions with AI. Furthermore, the sharing of experiences and knowledge within communities can facilitate decision-making processes. Professionals from the same occupation, for example, may pass on their experiences and insights through shared mediums, enabling others to make informed decisions in their respective fields.

Studying people as actors within communities presents a promising avenue for future research. Exploring the dynamics of communal knowledge, shared experiences, and their influence on human-AI collaborations can provide valuable insights into the societal, cultural, and contextual factors that shape people's interactions and partnerships with AI.

3.5.2.5 Discussions of Gaps and Implications

Indeed, the category of communal traits and experiences was relatively underexplored within the scope of our defined categories. It is notable that the majority of studies focusing on communal elements were primarily within the domain of AI fairness research. While communal aspects are certainly integral to fairness considerations, it is important to extend the exploration of communal traits and experiences to other human-centered domains of AI research, particularly those related to human-AI collaborations and interactions.

Studying communal knowledge, shared experiences, and community-level influences can provide valuable insights into the socio-cultural dynamics that shape human-AI collaborations. By incorporating these communal elements into research on human-AI interactions beyond fairness, we can gain a deeper understanding of how people's collective experiences, beliefs, biases, stereotypes, and values impact their interactions with AI systems. This knowledge can inform the design of AI technologies that are more contextually sensitive, inclusive, and aligned with the needs and expectations of different communities.

3.5.3 Current Usage

The cumulative experiences and perceptions individuals develop with AI systems can have long-term and enduring effects on human-AI collaborations, as we have discussed in the preceding sections. However, it is equally important to consider the impact of short-term experiences, perceptions, and behaviors that arise during interactions with AI systems, as they can significantly influence the outcomes of human-AI partnerships. In this section, our focus shifts to these short-term factors and explores them through two distinct categories: static experiences and evolving experiences.

3.5.3.1 Short-term Experiences

Within this section, we examine a collection of papers that incorporated controls, measurements, or observations of short-term experiences and behaviors, particularly during usage sessions. These studies are categorized into six distinct categories, allowing us to delve into the nuanced aspects of short-term influences in human-AI interactions.

3.5.3.1.1. (SS1) Cognitive Processes & Biases In addition to the presence of long-term biases, individuals have the potential to develop or activate biases based on their interactions with AI systems and the performance of the AI itself. Researchers have found evidence of confirmation bias as a result of human-in-the-loop [173, 166, 101]. For instance, when studying which explanation modalities users of varied level of domain experience preferred in a news recommender system, Szymanski et al. [173] found that novice users were more prone to confirmation bias, often relying on heuristic reasoning when analyzing the output data. In contrast, experienced users demonstrated reduced susceptibility to confirmation bias, likely due to their inclination towards deductive reasoning.

Another recognized short-term bias is anchoring, which can be influenced by the order in which people encounter key information and outputs, commonly known as "first impressions" [129, 130]. It can also manifest as individuals fixate on information received early on or rely on what is more readily available to them, known as availability bias [146, 70, 35]. For instance, Poursabzi-Sangdeh [146] discovered evidence that individuals become anchored to information that is more readily available to them through the interface or based on their recent predictions. In the subsequent chapters, we will delve deeper into anchoring bias by intentionally controlling people's exposure and timing of errors, aiming to gain a deeper understanding of its influence on human-AI collaborations.

3.5.3.1.2. (SS2) Reliance & Aversion In certain studies, researchers have quantified user reliance on algorithmic recommendations and predictions by assessing the extent to which users adhere to the advice provided, which is an indication of the utility and effectiveness of the model. Reliance is often measured using the Weight-of-Advice (WoA) metric [146, 136, 137], with its calculations varying depending on the specific task. For instance, Poursabzi-Sangdeh [146] calculated WoA by measuring participants' estimation of the ground truth before and after seeing the model's outcomes. Other researchers have employed questionnaires administered during or after the study to gauge participants' reliance on AI systems [85]. Our analysis focused

specifically on papers that utilized explicit measures of reliance, as other measures like user agreement and alignment with AI outcomes are considered indirect or implicit indications of reliance.

Contrasting to reliance on AI is algorithm aversion; a pertinent yet relatively understudied behavior, which refers to the phenomenon where individuals are less inclined to accept advice from an algorithm. This aversion stems not from the nature of the advice itself but rather from the source of the advice. In other words, algorithmic aversion manifests when individuals are more receptive to recommendations from human sources but exhibit reluctance or hesitation in accepting the same recommendations when they originate from an algorithm. Within our paper set, Panigutti et al. [137] documented instances of algorithm aversion among some participants consisting of medical professionals. The study focused on AI explainability and its influence on user trust and behavioral intentions in estimating patients' likelihood of experiencing acute myocardial infarction (MI). Their findings revealed that certain healthcare professionals exhibited algorithm aversion, which stemmed from concerns about potential job displacement by AI systems.

3.5.3.1.3. (SS3) Trust In HCAI and human-AI collaborations domains, the concept of user trust holds significant importance and has been extensively measured and studied in the recent years, as seen in our paper subset. User trust encompasses various dimensions, including trust in the algorithm itself, trust in the outcomes generated by the AI system, and/or trust in the explanations or other characteristics of the system provided. Understanding the dynamics of trust is crucial as it greatly influences the types of usage behaviors exhibited by individuals in their interactions with AI systems. As trust is a dynamic construct that can evolve and become calibrated over time, it is best measured through longitudinal assessments to capture its changes and trajectory. However, it is important to note that the papers in this category specifically focus on measuring accumulative trust. Instead of examining trust over time, these studies employ one or two measures to assess trust during or after the study period. While this approach provides

valuable insights into the overall level of trust experienced by users, it does not capture the nuanced changes and fluctuations in trust that occur over an extended duration, which is a limiting factor.

To address the challenges of measuring trust, many researchers in the field of human-AI collaborations rely on preexisting questionnaires as a standardized and validated approach to assess this complex construct (e.g., [97, 41, 98, 23, 129]). Occasionally, users are asked to self-report their level of trust in the system/model/outcome [137, 85]. Analyzing and comprehending trust can be enhanced by considering the appropriateness of trust and the trustworthiness of the system, enabling researchers to assess whether users tend to over-trust or under-trust the system [193]. In summary, although static trust has been extensively studied in recent years and is a recurring theme in approximately 40% of the papers in our dataset, it remains an aspect that is not yet fully comprehended, while the lack of objective measurements for trust continue to present challenges. A promising initial step toward improving our understanding is to explore the measurement of trust over time, as exemplified by certain studies that will be discussed in section (SE1).

3.5.3.1.4. (SS4) Mental Models Understanding users' mental models of intelligent systems becomes crucial as it influences their behaviors and interactions with these systems. The real-time usage of an AI system has a direct impact on how users develop short-term mental models of the system, which in turn can shape their immediate interactions, decision-making, and ultimately, their long-term perceptions of AI. While mental models have been extensively studied in fields such as psychology, identifying objective measures to extract these models has proven challenging, particularly when seeking methods that are efficient and impose minimal cognitive load on participants. One of the more common approaches of measuring mental models in HCAI is through prediction tasks, wherein individuals are prompted to anticipate the actions or decisions the AI model would make in specific scenarios. The study conducted by Poursabzi-Sangdeh [146] serves as an excellent example of using prediction tasks to explore

individuals' abilities to simulate and predict a model's decision-making capabilities. Through a series of experiments, they aimed to gain insights into people's understanding of the model, their proficiency in predicting its behaviors, as well as their capacity to detect and correct errors. Other studies have also opted for this approach as part of their study design(e.g., [7, 128, 135, 56]). The approach of using prediction tasks is particularly valuable in assessing individuals' comprehension of the strengths and limitations of AI models. Other studies used questionnaires to measure people's mental model and understanding of the algorithm []. However, while questionnaires can be valuable in assessing certain aspects of understanding, they may not provide a focused assessment of individuals' ability to differentiate between the weaknesses and capabilities of a model, the way prediction tasks do.

One limitation of both the prediction task and questionnaire approaches is that mental models are subject to evolution and can significantly shift with increased usage or other contextual factors. Consequently, treating mental models as static and short-term factors may not fully capture their dynamic nature. To gain a more comprehensive understanding, it is essential to acknowledge and account for the evolving nature of mental models over time and in response to various influences. We will discuss some of the studies that focus on mental model dynamicity in the next section (SE2).

3.5.3.1.5. (SS5) Perceptions & Impressions Intelligent system users also develop perceptions and impressions of the system and AI as they engage in ongoing collaborations and interactions. Researchers often measure these perceptions to gain insights into the effectiveness of human-in-the-loop interactions. One common approach is to assess people's overall perception of the model's accuracy or performance (e.g., [140, 166, 128, 130, 89]). This measurement focuses on individuals' subjective and holistic evaluations of how well the system is performing and delivering the expected outcomes, and provides valuable information about users' satisfaction and confidence in the system. However, measuring user perceptions of accuracy is often confused with and seen as a possible measurement for mental models. It is important to note that while

perception of accuracy is related to mental models, it is just one aspect of the broader construct. Mental models are influenced by a range of factors beyond perception of accuracy, such as system feedback, user interactions, training, and prior knowledge. In our analysis, we encountered papers that measured other types of perceptions/impressions. For instance, Guo et al. [59] measured people's perception of control; Cai et al. [23] investigated participants' perception of the recommender system by measuring perceptions of trust[worthiness], recommendation quality, and conversational interactions; and Kouki et al. [89] measured perceptions of explanation persuasiveness.

3.5.3.1.6. (SS6) Other Attitudes Other short-term factors and attitudes studied include user's level of confidence, either in the system and its outputs or in their own assessments of the system, and user agreement with the system. For example, in the study by Alqaraawi et al. [7], participants were asked to provide their confidence in their own predictions of a model's outputs as part of a prediction task. This allowed researchers to assess the effectiveness of saliency maps as explanations and gain insights into participants' confidence in their mental models. Studying people's self-reported confidence, when coupled with other study measures, can reveal interesting (and often hidden) patterns of behaviour or help result interpretations. In the study by Heimerl et al. [62], participants' mental models of the proposed XAI system, NOVA, were evaluated across three explainability levels. Surprisingly, even in the absence of explanations, participants developed elaborate interpretations of the neural network's performance and displayed significant confidence in their assessments. This unwarranted confidence was found to be influenced by individuals' high computer self-efficacy, indicating that in tasks from more familiar domains, individuals with higher computer self-efficacy tend to attribute their own assumptions to that of the algorithms'.

3.5.3.1.7. Discussions of Gaps and Implications Within the categories we explored, algorithmic reliance and aversion were identified as areas with relatively limited research attention in the context of human-AI collaborations. Both algorithmic reliance and aversion are

measurable constructs that can significantly influence people's behaviors towards AI systems. While algorithm aversion is a well-known phenomenon in other fields of machine automation, its study in the context of human-AI collaborations has been comparatively less explored. Moreover, many researchers use trust and reliance interchangeably, while these two are clearly not the same construct. For example, people may decide to rely on an algorithm despite not fully trusting it. It is important to dedicate more efforts to understand and examine algorithm aversion and reliance, particularly by finding better ways to measure it and moving beyond solely focusing on measuring trust.

Additionally, it is worth noting that current studies often rely on measuring mental models through means such as questionnaires. While questionnaires can capture certain aspects of user mental models, they may not fully capture the complexity and nuances of these models. Future research can explore alternative methods beyond questionnaires and prediction tasks to gain a more comprehensive understanding of users' mental models. By employing innovative techniques, such as cognitive mapping, interviews, or cognitive task analysis, researchers can delve deeper into the intricacies of mental models in the context of human-AI collaborations. This will allow for a more holistic and nuanced understanding of how users perceive, interpret, and interact with AI systems.

3.5.3.2 Short-term Evolution & Changes over time

A subset of studies in our analysis employed a longitudinal measurement approach, focusing on capturing short-term and transient factors during the usage and over time. In this section, we will delve into these studies and explore their findings.

3.5.3.2.1. (SE1) Trust In five papers from our analysis, researchers focused on exploring changes in trust over time as a means to gain insights into the dynamic nature of people's trust in AI. By examining how trust evolves and calibrates based on user's observations and perceptions of model performance, these studies aimed to understand the factors influencing changes in trust. Notably, our previous study [129] specifically measured changes in trust and examined trust

calibration over time. This study stands out as the only one in our analysis that investigated the temporal dynamics of trust and how it evolves over time, which will be presented in Chapter 5.

Some researchers [193, 189] sought a different approach to measure trust over time, which was based on the alignment of people's decisions in each trial with the model's decisions. For instance, Yang et al. [193] defined four types/levels of trust, which were appropriate trust, over-trust, under-trust, and self-reliance, and calculated the accumulate decision-alignment for each level. So despite having 4 trust values, these measure still present trust over time. Another common approach is to measure and report changes in trust by assessing it prior to and after a task, as seen in studies by Yang et al. [193], Papenmeier et al. [139], and Nourani et al. [129]. Note that approaches are agnostic to the measure itself; i.e., trust can be measured implicitly, explicitly through questions over the course of the study, or via questionnaires before and after the task.

3.5.3.2.2. (SE2) Mental Models Similarly, some studies considered changes of user mental models over time as part of their experiment design. For example, Bansal et al. [11] measured the evolution of mental models over multiple trials and found that participants' mental models refined as they interacted more with the model. This suggests that continued interactions and observing model errors over time can help users learn about the model's weaknesses and adjust their mental models accordingly. Another study by Omeiza et al. [135] examined changes in mental models before and after participants read prompts depicting scenarios in the domain of autonomous driving, aiming to understand their preferred types of explanations.

3.5.3.2.3. (SE3) Task Performance Studying changes in user-machine task performance over time can provide valuable insights into the effectiveness of collaboration and other related factors. However, we found limited evidence of researchers examining this factor among the papers in our analysis, and even the existing studies did not strongly focus on performance itself. For example, Das et al.[31] asked participants to rate their perceived performance improvement over time, while Bansal et al.[11] observed differences in AI-human team performance based on

the availability of explanations when studying mental models over time. Future work in this area can play an integral role in enhancing our understanding of human-AI team performance and task accuracy.

3.5.3.2.4. (SE4) Perceptions Some studies have explored people's perceptions of AI systems over time, especially by manipulating certain characteristics of the model, such as changes in accuracy or error behaviors. For example, Honeycutt et al. [71] investigated how changes in accuracy over time influenced people's perceptions of a human-in-the-loop approach, specifically when providing feedback to the model's mistakes. They discovered that regardless of the accuracy change, individuals began losing trust in the model when asked to provide feedback, which they attributed to the strong influence of availability bias, where errors are more readily recalled as participants are asked to address them. Other studies have examined changes in user acceptance of advice over time [157] and changes in perception towards the system over time [112].

3.5.3.2.5. (SE5) Learning Effects While many studies in the short-term category focus on examining human-AI collaborations within a single session, it is also valuable to consider short-term and transient experiences that occur over multiple sessions of interaction [25, 31, 148]. By studying these factors across multiple sessions, researchers can account for longer-term learning effects and capture the dynamics of user experiences over time. For example, Ramachandran et al. [148] investigated child-robot interactions across multiple sessions to explore changes in children's behavior and learning outcomes when using a tutoring robot. They discovered that children's help-seeking behaviors played a crucial role in enhancing engagement and improving their learning outcomes over time.

3.5.3.2.6. (SE6) Changes of Attitudes In addition to the factors we have previously discussed, researchers have also examined the changes in users' attitudes over time. For instance, studies have explored changes in attitudes toward automated data science [185], adjustments of expectations of AI systems over time [164], changes in stickiness over time (referring to the

likelihood of users to continue using the system in the future)[25], and the extent to which individuals update their prior beliefs about AI after observing its performance[146].

These measures were observed sporadically in our analysis, prompting us to consolidate them into a single category. Nevertheless, each of these measures has the potential to be further explored to gain a deeper understanding of human-AI collaborations. Examining them in more detail can provide valuable insights into the dynamics of these collaborations and contribute to the overall understanding of human-AI interaction.

3.5.3.2.7. Discussions of Gaps and Implications Understanding human-AI partnerships and human-in-the-loop workflows in real-world scenarios requires controlling, measuring, and exploring usage behaviors, perceptions, and attitudes over time or during multiple sessions. However, this research direction remains relatively understudied in the field of human-centered AI. Admittedly, one of the main challenges is the difficulty of measuring variables over time without introducing distractions, confounding factors, or biases, while also minimizing participant fatigue

The current research landscape lacks objective and implicit measures that can be collected in the background to study usage behaviors over time, which presents a significant gap in the field. Additionally, there is a need for methods and techniques that can dynamically utilize calibration data to address the risks associated with incorrect assumptions and misunderstandings. For example, by monitoring people's mental models and trust calibration over time, designers can adapt model or system behaviors in real-time to enhance understanding and enable users to discern when to trust the model and when not to. Similarly, conducting studies over multiple sessions can offer unique insights and enable the measurement of factors that would otherwise remain hidden. However, the current body of work in human-centered AI is limited in terms of this study design, and there is a need for future research to address this gap. By adopting a multi-session approach, researchers can uncover nuanced findings and gain a more

comprehensive understanding of the dynamics and complexities involved in human-AI partnership that is closer to the human-in-the-loop in the wild.

3.5.4 Borderline and Interactions

Before concluding this section and our analysis, it is important to acknowledge the existence of experimental design elements that are time-dependent but do not fit neatly within the categories we have discussed so far. We describe and classify these two into two groups: (1) borderline elements and (2) interaction design.

Certain study elements and outcomes, such as automation bias, can exhibit characteristics of both long-term, stable factors and short-term, transient factors. We refer to them as borderline elements. The interpretation of automation bias may vary depending on the context in which it is observed. In some cases, it could be attributed to long-standing cognitive biases or pre-existing attitudes toward AI. In other cases, it may arise as a result of recent and specific interactions with the AI system. The distinction between the long-term and short-term aspects of automation bias depends on the specific study context and the factors influencing its occurrence. Another example is task familiarity, which can encompass both long-term past familiarity with a specific task domain and short-term effects of using a particular system to perform specific tasks over time. For example, in the study by Guo et al.[59], they measured participants' familiarity with playing Tic-Tac-Toe, which reflects their long-term past experiences with the game. In contrast, Chiang et al.[28] examined task familiarity in relation to the types of interactive tutorials provided to users, which represents a short-term effect of acquiring familiarity with a specific task through system-guided interactions.

Most of the studies included in our analysis often focused on either long-term or short-term factors and outcomes, without extensively exploring the interplay between the two. While some studies measured and observed these elements as separate measures, there is a need for research that systematically examines the interrelationship between long- and short-term factors. By controlling or comparing the interplay between these elements simultaneously, researchers can gain a deeper understanding of how they influence each other and contribute to overall human-AI

collaborations. Only 23% of the papers in our set studied the interaction effects between two or more long-term and short-term variables. These papers studied the interplay between the effects of expertise on changes of short-term and over time behaviours [157, 193, 173]; personal differences & experiences on trust, understandability, and overtime behaviours [148, 41, 89, 139]; personal beliefs and perceived stereotypes [184]; and cognitive biases and short-term experiences [129, 130, 106]. The current state of research on these topics is promising, but is not enough. Future research in the field of human-AI collaborations should strive to investigate the interplay between long and short-term differences and experiences, as well as their effects on various outcomes. By examining these factors in relation to one another, researchers can gain a more comprehensive understanding of the underlying mechanisms and dynamics at play in human-AI interactions. This holistic approach will provide valuable insights for designing AI systems that are sensitive to individual differences, promote effective collaborations, and enhance overall user experiences.

3.6 Conceptual Model of User's Past Experiences & Differences

Based on our literature review and analysis, we propose a conceptual model of user's past experiences and differences, which provides an organized overview of how people's usage sessions are influenced by their stable or transient experiences and traits. In this model, when someone begins using an AI system, they start developing mental models, perceptions, attitudes, and experiences with the system. These initial interactions shape their understanding and expectations. Over time, as they continue to engage with the AI system, they calibrate and modify their perceptions and behaviors based on their ongoing experiences and observations of the system's performance. Furthermore, people's stable and internalized prior experiences and personality traits come into play, contributing to their perceptions and current usage behaviors. These stable elements, which can include past experiences, beliefs, and traits, are activated and integrated with the ongoing usage session, influencing the user's decision-making process and interactions with the AI system. Figure 3-4 provides a high-level overview of this conceptual model, illustrating the dynamic interplay between stable and transient experiences, traits, and

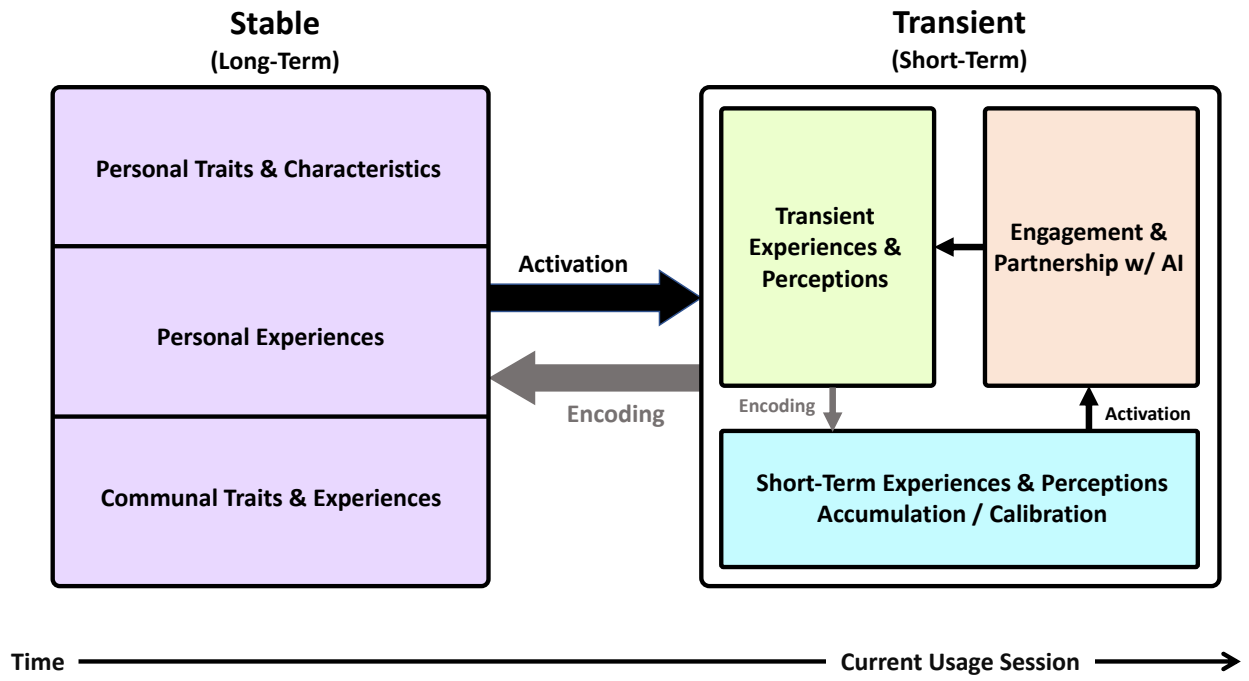


Figure 3-3. An overview of the conceptual model of user's past experiences. The model works based on stability, where the left box demonstrates long-term or stable factors while the one on the right shows transient and short-term factors.

usage session. It highlights the interconnectedness of these factors in shaping the user's overall experience and behavior during human-AI collaborations.

Let us delve into a case example to illustrate the conceptual model in practice. Meet Lucy, a biology student working on her senior project about the impact of pollution on aquatic ecosystems. Seeking guidance, her friend recommends a new conversation AI chatbot that can assist her in understanding the topic and initiating her research. Intrigued and in need of a solution, Lucy decides to give it a try.

As Lucy engages with the chatbot for the first time, her interactions are influenced by various stable elements: her background knowledge in biology, her limited experience with AI, her personal traits and cognitive characteristics, and her expectations of the chatbot's capabilities. Through a series of exchanges, Lucy begins forming a preliminary mental model of the chatbot. She discerns its strengths and limitations, noting instances where the AI's suggestions may conflict with her existing knowledge from her coursework. While she appreciates the assistance

the chatbot provides, she also harbors concerns about the potential future implications of AI technology on her career.

In subsequent interaction sessions with the chatbot, Lucy adjusts her expectations based on her prior experiences. She fine-tunes her prompts to elicit the desired responses and tailors her interactions to obtain optimal results from the chatbot. Lucy's collaboration with the AI chatbot is an iterative process that continuously evolves based on a variety of factors. Her past experiences with the system, along with her stable traits, experiences, and engagements in each interaction, collectively shape the outcomes of their collaboration.

As Lucy engages with the chatbot, her prior experiences influence her interpretations of the outcomes. She critically evaluates the suggestions and outputs provided by the AI, considering the alignment with her existing knowledge and understanding from her biology studies. These evaluations inform her decision-making process, determining whether she accepts the chatbot's outputs as-is or seeks further refinement. Additionally, her trust and familiarity with the system develop over time, influenced by her perceptions of the chatbot's accuracy, reliability, and consistency. This trust can influence the speed at which she relies on the AI's recommendations and the extent to which she feels confident in the system's capabilities.

Through this ongoing collaboration, Lucy's experiences and interactions with the chatbot continually shape her understanding and expectations. She leverages her accumulated knowledge and insights to optimize her future engagements, refining her prompts and adjusting her approach to enhance the outcomes of their collaboration. However, there may also be instances where unforeseen challenges arise or misinterpretations occur, leading to potential disruptions in the collaborative process.

3.6.1 Implications for Researchers and Practitioners

The conceptual model, when considered in the light of our literature review, can reveal interesting research gaps and avenues for future work. We present the empirically-backed version of the conceptual model in Figure 3-4. Researchers can examine human-AI collaboration by focusing specifically on the topics on each presented box/category individually, or consider

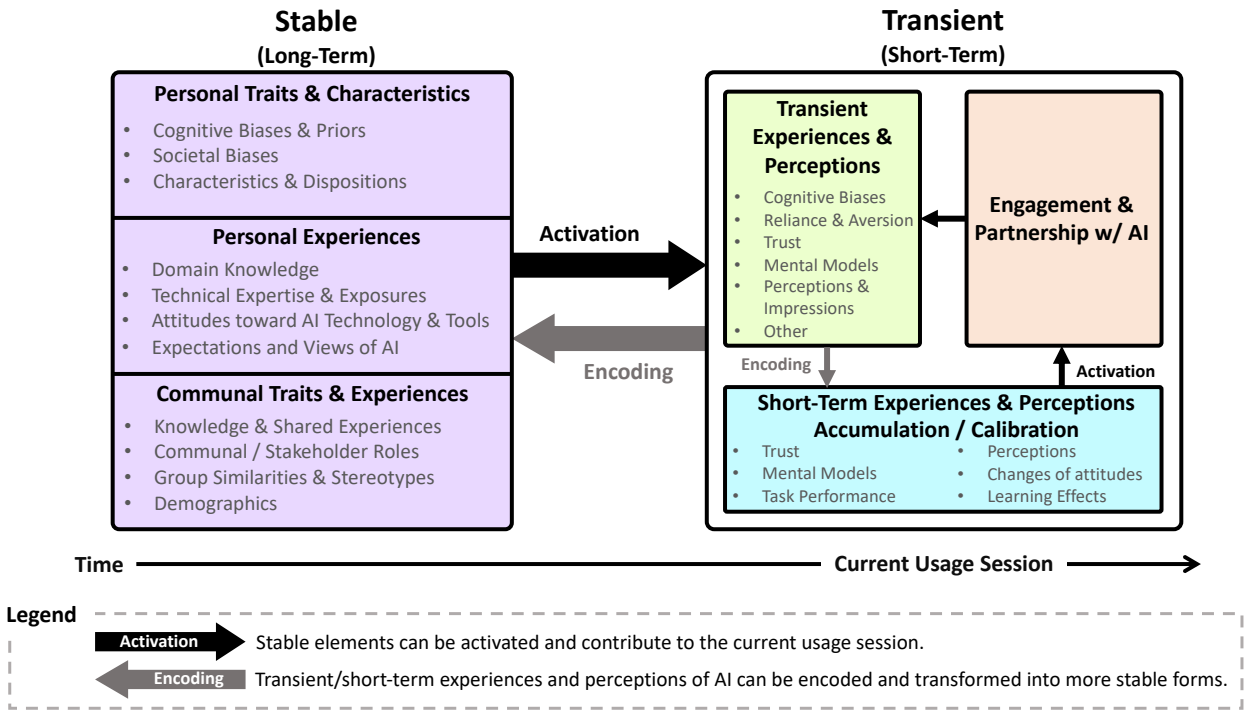


Figure 3-4. A detailed view of the conceptual model of user's past experiences can be derived from the results of the literature review, as shown in Table 3-4. This comparison allows us to examine the current coverage of topics and identify research gaps in the field.

controlling/measuring a few factors from each box and in association with one another. An intriguing perspective that warrants attention is the interplay between stable factors and transient factors within the conceptual model. Surprisingly, this aspect has received relatively less attention compared to other types of elements. The majority of studies have primarily focused on understanding how stable factors activate and influence transient factors. However, investigating the dynamic relationship and reciprocal influence between stable and transient factors holds great potential for advancing our understanding of human-AI collaborations. A notable gap in the field of human-AI collaborations is the limited research on how transient and short-term factors transform and encode into long-term and stable elements. This is primarily due to the inherent challenges associated with studying individuals' long-term perceptions and experiences across multiple sessions over an extended period, while effectively controlling for confounding variables. However, addressing this research gap is crucial for gaining a comprehensive understanding of the

long-term effects and dynamics of human-AI collaborations. Nevertheless, focusing on studying the encoding and activation of usage elements and experiences during short-term sessions presents a more feasible approach. By examining these transient factors, researchers can uncover intriguing patterns of behaviors and identify key variables that can be controlled and manipulated.

When designing AI systems, practitioners should be mindful of both stable and immediate patterns of behaviors exhibited by users. By considering these patterns and incorporating them into the design process, practitioners can leverage them to enhance the collaboration between humans and AI systems or find ways to address potential challenges. The conceptual model we have presented provides a framework for understanding the role of people's past experiences and differences in human-AI collaborations, serving as a valuable tool for guiding the design and understanding of AI systems that prioritize user-centered considerations.

3.7 Limitations and Future Work

Every research endeavor has its inherent limitations, and our work is no exception. One of the challenges we encountered was the process of extracting relevant literature on human-AI collaboration from diverse communities in a consistent yet non-exhaustive manner. The observed variations in the number of examples across different categories may be attributed to selection bias and the limitations associated with specific keyword choices. It is important to note that our study does not claim to provide a comprehensive or structured literature review. Rather, our aim was to gather sufficient examples to identify important and noteworthy research trends in the field.

We acknowledge that selecting distinct yet inclusive categories that adequately capture the diverse range of stable and transient factors was challenging. Some categories may exhibit overlap, and the level of granularity in distinguishing between them may not be as precise as desired. This limitation stems from the inherent complexity of studying human psychology, cognition, lived experiences, and societal roles. The multifaceted nature of humans and their interactions makes it difficult to create a rigid categorization that perfectly encapsulates every aspect. However, our model serves as a framework to understand and explore the various

dimensions of human-AI collaboration, while recognizing the need for ongoing refinement and adaptation as our understanding of these factors evolves.

In future work, we aim to conduct a more comprehensive examination of the literature to enhance and expand upon this conceptual model, deepening our understanding of human-AI collaborations. This will involve exploring additional research studies across various domains and disciplines to capture a broader range of stable and transient factors. Furthermore, we seek to leverage this conceptual model as a framework for identifying gaps in our knowledge and understanding of human-AI collaborations. By analyzing the existing literature and identifying areas that have received limited attention, we can highlight avenues for further research and exploration, as we discussed earlier.

In the upcoming chapters, we present user studies that serve as illustrations of how this conceptual model can be applied. These studies will demonstrate the practical utility of considering past experiences and personalities, both stable and transient, in understanding and predicting people's behaviors and usages of AI systems. By examining these examples, we aim to showcase the significance of incorporating the factors highlighted in our conceptual model into the design and evaluation of AI systems.

CHAPTER 4
ANCHORING BIAS AFFECTS MENTAL MODEL FORMATION AND USER RELIANCE IN EXPLAINABLE AI SYSTEMS *

According to the conceptual model presented in Chapter 3, it is essential for AI system designers to carefully consider the impact of past influences on users' usage behaviors. These influences can be categorized as either stable/long-term or transient/short-term. In this study, we aim to investigate the effects of short-term experiences, specifically focusing on anchoring bias formations, on the dynamics of human-AI partnerships. Anchoring bias refers to the cognitive phenomenon in which individuals tend to heavily rely on the initial information they receive when making subsequent judgments or decisions. In the context of using an AI tool for the first time, users may develop anchoring biases based on their initial interactions and exposure to the model's mistakes or inaccuracies. These short-term experiences can significantly shape users' perceptions and subsequent behaviors throughout their usage session. Since user's ability to trust and rely on an algorithm is deeply rooted in such an understanding [115, 128], it is important to study mental models' formation and quality with intelligent systems. In this chapter *, we present an empirical user study, with two high-level goals: 1) to understand how transient prior experiences affect people's mental models, usage, and confidence in the model, and 2) to investigate the interplay between explainability and short-term experiences, and whether explainability is a consistent solution to improving mental models.

4.1 Explainable System

4.1.1 System Context

For this study, we sought an open-ended scenario where users could explore the system and build a mental model of how it works. With some intelligent systems, errors can be tolerated to some extent and they may not be fatal. That is why it might seem unnecessary for the users to build mental models of the system. However, some systems naturally require a human agent to monitor the outcomes and predictions rather than automatically accepting failures without worrying about the consequences. Examples of such systems, and our system of choice, include

*This chapter is based on my published work, which has also received a best paper, honorable mention award: Nourani, Mahsan, Roy, Chiradeep, Block, Jeremy E., Honeycutt, Donald R., Rahman, Tahrima, Ragan, Eric D. and Gogate, Vibhav, 2021, April. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In 26th International Conference on Intelligent User Interfaces (pp. 340-350).

video activity recognition systems, where a model can be trained to automatically detect activities that take place in the videos. In real-world scenarios, activity recognition has many use-cases and can be critical due to physical limitations and time constraints. Some examples include fire detection [91], airport security [176], smart hospitals [174, 194], and elderly care [84]. Since we desired a task where users are novices and do not require any certain expertise or professional training, we chose a cooking video scenario where the system was designed to identify cooking-related tasks in a kitchen. In the rest of this chapter, we briefly describe the model and interface we used for the system we designed for our experiment.

4.1.2 XAI Model

The XAI model used in this study was trained on a pre-annotated dataset of cooking videos called the TACoS dataset [154]. Note that the development of the XAI model is not a part of the contributions presented in this paper, as the model was only used to serve the goals of the experiment while using a real explainable model for the system. More details on the specifics of the model can be found in our previous work [155]. Here, we provide an overview of the model to help readers understand the basis for the model capabilities and explanations.

In the TACoS cooking videos [154], each frame of each video had a set of labels (which we call ground labels) that summarized the activity taking place in the video (for example, {"wash", "carrot", "sink"} in frames where a carrot was being washed in the sink). The problem was formulated as a multi-label classification problem where given each frame of the video, the model had to assign the correct labels to it. Each label was modeled as a binary random variable where 0 and 1 indicated that the label was off or on respectively. We implemented a two-layer architecture where the first layer comprised a deep neural network based on GoogleNet [172] that converted each frame into a set of noisy labels and the second layer used a dynamic version of a tractable probabilistic model called a cutset network [147] that modeled a conditional probability distribution of the ground (true) labels given the noisy labels from the neural network, i.e., $P(G^{1:t}|E^{1:t})$ where $G^t = \{G^t_1, ..., G^t_n\}$ is the set of ground labels at frame t and $E^t = \{E^t_1, ..., E^t_n\}$ is the set of corresponding noisy evidence labels. The top layer was designed as an "explanation"

layer in order to (1) remove the noise from the GoogleNet labels and (2) model the temporal relationships between the ground (true) labels. The model was trained on 30 videos with a vocabulary of 35 labels. Explanations were computed on the final trained model by formulating them as two standard probabilistic inference queries: posterior marginal (MAR) and top-k most probable explanation (MPE). The MAR query seeks to estimate the probability of the true label given noisy labels obtained from GoogleNet while the top-k MPE query seeks to find the top k most likely assignments to the true labels.

4.1.3 Main Interface

We designed a video activity searching tool to allow users to build specific queries and sort the videos from the dataset. In this tool, we define each activity using three component types: Action, Object, and Location. Figure 4-1 shows the overview of the interface. The top of the screen has a simple query builder where users can input specific component combinations or select a generic form (e.g., any action). After searching, the interface would organize the videos into two lists based on whether the model found the searched activity in each video or not. The XAI system showed thumbnails for each video to distinguish them from the other videos in the list. Each video was assigned an id number and day of the week to help users track how the system responded.

4.1.4 Explanation Interface

By clicking on a thumbnail, a modal overlay would open where users could watch the video and see the model explanations to examine why the video was categorized as a match (or non-match) for the query. Figure 4-2 shows the three explanation elements for each video that aimed to assist the users in understanding why the model matched the query with the video. Directly under the video progress-bar (Figure 4-2.C) was a series of video segments that highlighted the most relevant set of frames used by the model to answer the current query. Clicking a video segment updated the information presented in the other two explanation elements: (1) The detected combinations (Figure 4-2.D) listed the top 3 queries that the model associated with the currently-selected video segment and (2) the detected components

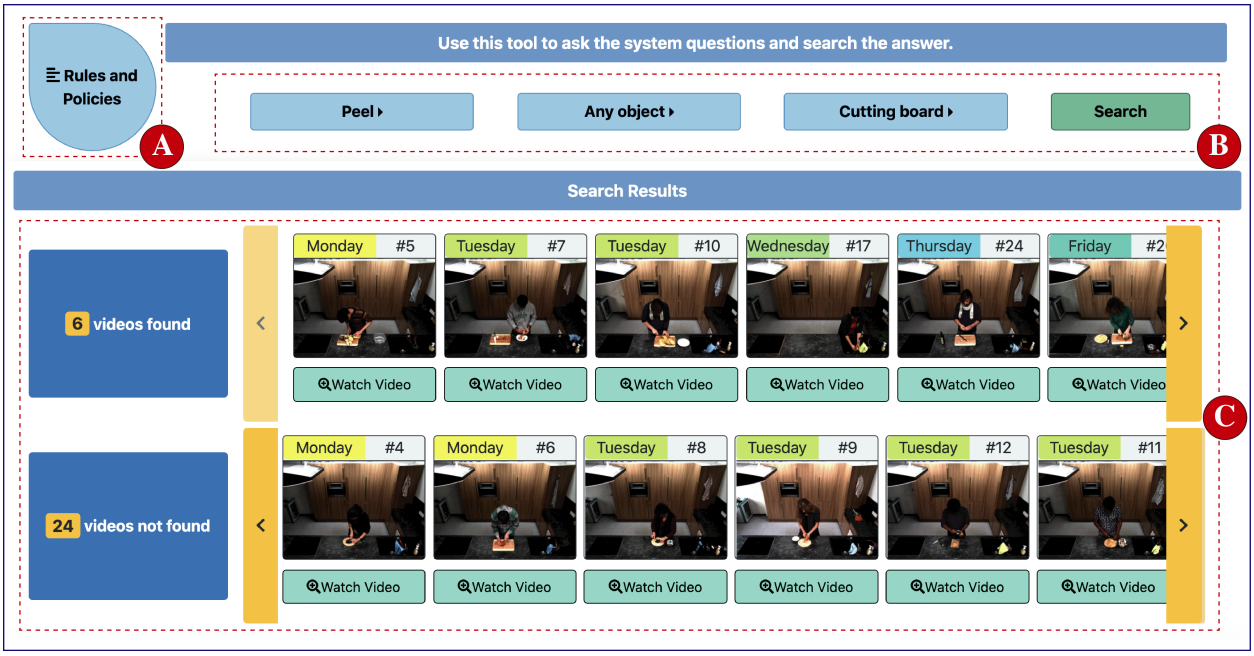


Figure 4-1. The main overview of the user interface. By clicking in the top left corner (A), a panel opens from the left side of the screen that includes a list of policies. Here, users recorded the kitchen's compliance with each statement. (B) Users selected components from three drop-downs to build a query and search for it among the videos. (C) The search sorted the thumbnails into two categories: matching and non-matching videos. By showing a thumbnail preview of each video, their assigned unique ID, and their corresponding weekday, users could select watch video to inspect and explore more.

(Figure 4-2.E) showed the model's confidence about the activity components detected separately in this video segment

4.2 User Experiment

We conducted a human evaluation to understand how first impressions of intelligent systems can influence user mental models, as well as task performance and reliance on the tool. We also sought to learn whether explanations can help bypass the biases formed in the earlier encounters with model predictions. In this section, we describe our experiment design in more detail.

4.2.1 Research Goals and Hypotheses

For this study, we were primarily motivated to understand the role of first impressions on a user's mental model formation. As one of the main motivations behind XAI research is to improve user understanding and mental models of intelligent systems [58], we deemed to test

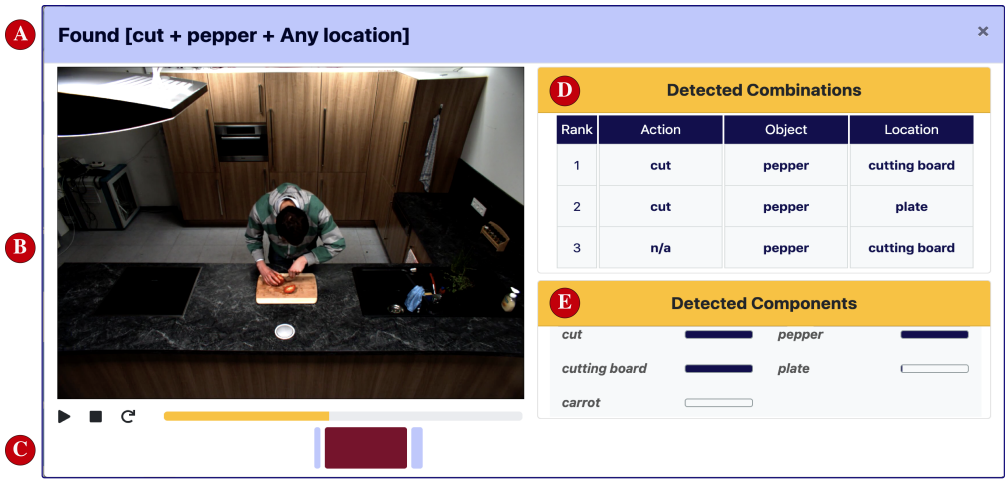


Figure 4-2. When clicking on the watch video button in the main interface, as seen in Figure 4-1, participants would see a modal to allow them to watch the video. (A) showed the selected query and whether the query is found or not found in the video (B). If they were in the explanation presence, they were shown all the video segments that were used to come up with the answer (found/not found) under the progress bar (C). They were able to click on each of the available segments to see the model justification based on the relevant activities found in the segment (D), as well as the system's confidence score in all the components it detected within the selected segment (E).

whether and how the addition of explanations can affect user mental models, given that users might have formed initial biases in their assumptions towards the system. Therefore, we designed a policy-verification task, where the system described in Section 4.1 was used to verify whether a set of kitchen guidelines and policies are being followed by the people performing cooking activities. This was a task, exploratory enough to allow users to freely test and observe various system predictions to build a mental model of both system weaknesses and strengths. Moreover, with an open-ended and real-world scenario, we are able to generalize our findings to other intelligent decision aids. We designed a study where participants observed the same set of policies, while we controlled that earlier in the usage, some observed policies that expose system weaknesses while others observed the policies that exposed system competencies. Also, with each order, some participants were provided explanations while others were not. By comparing these conditions, our evaluation explored how users' interpretation of the same system may be different

based on their experience of system performance with or without the addition of explanations. These goals and research question are summarized in the following set of hypotheses:

- H1: Encountering model weaknesses early-on will lead to less usage and reliance compared to encountering model strengths early.
- H2: Positive first impressions can improve user mental models while negative first impressions can impair them.
- H3: Regardless of the order of encountering model weaknesses and strengths, model explanations help decrease or eliminate the effect of anchoring bias on user reliance on the system.
- H4 The addition of explanations will significantly improve user task-performance and mental models by increasing their understanding of AI system weaknesses and competencies.

4.2.2 User task

Using the XAI system described in Section 4.1, we sought an exploratory task to allow the participants to use and experience the system and build a mental model of it. As we were also considering a task that did not require any expertise or professional training, we used a kitchen policy scenario, where participants were given a set of kitchen rules and policies and were asked to determine, using the system, which of the policies were being followed by the kitchen staff.

We generated intricate policies that generally required users to build and test multiple queries in order to encourage further use of the intelligent system. Each policy was designed to either expose model weaknesses (i.e., components that were misidentified or remained unidentified) or model strengths (i.e., those components known to be consistently identified correctly). Due to this design, we ended up with 4 policies focused on system weaknesses and 4 policies focused on system strengths. Additionally, we used one policy as attention check, which was unique since it was not ubiquitously followed by the kitchen staff, but would sound logical to users not watching the videos: “Employees wash their hands immediately after entering the

kitchen". Ultimately, participants received nine policies to interpret and were asked to determine their truthfulness in a set of thirty cooking videos. Policies were simple statements of fact that used components available in the query builder, like "Employees must not use pineapples more than 3 days a week" or "Carrots are only cut on rectangular cutting boards". Additionally, since the post-task questionnaire asked users to report on their mental models and usage of the system, we repeated components in multiple policies to increase memorability and to support user understanding.

The interface included a list of policies (a hidden panel on the left side of the screen until the participants decided to open them by pressing the "Rules and Policies" on the top left corner of the screen, as seen in Figure 4-1.A), and participants indicated if each was met with yes and no buttons.

4.2.3 Conditions

To address our goals and hypotheses, we designed a 2x2 between-subjects user study with two independent variables: (1) policy order and (2) explanation presence. Participants were assigned one of the four conditions randomly and everyone completed the same task. We controlled the order of observing policies so that some participants were exposed to system weaknesses first while others were exposed to system strengths first. We also maintained that the attention check policy would always remain in the middle of the list of policies. Ultimately, all participants observed the same set of policies, but with varying order. In pilot testing, we observed that participants consistently examined each policy in sequence starting from the top of the list, so we relied on this behavior to control for the policy order factor. We also updated the system interface described in sections 4.1.3 and 4.1.4 to match the assigned condition. We changed the video thumbnails to show the most relevant frame for the with explanations conditions and the middle frame for the no explanations conditions. Also, while those in the with explanations conditions observed all the three explanation elements within the explanation interface, the participants in the no explanations conditions were only provided with the video player (i.e., only elements (A) and (B) in Figure 4-2).

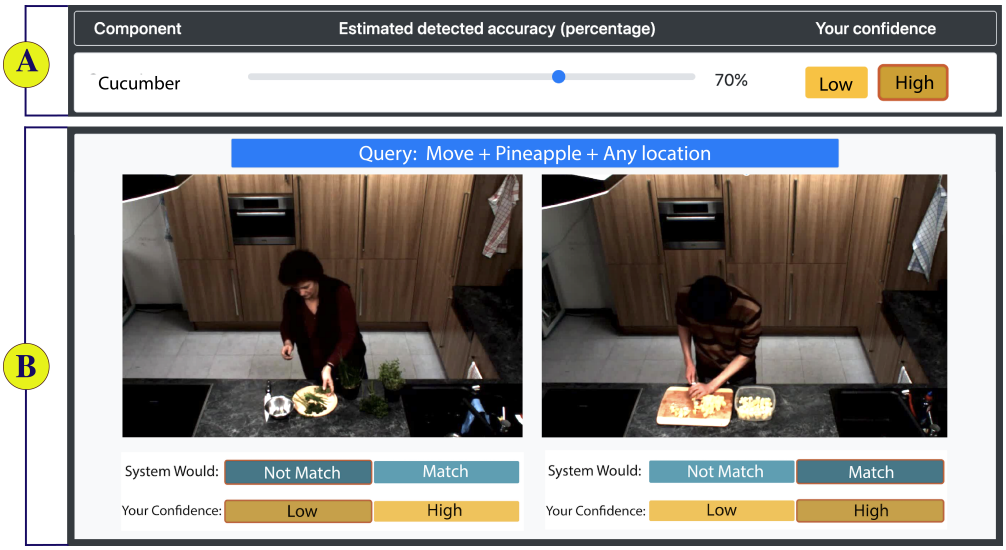


Figure 4-3. Examples of the mental model questions for the user study. (A) The user estimated the accuracy for cucumber was 70% and had a high confidence in their estimation. (B) Frame-query estimation where the user guessed whether the system matched each frame to the query and rated their confidence in their response.

4.2.4 Measures

In addition to interaction logs, we asked participants to complete four post-task questionnaires designed to quantify and explore the limits of users' perception of the system's strengths and weaknesses (i.e., their mental models), as well as usage and reliance. We selected two types of questions for assessing mental models. The first, as shown in Figure 4-3.A, asked users to estimate the detection accuracy for eight activity components we selected that appeared in the policies frequently. Some of these components were from model weaknesses (e.g., pineapple) and some of them were from model strengths (e.g., carrot). Estimation of accuracy is an established known method for estimating general user understanding of model performance and mental model of system capability (e.g., [115, 129, 70]). With a slider, users indicated how accurately the system detected each component (0–100%) and also marked their confidence (low or high) in their answer. In the second question, as seen in Figure 4-3.B, the participants were given an activity query with a set of 4 video thumbnails and were asked to predict whether the system would categorize each thumbnail as a match or non match using their mental model of the system. They were also asked to rate their confidence in their prediction (low or high). We

provided three queries, each with four assigned thumbnails, making a total of 12 frame-query predictions per participant. This measure was inspired by prediction tasks which are another established method in assessing and measuring the user's mental model of AI/XAI systems [115, 67].

We then asked the participants to rate both usage and helpfulness for each interface element on a 5-point Likert scale. These measures were adjusted for participants based on their explanation condition (i.e., they were only asked about components they saw). Finally, they rated their estimation of the model's overall accuracy in percentage, as well as answering a few free-response questions describing any noticeable weaknesses or feedback to the researchers.

4.2.5 Procedure

In a single online session, participants completed the following, as summarized visually in Figure 5-3. The research was approved by the organization's institutional review board (IRB). All participants took about 20 minutes to verify all the policies. After observing the study's informed consent, participants were asked to complete a brief demographic background questionnaire.

Participants were then introduced to their task via video tutorial that described the task as well as how to form a query by providing an example. To help participants understand the task better, we designed a tutorial video, introducing a hypothetical restaurant owner who asks the participants to use the intelligent tool and verify whether the kitchen rules are being followed by her employees by inspecting the surveillance footage from the past week. Participants were informed that one food was prepared by one chef per video and that there were six videos per day of the week (i.e., 30 videos in total). The tutorial then described how to use the tool and how the task can be achieved. To avoid learning effects, the tutorial used an extra policy to demonstrate the interface functions. We created two versions of the video for each of the with explanations and no explanations conditions. We also included a summary of the tasks and important considerations on the main page under the query building tool for users to refer to during the study.

After the tutorial, the main task had participants verify nine relevant kitchen policies listed in a sidebar. After answering all nine policies, the participant continued to the post-study

questionnaire to evaluate their mental model and understanding of model weaknesses and strengths (more detail provided in Section 4.2.4).

4.2.6 Participants

We recruited a total of 116 participants from the university graduate and undergraduate students to complete the study online for class credit. The participants consisted of 78 males and 38 females. After carefully investigating the responses, we removed a total of 6 participants since they did not pass the attention check. Of the 110 remaining participants, 54 observed explanations: 28 of whom saw strong policies first and another 26 observed the weak policies first. Of those provided no explanations, 29 observed strong policies first while the remaining 27 initially saw weak policies. All participants were compensated, including those who did not pass the attention check.

4.3 Results

In this section, we present the measures of our study and provide an analysis of the results.

Before performing data analysis, two steps were taken to avoid certain problems caused by performing an online study. To ensure the quality of participant responses without having a researcher present during the study sessions, we added an attention check policy and removed all of whom did not pass the test. Additionally, to account for some participants taking breaks during the task, we adjusted the task completion time by not counting any period of inactivity longer than five minutes. For each of our measures, we used a two-way factorial ANOVA for the main effect and Tukey HSD post-hoc testing for significant interaction effects, when applicable.

4.3.1 User-task Performance

First, to test our hypothesis about user-task performance, we tested both task time and task error to test. Task time is defined as the amount of active time spent on the policy review task. Task error was measured as the proportion of policies that the participant answered incorrectly. No significant effect was found for explanation presence. However, participants in the weak first conditions had significantly less error in their answers to the policy questions than participants in the strong first conditions, with $F(1,106) = 6.55$, $p < 0.05$, $\eta_p^2 = 0.058$. No evidence of an

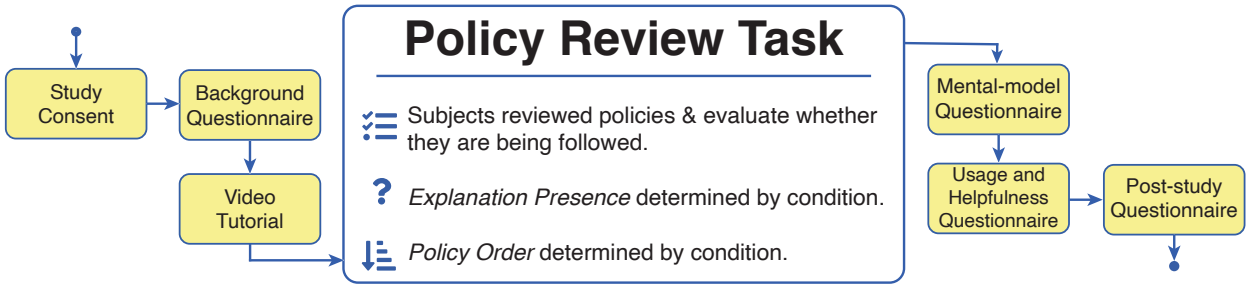


Figure 4-4. An overview of the user study procedure.

interaction effect between explanation presence and policy order was observed. Additionally, no significant effects were observed on task time. Figure 4-6.a shows the distribution of the task-error results across the conditions.

4.3.2 Component Accuracy

After completing the policy-review task, participants were asked to estimate the model's detection accuracy (percentage) for several components as described in Section 4.2.4. An example question for this measure is shown in Figure 4-3.A. We selected these components so that five corresponded to system weaknesses (low model accuracy) and four to system strengths (high model accuracy). We compared the participants' perceived accuracy of each component with the system's actual accuracy for that component. Since our task and interface primarily had participants focusing on the matches returned by the system, we selected the system's positive predictive value of each component as the metric for system accuracy. Additionally, we only considered system performance on the videos that were used in the task.

For analysis purposes, we used the average error in percentage for both weaknesses and strengths for each participant separately, i.e., two metrics per participant. A similar approach was used for the confidence scores. The reason for this decision was to be able to compare the user's mental model of both system weaknesses and strengths and understand how each independent variable affected this understanding. We will discuss each of the two separately below:
Weakness Detection: For components that corresponded to system weaknesses, the statistical tests did not indicate significant differences across the conditions for neither the accuracy nor confidence.

Strength Detection: For components that corresponded to system strengths, participants who observed weaknesses first significantly underestimated the model's detection accuracy compared to those who saw strengths first, with $F(1,106) = 6.24$, $p < 0.05$, $\eta_p^2 = 0.056$. Additionally, participants who observed weaknesses early-on were significantly less confident about their estimations compared to those who saw strengths early, with $F(1,106) = 3.94$, $p < 0.05$, $\eta_p^2 = 0.036$. We did not observe any significant effect based on explanation presence on the user's strength-components' accuracy estimation or the confidence in their estimations. Figure 4-5.a and 4-5.b show participant responses and their confidence across the conditions, respectively.

4.3.3 Frame-Query Prediction

Additionally, we asked participants to predict what output the system would have on a given frame-query pair, as observed in Section 4.2.4. An example of this prediction question can be seen in Figure 4-3.B. We did not observe any significant differences among the conditions for the prediction accuracy. The mean prediction accuracy was $M = 0.599$ with a standard deviation of $SD = 0.127$ for participants with explanations and $M = 0.601$ with a standard deviation of $SD = 0.148$ for participants without explanations. This shows that users' estimations were barely better than guessing. However, a significant effect was observed on the confidence participants had in their responses. Participants with explanations were significantly more confident in their predictions than those without explanations, with $F(1,106) = 4.12, p < 0.05, \eta_p^2 = 0.035$. There was also a significant interaction effect between explanation presence and policy order with $F(1,106) = 5.20, p < 0.05, \eta_p^2 = 0.047$. A Tukey multiple comparison test showed the following significant interactions: Among the participants with no explanations, those who observed strong policies first were significantly more confident than their counterparts ($p < 0.05$). Participants with system explanations and strong policies first were more confident than those with no explanations and weak policies first ($p < 0.05$). Finally, of the participants who observed policies reflecting weaknesses early on, those who had system explanations were significantly more confident than those without explanations ($p < 0.01$). Figure 4-5.c shows the confidence of the participant's responses among the conditions.

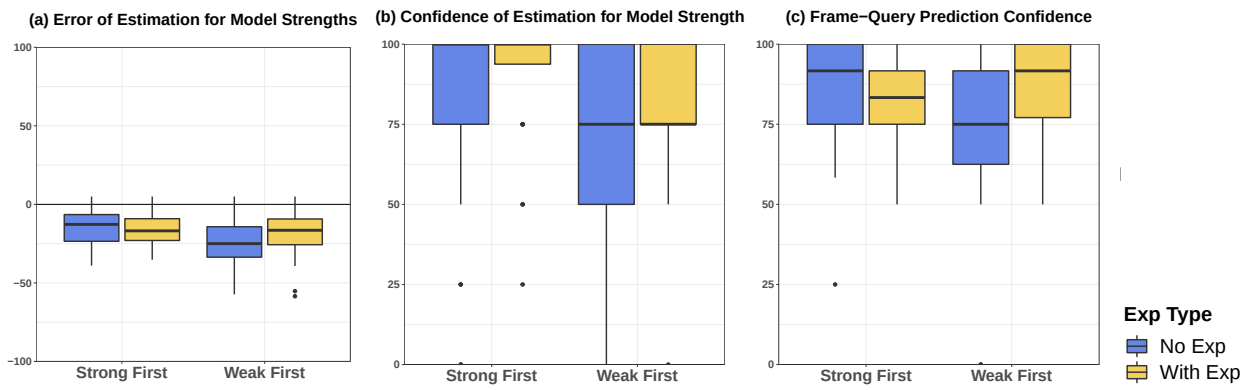


Figure 4-5. Mental model metrics. (a) Participants' error of estimation for component accuracy (below 0 is underestimation). (b) Percentage of components for which participants rated as being confident in their estimation. (c) Percentage of frame-query pairs for which participants felt confident in their predictions. The last two plots are based on strength-detection (as described in Section 4.3.2)

4.3.4 Explanation Usage and Helpfulness

After finishing the mental model questions, we asked the participants to report their usage of different interface components and how helpful they found them during their interaction period. Particularly, we were interested in the responses from those in the with explanations conditions about the provided system explanations; i.e., video segments (Figure 4-2C), detected combinations (Figure 4-2D), and detected components (Figure 4-2E). Both usage and helpfulness were measured through a 5-point Likert scale. To run a more accurate analysis based on these three explanation types and policy order, we defined explanation type as a new independent variable for the analysis, and then performed a two-way independent ANOVA on explanation usage and explanation helpfulness. The results show participants who encountered weaknesses first reported a significantly lower rate of usage of system explanations than participants who encountered strengths first, with $F(1,156) = 4.76$, $p < 0.05$, $\eta_p^2 = 0.030$. Additionally, we found that regardless of policy order, participants strongly preferred the video segments (Figure 4-2C) in terms of both helpfulness and self-reported usage, with $F(2,156) = 9.77$, $p < 0.001$, $\eta_p^2 = 0.111$ for explanation helpfulness and $F(2,156) = 16.70$, $p < 0.001$, $\eta_p^2 = 0.176$ for self-reported explanation usage. We also analyzed user behavior—captured through interaction logs—to

understand the usefulness of explanations by measuring how many queries participants performed on average for each policy. Participants who had system explanations completed the policy review task with significantly fewer queries per policy than participants who did not have system explanations, with $F(1,106) = 4.94, p < 0.05, \eta_p^2 = 0.045$. No effect of policy order was observed for the number of queries made. Figure 4-6 shows the self-reported usage and helpfulness of the different explanation types and the number of queries performed based on condition.

4.4 Discussion

Our results demonstrate significant effects of first impressions on mental model formation, user reliance, and usage of the intelligent systems. In this section, we discuss the general indications of our results as well as their limitations and provide implications for system designers and opportunities for future work.

4.4.1 Interpretation of the Results

Participants in the strong first conditions had significantly more user-task error compared to those in weak first conditions. While this might seem counter-intuitive, it can be explained when compared to the findings from usage and helpfulness, as those who encountered system strengths earlier used explanations significantly more and found them to be significantly more helpful in the task compared to those who encountered weaknesses early. This indicates that observing strengths first can cause users to rely on the system more than they should (i.e., automation bias), while seeing weaknesses in the beginning can prevent this problem.

On the other hand, users in the weak first condition had problems forming their mental models of the system competencies and strengths. They significantly underestimated the system capabilities while also having less confidence in their estimations. These users are skeptical of system strengths but not confident in their skepticism because the weaknesses they observed earlier obscured their judgment of the system capabilities. This causes them to rely more on themselves rather than the model, leading to more confusion when shaping their mental model.

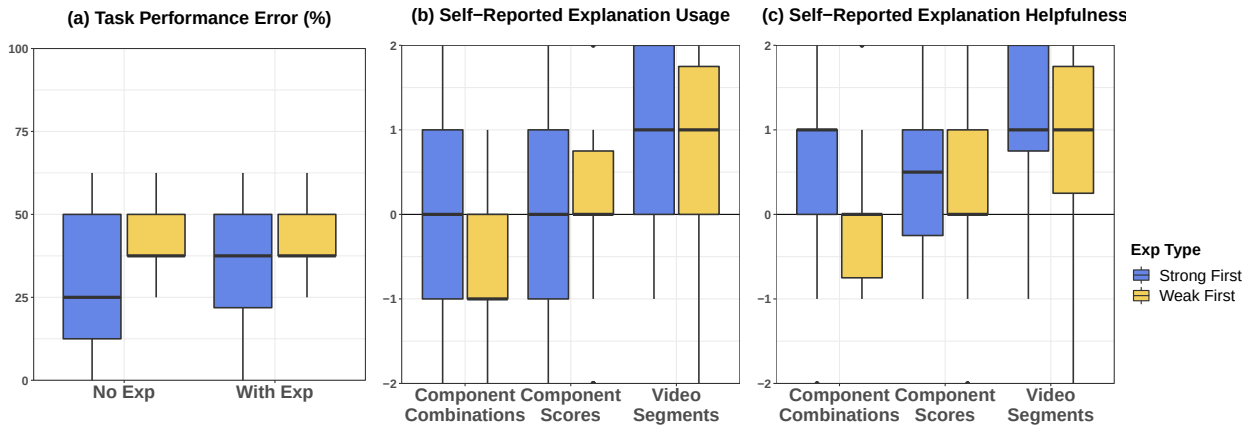


Figure 4-6. Reliance and Usage metrics. (a) Participant error on the policy task (Percentage). (b) Responses to the question "How much did you use this element?". (c) Responses to the question "How helpful did you find this element?". The last two were measured on a 5-point Likert scale, with higher values indicating a higher rating of helpfulness and usage.

We designed the frame-query prediction task to measure the user's granular mental model based on the specifics of the system. Though we did not observe any significant effects on the user's prediction, we did observe significant effects for the user's confidence in their prediction. Participants were more confident about their mental models when explanations were present. However, given that the mean for their original predictions were consistently around 50% in all the conditions (which is similar to guessing), we can conclude that these relatively high reported confidence scores are overconfidence. Our interaction effects show that without explanations, users in the strong first condition were more confident about their mental model, which we suspect is due to their automation bias, as discussed before. However, we observe that with explanations, users—regardless of their policy order—were more confident about their estimations compared to without explanation condition in weak first order. This might indicate that users can experience overconfidence in their mental model either when explanations are present or when strengths are observed earlier. However, we observed this overconfidence and overreliance through multiple tests for strong first order, showing that the order effect plays a more important role on a user's mental model than explanation presence (this can be supported by our results related to user-task

error: users in weak first condition made fewer errors regardless of their explanation condition). This suggests that explanations alone cannot solve the strong bias created by first impressions.

Overall, these results suggest that unlike the general belief that model explanations can increase user understanding, they might not necessarily be beneficial. Explanations might cause a misbelief in the users that they understand how the model works when, in fact, they do not. As shown by previous research in psychology, overconfidence (in this case, in the form of overprecision) can have serious consequences [118, 47]. Similarly, previous research suggests overreliance can cause several problems [19, 157], and our results provide a clear example of users making more errors due to automation bias. First impressions have strong influences on human's minds towards information [175], and as shown by our results, they can be strong against automated systems as well. We would encourage future research into mitigating such biases, as they can have lasting effects on users' minds. More intensive and meaningful user studies are needed with realistic systems—as other researchers (e.g., [113, 42, 1, 114]) have also argued—to expose such biases and find techniques to (1) make users aware of their biases, (2) prevent users from forming new biases, and (3) help users rectify their own misconceptions and inaccuracies in mental models.

4.4.2 Implications for Intelligent System Designers

With more complex and exploratory systems, the role of instructions and guided training becomes more inevitable; that is, allowing the users to use the system without interventions might affect how their mental models are shaped. With more critical tasks, it might be beneficial for the system to guide the formation of the mental model early-on to help users develop a more accurate foundational understanding of the system before actually using it in practice to make important decisions. Through this initial training phase, designers can control what kind of predictions users observe and in what order they are observing them. These decisions are task-dependent and can be made based on the priorities in that system. For instance, if sacrificing human-task accuracy (due to errors made from automation bias) to encourage the formation of more accurate mental models is acceptable, the introduction might focus more on showing system strengths earlier in

the usage. Designers might also choose to sacrifice the mental model formation since they want to limit the number of mistakes made by the users, and thus, they can focus on highlighting more errors earlier in the usage. However, most designers might strive for the best of both worlds: limit the user mistakes by avoiding automation bias while allowing users to maintain an appropriate mental model of the system. Based on our findings, users who observed strengths earlier made more errors but formed a better mental model of the system strengths. Considering this finding, in the initial training, designers can guide users' early observations toward model strengths but also intervene and show errors occasionally to balance users' attention with errors as well. When errors are shown, designers can focus more on explaining why they happen. This can be done by altering explanation type, scope, and focus and differentiating it from the explanations provided for the correct predictions. Note that this is only possible in guided training as designers know what instances are correct and which are wrong.

Theoretically, a higher-level explanation could help users scaffold more accurate mental models by first introducing how the system works before using the instance-level explanations. Previous research suggests that global visualization and explanations can help users form a more appropriate perception of how the model works [115]. Allowing users to explore and understand how the model works on a higher level might help users form a mental model before encountering the intelligent system for the first time. Future research needs to test the extent of information sufficient for global visualizations for mental model formation, and whether this approach is effective for avoiding ordering and anchoring biases when using instance-level models. Finally, designers need to consider the effect of first impressions when designing explainable interfaces and be aware that the sole addition of explanations cannot circumvent bias formation. Comparing various types of explanations against one another (e.g., why and how explanations [99, 42, 1]) to understand which method works better against certain biases, or incorporating multiple explanation scopes within one interface might allow users to decide what they want to explore to understand the model decisions better. For example, with an analytical tool, a user can look for

different types of information and explanations from the model when encountering errors to improve their understanding of the model.

4.4.3 Limitations and Conclusion

In this research, we studied how ordering biases can affect a user's mental model and reliance formation in intelligent systems and what role explanations play with such biases. Our study presents novel findings that highlight the importance of users' first impressions on their formed mental model of the intelligent system. The results demonstrate that when encountering system strengths earlier in the usage, users built a better mental model of the system strengths as they used the system explanations more frequently. But, positive first impressions can lead to automation bias and more errors as the user is overconfident in not only the model's strengths but also the weaknesses of the system; and they generally over-rely on the system. In contrast, when encountering system weaknesses early-on, users tend to rely more on themselves and make fewer errors; likely because they develop a mental model that is skeptical of the system strengths due to their negative first impressions.

Our focus was on a machine learning technique that produces high-level explanations with a novice-friendly explanation interface (e.g., instead of using probabilities, we showed visual bars). While we believe our results can generalize for various real-world systems incorporating this class of explanations, these results might not generalize for low-level, more technical explanations. Future research needs to test and compare ordering bias with these explanations as well. Further, since our system employed instance-level and local explanations, additional research is needed to assess whether these results hold for higher-level, global intelligent systems.

Due to the nature of the design for our query-building tool, when users searched for an activity, we divided the video into two categories of matched and not matched based on whether the system detected the activity within each video. The detection is of course not always correct, i.e., a system might categorize a video as a match when the activity did not take place in the video (false positive error) or categorize a video as a mismatch while the activity is in fact taking place in the video (false negative error). For most of the activities, the number of matched videos was

smaller than the number of not matched videos, and thus, users needed to explore and view fewer videos to detect false positives. Since it was easier to determine false positives, we expect that the participants would fail to catch lots of false negative errors, i.e., the videos that the system failed to match for the query. As a result, some system weaknesses were harder to identify, potentially leading to improper mental models of system weaknesses. We suspect that this is the reason the study could not find evidence of differences between the conditions based on a user's mental model of the model's weaknesses. Future research may benefit from refined evaluations focusing on both error types to test user's mental model formation for both strengths and weaknesses.

CHAPTER 5
THE ROLE OF DOMAIN EXPERTISE IN USER TRUST AND THE IMPACT OF FIRST IMPRESSIONS WITH INTELLIGENT SYSTEMS *

The conceptual model of of user's past experiences (presented in Chapter 3) describes the importance of studying stable and transient user past experiences and differences that can lead to different usage behaviours. In Chapter 4, we demonstrated how recent past experiences (through the lens of anchoring bias as an evident example of transient past) affect people's mental model formations, confidence, and task performance. In this Chapter*, however, we study user trust formations and calibrations based on the interactions between transient and stable past experiences and differences. Similar to the previous chapter, in this study, we investigate the impact of temporary experiences on human-AI collaborations by controlling people's first impressions based on their anchoring bias. Additionally, we incorporate stable past experiences as a control factor, through considering and recruiting based on participants' domain expertise.

5.1 User Experiment

We conducted a user study with a simulated multi-class image classification scenario to understand how domain knowledge and order of observing system errors can affect user trust. In this section, we discuss the study design, goals, and participants in more detail.

5.1.1 Research Goals and Hypotheses

The primary motivation of this study was to understand how first impressions of an intelligent system can affect user trust, and whether and how domain expertise can help bypass the influences of these first impressions. We focused on systems with local outputs and explanations, i.e., systems that show one output at a time to their users – e.g., [151] – rather than intelligent systems at a global scope where users see a representation of how the model works on a high level – e.g., [69]. Considering these systems, our goal was to understand how user domain expertise affects the formation of first impressions, changes of trust over time, and estimation of system accuracy. To address this research question, we summarized the following set of hypotheses:

*This chapter is based on my published work: Nourani, Mahsan, King, Joanie, and Ragan, Eric, 2020, October. 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 (Vol. 8, No. 1, pp. 112-121).

- H1: Ordering bias only affects first impression formation in users with domain expertise, whereas novice users are more prone to automation bias due to having constantly high trust on the system.
- H2: Users with domain expertise and positive first impressions have a higher overall trust on the system compared to those with negative first impressions.
- H3: Users with domain expertise and positive first impressions will adjust their trust over time based on their observation of system errors, while those with negative first impressions will continue mistrusting the system, regardless of their observation of the system performance.

To test these hypotheses, we controlled two different orders of presenting system outputs: (1) Participants observed all the correct predictions in the beginning and all the mis-predictions at the end (i.e., a positive first impression) and (2) observations follow the opposite order (i.e., a negative first impression). Note that the only difference in these conditions was the order of presenting the output, while the accuracy and observed trials remained the same. Figure 5-2 shows an example of an image with its corresponding label and explanation.

5.1.2 Experimental Design

To test our hypotheses, we designed a user study where participants were asked to review a set of images from a multi-class simulated classification scenario. Based on our research questions and goals, we sought an image classification domain where background knowledge is not a requirement, while having it could help a lot in completing the task. We chose entomology—the study of insects and other terrestrial arthropods—as a domain where novice users can partially identify system errors, but domain experts are expected to excel at this task. The arthropod review task consisted of a set of 40 trials with a classification accuracy of 47.5%, and was designed to allow the participants to experience and use the classification system over time in order to measure their level of trust.

(a) Desert Tarantula

(b) Ecuadorian Lubber Grasshopper

(c) Rhopalid
(scentless plant bug)

(d) Metalmark Butterfly
(Calephelis sp.)

Figure 5-1. Four examples of raw images from the study dataset. (a) and (b) are examples of easy-to-detect images while (c) and (d) are examples of hard-to-detect images.

We defined two independent variables for the study: domain knowledge and order of observing correct outputs. First, domain knowledge refers to a user's level of familiarity with and knowledge of entomology, for which we defined two levels: novice and experienced. Second, we controlled the order of observing outputs in two different manners. For the correct first level, all the correctly-classified outputs were observed in the beginning while all the mis-classifications were shown afterwards. The reverse order was provided for the wrong first level.

The study followed a 2x2 between-subjects design, where each participant from novice and experienced group completed and observed the trials in one of the two defined orders. Since the subjects were exposed to the same set of trials, only with a different order, we incorporated a

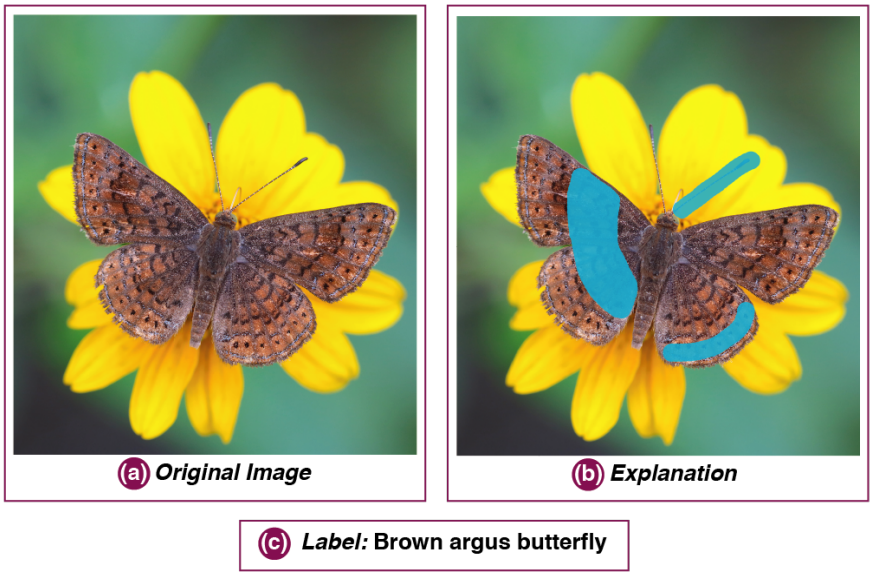


Figure 5-2. An example of what participants observed in the study. (a) the original image (which is a Calephelis sp.); (b) a blue saliency map to explain which regions of the image were used by an expert to determine the name and species of the arthropod; (c) the image classification label, which in this case is incorrect.

between-subjects design over a within-subjects design in order to avoid biases and learning effects.

5.1.3 Dataset

For the purposes of the study, the experiment's data used 40 high-quality macro images of different arthropods, photographed by an entomologist on the research team. Different regions in the world have bug species that are specific to each area and might not be found in other places. Since we were running the study in the US and our target entomology participants were mostly familiar with the arthropods in this country, all the selected arthropod images were from arthropod families that are found in the United States. Figure 5-1 shows examples of the raw images used in the study.

As our study was designed for both novice and experienced users, we designed the image set to contain a mix of both easy-to-detect and hard-to-detect arthropods in order to make the task fair for the novices as well. After selecting the images for the study, our expert entomologist generated a textual classification label for each image. The label contained the name of the arthropod and, in some cases, the family and species of the arthropod in brackets. For each image,

our expert also created high-fidelity explanations in the form of saliency maps on top of the image. These saliency maps were chosen as portions of the image that the expert would use herself to detect the bug in the image. However, to address our goals and hypotheses of the study, we selected the classification accuracy of the simulated system as 47.5%, i.e., 19 images included correct labels and 21 images included false labels.

5.1.4 Participants

We recruited a total of 116 participants for this study, with 48 females, 61 males, and 7 others (non-binary, non-listed, or unknown). For the purposes of this study, we distinguish two groups of participants: 1) people who had at least 1 year of university coursework in entomology (i.e., the experienced), and 2) people with little or no familiarity with entomology (i.e., the novices). These two groups were recruited separately. The novice participants were recruited from undergraduate and graduate level university students, most of whom studying in computing majors. The experienced subjects were university students and practitioners in entomology or related fields. Among these participants, 71.23% held or were pursing a graduate degree. To help verify participants were considered in the appropriate group for expertise level, participants self-reported their level of familiarity with entomology as well as their occupation or major. Familiarity was measured through a seven-point Likert scale from 1 to 7 for no knowledge to expert, respectively. Since this self-reported measure is subjective, novices might overestimate their knowledge, whereas experts might underestimate it [9, 46]. A two-way factorial ANOVA found significant differences between these groups, with $F(1,107) = 712.99$, $p < 0.001$, showing the domain-experienced group significantly rated their familiarity higher than novices.

5.1.5 Study Procedure and Measures

The user study was conducted online through a custom web application and took roughly 20 minutes. Participants were asked to complete the study in a single session using a preferred web browser on a desktop computer. The study was approved by the organization's Institutional Review Board (IRB). Figure 5-3 shows the overall procedure of the study.

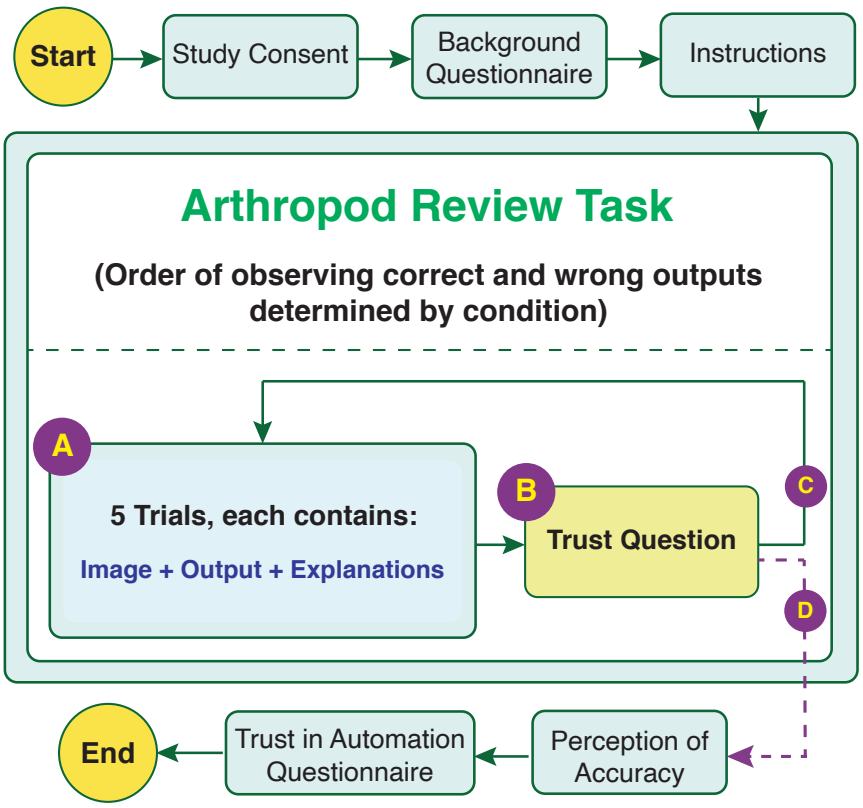


Figure 5-3. An overview of the study procedure. The Arthropod Review Task starts with (A), consisting of 5 trials, and continues to a trust question (B). By following path (C), participants iterate through (A) and (B) seven more times. After answering to the trust question for the 8^{th} time, subjects continue to the post-study questionnaire through path (D).

The participants filled out a background questionnaire about demographic information, education, occupation, and familiarity with machine learning and entomology. They were shown instructions about the study and the task. After reviewing the instructions, participants started the arthropod review task, where they reviewed 40 trials. Each trial consisted of: 1) an image of an arthropod, 2) a textual label for the classification of the image (which might also include the family and species of the bug), and 3) a feature explanation for the classification in form of a saliency map (as described in section 5.1.3).

For each trial, participants were required to rate their agreement with two statements on a 5-point Likert scale, as seen below. In order to answer these questions, participants were advised

to use their best judgement for identifying the arthropod in each trial, and in case the bug was unfamiliar, they were advised to refer to the provided explanations.

1. I believe the highlighted explanation is appropriate with regards to the system answer.
2. I am confident that the system answer is correct.

These questions were meant to focus user attention to the label and explanation of each image before moving on and to build an understanding of how the classifier works. After every five trials, participants were asked to report their level of trust on the system based on their observations up to that point. Trust was rated on a 7-point Likert scale from 1 (distrust) to 7 (trust).

After the arthropod review task, participants answered a questionnaire on trust in explainable AI [67] and estimated the system accuracy in percent. They also answered two free-response questions, asking them to explain how the system accuracy and their trust changed over time.

5.2 Results

We analyzed study results for the presented metrics based on the data collected from the arthropod review task and post-study questionnaire. For simplicity, we use NovCorrectFirst and NovWrongFirst condition names for novice participants, as well as ExpCorrectFirst and ExpWrongFirst condition names for experienced participants, with correct first and wrong first order trials, accordingly. For analysis, we used a two-way factorial ANOVA to test the main effects and Tukey HSD tests for posthoc pairwise comparisons.

5.2.1 Data Pre-processing

For quality verification, we removed the results from five participants due to data collection errors or evidence of lack of appropriate attention judged by their responses to the final open-ended questions. This left us with data from 111 participants, with NovCorrectFirst, NovWrongFirst, ExpCorrectFirst, and ExpWrongFirst having 28, 28, 28, and 27 data points, accordingly.

Table 5-1. Summary of results for average trust and change of trust. We used a two-way factorial ANOVA test for the main effect and a Tukey HSD test for pairwise comparison. For the post-hoc results, bold texts represent the conditions with (a) higher trust and (b) more change.

(a) Average Trust Rating	Main Effect
	Domain Knowledge: $F(1,107) = 39.07, p < 0.001$ *
	Order of Trials: $F(1,107) = 17.66, p < 0.001$ *
	Interaction Effect: $F(1,107) = 26.14, p < 0.001$ *
	Post-Hoc Test
	ExpCorrectFirst vs. ExpWrongFirst ($p < 0.001$) *
	NovWrongFirst vs. ExpWrongFirst ($p < 0.001$) *
	NovCorrectFirst vs. ExpWrongFirst ($p < 0.001$) *
(b) Change of Trust Rating	Main Effect
	Domain Knowledge: $F(1,107) = 17.58, p < 0.001$ *
	Order of Trials: $F(1,107) = 39.02, p < 0.001$ *
	Interaction Effect: $F(1,107) = 39.02, p < 0.001$ *
	Post-Hoc Test
	ExpCorrectFirst vs. ExpWrongFirst ($p < 0.001$) *
	ExpCorrectFirst vs. NovCorrectFirst ($p < 0.001$) *
	ExpCorrectFirst vs. NovWrongFirst ($p < 0.001$) *

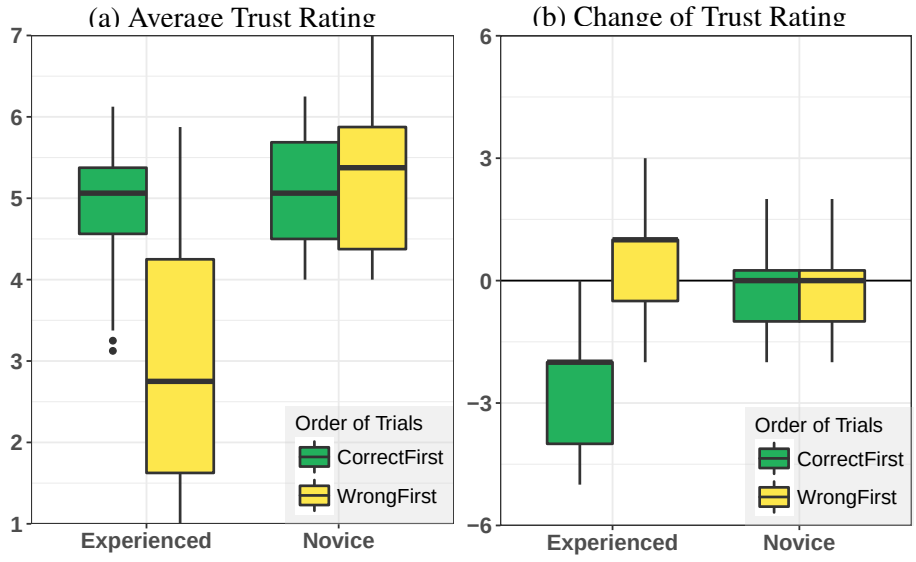


Figure 5-4. Results from self-reported trust in the arthropod review task. (a) average of the 8 self-reported trusts, and (b) difference between the first and the last reported trust. Negative values indicate trust over time declining and positive points show increasing.

(a) Changes of trust over time for novice participants. (b) Changes of trust over time for those experienced.

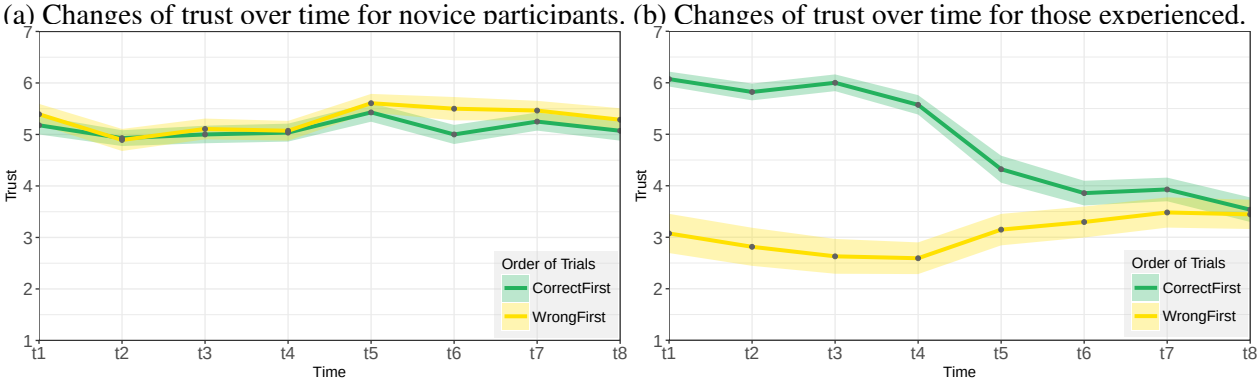


Figure 5-5. Average of participants' self-reported trust after every 5 trials in the main task. The y-axis indicates level of trust, where 1 indicates distrust and 7 indicates total trust. The ribbon around each line shows the standard error of the mean. The x-axis shows the time-step when trust question is asked (every 5 trials).

5.2.2 Average Self-reported Trust

We tested the effects of background domain knowledge and order of observing system correctness on user self-reported trust. Participants rated their trust in the system eight times during the arthropod review task. To address our first two hypotheses (H1 and H2), we calculated the average trust for each participant and compared them to find differences across the conditions. Table 5-1a and Figure 5-4a show the summary of the results and distribution of this data. The findings provide strong evidence that the average trust was significantly affected by domain knowledge and order of observing correct outputs. However, a significant interaction effect indicates these factors are interdependent. The pairwise comparison reveals that the order of observing correct system outputs is only significantly affecting trust for knowledgeable participants, which aligns with H1. Moreover, users with domain expertise and positive first impressions report significantly higher trust compared to those with negative first impressions (H2).

5.2.3 Changes in Trust over Time

To test our third hypothesis (H3), participants rated their trust throughout the study so we could track how it evolves over time. For each condition, we calculated the average trust of all participants per time-step, resulting in one trust value at each of the eight time-steps. Figures 5-5a

and 5-5b show line charts of changes over time for the novices and the experienced groups, respectively.

In order to statistically compare the magnitude and direction of change in trust over time across the conditions, we calculated the difference of initial trust and final trust. With this measure, negative values indicate declining trust and positive values indicate increasing trust. Table 5-1b and Figure 5-4b show the results for this comparison. Experienced participants in the correct first condition, had a significantly larger change-of-trust than those in the wrong first condition. The direction of this change was negative, indicating a loss of trust. To understand these results further, refer to Figure 5-5b. Participants from the correct first condition start with higher trust, which decreases over time; this is expected as the system accuracy lowers with time. In contrast, those from the wrong first condition start off with lower trust due to their negative first impressions, and the magnitude of their change-in-trust is significantly less than their counterparts', slightly going up while remaining relatively low. We did not observe any significant differences for novice participants.

We further reviewed the open-ended responses from the experienced participants to analyze how their trust in the system changed over time to understand the trends and themes based on first impressions. A summary of the main observations is presented in Table 5-2. We expected experienced participants to have a proper understanding of how and when the system accuracy changed. According to Yu et al. [196], this understanding should reflect in their change of trust. Thus, we counted the number of participants whose comments indicated their assessments

Table 5-2. Most common qualitative themes identified for how trust changed over time for participants from the experienced condition. The themes were retrieved from an open-ended question at the end of the study where participants were asked how their overall trust changed over time.

Ordering Condition	Common Themes		
	Strong Distrust	Trust Decreased	Trust Increased
CorrectFirst	0	21	0
WrongFirst	11	5	8

matched our expectation; that is, a decrease in trust for correct first and an increase in trust for wrong first participants. In total, 21 out of 28 experienced participants in the correct first condition stated that their trust was high in the beginning but lowered over time. From those in the wrong first condition, only 8 out of 27 indicated a slight increase in their trust. However, different themes were observed among these responses. For example, one participant who detected a slight increase in accuracy noted:

> "Strangely, the system accuracy got much better towards the end, but by then I distrusted the system's [outputs, as] the misidentifications it made were too scandalous."

Similarly, 11 out of 27 participants stated that regardless of the change in their trust, they did not trust the system at all. For instance, one participant noted:

> "I didn't trust the system in the beginning and as the test continued, I only became surer that my distrust was the appropriate response."

Moreover, 5 participants mentioned that their trust decreased over time, which is unexpected since the system grew more accurate towards the end.

These observations—backed by the statistical analysis for changes of trust—support our hypotheses (H1 and H3) that first impressions of a system with local scope only matter when the user has background knowledge of the domain. Positive first impressions provide a chance for users to build trust and not give up on the system when it makes mistakes, while negative first impressions can cause an overall distrust in the system.

5.2.4 Post-Study Questionnaire

After the arthropod review task, participants estimated the accuracy of the classification system. It is important to keep in mind that all participants (regardless of the condition) observed the same simulated classification results with the same controlled accuracy across observed instances. The only difference was the order of observing the correct classifications. We assessed the error of each participant's perceived system accuracy by calculating the difference between

Table 5-3. Summary of results for error of perceived accuracy and trust questionnaire. We used a two-way factorial ANOVA to test the main effect and a Tukey HSD test for pairwise comparison. (a) for the post-hoc results, the condition with bold text shows higher overestimation of accuracy.

(a) Error of Perceived Accuracy	Main Effect
	Domain Knowledge: $F(1,107) = 76.88, p < 0.001$ *
	Order of Trials: $F(1,107) = 14.48, p < 0.001$ *
	Interaction Effect: $F(1,107) = 13.13, p < 0.001$ *
	Post-Hoc Test
	ExpCorrectFirst vs. ExpWrongFirst ($p < 0.001$) *
(b) Trust in XAI questionnaire	Main Effect
	Domain Knowledge: $F(1,107) = 87.84, p < 0.001$ *
	Order of Trials: $F(1,107) = 3.75, p = 0.055$ (NS)
	Interaction Effect: $F(1,107) = 2.10, p = 0.149$ (NS)

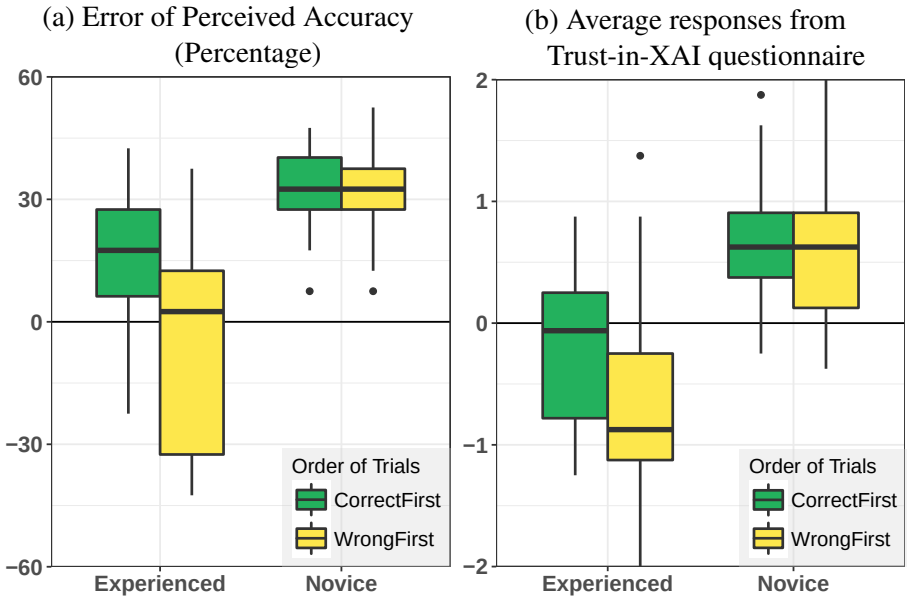


Figure 5-6. (a) Error of user-estimation of system (difference between user estimations and the actual observed system accuracy). Positive numbers represent overestimation and negative numbers represent underestimation of system accuracy. (b) Average of user responses to trust questionnaire results. Higher numbers indicate higher level of trust.

the estimated accuracy and actual observed system accuracy. In addition, participants answered a set of questions about trust in automation and explainable AI systems [67] through a 5-point Likert scale, where higher values indicate higher trust in the intelligent system. We calculated the average of all questionnaire responses into one single score per participant and analyzed them to

test for differences among conditions. Table 5-3 and Figure 5-6 show a summary and distribution of the results.

The results show that novices significantly overestimated the system while experienced participants may underestimate or overestimate the accuracy based on the order of observing the correct trials. The results from the trust questionnaire demonstrate that novice participants tended to trust the system significantly more than those with domain experience, which aligns with our first hypothesis (H1).

5.3 Discussions

Overall, our results demonstrate the importance of first impression formation with users with domain expertise and how it affects their trust. In this section, we discuss the results more generally, their importance to system designers, and possible future directions.

5.3.1 Interpreting the Results

Our first goal in this study was to understand whether first impression formation is influenced by domain expertise. Different implicit and explicit trust measures clearly indicate that novice users over-trusted the intelligent system although its overall accuracy was low. However, since domain experienced users tend to be more skeptical due to their knowledge, their overall trust depended on their early observations of the system performance. Domain experts' perception of system accuracy also varied by these first impressions, while novices always overestimated the accuracy. Previous research by Papenmeier et al. [138] demonstrated users cannot be tricked into trusting a low accuracy intelligent system with high fidelity explanations. Our work builds on their findings and shows that given a domain-specific task where domain knowledge might be beneficial, novice users are prone to trusting such systems as they do not have enough knowledge to detect errors.

When comparing changes of user trust over time, domain experienced users showed different trends of trusting the system depending on the order of observed errors. Experienced users with positive first impressions had a significantly higher magnitude of change in their trust compared to those with a negative first impression (Figure 5-4b). This observation indicates that

with positive first impressions, expert users tend to adjust their trust, whereas those with negative first impressions start with lower trust which stays low throughout the usage.

As previous work in automation bias and psychology shows, it is easier to lose trust than to reestablish trust [65]. Starting with a system with good performance, experts are likely to form trust initially and to adjust their trust over time—also known as swift trust [65]. However, our results show that expert users with negative first impressions lose trust in the system and stated that they are not willing to use it in the future (see Table 5-2). This has implications on real-world systems, as users might not continue using a system they do not trust [36].

5.3.2 Implications for Intelligent System Designers

This study presents important findings for intelligent system designers who are involved with designing domain-specific systems with various target users. Designing one system for all users is indeed tricky and requires certain considerations with user domain knowledge. Failing to account for such considerations can cause various problems such as under-reliance and over-reliance. Our results indicate that while experienced domain users tend to be more skeptical of the system, novices might not be able to catch system problems and suffer from automation bias.

System designers can incorporate techniques to help fill the knowledge gap for novice users and utilize techniques to guide domain experts into observing system performance, e.g., by using a more detailed explanation interface to provide more information. Rather than allowing users to use a system with zero understanding or a poor mental model of how the system works, designers can incorporate introductory sessions to alert users about system strengths and shortcomings so that users can decide whether and when they should trust the system's outputs. Alternatively, designers can provide a high-level overview of the model with key information (e.g., system accuracy and known weaknesses) that can influence impression formation. Additionally, one approach for consideration could be showcasing examples of both correct and incorrect predictions or outcomes in the beginning of usage to help experts develop first impressions, covering the variability of system capabilities to reduce the risk of encountering an

unrepresentative sample by chance. Further research would need to explore the implications of such an approach. The showcasing method could also consider attempts to more strongly encourage users to review explanations for both correct and incorrect examples. Our results indicate that users tend to check the explanation for further information when they encounter errors, but not necessarily when they perceive the system to be correct. For novice users, however, errors need to be shown and explained to circumvent automation bias.

5.3.3 Limitations and Future Work Opportunities

This study contributed novel findings regarding how domain expertise can affect first impression formations and user trust. Our results show that novice users trusted the system regardless of the order of observing system errors and the overall low system accuracy. One possible explanation for this observation is novice users' inability to identify errors. A user's assigned level of expertise might vary with the domain and task at hand. While some systems consider novices as students or inexperienced domain knowledgeable users, we expected novices to have little or no domain knowledge at all. Thus, our study results from domain experts might also be observed for novices if they have the ability to judge system errors correctly.

Measuring trust is tricky, and any chosen evaluation methodology will have limitations. Directly asking subjects to rate their trust might bias them about the purposes of the study and hence, affect their response. Another issue relates to the use of Likert-scales for self-reported trust estimations and whether participants are able to differentiate the values in their response. Although 7-point Likert-scales are generally reliable for ordinal self-reported measures [133], we cannot control nor can we precisely know how each participant differentiates each point (e.g., value 5 from 6). In our study, we looked to qualitative data and free-response questions to help address this limitation and provide a better understanding for the collected explicit quantitative trust metrics. While the qualitative and quantitative results align, our study is limited in its ability to dissect specific characteristics of the observed trust and mistrust due to the open-ended nature of the free response data collection.

To maintain experimental control for our study, we selected a task where there is a definition for correct predictions. In other words, for these tasks, there is a concrete distinction between when the system makes a mistake and when not. To achieve this, the study is based on a multi-class image classifier with visual explanations. While such tasks are quite common in different domains and our results can be generalized for such intelligent systems, future work is required to verify if these findings hold for more exploratory and complex tasks. Specifically, for these tasks, errors might be challenging for users to detect, while they can also be difficult or impossible to define. For example, missing information or missing values can cause a model to predict different outcomes, all of which may be correct based on different hypothetical values for the missing information. As another example, recommendation systems strive to help users by making suggestions, but how these suggestions fit a user's needs is not easy to assess, and "false suggestions" are often impossible or difficult to define. Future research can extend the study of our research questions to such alternative analytical systems and tasks.

Finally, our findings show strong impression-formations for expert users based on instance-level observations of system performance. The presented study used a system with simple representations of model outputs. Future research of higher-level representations or visualizations can investigate how our findings generalize to contexts that allow deeper expert analysis of the model as-a-whole.

CHAPTER 6
THE EFFECTS OF PEOPLE'S DIFFERENCES AND AI ATTITUDES ON USAGE BEHAVIOURS

In Chapter 4, we examined the impact of transient past experiences, specifically anchoring bias, on people's usage behaviors and perceptions of an AI system. Through our conceptual model and literature review analysis, we demonstrated how these temporary experiences can significantly influence individuals' interactions with AI systems, highlighting the importance of considering the effects of transient factors in system design and user experience. In Chapter 5, we extended our exploration to focus on the role of long-term, stable past experiences, particularly domain expertise, in shaping people's usage behaviors. By studying the interaction between stable and transient elements, we gained valuable insights into how these different types of experiences influence individuals in varying ways. This research further emphasized the complexities and challenges involved in designing AI systems that cater to users with diverse backgrounds and experiences.

While these two studies present important implications of accounting for long-term and short-term effects such as domain expertise and anchoring bias with intelligent systems, with real users and in the wild, there are often countless experiences and personal differences that can result in myriad of behaviours. To be able to solve the problems caused by these various behaviours, an critical first step is to understand what these behaviours are and how they are connected. AI systems are not and should not be approached as a one-size-fits-all problem and as such, solutions to these challenges need to be mindful of people's differences in behaviours. Personalization is one solution to mitigating potential effects of long-term and short-term past experiences and differences; i.e., each user can see the version of the system that is adapted to specifically match them and their behaviours specifically introduced by them. Given that it is almost impossible (or unrealistic) to personalize systems based on every potential individual that might use the system, user profiling could be a promising approach to categorizing a new user into the best possible personalized experience possible.

The goal of this experiment is to examine how people's anchoring bias (transient past influence) interacts with various factors from the long-term past experiences and personal

differences simultaneously. This is a direct extension to the study from Chapter 5. In the previous study, we only examined one stable past factor that was easier to categorize (novice vs. experienced in the domain). However, in reality, various different types of past experiences contribute towards current thinking and we would expect many possible combined categories for different groups of users. The ultimate goal in this chapter is to investigate the interplay between transient and multiple stable factors (presented with one category with multiple levels). This categorization may serve as a form of user profiling, which would allow system designers to predict usage behaviours prior encountering with an AI system to mitigate biases and preconceptions toward the system. Through a user study, we explored the possibility of using instruments to collect some of these differences to be able to predict people's usage and anchoring behaviours as a first step to user profiling as a solution to anchoring bias. In this chapter, we describe the user experiment designed to address these questions, as well as the results and implications for intelligent systems.

6.1 User Experiment

6.1.1 Research Goals and Hypotheses

As seen in the conceptual model from Chapter 3, there are many known and unknown categories of long-term past experiences that can affect human-AI partnership. Some of the long-term effects might originate from people's prior exposures to AI/ML technology, whether from their firsthand experiences with the AI technology or through implicit learning* and anecdotal knowledge†. Here, our focus is not on how they obtained these presumptions about AI, but rather how these presumptions affect usage behaviours once they manifest into measurable feelings, attitudes, or perceptions.

For the purposes of this study, we select three types of long-term past experiences, including implicit knowledge of AI (non-expert knowledge), individual characteristics (personality traits), and attitudes and "feelings" toward AI (e.g., anxiety, fear, acceptance, and positive/negative attitudes) along with anchoring bias as a short-term effect. By considering

*Implicit learning refers to knowledge acquired without conscious awareness or deliberate learning attempts [149].
†Anecdotal knowledge refers to knowledge gained through personal experiences or stories from others.

multiple factors, we aim to study how people's differences in terms of individual characteristics and prior experiences affect their usage behaviours, particularly when they use a system for the first time and might form first impressions through short-term usage. In particular, this experiment is motivated by the following research questions:

- RQ1: How do people's preexisting perceptions of AI affect their usage behaviours with an intelligent system?
- RQ2: Which type of long-term past experiences has a stronger influence on people's perceptions (implicit knowledge vs. individual characteristics vs. attitudes and feelings towards AI)?
- RQ3: Can we categorize people into distinct groups by profiling these measured long-term past experiences? If so, how would people's first impressions and usage behaviours change based on their assigned categories?

The first two research questions aim to study multiple long-term past factors simultaneously and allows us to understand them in the light of the others. The last research question is a first step towards user profiling; once we measure the people's differences based on the long-term past factors, we can cluster people into their most relevant category/profile (i.e., unlike in Chapter 5; it will not be a more clear classification where we know people's classes, but rather by clustering people into groups where members are most similar). Using these clusters, we will study whether people from each cluster show similar usage behaviours towards the AI system, and what is the interplay between their assigned clusters and first impressions with the system. Our goal is to determine whether such relationships exist so that in the future we may predict which behaviours a person might show towards the AI system simply by measuring those past-experiences.

6.1.2 The Explainable System

This experiment used the same XAI system built and introduced in Chapter 4, Section 4.1 as the study platform. 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.

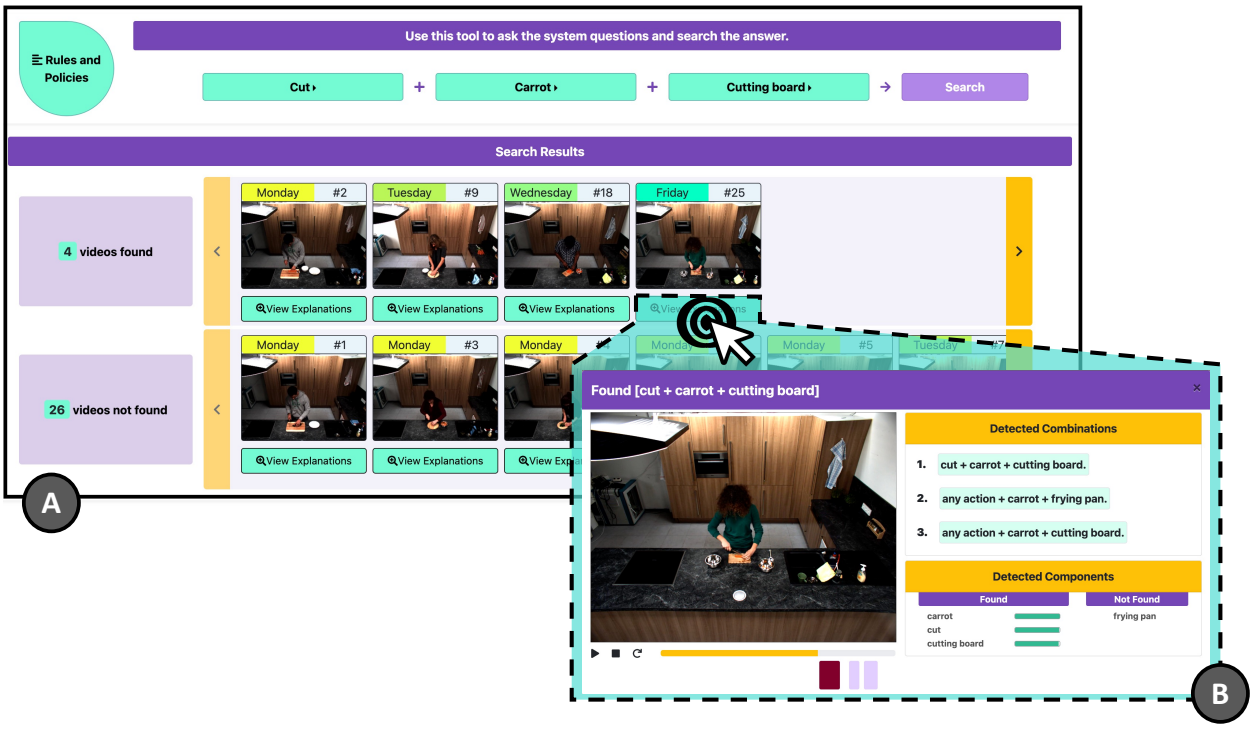


Figure 6-1. The XAI application for the study was the same as presented in Chapter 4. (A) shows the query-building tool, which would allow users to search activities within videos. By clicking in the top left corner, a panel opens from the left side of the screen that includes a list of kitchen policies. The search results are sorted based on whether the video was matched with the query or not. Each video is represented with a thumbnail preview, a unique number, and their corresponding weekday. By clicking on View Explanations, a modal would pop-up that includes a video player and the explanations for why the video was determined as a match or non-matched (B).

Activities are defined as actions done to/with objects in/on a location. The XAI system is designed to allow users to query the videos by making an {action, object, location} combination, where at least one of the three components is a real item from the vocabulary. For instance, both {cut, carrot, cutting board} and {cut, any object, any location} are valid queries while {any action, any object, any location} is not.

By providing three dropdown-style buttons, the interface would allow users to build all possible combinations of components to make a query and search for all the videos in the system ($N = 30$) to determine in which of them the activity is detected. The interface would present the hits and misses in two categories with a horizontal list, ordered based on the video number of day of the week that video occurred. Users can then decided whether they want to inspect a video for

more detail, in which case a pop-up modal would provide a video player, with post-hoc explanations for why the video categorized as a hit or miss for that query. These explanations include the segments (i.e., a duration of frames) of the video that was used to determine the answer for that query, as well as the top 3 detected activities and the individual activity components found within the selected segment. These segments were overlayed through square-shaped buttons in a temporal design under the video progress bar, and clicking each of these segments would reveal the unique information for that segment (i.e., top 3 activities and individual activity components). Figure 6-1 provides an overview of the XAI system.

6.1.3 Experimental Design

To address our research questions, we conducted a user study using the explainable system discussed in Section 6.1.2. This study was similar in many ways to the one presented in Chapter 4, with some alterations introduced in the study variables and procedure. In this section, we describe the design of the current study in more detail while providing comparisons to the study from Chapter 4.

6.1.3.1 Study Task

To study the effects of users' different past experiences on their usages of the AI system, we sought for a task that involves elements of human-AI collaboration with a real AI system that is complex and open-ended enough to allow for measuring users' task-performance and their mental models. As such, we utilized the XAI system with the same policy-verification task as the previous study from Chapter 4, where participants were tasked to verify whether the kitchen employees are following a set of rules and policies imposed by the restaurant. We designed hypothetical policies requiring people to build and test multiple queries in order to confirm each one. This would be beneficial in two ways: (1) they would be able to observe enough model performance to be able to build a mental model of the system, and (2) they would allow us to control when they would be exposed to the model weaknesses and strengths to some extent. To achieve the latter, we designed four policies that would expose participants to the model weaknesses and four policies that would expose them to the model strengths.

We presented these policies as a list, hidden overlay side panel in the XAI interface that could be accessed by clicking on a button "Rules and Policies" that was located on the top-left corner of the screen. Each policy was created as a true or false statement that needed to be assessed with a Yes or No response to whether they believe the policy is being followed or not.

6.1.3.2 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 allows influences people's first impressions (short-term past experience) of the XAI algorithm. Regardless of their 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, while the other ordering moved these "problem" policies to the bottom of the list.

This approach is similar to the user study in Chapter 4. However, unlike that study, the new study did not vary presence or absence of explanations; all participants had access to full explanations for the policy-review study task in this study.

6.1.3.3 User Study Measures

We collected participants' responses to the policies as well as all the interaction logs in the main study (with examples including each query they search, each video they view, when they answer a policy, and time for each interaction). Furthermore, similar to the study in Chapter 4, the participants completed a post-study task were they rated their perception of accuracy for nine objects from the model's activity components on a scale of 0–100%, as well as a binary rating confidence of their confidence in their answer (High vs. Low). This was designed to measure their understanding of the XAI's weaknesses and strengths after a period of usage. This question set included 4 components reflect model strengths (high detection accuracy) and 5 associated with 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.

6.1.3.4 Questionnaires

With the study aimed to capture the long-term influences of AI that could lead to differences in usage behaviours with real AI systems, we opted to utilize questionnaires from the existing body of work to capture a sample of those influences. Questionnaires are common data collecting approaches for measuring people's differences (in this case, differences based on pre-existing expectations and attitudes toward AI technology) and grouping them accordingly, making them a suitable data collection means for our purpose. We selected six different instruments that would allow us to capture people's individual differences and experiences towards AI. Here, we describe each questionnaire and why it was chosen. Table 6-1 provides an overview of these questionnaires. The questionnaires were also included in Appendix A for transparency.

(A) AI Literacy Questionnaire: This questionnaire by Laupichler et al. [95] allows for measuring people's level of AI literacy, which refers to their ability to critically evaluate the competencies of the AI technology to be able to use them in their day-to-day lives and tasks [104]. This instrument consists of 38 subjective statements starting with "I can...", for each of which the participants would rate their level of agreement through an 11-point Likert-scale. The questionnaire includes important questions that would allow for measuring people's level of knowledge in AI/ML concepts, from both technical and societal perspectives, such as their ability to differentiate between the instances of AI shown in games and movies and the current usages of this technology.

(B) AI Anxiety Questionnaire: This questionnaire by Li et al. [96] provides a set of statements to capture people's anxiety towards AI, characterized by multiple sub-categories (e.g., people's anxiety towards ethical implications, bias behaviour, and privacy and violation) . The questionnaire includes 20 factual statements with 7-point Likert-scale agreement levels.

(C) AI Attitudes Questionnaire: Proposed by Schepman et al. [158], this questionnaire measures people's positive and negative attitudes towards AI systems. Similar to the other questionnaires, it includes statements ($N = 20$) which participants have to rate their agreements with, on a 5-point Likert-scale. This questionnaire results in two seperate measures that would reflect both positive and negative attitudes.

(D) Attitudes Towards AI (ATAI) Questionnaire: Similar to (C), ATAI captures people's attitudes towards AI; however, instead of general positive and negative categories, it measures people's fear and acceptance of AI. This measure was developed by Sindermann et al. [163], with only five statements rated on an 11-point Likert-scale.

(E) The Big Give Personality Traits Assessment Questionnaire: While the other four questionnaires focus on long-term experiences that were resulted by people's exposure to AI/ML systems in the past, other long-term individual differences affect their view of the world. Specifically, people's personality traits might influence how they interact with and use AI systems. The Big Five questionnaire, originally proposed by McCrae et al. [109], is a common approach to measure people's traits in 5 different categories: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to experience. In other words, it results in 5 separate measures per participant. While the original questionnaire includes 44 statements to be rated, we used a 15-statement subset that is shown to be as robust across all survey methods except for phone interviews [93]. Given the inclusion of other questionnaires in the user study, we opted for this choice to minimize survey fatigue. The general structure of The Big Five instrument consists of descriptive phrases that complete the following sentence "I see myself as someone who... "; e.g., "I see myself as someone who worries a lot". Participants were asked to rate each statement on a 7-point Likert-scale (a more nuanced alternative to the 5-point scale used by the original Big Five survey).

6.1.3.5 Procedure

We conducted the study online and through a web-based interface. The study was approved by the Institute's Review Board (IRB). Participants were asked to complete the study in a single

Table 6-1. Overview of the questionnaires used for the user study.

Questionnaire	Authors (abbreviated)	Scale	Items #	Measure #
(A) AI Literacy	Laupichler [95]	11-point Likert-scale	38	1
Description: Designed to capture people's ability to critically evaluate AI technology competencies to use them in daily lives.				
(B) AI Anxiety	Li [96]	7-point Likert-scale	20	1
Description: Designed to capture people's anxiety toward AI technology, with questions covering multiple sub-categories, such as anxiety toward privacy violations.				
(C) AI Attitudes	Schepman [158]	5-point Likert-scale	20	2
Description: Designed to capture people's positive and negative attitudes toward AI systems.				
(D) Attitudes Toward AI (ATAI)	Sindermann [163]	11-point Likert-scale	5	2
Description: Designed to capture people's broad fear and acceptance of AI technology.				
(E) The Big Five Personality Traits	Lang [93]	7-point Likert-scale	15	5
Description: Designed to measure people's personality traits in five different categories of extraversion, agreeableness, conscientiousness, neuroticism, and openness.				

session on a computer device. The study took approximately 30–40 minutes to complete. Figure 6-2 provides an overview of the study procedure.

After consenting to participate in the study, participants completed the questionnaires introduced in Section 6.1.3.4, that were chosen to measure their pre-existing perceptions and attitudes towards AI as well as their personality traits. We designed the interface to include each questionnaire in a separate page, with a progress bar showing the percentage of completion, to give the participants an estimate of their progress. All the participants observed the questionnaires in the same order. We selected this unique ordering based on the following rationale: (1) we aimed to minimize the effects of survey fatigue by limiting the number of questions per page as they progressed through the last questionnaire; (2) we aimed to minimize 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) we aimed to minimize potential biases caused by seeing negative or positive statements as the first questionnaire. For this reason, we chose the following order:

1. AI Literacy: The longest (38 items) and the most neutral questionnaire;
2. AI Anxiety: Second longest (20 items) and most negative questionnaire;
3. AI Attitudes: Second longest (20 items), but includes both positive and negative statements;
4. Attitudes Towards AI (ATAI): The shortest (5 items), but includes both positive and negative statements;
5. The Big Five questionnaire: While not the shortest (15 items), it includes questions irrelevant to AI and can be seen as a distraction to help prevent availability bias.

Upon completing the questionnaires, participants were directed to watch a short video tutorial about the video-query tool and the policy verification task they had to complete. Similar to the study in Chapter 4, the video tutorial was based on a hypothetical restaurant where the owner needed help to verify whether the policies in the kitchen were being followed by the kitchen staff, by using the system with videos from the prior week. The video provided necessary details about the task and an example of how they can verify policies. This example policy was not included in the original policy set used for the main user study. A summary of the task and key information from the tutorial was also made available on the main interface to be used as a refresher during the task.

Once they went through the tutorial, participants were assigned to one of the two study conditions and completed the policy review task, where they were asked to verify 9 kitchen policies listed on the left side panel of the interface. Note that the order of these policies were altered based on the assigned condition (as described in Section 6.1.3.2), while the 5th policy was the same policy and was designed to serve as attention check.

After completing the policy review task, participants responded to model understanding questions (i.e., mental model questions as described in Section 6.1.3.3), followed by post-study questionnaire and background questionnaire. To prevent the study from getting too long, 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, we included the ATAI questionnaire ($N = 5$) once more before the background questionnaire.

6.1.3.6 Attention Check

To ensure the quality of the crowdsourced data, we included multiple attention check factors in the pre-task questionnaire and the main task. For questionnaires with $N \geq 20$, we added one attention check statement per roughly every 20 statements. This resulted in 4 total attention checks for the pre-task questionnaires. We chose the statements to reflect common knowledge facts. However, given that the responses were Likert-scales to measure agreement level, we chose the statements in a way that makes answering to them a bit more challenging. For instance, "The Statue of Liberty is located in New York City. Please select (0) Strongly Disagree". Here, a participant had to respond Strongly Disagree despite the statement being true. Furthermore, the 5th policy in the policy-verification task was used as an attention check, where we chose a statement that is not always followed by all kitchen staff, but seemed logical enough to be followed by them from the perspective of someone who did not use the tool: "Employees wash their hands immediately after entering the kitchen".

6.1.3.7 Participants

We recruited a total of 140 participants to complete the study. Participants were recruited through multiple channels, including advertisement on our university's crowdsourcing platform and through word-of-mouth referrals. They consisted of 42 females, 93 males, and 3 non binary (while 2 preferred not to disclose their gender), with a majority ($> 96\%$) reporting their age between 18 to 34 years old. The participants were randomly assigned to a condition. There were 67 in the weak first condition and 74 in the strong first condition.

To get a better understanding of our participant population, we asked them to self-report their level of knowledge in AI and computer science (ordinal scale between 1 to 5, adapted from Ehsan et al. [48]). For level of knowledge in computer science ($M = 2.97, Sd = 0.73$), a one-way ANOVA test did not reveal a significant difference between groups:

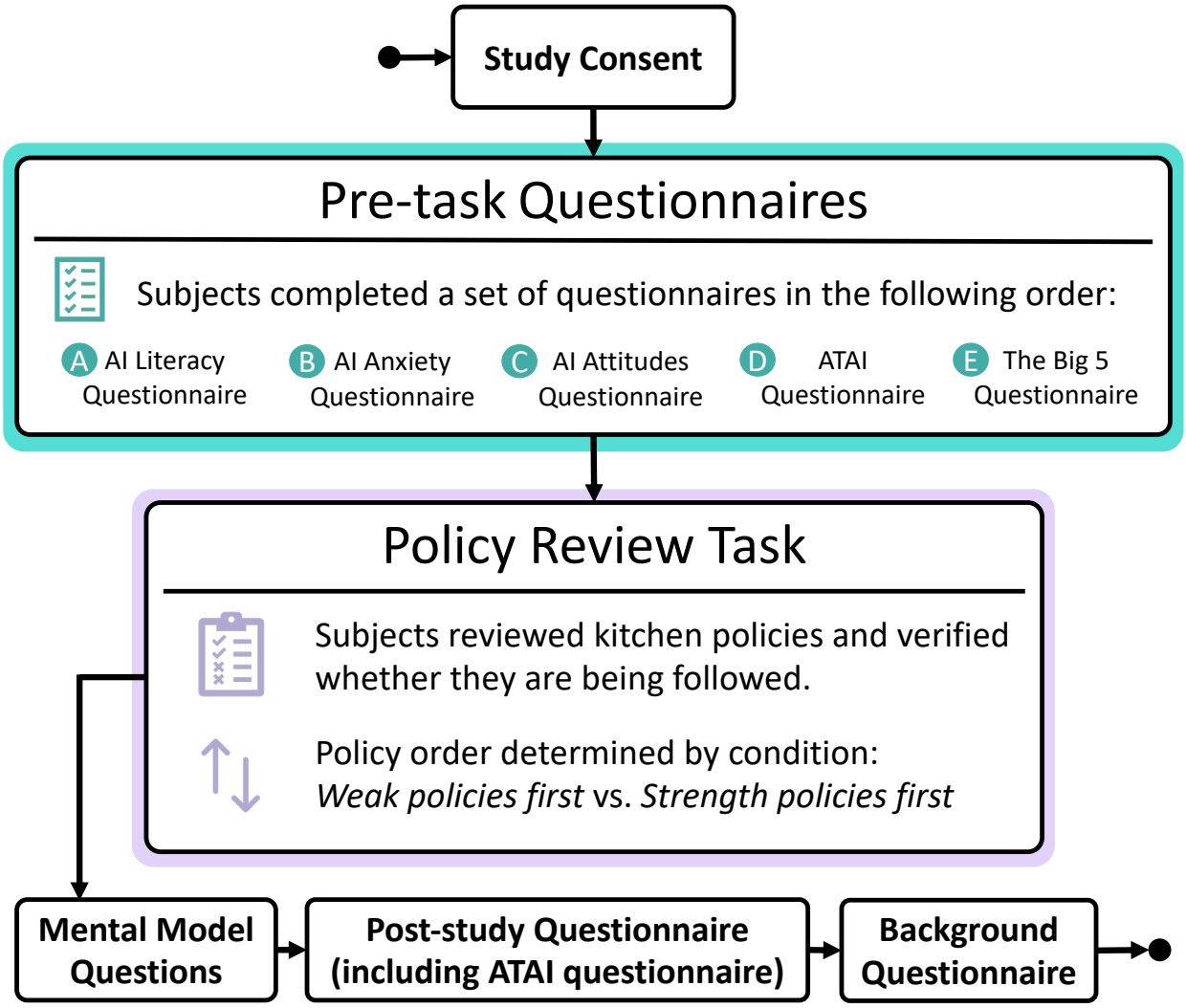


Figure 6-2. An overview of the study procedure. The pre-task questionnaire includes (A) AI literacy questionnaire [95], (B) AI anxiety questionnaire [96], (C) AI attitudes questionnaire [158], (D) Attitudes Towards AI (ATAI) quesionnaire [163], and (E) a shorter version of the Big Five personality traits questionnaire [109]. Note that all the participants completed these questionnaires in the same order and prior to the policy review task, with the exception of ATAI questionnaire, which was repeated in the post-study questionnaire to measure whether using the XAI system could sway people's attitudes toward AI.

$F(1,138) = 3.59, p = 0.06 (NS)$. Similarly for AI knowledge ($M = 1.7, Sd = 0.99$), no significant differences were found across the conditions: $F(1,138) = 0.25, p = 0.62 (NS)$[‡]. Additionally, 65.7% of the participants stated they have not taken any AI/ML related courses. The university

[‡]These statistics were computed before performing data cleaning and quality checks, which led to the removal of some data points. We conducted a subsequent analysis on the finalized set of participants, and the results were consistent with our initial findings.

students were compensated by receiving up to 1 extra credit towards a participating course, whereas those recruited through word-of-mouth voluntarily participated without receiving any form of compensation.

6.2 Results

6.2.1 Data Preparation: Cleaning, Quality Assurance, and Measure Calculation

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 data points that (a) did not meet the required formatting or (b) failed to pass explicit and implicit attention checks. To test the latter, we considered task length (total time spent on finishing the main task), number of queries built, and the correctness of the responses to all the four attention check questions. We set minimum values for these measures, and only those who passed all the criteria were included in the final analysis. The exclusion criteria was as follows:

- if they spent less than 5 minutes on the main task, OR
- if they build less than 9 queries (i.e., they built at least one query per policy), OR
- if they answer any of the attention checks incorrectly.

Upon testing all of this criteria, we noticed a pattern. All the participants who failed the attention checks also failed the first two criteria. This observation verify the effectiveness of our attention check questions in eliminating people who were not paying attention. The only exception were two participants who passed all the criteria, except one attention check question (out of 4) from the pre-study questionnaires. This was one of the two attention checks from the first questionnaires where the participants were asked to respond Strongly Disagree despite the statement being correct. The two participants were responded Strongly Agree to this question. Given that our implicit attention check metrics (such as study time, number of queries searched, and number of videos watched) and the rest of the explicit attention check questions were all reflecting their attentiveness to the task, we decided to make an exception to the previous rules and include these participants in the final analysis. This resulted in excluding 17 and 19 total

results from the weak first and the strong first conditions, respectively, resulting in data from 105 participants used for the statistical analysis (weak first: 51, strong first: 54). Note that the discrepancy in the number of collected and omitted data points was purely a result of random condition assignment and the varying levels of participant engagement and attentiveness across the groups.

Considering the study's online format, it is possible that participants had idle periods where their attention is not actively directed towards the study. Although occasional lapses in attention occur naturally, both in the online study and the real-world settings, we aimed to minimize their impact and ensure consistent data analysis within the group. To achieve this, we defined time per trial as the duration between answering two policies. This approach allows us to calculate and compare the average trial time per participant, rather than relying solely on the overall time difference between starting and completing the main task. By normalizing the data in this manner, we can account for variations in attention and maintain a more accurate analysis. We then filtered out potential outliers by removing trial times that fell outside the $\pm 1.5 \times IQR$ range for each condition. Any outliers were replaced with the average time for that trial and condition, excluding the outlier itself. Subsequently, we calculated the average trial time across the nine trials per participant.

To understand whether and how the XAI's predictions and explanations affected participants' task performance, we defined and calculated an approximate user agreement measure based on how they respond to the policies. For policies that expose model weaknesses, following the model's answers to the searched queries could lead to mistakes when verifying the policy. On the contrary, for policies that expose model strengths, following the model's advice would likely lead to correct responses to the policy. Thus, we measure user agreement per participant by calculating the percentage of times they either correctly verified a strength policy or wrongly assessed a weakness policy. For this measure, we excluded the 5th policy, which was used as the attention check. We acknowledge that this is an approximation and participants' errors or accuracy might be caused by an assortment of factors due to the exploratory nature of the task.

Nonetheless, this approximation provides a reasonable measure of the alignment between users' assessment and the advice offered by the model.

Similar to Chapter 4, we divided participants' mental models into two separate factors to measure their perceptions of model weaknesses and strengths separately. Among 9 estimation tasks post-study, 5 were selected to reflect model strengths and 4 were selected to reflect model weaknesses. we converted these estimations into error of estimation by subtracting the ground truth for that reported accuracy, so that positive values show overestimation and negative values, underestimation. The averaged percentage of error per each category was then calculated accordingly. We also computed their self-reported confidence in their mental model based on the same categorization for both model weaknesses and strengths.

As the final step, we calculated results for each questionnaire following the suggested analysis and guidelines provided by the questionnaire designers. This approach resulted in one or two (or five, in the case of the Big Five questionnaire) measures that accurately capture the intended constructs. This required reversing or grouping certain questions together before calculating the average value for that measure per participant. We also ensured that each measure was presented in a way that higher values correspond to higher levels of the measured construct. Below is the summary of the measures and the code names assigned to them for further analysis:

- AI literacy (Non-eXpert's AI literacy–NXAI) resulted in 1 measure, with higher values meaning a more literate person.
- AI anxiety (AIANX) resulted in 1 measure, with higher values meaning higher anxiety towards AI.
- AI attitudes (ATT) resulted in 2 measures, PosATT and NegATT, with higher values meaning a generally more positive or more negative attitude towards AI.
- Attitudes towards AI (ATAI) resulted in 2 measures, ATAI-Fear and ATAI-Acceptance, with higher values representing more fear and more acceptance. However, this is the only questionnaire that was shown both before and after (optional) the study. As a result, we

obtained 4 total measures for this questionnaire. Since the post-study version of this questionnaire was optional, we encountered missing values for some participants. We used Mean Imputation (MI) [77] for handling the missing values for the analysis in these cases.

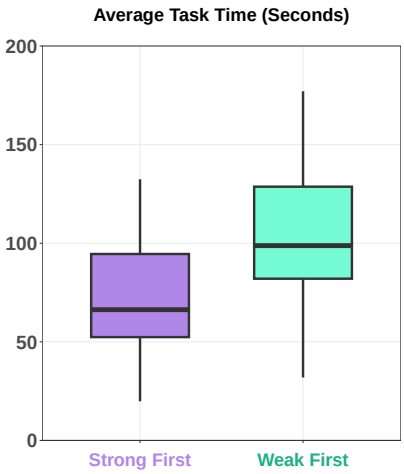
- The big five (BGFVE) generated 5 measures, Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Higher values indicate a greater inclination towards the corresponding personality trait, while lower values indicate the opposite. For instance, in the case of extraversion, higher values indicate a more extroverted personality, while lower values suggest a more introverted disposition.

6.2.2 Effects of Anchoring Bias on Usage Behaviours

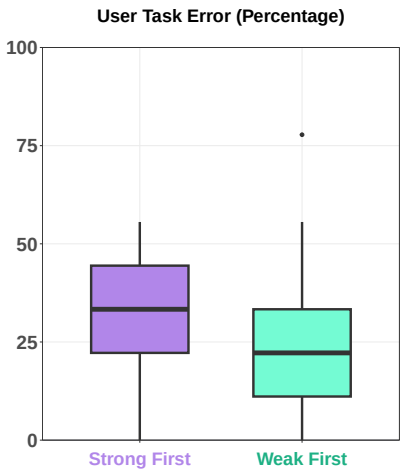
As the first step to the analysis, we examined how study conditions (anchoring bias) affect usage behaviours. Since the study task and experimental design for controlling anchoring bias is a replication of the study from Chapter 4, this was a critical step to ensure the first impressions are in fact formed based on how the participants' were anchored. We repeat the note that unlike the study from Chapter 4, we did not control explanation presence as an independent factor in the new experiment.

We tested for the differences between the control groups for average time per trial (seconds), task error (percentage), user agreement, mental model of weaknesses & strengths, and users' confidence in each of the measured mental models.

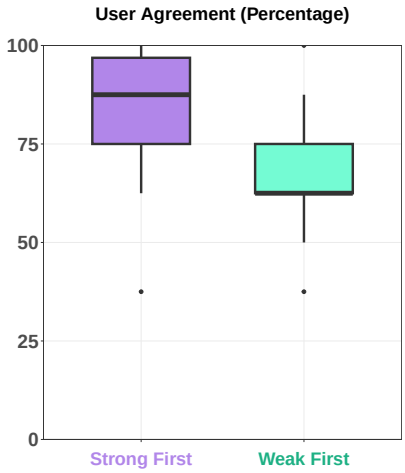
To determine the appropriate analysis approach for the measures, we assessed the normality of the data using a Shapiro-Wilk test and also visually examined the data distribution via histograms. If the data exhibited a normal distribution, we employed a one-way ANOVA to examine the differences between the conditions. In cases where the data did not meet the criteria of normality, we utilized the Kruskal-Wallis rank sum non-parametric test instead. Figures 6-3 and 6-4 provide a summary of the findings as well as the distribution of the measures per condition via box-and-whiskers plots.



(a) Distribution of average time per task in seconds. Lower values mean less time was spent on each task on average.



(b) Distribution of the error in verifying the policies. Lower values show participants made less errors in the task.

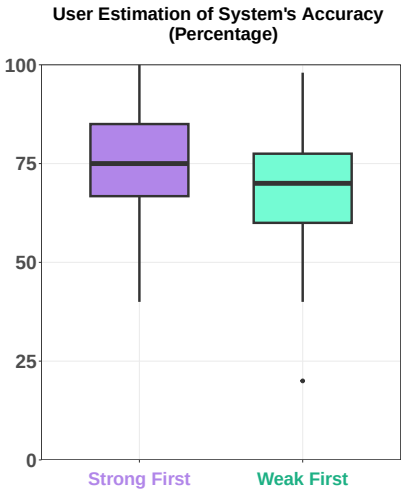


(c) The percentage of times user followed model's advice when verifying policies. Higher values show higher agreement.

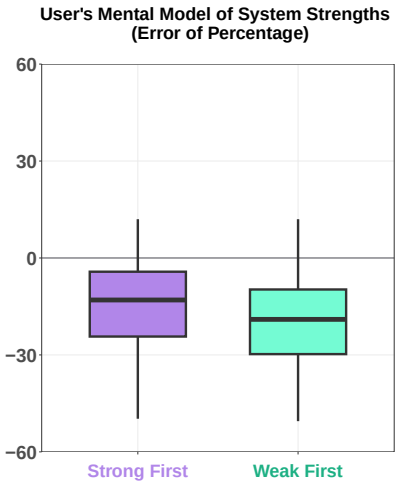
Measure	Unit	Main Effect & Comparison
(a) Average Task Time	Seconds (s)	$F(1,103) = 23.02, p < 0.001$ * $weakfirst > strongfirst$
(b) Task Error	Percentage (%)	$F(1,103) = 6.24, p < 0.05$ * $weakfirst < strongfirst$
(c) User Agreement	Percentage (%)	$\chi^2(1,103) = 25.01, p < 0.001$ * $weakfirst < strongfirst$

(d) Summary of results for the main task (user-machine task performance and agreement). F represents a one-way ANOVA test and the χ^2 represents a non-parametric Kruskal-Wallis rank sum test. For each measure, the winner of the comparison—the condition that achieved a more desired behaviour in a certain measure—is represented in bold. For example, for the second measure, the winner is the condition that makes less error. The greater-than symbol (<>) was also used to show the condition that had the higher values in a given measure. We consider $p < 0.5$ significant (*).

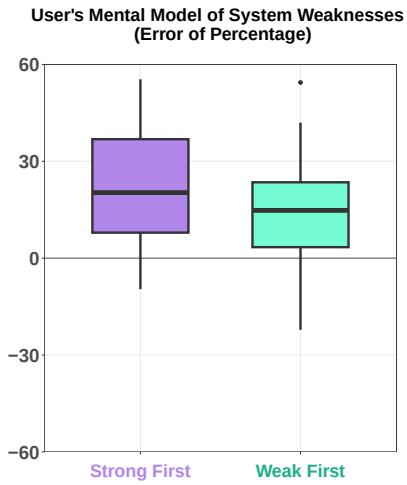
Figure 6-3. The distribution and the summary of findings for the measures from the main task.



(a) Distribution of participants' self-reported estimation of system accuracy. Higher values mean they perceived the model to be more accurate.



(b) Error in participants' perception of model strengths. $+/-$ values represent over/under estimations, respectively.



(c) Error in participants' perception of model weaknesses. $+/-$ values represent over/under estimations, respectively.

Measure	Unit	Main Effect & Comparison
Estimation of Accuracy	Percentage (%)	$\chi^2(1,103) = 6.59 p < 0.01$ * $weak first < strong first$
Perception of Strengths	Percent. Err. (%)	$F(1,103) = 3.27, p = 0.07$ (MS) $weak first < strong first$
Conf. in Strengths Perception	Percentage (%)	$\chi^2(1,103) = 0.05, p = 0.8$ (NS)
Perception of Weaknesses	Percent. Err. (%)	$\chi^2(1,103) = 6.88, p < 0.01$ * $weak first < strong first$
Conf. in Perception of Weaknesses	Percentage (%)	$\chi^2(1,103) = 0.01, p < 0.9$ (NS)

(d) Summary of the results for the post-study mental model questions. F represents a one-way ANOVA test and the χ^2 represents a non-parametric Kruskal-Wallis rank sum test. For each measure, the winner of the comparison—the condition that achieved a more desired behaviour in a certain measure—is represented in bold. For example, for the second measure and fourth measures below, the winner has a more accurate perception of the model (i.e., better mental models). The greater-than symbol (<>) was also used to show the condition that had the higher values in a given measure. Note that conf. is short for confidence. We consider $p < 0.5$ significant (*), $0.05 < p < 0.1$ marginally significant (MS) [39], and $p > 0.1$ not significant (NS).

Figure 6-4. The distribution and the summary of findings for the measures from the main task.

In summary, our results from "partially" replicating the study in Chapter 4 support our hypothesis that first impressions of AI systems caused by anchoring bias leads to differences in usage behaviour and mental models based on the direction of anchor.

Positive anchors fostered increased reliance on the model, resulting in faster human-machine task performance. However, this heightened dependence also led to a higher frequency of mistakes. Based on our findings, participants who were positively anchored formed positive first impressions of the XAI system, which led them to agree with the model's output significantly more, even when the model's advice might be wrong. This is evident from the results from task error and time, as these people were significantly faster at verifying each policy while significantly making more erroneous judgments. In other words, it seems that positive first impressions can impair people's judgment of a system they use for the first time, leading to automation bias.

Meanwhile, negative anchors led to more self-reliance, which increased response time but did not effectively limit the mistakes to a minimal level. Those with negative first impressions were also showing undesired behaviours. They spent significantly more time on the task (almost 1.5 times more on average), which could indicate their self-reliance, meaning they spent more time watching the videos than relying on the system's automation. Upon investigating the interaction logs, we observed that this group clicked on (and viewed) significantly more videos than those from the other group ($F(1,103) = 4.55$, $p < 0.05$), supporting our hypothesis that people with negative first impressions under-relied on the model. This self-reliance (1) led them to significantly disagree with the model's advice, i.e., there were significantly less instances where their correct/incorrect policy verification matched with the model's mistakes or accurate predictions; and (2) making significantly less mistakes in general. However, despite the latter case, they still did make mistakes. This is evident when comparing the mean and standard deviation for task error for the weak first condition ($M = 24.83\%$, $SD = 16.86\%$) to the those of the strong first condition ($M = 32.92\%$, $SD = 16.29\%$). We have observed that the presence of negative anchors resulted in participants spending more time on the task. However, self-reliance

was not enough to limit participants' mistakes to a bare minimum. In conclusion, both positive and negative anchoring biases can introduce unanticipated challenges in human-AI collaborations.

Additionally, participants who were exposed to weaknesses early-on perceived the XAI system to be significantly less accurate. Note that this is despite both groups having observed the same XAI model. Analyzing the mental model of weaknesses and strengths also show interesting findings. These measures are calculated as error of estimation, meaning positive values show overestimation while negative values represent underestimations. Here, overestimation means participants overestimated the accuracy of certain activity components, i.e., if the results show overestimation of weaknesses, it means the participants overestimated the model's accuracy in determining the objects that the model, in fact, has problem identifying. Both over and underestimation reflect a mismatch in user mental models of the system, and the values closer to zero reflect higher accuracy mental models. In our study, participants from strong first condition had a more accurate mental model of strengths (marginally significant), while they also had significantly less accurate mental models of system weakness (i.e., they overestimated the models capabilities where it is weak). On the contrary, the weak first participants were significantly underestimating model capabilities while had a more calibrated mental model of weaknesses. Overall, these results support the findings from our previous studies in chapters 5 and 4, in that Positive first impressions not only generate more positive perceptions of an AI model but also, despite the potential for automation bias, lead users to exhibit greater acceptance and forgiveness towards model mistakes. Conversely, negative first impressions foster skepticism towards the model's capabilities and encourage a greater reliance on one's own judgment, which can be counterproductive in AI-supported decision-making.

Irrespective of how first impressions influence individuals' judgment of the AI system and their task performance, it is crucial to recognize that anchoring bias can result in harm and unpredicted behaviors. Our findings underscore the significance of addressing people's anchoring bias from the outset of system usage, as well as the necessity to develop methods for timely detection and mitigation. It serves as a reminder that designing XAI systems should avoid a

one-size-fits-all approach and instead, consider designing them mindful of individual differences. This serves as a fundamental motivation for this study and forms the basis for further analysis in the subsequent sections.

6.2.3 Questionnaire analysis: Correlations Among Questionnaire Measures

As discussed in Section 6.2.1, each pre-study questionnaire resulted in one or more aggregated measures. In summary, we ended up with 5 measures for the Big Five questionnaire (Openness, Extraversion, Neuroticism, Agreeableness, and Conscientiousness); 1 measure for non-expert's AI literacy (NXAI); 1 measure for AI anxiety (AIANX); 2 measures for positive and negative AI attitudes questionnaire (PosATT and NegATT); and 2 measures for attitudes towards AI questionnaire (ATAI_Fear and ATAI_Acceptance). Overall, a total of 11 measures were extracted per participant for the pre-study questionnaire for further analysis. We also included 2 extra measures from the post-study background questionnaire, which were self-reported measures for AI/ML and Computer Science familiarity.

A Pearson correlation coefficient was computed to assess the linear pairwise relationship between these 13 measures. The results can be seen in Table 6-2. Our analysis reveals that three measures are strongly, positively, and significantly correlated with one another:

1. ATAI_Fear and AIANX: $r(11) = 0.933$, $p < 0.001$;
2. AIANX and NegATT: $r(11) = 0.890$, $p < 0.001$;
3. ATAI_Fear and NegATT $r(11) = 0.914$, $p < 0.001$.

This finding indicates that there is a strong association among the more negative attitudes and feelings towards AI. Similarly, we found a significant, moderate, positive correlation between acceptance of AI and positive attitudes ($r(11) = 0.850, p < 0.001$). Our test also reveals a significant, weak, and negative correlation between acceptance of AI (ATAI_Acc and NegATT ($r(11) = -0.884, p < 0.001$). That is, higher acceptance of AI is correlated with lower negative attitudes towards AI. Finally, we observed a high, significant, and positive correlation between

Table 6-2. The pairwise Pearson correlation across all the questionnaire measures. Strong correlations (> 0.5) are highlighted with shades of green, with the darker colors demonstrating stronger correlations. Also, significant correlations (i.e., where p-value is < 0.001) are marked with an asterisk (*). $+/-$ depict the direction of the correlation, where negative values represent an opposite correlation. We rounded each correlation to two decimal points. As such, a zero correlation shows a very small but possibly non-zero correlation. Note that the correlation results in a symmetric matrix, so we only display one half of the matrix to avoid redundancy.

	ATAI_Acc	ATAI_Fear	Neurotic	Extrovert	Open	Agreeable	Conscien.	PosATT	NegATT	AIANX	NXAI	mlFam	csFam
csFam	0.16	-0.08	-0.35	0.09	0.30	0.08	-0.05	0.18	-0.03	-0.20	0.44	0.45	1.00
mlFam	0.35	-0.02	-0.01	0.10	0.20	-0.10	-0.06	0.20	-0.06	-0.09	0.71*	1.00	
NXAI	0.33	-0.13	-0.10	0.18	0.38	0.03	-0.12	0.29	-0.09	-0.14	1.00		
AIANX	-0.16	0.67*	0.21	-0.20	-0.09	-0.04	0.06	-0.27	0.67*	1.00			
NegATT	-0.49*	0.70*	0.17	-0.01	0.02	0.02	-0.16	-0.29	1.00				
PosATT	0.56*	-0.17	-0.21	0.09	0.04	-0.11	0.04	1.00					
Conscien.	0.15	-0.01	0.13	0.00	-0.09	-0.07	1.00						
Agreeable	0.00	-0.11	-0.10	0.07	0.17	1.00							
Open	0.17	-0.12	-0.03	0.31	1.00								
Extrovert	0.008	-0.16	-0.02	1.00									
Neurotic	-0.12	0.20	1.00										
ATAI_Fear	-0.22	1.00											
ATAI_Acc	1.00												

NXAI and mlFam, meaning that the people's perceptions of their own knowledge in AI/ML correlates with the score they earned from the non-expert AI literacy questionnaire.

It is crucial to note that the correlations tested and reported here do not imply causation. For example, while we can observe similar trend of responses across measures that measure negative feelings and attitudes towards AI, we can not conclude that AI anxiety leads to negative attitudes

towards AI and vice versa. Further studies are required to explore potential causation relationships across these instruments and measures.

6.2.4 Questionnaire Analysis: Linear Regression Model for Outcomes

To further understand the how strongly the questionnaire measures affect the usage behaviours measured during the main study task (i.e., the study dependent variables), we built linear regression models of the measures for each study outcome. The goal was to determine which of the questionnaire measures significantly influenced each outcome. The coefficients for all these linear regression models for each dependent variable can be seen in Table 6-3. These results provide evidence to support our hypothesis that long-term past experiences and individual differences can affect human-AI collaborations simultaneously. For instance, people's self-reported estimation of accuracy was significantly affected by AI acceptance and positive attitudes towards AI, as well as conscientiousness personality trait ($p < 0.05$). Here, we must again emphasize that while the multivariate linear regression shows which measures affect each dependable outcome, this test is not sufficient to determine how the measures affect the measures. Moreover, this analysis does not consider the controlled condition (short-term anchoring bias). In the next sections, we will analyze how study outcomes are influenced by both pre-study questionnaire measures and the study conditions.

6.2.5 Questionnaire Analysis: Clustering and User profiling

The primary goal of this study was to investigate the feasibility of utilizing pre-task instruments as predictive tools for determining a new user's potential usage behavior and understanding of an AI system based on their long-term past influences. This research addresses an important problem, as it recognizes that designing AI systems to leverage/mitigate individuals' biases and prior beliefs is insufficient without considering the influence of individual differences and long-term past experiences. Researching this problem allows us to develop design approaches for AI systems that are tailored and effective for different individuals.

To predict usage behaviors and customize the design of the intelligent system accordingly, it is necessary to categorize users into similar groups based on their questionnaire responses. In this

Table 6-3. The Multivariate Linear Regression coefficients for each questionnaire measure and the final outcome. The coefficients show which of the independent measures correlate to a given dependent variable (study outcome), and in which direction. While bigger absolute values show stronger correlations, they do not necessarily indicate whether the coefficient is significantly different than the others. As such, table is coded in three different ways: (1) we use asterisks (*) to indicate coefficients that significantly influence the outcome ($p < 0.05$), (2) we highlighted the strongest coefficient using a bold font, and (3) we use color to code the direction of the significant and/or strong coefficients, where cyan demonstrate positive and magenta represents the negative correlations. Note that only in some cases, all these three fall in the same cell.

Outcomes / Measures	Agreement	Time	Error	EstAcc	StrenMM	WeakMM	StrenMM Confidence	WeakMM Confidence
ATAI_Acc	1.13	-6.25	-1.99	5.89 *	5.38 *	5.04	15.40	12.84 *
ATAI_Fear	-3.91	0.40	-2.32	-1.17	-0.72	-1.33	-2.75	-1.79
Neurotic	-0.41	-2.85	1.86	-0.73	3.81 *	0.89	0.59	2.77
Extrovert	-1.21	4.11	-3.53	-1.53	0.83	0.47	-2.25	-1.77
Open	-0.67	-1.21	3.75	1.99	1.68	1.92	2.54	1.05
Agreeable	-0.18	8.30 *	-3.01	0.29	2.38	1.19	1.29	4.74
Conscien.	0.98	6.03	-0.99	3.02 *	-0.08	1.32	3.69	6.11
PosATT	-2.07	5.16	0.06	-3.79 *	-0.88	-2.24	-9.07 *	-7.58
NegATT	5.15 *	-5.76	2.44	2.39	4.81	3.97	9.41	9.91
AIANX	-1.59	2.77	0.04	1.57	-1.41	2.49	-3.07	-1.99
NXAI	-0.96	-4.39	-1.92	0.39	4.14	6.07	-5.13	5.76
mlFam	1.18	0.96	1.64	-0.37	-3.98	-4.48	2.68	-8.82
csFam	-0.89	-1.07	-0.61	0.28	1.61	-0.19	-0.87	-0.73

work, we select group-level over individual-level analysis, as individual responses tend to be more granular and the data may contain deviations and noise on a per-person basis. Furthermore, considering practical constraints in real-life scenarios, it may not be feasible to modify the design of the system at an individual level. Therefore, working with groups provides a more practical approach for implementing design alterations. By grouping participants based on their responses, we can identify common patterns and trends that inform the design adaptations required to optimize the system's performance.

In this study, we used several questionnaires to capture different long-term constructs. For grouping people, we considered a bottom-up, exploratory approach without knowing either ground truths for the groups or the number of groups. With each questionnaire measure serving as one feature, this is an unsupervised clustering problem.

We used k-Means clustering algorithm to group participants. For the algorithm to operate effectively, it was necessary to elucidate several crucial aspects: (1) find the number of clusters (K) that would be suitable for our sample size and (2) ensure the data is scaled similarly across all the measures (via the scale() function in *R*). Despite some of the measures being correlated, we decided not to remove any of them (i.e., through feature-selection) as the actual questionnaires were capturing different constructs and each of them might contribute differently to the final clustering. However, for presentation purposes, we used feature selection techniques to identify the top features that contributed significantly to the clustering outcomes, aiming to extract the most relevant information and simplify the interpretation of the resulting clusters.

To determine the appropriate number of clusters (K), we adopted a comprehensive approach that considered statistical methods and data-specific considerations.

To maintain adequate statistical power, we set a constraint of $K \leq 5$ for our study analysis. This decision was driven by the understanding that higher values of K could result in smaller clusters with nuanced distinctions. With our dataset consisting of 105 data points, it was crucial to strike a balance between capturing meaningful variability and maintaining a sufficient number of data points within each cluster. By adhering to this constraint, we aimed to avoid excessively fragmented clusters that might hinder the identification of substantial patterns or insights.

To pinpoint the best number of clusters, we further employed two commonly used techniques, namely the average silhouette and elbow methods, to visualize and analyze the best number of clusters for our questionnaire measures. The average silhouette method indicated K=2 as the optimal choice for clustering, whereas the elbow method suggested that the elbow point fell between K=2 and K=3. Considering the objective of capturing diverse user profiles, we ultimately

selected K=3 as the number of clusters for the K-Means algorithm. This decision aimed to ensure a meaningful representation of distinct user groups within the data.

After performing the K-means clustering algorithm, we performed a multivariate ANOVA (MANOVA) statistical analysis to test whether the three clusters are significantly different. We chose this test to take all the features (i.e., questionnaire measures) into account simultaneously when determining the differences across the clusters (here, cluster serves as independent variable in our test). The MANOVA test finds evidence that there is a significant difference between the three clusters, with $F(1,103) = 17.39$, $p < 0.001$. In other words, our test shows evidence that the clusters are separated. However, MANOVA only tests for the main effect and does not account for which of the clusters are separated.

To conduct a pairwise comparison between clusters, we ran a Linear Discrimination Analysis (LDA)[§]. LDA calculates two coefficients, LD1 and LD2, based on the input data. By leveraging a two-dimensional visualization of the data points based on these derived variables (as seen in Figure 6-5), we are able to visually evaluate the extent of separation between the clusters. Through visual examination of the clusters, it is evident that there is a substantial degree of separation among them, even though there are a limited number of outlier data points that were not assigned to their respective clusters. Notably, cluster 1 exhibits a significant distinction from the other two clusters, while clusters 2 and 3 display a high level of separation, although not entirely distinct from each other.

This data analysis resulted in 15, 43, and 47 data points in clusters 1, 2, and 3, respectively. It is important to note that the imbalance in cluster sizes may be caused by both the sampling of the data and the inherent differences and trends among individuals from the diverse groups. While the size discrepancy can potentially be minimized with a larger sample size, it is also plausible that it reflects genuine variations and patterns across different population groups.

To better understand and interpret the results based on the clusters, we sought after meaningful labels that represent most of the data points within a cluster. Considering clustering as

[§]This method is used and supported by data science practitioners. See the URL for an example: https://www.r-bloggers.com/2021/11/manovamultivariate-analysis-of-variance-using-r/

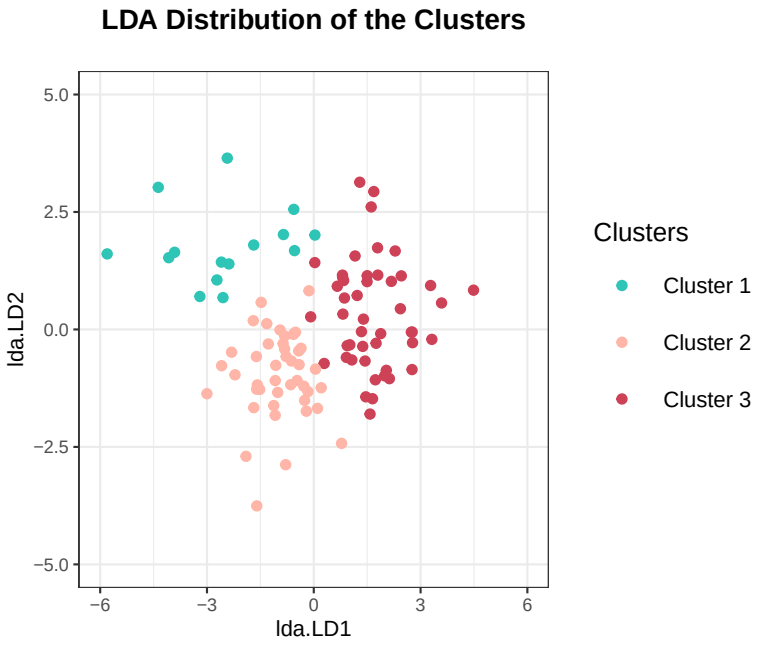


Figure 6-5. The distribution of the data points and clusters based on the post-hoc clustering analysis, done through Linear Discrimination Analysis (LDA), to test how separated the final three clusters are based on the calculated LD1 and LD2 coefficients. By comparing the grouping of the data visually, we can conclude that the clusters are highly separated despite having a few outlier data points that were not accurately grouped, specifically between clusters 2 and 3.

Table 6-4. We conducted a one-way ANOVA to examine the significant differences among the features across the clusters, aiming to identify the influential factors for clustering. Only the statistically significant results are presented, based on the pairwise comparisons of each feature across the clusters. For instance, cluster 3 exhibited significantly high acceptance attitudes towards AI compared to the other two clusters, leading us to assign the label High to ATAI_Acc for cluster 3 (column 1, row 3). By leveraging these significant comparisons, we derived meaningful labels for each cluster that capture their distinct characteristics and patterns.

	ATAI_Acc	ATAI_Fear	Neurotic	Extrovert	Open	Agreeable	Conscien.	PosATT	NegATT	AIANX	NXAI
Cluster 1	-	High	-	-	-	-	-	-	High	High	-
Cluster 2	-	Average	-	-	Low	-	-	-	Average	Average	Low
Cluster 3	High	Low	Low	High	-	-	-	High	Low	Low	-

the independent factor with three levels, we tested each feature using a one-way ANOVA for the main effect, as well as a Tukey HSD post-hoc test for pairwise comparison across the clusters.

This analysis allows us to determine which features significantly contributed to each cluster and how. Table 6-4 provides an overview of this analysis. Based on our analysis, the least contributing factors across all the clusters were agreeableness and conscientiousness. However, the remaining nine questionnaire measures demonstrated significant influence on at least one of the clusters, and were utilized to assign appropriate labels to each group:

1. AI Skeptics (cluster 1)—People in this group exhibit strong fear, anxiety, and higher negative attitudes towards AI.

2. AI Ambivalent Individuals (cluster 2)—This group includes people with lowest openness personality traits and AI literacy. These people have moderate levels of anxiety, fear, and negative attitudes towards AI.

3. AI Positive Individuals (cluster 3)—A group who show the highest positivity in their attitudes towards AI: they are more accepting and less fearful, anxious, and negative towards AI. This group comprises more extroverted and less neurotic people, which might indicate these people are overall less anxious than their counterparts.

Later in this chapter, we use these labels to discuss our results in a more meaningful way.

6.2.6 Clustering vs. Anchoring Bias in Outcome Prediction

To test whether and how user profile (cluster) influences usage behaviours, we consider it as the second independent variable of our experiment for further analysis. The goal was to study the interplay between anchoring bias and people's background profiles on user perceptions and collaborations with an AI algorithm. This ultimately allows us to understand the implications of grouping people based on their long-term differences prior to usage and incorporating group-specific techniques to mitigate pre-conceptions and biases towards intelligent systems.

We use a 2-way independent ANOVA to test the differences across the 2×3 comparison, and a Tukey-HSD post-hoc pairwise comparison for each of the seven dependent variables. In this section, we report and discuss the results of this analysis, only for the measures where we observed significant differences based on user profile and anchoring bias.

Main Task Usage Behaviours: Our results from the main task (user performance and agreement) only show significant main effects based on the anchoring bias, which align with our baseline analysis in Section 6.2.2. Particularly, regardless of their background usage profiles, participants from strong first condition were significantly faster ($F(1,99) = 13.53$, $p < 0.001$), made significantly more error ($F(1,99) = 6.37$, $p < 0.01$), and had higher rates of agreement with the model ($F(1,99) = 14.58$, $p < 0.001$). We did not observe interaction effects or differences based on user profiles.

User Perceptions and Mental Models: Despite the lack of detected relationships between profile clusters and task performance, the findings pertaining to the mental models and user perceptions had more interesting results. Notably, these results exhibited discernible patterns within our identified clusters (i.e., user profiles based on long-term effects), further enhancing their significance and potential implications. Particularly, the two-way ANOVA show a significant main effect based on clusters for participants' perception of weaknesses: $F(2,99) = 3.52, p < 0.05$). Profile 1 participants exhibited a substantial tendency to overestimate the accuracy of the model, particularly in cases where its performance was objectively weaker. The pairwise comparison revealed this behaviour consistently compared to participants from profile 2 ($p < 0.001$) and profile 3 ($p < 0.05$). The main effect also showed a marginally significant effect based on the anchoring bias $F(1,99) = 3.88$, $p = 0.051$), where participants in the weak first had more accurate perception of XAI weaknesses than their counterparts.

There were no significant interaction effect observed across the conditions for this measure. By comparing the distributions for profile 1, we notice a bigger difference between positive and negative anchors, which may indicate people from this background profile might showcase different anchoring behaviours. However, it is important to note that the current results do not provide sufficient evidence to support this claim. Further research and investigation would be necessary to directly explore and validate this hypothesis.

The tests also revealed a significant main effect based on user profile on participants' self-reported confidence in their perceptions of the weaknesses, with $F(2,99) = 4.87$, $p < 0.01$.

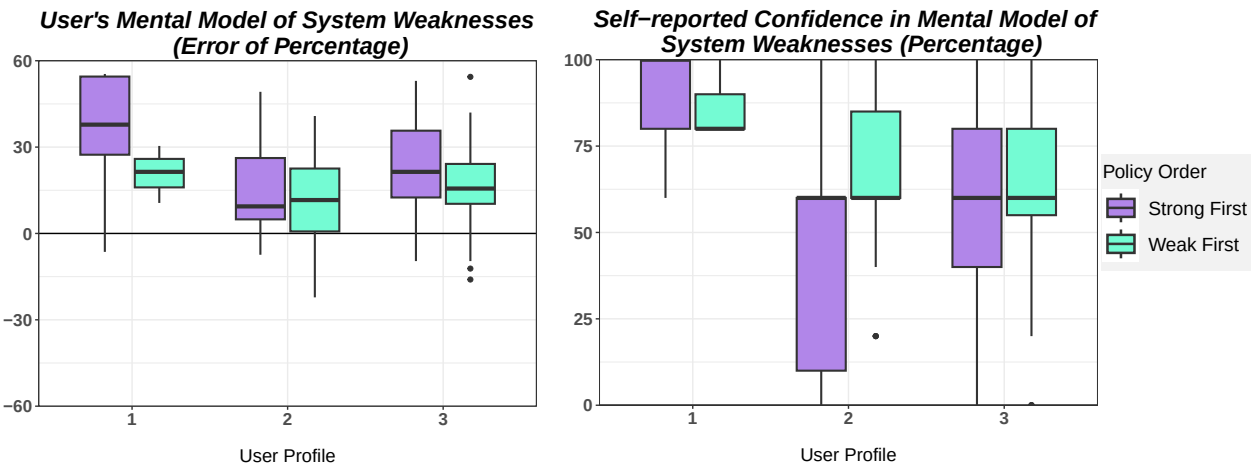
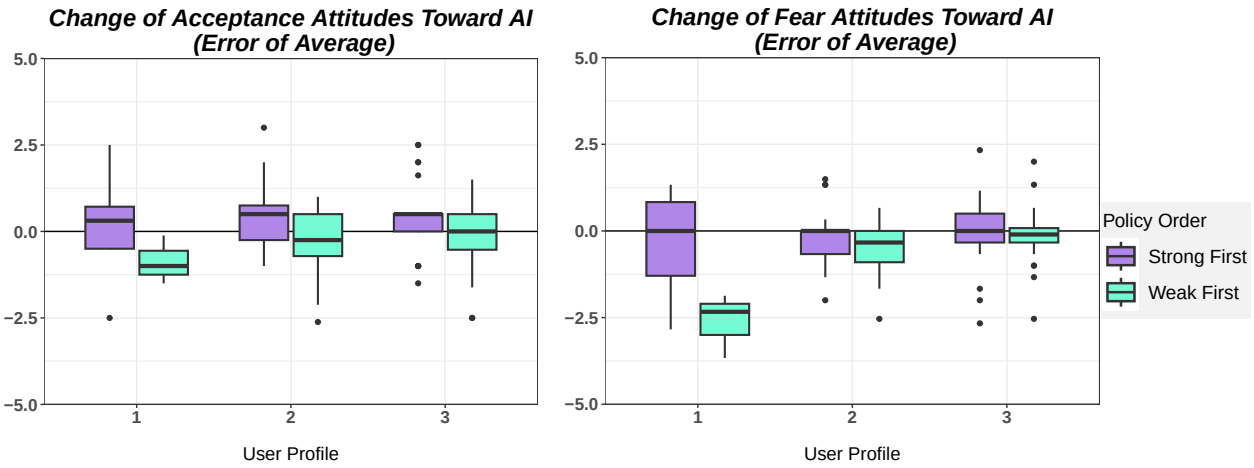


Figure 6-6. The distributions of the mental model of weaknesses measures based on the two-way analysis of anchoring bias and user's background profile. The chart on the left shows the estimation of weaknesses, where positive/negative values determine over/underestimation, and the values closer to zero show higher alignment of mental models and the actual model performance in weak cases. The chart on right shows people's confidence in their own estimation of the mental model of system weaknesses. Higher values show more confidence, and when considered together with the mental model alignment, they could represent under/overconfidence. The two-way ANOVA shows participants from background profile 1 exhibit significantly worse mental models while being overconfident about their perceptions.

The pairwise comparison reveals that participants from profile 1 were significantly more confident in their perceptions of the XAI's weaknesses, compared to their counterparts from profile 2 ($p < 0.01$) and profile 3 ($p < 0.01$). No significant differences based on the anchoring effect and an interaction effects were found.

Change of Attitudes: To understand how people's prior attitudes towards AI might change after using a new intelligent system, we asked the participants to complete the ATAI questionnaire ($N = 5$) one more time post-study. Responding to the questionnaire at the end was optional to minimize the study length and survey fatigue. As a result, 14 participants opted out of completing this questionnaire (6 in strong first and 8 in weak first). In order to perform statistical analysis, we used the mean imputation technique (within each condition and for each question) to replace the missing values. We then calculated the change in attitudes in terms of acceptance and fear by



(a) Changes of responses to ATAI_Acc before and after the study.

(b) Changes of responses to ATAI_Fear before and after the study.

(c) Summary of results for change in fear and acceptance attitudes towards AI (ATAI_fear and ATAI_Acc). Negative values demonstrate decrease in that attitude while positive values show an increase in the value. We used a two-way factorial ANOVA to test the main effect and a Tukey HSD test for pairwise comparison. For the post-hoc results, the condition on the right-side of the comparison shows shows less magnitude of change.

Change in AI Acceptance	Main Effect
	Policy Order: $F(1,99) = 10.18, p < 0.01$ *
	User Profile: $F(1,99) = 0.38, p = 0.6$ (NS)
	Interaction Effect: $F(1,99) = 0.26, p = 0.7$ (NS)
Change in AI Fear	Main Effect
	Policy Order: $F(1,99) = 9.14, p < 0.01$ *
	User Profile: $F(1,99) = 4.81, p < 0.05$ *
	Interaction Effect: $F(1,99) = 3.51, p < 0.05$ *
	Post-Hoc Test
	$weakfirst - c1 < strongfirst - c1$ $(p < 0.05)$ *
	$weakfirst - c1 < strongfirst - c2$ $(p < 0.05)$ *
	$weakfirst - c1 < strongfirst - c3$ $(p < 0.05)$ *
	$weakfirst - c1 < weakfirst - c3$ $(p < 0.05)$ *
	$weakfirst - c1 < weakfirst - c2$ $(p < 0.06)$ (MS)

Figure 6-7. The distribution and summary of findings for change of attitudes pre/post the study.

computing the difference between ATAI questionnaire post and pre study, where the positive values show an increase in acceptance or fear, and the negative values show decrease in these attitudes after using the XAI system. We then conducted the two-way statistical analyses to understand how changes of attitudes differ based on anchoring and profile groups. Table 6-7c

provides a summary of these analyses. Our results show that people's acceptance of AI was significantly increased when they were positively anchored (i.e., strong first group): $F(1,99) = 10.18$, $p < 0.01$. Our findings also provide strong evidence that change in AI fear is significantly affected by both policy order (anchoring bias) and user profile. However, a significant interaction effect reveals that the relationship between anchoring bias and change of AI fear depends on the user profile (cluster). Unlike acceptance, anchoring effect only influences change in AI fear for participants in profile 1. The users from profile 1 who are negatively anchored express significantly more fear towards AI (in general) after using the XAI model compared to their counterparts in the same profile who were positively anchored. In fact, the fear of AI increased significantly in the weak first condition of profile 1 compared to all other participants from other anchoring conditions and user profiles. This finding provides strong evidence that a combination of long-term factors and first impressions of an AI algorithm can strongly contribute to how people perceive AI technology, which can have lasting effects on their future usage and interaction with this technology.

6.3 Discussion

This experiment was designed to explore using questionnaires to extract constructs related to long-term influences (experiences and individual differences) from people, to examine how these constructs affect people's usage behaviours when collaborating with AI systems. We aimed to study the potential of using these long-term past constructs to categorize people into meaningful groups using unsupervised methods, i.e., building user profiles that represent certain types of users. We were primarily looking into how people from different profiles get anchored differently towards their usage and understanding of the intelligent system. This is along the lines of the work presented in Chapter 5, were we studied the interplay between domain experience (as a long-term past experience) and anchoring bias (short-term experience). However, in this study, we look at the potential of extracting and summarizing multiple long-term factors simultaneously as user profiles to understand the interaction between user profiles and short-term anchoring effects. This work was motivated to study user profiling by taking long-term past influences into

account, as having such grouping methods could potentially allow the AI designers and practitioners to find techniques to mitigate undesired behaviours and biases (such as anchoring effect) towards AI system through penalization (based on the groups). In this section, we discuss the findings and implications, as well as the study limitations and opportunities for future work.

6.3.1 Summary, Interpretation, and Implications of the Results

Our results from the main study task analysis show that different trends based on anchoring group. People with positive first impressions: (1) were more efficient (faster) while making more mistakes as a result of strong agreements with the AI system (automation bias) and (2) had higher estimation of the system accuracy, and were able to build a more accurate mental model of AI strengths. In contrast, people with negative first impressions: (1) were exhibiting higher self-reliance that resulted in making less mistakes (while the mistakes were not minimal), but that lead to less agreement with the model and taking almost 1.5 times longer (on average) to complete the task; and (2) had lower estimation of the system accuracy, and were able to build a more accurate mental model of AI weaknesses. These findings support our findings from previous Chapters 4 (the study which we replicated) and Chapter 5 that anchoring bias affects people's usage behaviours and perceptions of the system, making the task a suitable baseline for addressing our primary goals for the study and analyses.

We built multivariate linear regression models based on the study outcomes (dependent variables) and the measures extracted from the pre-study questionnaires that were extracting people's long-term past influences of AI (positive and negative attitudes towards AI, acceptance and fear attitudes towards AI, AI anxiety, and non-expert AI literacy) and the big five questionnaire to elicit personality traits (long-term individual differences). Our results showed that multiple types of long-term factors can simultaneously affect the outcomes of human-AI collaborations. For example, we found that people's non-expert AI literacy significantly and strongly affect the accuracy of their mental models of AI's weaknesses. However, this test only showed which of the questionnaire measures were correlated with the usage behaviours and mental model formations and not how.

Using K-means clustering algorithm, we grouped participants based on their questionnaire measures into three clusters, i.e., 3 user profiles. We used statistical and visual analyses to verify the profiles are separated, and found meaningful labels for them based on the significantly contributing measures in each cluster. This led to three user profiles labeled as AI Skeptics (profile 1), AI Ambivalent (profile 2), and AI positive individuals (profile 3). We performed the main study analyses again, but this time we used profiles as a second independent variable for a two-way comparison.

Our result demonstrate that participants from profile 1—i.e., the AI Skeptics—built a significantly worse mental model of the system weaknesses and were significantly overconfident of their mental models, compared to those from the other profiles. This is an interesting and unexpected observation, as there is a notable discrepancy between their expressed skepticism/fear of AI that was measured from the questionnaires and their actual usage behaviours. The AI Skeptics group fail to notice models' weaknesses (they overestimated the model's accuracy in cases where it is weak at) while showing strong overconfidence towards their perceptions. One potential justification is the lack of AI knowledge and literacy. Despite harboring skepticism towards AI, novice users may face challenges in accurately identifying model weaknesses and objectively assessing accuracy. In particular, their subjective perceptions of weaknesses can be inflated, driven by an expectation of perfection where a model is deemed strong only if it achieves flawless performance without errors (100% accuracy), while considering a 90% accuracy rate as low, despite it being relatively high in the AI/ML community. The correlation between mental models of weaknesses and non-expert literacy is significant, positive, and strong as found from our linear regression model (see Table 6-3), providing additional support for this hypothesis.

Our results indicated that with positive first impressions of the model, people's acceptance toward AI was positively shifted after interacting with the model. Another interesting finding was the interaction effect between user profile and anchoring group in changes of fear in AI after using an intelligent system. Among all the six analysis conditions, fear of AI strongly increased only for AI Skeptics who developed negative first impressions of the system (weak first condition). We

also observed the biggest magnitude of change in AI fear in strong first condition for profile 1. In other words, while the mean did not change significantly, the were more people who changed their responses (ATAI_Pre: $M = 6.80$, $SD = 1.82$; ATAT_Post: $M = 6.56$, $SD = 2.43$), compared to other conditions. This could indicate that AI Skeptics were more likely to change their fear after using the XAI system when they were positively anchored. When intelligent systems provide a positive first impression, people's attitudes towards AI technology as a whole can be shifted towards less fear and more acceptance. While this is an interesting observation, future studies are indeed necessary to understand the extent and cause of such behaviours.

Our study provides supporting evidence, as a proof-of-concept, that profiling users based on questionnaires that capture their long-term past influences, experiences, and differences prior to usage can be beneficial to designing AI systems to mitigate the added usage challenges and behaviours caused by these differences. Particularly, our results provide empirical evidence that user profiling can be a beneficial approach to bias (and/or unanticipated usage) mitigation instead of seeking after one-size-fit-all solutions. People carry a baggage of long-term past experiences and differences, adding complexities to human-AI interactions and collaborations. Taking anchoring bias as an example, this and the studies in the other chapters provide supporting evidence that not all the people get anchored similarly, and various types of anchoring depends on other long-term factors such as domain expertise, AI familiarity, personality traits, and attitudes and perceptions of AI technology, to name a few. Designers and practitioners need to come up with approaches (specific to the domain and task of their intelligent system) to (1) elicit user differences based on long-term influences and factors, (2) seek suitable user profiling methods based on these factors, (3) predict how new users might interact with the AI system by assigning them to the most matching profile , and (4) design anchoring mitigation techniques based on the profile, which can be an iterative refining process.

For instance, incorporating guided interactions during a tutorial phase can serve as a potential solution to mitigate anchoring bias. Rather than allowing users unrestricted access to the system, designers can strategically introduce controlled encounters with errors early on, before

granting full access to the tool. However, user profiling may indicate that anchoring effects are more pronounced within specific groups of individuals. By identifying these groups through user profiling, we can anticipate the potential hazards of first impressions and tailor guided interactions specifically for these groups. This approach allows for a more targeted and effective intervention to address anchoring biases and enhance users' overall experience and perception of the system.

6.3.2 Limitations and Future Work

Our experiment was a first step to study utilizing questionnaires prior to first-time usage of intelligent systems to extract people's long-term internalized differences and past experiences and group them into similar user profiles. Like all other designed experiments, our study has limitations and potential confounds. One of the main challenges was designing an online study while maintaining the quality of the data. Online studies are more accessible and offer broader participant pools compared to in-person studies. However, people are naturally more likely to get distracted or experience interruptions that would influence their attention. Due to this known problem, using attention check mechanisms and limiting the length of the study to minimize survey fatigue can be effective approaches to improve the quality of data.

One challenge was choosing questionnaires that introduce unique questions, while limiting the size. Although we managed to find smaller versions for certain questionnaires (e.g., the Big Five), we chose not to modify others due to the lack of known alternatives and our commitment to maintaining the integrity of the questionnaire measures. This resulted in a total of 98 questions prior to the study, with additional questions administered post-study. Despite the median study duration being only 35 minutes, indicating a reasonable time commitment, it is worth noting that some participants from the online crowd may have found the study tiring. Moreover, our recruitment strategy targeted university-affiliated individuals and relied on word-of-mouth, resulting in enhanced quality control compared to using public crowdsourcing platforms like Amazon Mechanical Turk. We speculate other means to control the engagement, such as incentivizing the participants based on their accuracy or conducting in-person studies might minimize the low quality data or other unseen confounding factors. Additionally, it is important

to acknowledge that our findings may have revealed different trends and patterns if we had (1) a substantially larger sample size, involving hundreds to thousands of participants, and (2) a more inclusive sampling approach, incorporating a diverse range of participants. For instance, we can expand the participant pool by recruiting people from different backgrounds, demographics, and degrees/occupations. These considerations present significant avenues for future research and exploration in the field of human-AI collaboration.

Finding a suitable unsupervised (clustering) approach is a non-trivial and exploratory task, posing challenges in finding the most optimal solution. We must emphasize that our goal here was not to propose an algorithmic technique for user profiling, but it was to demonstrate a proof-of-concept for the benefits of utilizing pre-usage questionnaires to elicit long-term past influences for the purposes of user profiling. Using K-Means was one of our many possible options, and we tried our best to adhere to its assumptions and requirements. we also employed evaluation methods to verify the extent of separation achieved among the clusters. Future work is needed to identify unsupervised algorithms for user profiling based on long-term past influences, that best suit domain, task, and the type of questionnaires. While we used questionnaires as potential data collection media, an important future direction is finding questionnaires that can capture multiple long-term past construct using shorter subset of questions in order to minimize survey fatigue. Moreover, other types of explicit and implicit instruments (as opposed to questionnaires) can be used to extract people's long-term constructs and experiences, which need to be identified in future work.

CHAPTER 7
CONCLUSIONS

In my dissertation, I present human-centered research to advance the fields of AI and human-centered AI, providing insights for challenges that arise when people from different backgrounds and experiences interact and collaborate with AI-supported tools and systems. In this section, I provide an overview of the contributions, implications and design recommendations for AI designers and practitioners, and discuss opportunities for future work.

7.1 Contributions

This dissertation thesis contributes novel research in the field of human-centered AI, offering valuable insights to enhance the ethical and responsible design of intelligent systems. Moreover, it illuminates unexplored avenues of research, contributing to the advancement of our knowledge in this domain. In summary, the key contributions of this work are as follows:

1. To promote the importance of factoring individual differences in when designing intelligent systems, I propose an empirical conceptual model of user's past experiences and differences that organizes the related work into two categories based on their stability. The conceptual model can be served as a road map with design considerations for human-AI collaborative systems and help researchers identify the present gaps in this field. This work was covered in Chapter 3.

2. In Chapter 4, I present a study to examine how transient factors influence current usage behaviors. In particular, we looked at people's differences based on their anchoring bias. Our results demonstrate important findings and implications on not only how anchoring bias affects people's mental model formations and confidence in the system, but also provides evidence to support our conceptual model and the importance of recognizing the short-term affects of using intelligent systems.

3. I further present another study in Chapter 5 to determine the interplay between stable and transient prior factors through comparing people's trust in intelligent systems w.r.t. domain expertise and anchoring bias. Our results demonstrate behavioral differences (e.g., trust)

based on the positivity of recent experiences and the differences in long-term experiences with the domain.

4. I present a third and final study to study the various stable and transient past experiences and factors that influence people-AI collaborations. I show how using questionnaires prior using for the first time can help elicit people's differences and organize them into similar groups in order to study how the role of long-term and stable perceptions and experiences of AI on usage behaviours and how they interact with short-term anchoring effects. This study is discussed elaborately in Chapter 6.

5. The findings of this work may help determine the need for capturing user's prior experiential differences prior to using an intelligent system in order to be wary of people's differences and long/short-term experiences when studying or designing collaborative intelligent systems.

7.2 Discussion on Design Implications & Future Work Opportunities

In this dissertation, I presented novel empirical research to underscore the significance of considering users' differences and past experiences when designing AI systems, as these factors can weigh in during the decision-making process and affect the collaborative outcomes. Moreover, interactions and encounters with intelligent systems can leave lasting impacts on people's expectations and attitudes towards broader AI technology and subsequent systems they may come across.

Our proposed conceptual model of user's past experiences (presented in Chapter 3) provides an abstract overview of what past experiences may affect usage behaviours and how–see Figure 3-3. This model, combined with the insights from our semi-structured literature review, serves as a valuable resource for researchers and practitioners to identify existing research gaps and formulate relevant research questions. We further expanded on the model by incorporating the emerging categories from our literature review, resulting in an extended version showcased in Figure 3-4. This detailed representation provides deeper insights into the various factors influencing user behavior based on prior experiences and differences. By cross-referencing the

categories from the conceptual model with the results from Table 3-4, researchers can gain a clearer understanding of the gaps in current research, which can in turn facilitate the formulation of targeted research questions to address those gaps.

For example, as demonstrated in the model, individuals' stable personal experiences, such as domain knowledge, can play a role in influencing their interactions with AI systems. However, a deeper understanding of how these experiences specifically impact people during usage is warranted. For instance, we can explore research questions such as: (1) How does users' domain knowledge influence the formation of trust in AI systems over time? (2) Do changes in users' mental models occur as a result of their domain knowledge interacting with AI systems, and if so, what are the patterns of these changes? (3) Are there any biases that tend to be formed in users' interactions with AI systems due to their domain knowledge, and how can these affect human-AI collaboration? By addressing these smaller research questions and examining their nuances, we can gain valuable insights into the broader picture of how users' stable personal experiences impact their transient experiences and interactions with AI systems. The study from Chapter 5 follows this framing approach and contributes significant findings to increase our understanding of the interactions between anchoring bias, domain knowledge, and trust formations and changes over time.

The current conceptual model serves as a valuable standalone contribution, facilitating an in-depth exploration of existing gaps and enhancing researchers' awareness of users' prior influences. Nevertheless, our plan is to extend this work in the future by proposing actionable guidelines for research question formulations, empowering researchers to identify gaps effectively. Additionally, our objective is to validate the practicality and efficacy of employing this model as a framework for designing AI systems, considering users' past experiences. To achieve this, we aim to engage in thorough discussions with AI designers and practitioners, which we plan to do as the next step of our research.

All the three studies in this dissertation particularly focused on one short-term, transient factor: anchoring bias. Among these studies, we controlled the anchoring bias by regulating the

order of participants' exposure to model mistakes and weaknesses, effectively controlling their first impression formations. This approach allowed us to examine an extreme scenario where all errors and weaknesses were presented either at the beginning or the end of the usage session. However, in real-world situations, designers have limited control over when errors may occur, which subsequently impacts the speed and timing of anchoring biases formation. Regardless, our studies demonstrated a trade-off between positive and negative first impressions, where neither of the two anchors consistently led to always desirable outcomes. This means that designers need to find solutions to account for anchoring bias when designing intelligent systems. Here, we propose a few recommendations for how these first impressions can be incorporated and accounted for in the design.

One possible approach to address the anchoring bias is to prevent the formation of first impressions altogether. By stopping or minimizing the problem at its source, designers can limit the consequences of anchoring bias on users' perceptions and usage behaviours with the AI system. Designers can implement engaging and interactive introductory sessions, such as tutorial phases, to gradually familiarize users with the system. By adopting this technique, designers can actively guide the users' understanding and control how their mental models and first impressions of the AI model are formed. This proactive approach helps in avoiding the risk of anchoring bias formation, which may occur if users are left to form their own impressions without proper guidance. Through the tutorial phases, designers can ensure that users receive accurate and comprehensive information, leading to more informed and less biased interactions with the intelligent system. Using explanations can potentially help in minimizing first impressions. Encountering errors might have less severe influences on users if the source of the error is more apparent to them as an additional context. For example, people may be more forgiving of errors attributed to a lack of training data, indicating the AI's limited knowledge in certain scenarios, compared to other types of errors. Designers can explore explainability techniques that provide additional context for the potential sources of error in each specific instance. Alternatively, they can choose explanations that offer insights into the model's reasoning on a more systematic

(global) level. Future research can potentially focus on studying and designing approaches for explaining and contextualizing errors. Utilizing such explanations can be particularly effective when applied selectively to instances where they have the potential to make users more forgiving of mistakes. By strategically providing explanations in these specific cases, designers can mitigate negative first impressions that might otherwise arise. Studying context-based explanations of errors is another potential avenue for future work.

Another approach to handling anchoring bias is to acknowledge its potential occurrence and accept that it may happen in certain cases. Instead of trying to prevent it entirely, designers can leverage the formed bias in their design process. In some cases, designers may choose to prioritize certain behaviors over others to achieve specific goals. For example, if the task allows for a higher tolerance of errors and the anchoring effect does not significantly impact the overall outcomes, designers might accept the bias as an accommodation to achieve other desirable outcomes. As evidenced by the studies conducted in this dissertation, a trade-off exists between usage behaviors influenced by positive and negative first impressions. Designers can leverage these trade-offs and make deliberate sacrifices in favor of one or the other to achieve more favorable outcomes in human-AI collaboration. For instance, by intentionally permitting early mistakes, designers can diminish potential automation biases and prompt users to trust their own judgments, fostering a sense of self-reliance, while simultaneously implementing other measures to prevent the overwhelming influence of negative first impressions. On the other hand, in certain scenarios, designers might choose to let users form positive first impressions, as this can contribute to improving users' mental models of the AI system. Combining these bias leveraging approaches with other bias mitigating techniques can be essential to enhance their effectiveness addressing anchoring bias and enhancing human-AI collaboration.

Finally, as explored in Chapter 6, personalization and user profiling can prove highly effective in anticipating how individuals might be anchored when using an AI system. By identifying users' profiles based on their past experiences and individual differences, designers can tailor bias mitigation approaches to match the specific profiles, resulting in personalized

experiences for different user groups instead of a one-size-fits-all approach. Although this dissertation lays the groundwork and provides evidence for this direction, future work should focus on developing various techniques to gather and categorize users' differences into profiles, as well as determining the most suitable bias mitigation or leveraging approaches for each specific profile.

APPENDIX
AI ATTITUDES AND PERSONALITY TRAITS QUESTIONNAIRES

A.1 AI Literacy Questionnaire

Questionnaire designed by Laupichler et al. [95]* to captured non-expert users' non-technical AI literacy. This questionnaire was measured on an 11-point Likert scale, with responses ranging from strongly disagree to strongly agree.

1. I can tell if the things I use frequently are supported by artificial intelligence.
2. I can name examples of technical applications that are supported by artificial intelligence.
3. I can explain the differences between human and artificial intelligence.
4. I can explain the difference between general (or strong) and narrow (or weak) artificial intelligence.
5. I can distinguish media representations of AI (e.g., in movies or video games) from realistic AI.
6. I can name and evaluate various weaknesses of artificial intelligence.
7. I can name and evaluate various strengths of artificial intelligence.
8. I can describe risks that may arise when using artificial intelligence systems.
9. I can describe rewards that can come from using artificial intelligence systems.
10. I can visualize potential future scenarios of artificial intelligence.
11. I can distinguish AI applications that already exist from AI applications that are currently still dreams of the future.
12. I can name examples of how computers make decisions.
13. I can explain how machine learning works.
14. I can describe how machine learning models are trained, validated, and tested.

*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, p.100126.

15. I can explain the difference between supervised learning and unsupervised learning (in the context of machine learning).
16. I can explain how reinforcement learning works on a basic level (in the context of machine learning).
17. I can explain how deep learning relates to machine learning.
18. I can explain what the term artificial neural network means.
19. I can critically evaluate the results of artificial intelligence applications in at least one subject area.
20. I can explain why data plays an important role in the development and application of artificial intelligence.
21. I can describe why humans play an important role in programming, choosing models, and fine-tuning artificial intelligence systems.
22. I can describe how some artificial intelligence systems can act in their environment and react to their environment.
23. I can explain how sensors enable computers to perceive the world.
24. I can name applications in which AI-assisted natural language processing/understanding is used.
25. I can describe the key ethical issues surrounding artificial intelligence.
26. I can explain what the term black box means in relation to artificial intelligence systems.
27. I can describe how algorithmic bias arises and what can be done about it.
28. I can critically reflect on the potential impact of artificial intelligence on the labor market.
29. I can give reasons why AI has become increasingly important since a few years.
30. I can explain how rule-based systems differ from learning systems.

31. I can assess if a problem can be solved with artificial intelligence methods (e.g., machine learning).
32. I can describe what artificial intelligence is.
33. I can describe the concept of explainable AI.
34. I can explain why data security must be considered when developing and using artificial intelligence applications & explain why data privacy must be considered when developing and using artificial intelligence applications.
35. I can describe the concept of big data.
36. I can give examples from my daily life (personal or professional) where I might be in contact with artificial intelligence.
37. I can explain what an algorithm is.
38. I can describe potential legal problems that may arise when using artificial intelligence.

A.2 AI Anxiety Questionnaire

Li et al. [96][†] designed this questionnaire to capture people's level of anxiety towards AI. Responses were measured through a 7-point Likert scale, ranging from strongly disagree to strongly agree.

1. I'm afraid that artificial intelligence (AI) will monitor my behavior.
2. I'm worried that AI will collect too much of my personal information.
3. AI's predictions of my preferences, such as well recommended ads or web pages, make me feel that my privacy is violated.
4. I am worried that AI will replace my work in the future.
5. I feel anxious working with AI that is smarter than me.
6. I'm worried that AI will replace many people's work.

[†]Li, J. and Huang, J.S., 2020. Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory. Technology in Society, 63, p.101410.

7. I do not think I would be able to perform well in professional courses in AI.
8. Understanding AI algorithms requires a high level of talent, which is difficult for me.
9. AI technology updates too quickly and is very difficult to learn.
10. AI may harm humans to achieve a goal, which gives me anxiety.
11. I worry that the control of AI by a few individuals will introduce great risks to the entire society.
12. The runaway of super AI will reduce the amount of time that humans stay on earth and will even result in human extinction, which is terrible.
13. I worry that humans have special feelings (such as love or adoration) for super AI.
14. I am disturbed that AI can deceive (for example, enticing people to buy goods).
15. I worry that AI will attain the same level of consciousness as humans.
16. The fact that AI that cannot tell the difference between humans and being conscious makes me uneasy.
17. AI has the same level of consciousness as humans, thus challenging the status of humans, which makes me anxious.
18. It's worrying if you do not know which part of AI has erred after AI makes a mistake.
19. I worry that people cannot figure out how AI makes decisions.
20. The responsibility for addressing operational failures in AI may be confusing.

A.3 Attitudes towards AI Questionnaire

Designed by Schepman et al. [158][‡], this instrument aimed to capture people's general positive and negative attitudes toward AI. The measures were designed on a 5-point Likert scale, varying from strongly disagree to strongly agree.

[‡]Schepman, A. and Rodway, P., 2020. Initial validation of the general attitudes towards Artificial Intelligence Scale. Computers in human behavior reports, 1, p.100014.

1. For routine transactions, I would rather interact with an artificially intelligent system than with a human.
2. Artificial Intelligence can provide new economic opportunities for this country.
3. Organizations use Artificial Intelligence unethically.
4. Artificially intelligent systems can help people feel happier.
5. I am impressed by what Artificial Intelligence can do.
6. I think artificially intelligent systems make many errors.
7. I am interested in using artificially intelligent systems in my daily life.
8. I find Artificial Intelligence sinister.
9. Artificial Intelligence might take control of people.
10. I think Artificial Intelligence is dangerous.
11. Artificial Intelligence can have positive impacts on people's wellbeing.
12. Artificial Intelligence is exciting.
13. I would be grateful if you could select agree.
14. An artificially intelligent agent would be better than an employee in many routine jobs.
15. There are many beneficial applications of Artificial Intelligence.
16. I shiver with discomfort when I think about future uses of Artificial Intelligence.
17. Artificially intelligent systems can perform better than humans.
18. Much of society will benefit from a future full of Artificial Intelligence
19. I would like to use Artificial Intelligence in my own job.
20. People like me will suffer if Artificial Intelligence is used more and more.
21. Artificial Intelligence is used to spy on people.

A.4 ATAI Questionnaire

Sindermann et al. [163][§] propose the Attitudes Towards AI (ATAI) questionnaire to capture people's general fear and acceptance sentiments of AI. An 11-point Likert scale (with responses ranging between strongly disagree to strongly agree) was used to measure these attitudes.

1. I fear artificial intelligence.
2. I trust artificial intelligence.
3. Artificial intelligence will destroy humankind.
4. Artificial intelligence will benefit humankind.
5. Artificial intelligence will cause many job losses.

A.5 The Big Five Personality Traits Questionnaire

This questionnaire, originally designed by McCrae and Costa [109] in 1987, captures people's personality trait across 5 different dimensions. That is, for each individual, the questionnaire provides varying levels of their Extroversion, Openness, Agreeableness, Conscientiousness, and Neuroticism. For the purposes of our study, we used a shorter version of this questionnaire by Lang et al. [93][¶] who showed it was similarly effective as the long version across all the data collection methods except for telephone interviewing. Compared to the original questionnaire, this short version used a scale with higher levels of granularity, i.e., a 7-point Likert scale, ranging from strongly disagree to strongly agree.

■ I see myself as someone who ···

1. ··· worries a lot.
2. ··· gets nervous easily.
3. ··· remains calm in tense situations.

[§]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, pp.109-118.

[¶]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, pp.548-567.

4. ⋯ is talkative
5. ⋯ is outgoing, sociable.
6. ⋯ is reserved.
7. ⋯ is original, comes up with new ideas.
8. ⋯ values artistic, aesthetic experiences.
9. ⋯ has an active imagination.
10. ⋯ is sometimes rude to others.
11. ⋯ has a forgiving nature.
12. ⋯ is considerate and kind to almost anyone.
13. ⋯ does a thorough job.
14. ⋯ tends to be lazy.
15. ⋯ does things efficiently.

REFERENCES

[1] Amina Adadi and Mohammed Berrada, *Peeking inside the black-box: A survey on explainable artificial intelligence (xai)*, IEEE Access 6 (2018), 52138–52160.

[2] Jonathan Aechtner, Lena Cabrera, Dennis Katwal, Pierre Onghena, Diego Penroz Valenzuela, and Anna Wilbik, *Comparing user perception of explanations developed with xai methods*, 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2022, pp. 1–7.

[3] Ankit Agrawal and Jane Cleland-Huang, *Explaining autonomous decisions in swarms of human-on-the-loop small unmanned aerial systems*, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 9, 2021, pp. 15–26.

[4] Yongsu Ahn and Yu-Ru Lin, *Fairsight: Visual analytics for fairness in decision making*, IEEE transactions on visualization and computer graphics (2019).

[5] E Alberdi, P Ayton, AA Povyakalo, and L Strigini, *Automation bias and system design: a case study in a medical application*, 2005 The IEE and MOD HFI DTC Symposium on People and Systems-Who Are We Designing For (Ref. No. 2005/11078), IET, 2005, pp. 53–60.

[6] José M Alonso, Alejandro Ramos-Soto, Ciro Castiello, and Corrado Mencar, *Explainable ai beer style classifier.*, SICSA ReaLX, 2018.

[7] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze, *Evaluating saliency map explanations for convolutional neural networks: a user study*, Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 275–285.

[8] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett, *Explaining reinforcement learning to mere mortals: An empirical study*, arXiv preprint arXiv:1903.09708 (2019).

[9] Claudio Aqueveque, *Ignorant experts and erudite novices: Exploring the dunning-kruger effect in wine consumers*, Food Quality and Preference 65 (2018), 181–184.

[10] Martin Atzmueller, Naveed Hayat, Matthias Trojahn, and Dennis Kroll, *Explicative human activity recognition using adaptive association rule-based classification*, 2018 IEEE International Conference on Future IoT Technologies (Future IoT), IEEE, 2018, pp. 1–6.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz, *Beyond accuracy: The role of mental models in human-ai team performance*, Proceedings of the AAAI conference on human computation and crowdsourcing, vol. 7, 2019, pp. 2–11.

[12] Bernard Barber, *The logic and limits of trust*, (1983).

[13] Richard Berk and Jordan Hyatt, *Machine learning forecasts of risk to inform sentencing decisions*, Federal Sentencing Reporter 27 (2015), no. 4, 222–228.

[14] Eric Bogert, Aaron Schecter, and Richard T Watson, *Humans rely more on algorithms than social influence as a task becomes more difficult*, Scientific reports 11 (2021), no. 1, 1–9.

[15] Christine L Borgman, *The user's mental model of an information retrieval system: An experiment on a prototype online catalog*, International Journal of man-machine studies 24 (1986), no. 1, 47–64.

[16] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki, *Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users*, 27th international conference on intelligent user interfaces, 2022, pp. 807–819.

[17] Clint A Bowers, Florian Jentsch, Eduardo Salas, and Curt C Braun, *Analyzing communication sequences for team training needs assessment*, Human factors 40 (1998), no. 4, 672–679.

[18] Clint A Bowers, Randall L Oser, Eduardo Salas, and Janis A Cannon-Bowers, *Team performance in automated systems*, Automation and human performance: Theory and applications (1996), 243–263.

[19] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan, *The role of explanations on trust and reliance in clinical decision support systems*, 2015 International Conference on Healthcare Informatics, IEEE, 2015, pp. 160–169.

[20] Carrie J Cai, Jonas Jongejan, and Jess Holbrook, *The effects of example-based explanations in a machine learning interface*, Proceedings of the 24th International Conference on Intelligent User Interfaces, ACM, 2019, pp. 258–262.

[21] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al., *Human-centered tools for coping with imperfect algorithms during medical decision-making*, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.

[22] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry, *" hello ai": Uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making*, Proceedings of the ACM on Human-Computer Interaction 3 (2019), no. CSCW, 1–24.

[23] Wanling Cai, Yucheng Jin, and Li Chen, *Impacts of personal characteristics on user trust in conversational recommender systems*, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–14.

[24] Ethem F Can and Aysu Ezen-Can, *The effect of data ordering in image classification*, arXiv preprint arXiv:2001.05857 (2020).

[25] Quan Ze Chen, Tobias Schnabel, Besmira Nushi, and Saleema Amershi, *Hint: Integration testing for ai-based features with humans in the loop*, 27th International Conference on Intelligent User Interfaces, 2022, pp. 549–565.

[26] Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Steven Wu, and Haiyi Zhu, *Soliciting stakeholders' fairness notions in child maltreatment predictive systems*, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–17.

[27] Zhiyong Cheng, Xiaojun Chang, Lei Zhu, Rose C Kanjirathinkal, and Mohan Kankanhalli, *Mmalfm: Explainable recommendation by leveraging reviews and images*, ACM Transactions on Information Systems (TOIS) 37 (2019), no. 2, 16.

[28] Chun-Wei Chiang and Ming Yin, *Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models*, 27th International Conference on Intelligent User Interfaces, 2022, pp. 148–161.

[29] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz, *I think i get your point, ai! the illusion of explanatory depth in explainable ai*, 26th International Conference on Intelligent User Interfaces, 2021, pp. 307–317.

[30] Mary Cummings, *Automation bias in intelligent time critical decision support systems*, AIAA 1st Intelligent Systems Technical Conference, 2004, p. 6313.

[31] Devleena Das and Sonia Chernova, *Leveraging rationales to improve human task performance*, Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 510–518.

[32] Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert A Lafrance, Nick Cramer, Kristin Cook, and Samuel Payne, *Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods*, IEEE transactions on visualization and computer graphics 23 (2016), no. 1, 271–280.

[33] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman, *The world is not enough: Trust in cognitive agents*, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 56 (2012), no. 1, 263–267.

[34] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco, *Impact of robot failures and feedback on real-time trust*, 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2013, pp. 251–258.

[35] Sanorita Dey, Brittany RL Duff, Niyati Chhaya, Wai Fu, Vishy Swaminathan, and Karrie Karahalios, *Recommendation for video advertisements based on personality traits and companion content*, Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 144–154.

[36] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey, *Algorithm aversion: People erroneously avoid algorithms after seeing them err.*, Journal of Experimental Psychology: General 144 (2015), no. 1, 114.

[37] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic, *A task-based taxonomy of cognitive biases for information visualization*, IEEE transactions on visualization and computer graphics 26 (2018), no. 2, 1413–1432.

[38] Alexandra Dimitroff, *Mental models and error behavior in an interactive bibliographic retrieval system.*, (1992).

[39] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan, *Explaining models: An empirical study of how explanations impact fairness judgment*, Proceedings of the 24th International Conference on Intelligent User Interfaces (New York, NY, USA), IUI '19, Association for Computing Machinery, 2019, p. 275–285.

[40] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan, *Explaining models: an empirical study of how explanations impact fairness judgment*, Proceedings of the 24th international conference on intelligent user interfaces, 2019, pp. 275–285.

[41] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra, *The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images*, Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 408–416.

[42] Finale Doshi-Velez and Been Kim, *Towards a rigorous science of interpretable machine learning*, arXiv preprint arXiv:1702.08608 (2017).

[43] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su, *Trust in automl: exploring information needs for establishing trust in automated machine learning systems*, Proceedings of the 25th international conference on intelligent user interfaces, 2020, pp. 297–307.

[44] Mengnan Du, Ninghao Liu, and Xia Hu, *Techniques for interpretable machine learning*, Communications of the ACM 63 (2019), no. 1, 68–77.

[45] David J Duke, Ken W Brodlie, David A Duce, and Ivan Herman, *Do you see what i mean?[data visualization]*, IEEE Computer Graphics and Applications 25 (2005), no. 3, 6–9.

[46] David Dunning, *The dunning–kruger effect: On being ignorant of one's own ignorance*, Advances in experimental social psychology, vol. 44, Elsevier, 2011, pp. 247–296.

[47] _____, *Confidence considered: Assessing the quality of decisions and performance*, Social metacognition (2012), 63–80.

[48] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al., *The who in explainable ai: How ai background shapes perceptions of ai explanations*, arXiv preprint arXiv:2107.13509 (2021).

[49] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann, *Bringing transparency design into practice*, 23rd International Conference on Intelligent User Interfaces, ACM, 2018, pp. 211–223.

[50] Eva Fourakis and Jeremy Cone, *Matters order: The role of information order on implicit impression formation*, Social Psychological and Personality Science 11 (2020), no. 1, 56–63.

[51] Jay Friedenberg and Gordon Silverman, *Mind as a black box: The behaviorist approach*, Cognitive science: An introduction to the study of mind (2006), 85–88.

[52] Xiaofeng Gao, Ran Gong, Yizhou Zhao, Shu Wang, Tianmin Shu, and Song-Chun Zhu, *Joint mind modeling for explanation generation in complex human-robot collaborative tasks*, 2020 29th IEEE international conference on robot and human interactive communication (RO-MAN), IEEE, 2020, pp. 1119–1126.

[53] Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie, *Explainable autonomy: A study of explanation styles for building clear mental models*, Proceedings of the 11th International Conference on Natural Language Generation, 2018, pp. 99–108.

[54] David Good, *Individuals, interpersonal relations, and trust*, Trust: Making and breaking cooperative relations (2000), 31–48.

[55] Harsh Goyal, Deepanshu Khandelwal, Aayush Aggarwal, and Piyush Bhardwaj, *Medical diagnosis using machine learning*, Bhagwan Parshuram Inst Technol 7 (2018).

[56] Nina Grgić-Hlača, Claude Castelluccia, and Krishna P Gummadi, *Taking advice from (dis) similar machines: The impact of human-machine similarity on machine-assisted decision-making*, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 10, 2022, pp. 74–88.

[57] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi, *A survey of methods for explaining black box models*, ACM computing surveys (CSUR) 51 (2018), no. 5, 1–42.

[58] David Gunning, *Explainable artificial intelligence (xai)*, Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2017), 2.

[59] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg, *Building trust in interactive machine learning via user contributed interpretable rules*, 27th International Conference on Intelligent User Interfaces, 2022, pp. 537–548.

[60] Wenkai Han and Hans-Jörg Schulz, *Beyond trust building—calibrating trust in visual analytics*, 2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX), IEEE, 2020, pp. 9–15.

[61] Adam W Harley, *An interactive node-link visualization of convolutional neural networks*, International Symposium on Visual Computing, Springer, 2015, pp. 867–877.

[62] Alexander Heimerl, Katharina Weitz, Tobias Baur, and Elisabeth André, *Unraveling ml models of emotion with nova: Multi-level explainable ai for non-experts*, IEEE Transactions on Affective Computing 13 (2020), no. 3, 1155–1167.

[63] Richards J Heuer, *Psychology of intelligence analysis*, Center for the Study of Intelligence, 1999.

[64] Kevin Anthony Hoff and Masooda Bashir, *Trust in automation: Integrating empirical evidence on factors that influence trust*, Human factors 57 (2015), no. 3, 407–434.

[65] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink, *Trust in automation*, IEEE Intelligent Systems 28 (2013), no. 1, 84–88.

[66] Robert R Hoffman, Gary Klein, and Shane T Mueller, *Explaining explanation for "explainable ai"*, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 62 (2018), no. 1, 197–201.

[67] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman, *Metrics for explainable ai: Challenges and prospects*, arXiv preprint arXiv:1812.04608 (2018).

[68] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau, *Visual analytics in deep learning: An interrogative survey for the next frontiers*, IEEE transactions on visualization and computer graphics 25 (2018), no. 8, 2674–2693.

[69] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau, *S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations*, IEEE transactions on visualization and computer graphics 26 (2019), no. 1, 1096–1106.

[70] Donald Honeycutt, Mahsan Nourani, and Eric Ragan, *Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy*, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 8, 2020, pp. 63–72.

[71] Donald R. Honeycutt, Mahsan Nourani, and Eric D. Ragan, *Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy*, Eighth AAAI Conference on Human Computation and Crowdsourcing, 2020.

[72] Eric Horvitz, *Mixed-initiative interaction*, IEEE Intelligent Systems (1999), 14–24.

[73] Wolfgang Jentner, Dominik Sacha, Florian Stoffel, Geoffrey Ellis, Leishi Zhang, and Daniel A Keim, *Making machine intelligence less scary for criminal analysts: reflections on designing a visual comparative case analysis tool*, The Visual Computer 34 (2018), no. 9, 1225–1241.

[74] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury, *Foundations for an empirically determined scale of trust in automated systems*, International journal of cognitive ergonomics 4 (2000), no. 1, 53–71.

[75] Oliver P John, Sanjay Srivastava, et al., *The big-five trait taxonomy: History, measurement, and theoretical perspectives*, (1999).

[76] Minsuk Kahng, Pierre Y Andrews, Aditya Kalro, and Duen Horng Polo Chau, *A cti v is: Visual exploration of industry-scale deep neural network models*, IEEE transactions on visualization and computer graphics 24 (2017), no. 1, 88–97.

[77] Graham Kalton, *The treatment of missing survey data*, Survey methodology 12 (1986), 1–16.

[78] Tharindu Kaluarachchi, Andrew Reis, and Suranga Nanayakkara, *A review of recent deep learning approaches in human-centered machine learning*, Sensors 21 (2021), no. 7, 2514.

[79] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan, *Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning*, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[80] Mark T Keane and Eoin M Kenny, *How case based reasoning explained neural networks: An xai survey of post-hoc explanation-by-example in ann-cbr twins*, arXiv preprint arXiv:1905.07186 (2019).

[81] Frank C Keil, *Explanation and understanding*, Annu. Rev. Psychol. 57 (2006), 227–254.

[82] Eoin M Kenny, Courtney Ford, Molly Quinn, and Mark T Keane, *Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies*, Artificial Intelligence 294 (2021), 103459.

[83] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi, *Reasons for physicians not adopting clinical decision support systems: critical analysis*, JMIR medical informatics 6 (2018), no. 2, e24.

[84] Zafar A. Khan and Won Sohn, *Abnormal human activity recognition system based on r-transform and kernel discriminant technique for elderly home care*, IEEE Transactions on Consumer Electronics 57 (2011), no. 4, 1843–1850.

[85] Chris Kim, Xiao Lin, Christopher Collins, Graham W Taylor, and Mohamed R Amer, *Learn, generate, rank, explain: A case study of visual explanation by generative machine learning*, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (2021), no. 3-4, 1–34.

[86] Youngwoo Kim and James Allan, *Unsupervised explainable controversy detection from online news*, European Conference on Information Retrieval, Springer, 2019, pp. 836–843.

[87] Ajay Kohli and Saurabh Jha, *Why cad failed in mammography*, Journal of the American College of Radiology 15 (2018), no. 3, 535–537.

[88] Moritz Körber, Lorenz Prasch, and Klaus Bengler, *Why do i have to drive now? post hoc explanations of takeover requests*, Human factors 60 (2018), no. 3, 305–323.

[89] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor, *Personalized explanations for hybrid recommender systems*, Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 379–390.

[90] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong, *Too much, too little, or just right? ways explanations impact end users' mental models*, 2013 IEEE Symposium on Visual Languages and Human Centric Computing, IEEE, 2013, pp. 3–10.

[91] Tai Yu Lai, Jong Yih Kuo, Yong-Yi Fanjiang, Shang-Pin Ma, and Yi Han Liao, *Robust little flame detection on real-time video surveillance system*, 2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications, IEEE, Sep 2012, p. 139–143.

[92] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan, *Towards a science of human-ai decision making: A survey of empirical studies*, arXiv preprint arXiv:2112.11471 (2021).

[93] Frieder R Lang, Dennis John, Oliver Lüdtke, Jürgen Schupp, and Gert G Wagner, *Short assessment of the big five: Robust across survey methods except telephone interviewing*, Behavior research methods 43 (2011), 548–567.

[94] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki, *The dangers of post-hoc interpretability: Unjustified counterfactual explanations*, arXiv preprint arXiv:1907.09294 (2019).

[95] Matthias Carl Laupichler, Alexandra Aster, and Tobias Raupach, *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 (2023), 100126.

[96] Jian Li and Jin-Song Huang, *Dimensions of artificial intelligence anxiety based on the integrated fear acquisition theory*, Technology in Society 63 (2020), 101410.

[97] Mengqi Liao and S Shyam Sundar, *How should ai systems talk to users when collecting their personal information? effects of role framing and self-referencing on human-ai interaction*, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–14.

[98] Mengqi Liao, S Shyam Sundar, and Joseph B. Walther, *User trust in recommendation systems: A comparison of content-based, collaborative and demographic filtering*, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–14.

[99] Q Vera Liao, Daniel Gruen, and Sarah Miller, *Questioning the ai: Informing design practices for explainable ai user experiences*, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–15.

[100] Brian Y Lim and Anind K Dey, *Investigating intelligibility for uncertain context-aware applications*, Proceedings of the 13th international conference on Ubiquitous computing, 2011, pp. 415–424.

[101] Martin Lindvall, Claes Lundström, and Jonas Löwgren, *Rapid assisted visual search: Supporting digital pathologists with imperfect ai*, 26th International Conference on Intelligent User Interfaces, 2021, pp. 504–513.

[102] Zachary C Lipton, *The mythos of model interpretability*, Queue 16 (2018), no. 3, 31–57.

[103] Tania Lombrozo, *Simplicity and probability in causal explanation*, Cognitive psychology 55 (2007), no. 3, 232–257.

[104] Duri Long and Brian Magerko, *What is ai literacy? competencies and design considerations*, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (New York, NY, USA), CHI '20, Association for Computing Machinery, 2020, p. 1–16.

[105] Niklas Luhmann, *Trust and power*, John Wiley & Sons, 2018.

[106] Shuai Ma, Mingfei Sun, and Xiaojuan Ma, *Modeling adaptive expression of robot learning engagement and exploring its effects on human teachers*, ACM Transactions on Computer-Human Interaction (2022).

[107] Poornima Madhavan and Douglas A Wiegmann, *Similarities and differences between human–human and human–automation trust: an integrative review*, Theoretical Issues in Ergonomics Science 8 (2007), no. 4, 277–301.

[108] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere, *Explainable reinforcement learning through a causal lens*, arXiv preprint arXiv:1905.10958 (2019).

[109] Robert R McCrae and Paul T Costa, *Validation of the five-factor model of personality across instruments and observers.*, Journal of personality and social psychology 52 (1987), no. 1, 81.

[110] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Frederick Tung, and Leonid Sigal, *Where and when to look? spatio-temporal attention for action recognition in videos*, arXiv preprint arXiv:1810.04511 (2018).

[111] Stephanie M Merritt, Deborah Lee, Jennifer L Unnerstall, and Kelli Huber, *Are well-calibrated users effective users? associations between calibration of trust and performance on an automation-aided task*, Human Factors 57 (2015), no. 1, 34–47.

[112] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert, *To explain or not to explain: the effects of personal characteristics when explaining music recommendations*, Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 397–407.

[113] Tim Miller, *Explanation in artificial intelligence: Insights from the social sciences*, Artificial Intelligence 267 (2019), 1–38.

[114] Tim Miller, Piers Howe, and Liz Sonenberg, *Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences*, arXiv preprint arXiv:1712.00547 (2017).

[115] Sina Mohseni, Niloofar Zarei, and Eric D Ragan, *A survey of evaluation methods and measures for interpretable machine learning*, ACM Transactions on Interactive Intelligent Systems (2018).

[116] _____, *A multidisciplinary survey and framework for design and evaluation of explainable ai systems*, arXiv preprint arXiv:1811.11839 (2019).

[117] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, *Methods for interpreting and understanding deep neural networks*, Digital Signal Processing 73 (2018), 1–15.

[118] Don A Moore and Paul J Healy, *The trouble with overconfidence.*, Psychological review 115 (2008), no. 2, 502.

[119] Kathleen L Mosier and Linda J Skitka, *Automation use and automation bias*, Proceedings of the human factors and ergonomics society annual meeting, vol. 43, SAGE Publications Sage CA: Los Angeles, CA, 1999, pp. 344–348.

[120] Bonnie M Muir, *Trust between humans and machines, and the design of decision aids*, International journal of man-machine studies 27 (1987), no. 5-6, 527–539.

[121] Chelsea M Myers, Evan Freed, Luis Fernando Laris Pardo, Anushay Furqan, Sebastian Risi, and Jichen Zhu, *Revealing neural network bias to non-experts through interactive counterfactual examples*, arXiv preprint arXiv:2001.02271 (2020).

[122] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli, *Toward involving end-users in interactive human-in-the-loop ai fairness*, ACM Transactions on Interactive Intelligent Systems (TiiS) 12 (2022), no. 3, 1–30.

[123] Clifford Nass and Kwan Min Lee, *Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction.*, Journal of experimental psychology: applied 7 (2001), no. 3, 171.

[124] Mario Popolin Neto and Fernando V Paulovich, *Explainable matrix—visualization for global and local interpretability of random forest classification ensembles*, IEEE Transactions on Visualization and Computer Graphics (2020).

[125] Raymond S Nickerson, *Confirmation bias: A ubiquitous phenomenon in many guises*, Review of general psychology 2 (1998), no. 2, 175–220.

[126] Donald A Norman, *Design rules based on analyses of human error*, Communications of the ACM 26 (1983), no. 4, 254–258.

[127] Mahsan Nourani, Donald R Honeycutt, Jeremy E Block, Chiradeep Roy, Tahrima Rahman, Eric D Ragan, and Vibhav Gogate, *Investigating the importance of first impressions and explainable ai with interactive video analysis*, Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2020, pp. 1–8.

[128] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan, *The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems*, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7 (2019), no. 1, 97–105.

[129] Mahsan Nourani, Joanie T. King, and Eric D. Ragan, *The role of domain expertise in user trust and the impact of first impressions with intelligent systems*, Eighth AAAI Conference on Human Computation and Crowdsourcing (HCOMP), 2020.

[130] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate, *Anchoring bias affects mental model formation and user reliance in explainable ai systems*, 26th International Conference on Intelligent User Interfaces, 2021, pp. 340–350.

[131] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric D Ragan, and Vibhav Gogate, *Anchoring bias affects mental models and user reliance in explainable ai systems*, Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI), 2021.

[132] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D Ragan, Nicholas Ruozzi, and Vibhav Gogate, *Don't explain without verifying veracity: An evaluation of explainable ai with video activity recognition*, arXiv preprint arXiv:2005.02335 (2020).

[133] TRF Oaster, *Number of alternatives per choice point and stability of likert-type scales*, Perceptual and Motor Skills 68 (1989), no. 2, 549–550.

[134] James M Oglesby, Kimberly Stowers, Kevin Leyva, Aaron Dietz, Shirley Sonesh, Shawn Burke, and Eduardo Salas, *Assessing human-automation system safety, efficiency, and performance: Developing a metrics framework*, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 58, SAGE Publications Sage CA: Los Angeles, CA, 2014, pp. 1149–1153.

[135] Daniel Omeiza, Konrad Kollnig, Helena Web, Marina Jirotka, and Lars Kunze, *Why not explain? effects of explanations on human perceptions of autonomous driving*, 2021 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO), IEEE, 2021, pp. 194–199.

[136] Jeroen Ooge, Shotallo Kato, and Katrien Verbert, *Explaining recommendations in e-learning: Effects on adolescents' trust*, 27th International Conference on Intelligent User Interfaces, 2022, pp. 93–105.

[137] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi, *Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems*, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–9.

[138] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert, *How model accuracy and explanation fidelity influence user trust*, arXiv preprint arXiv:1907.12652 (2019).

[139] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert, *It's complicated: The relationship between user trust, model accuracy and explanations in ai*, ACM Transactions on Computer-Human Interaction (TOCHI) 29 (2022), no. 4, 1–33.

[140] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert, *How accurate does it feel?–human perception of different types of classification mistakes*, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–13.

[141] Raja Parasuraman and Victor Riley, *Humans and automation: Use, misuse, disuse, abuse*, Human factors 39 (1997), no. 2, 230–253.

[142] Alyssa M Pena, Ehsanul Haque Nirjhar, Andrew Pachuilo, Theodora Chaspari, and Eric D Ragan, *Detecting changes in user behavior to understand interaction provenance during visual data analysis.*, IUI Workshops, 2019.

[143] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen, Siddharth Suri, and Ece Kamar, *What you see is what you get? the impact of representation criteria on human bias in hiring*, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 7, 2019, pp. 125–134.

[144] Bjorn Petrak, Katharina Weitz, Ilhan Aslan, and Elisabeth Andre, *Let me show you your new home: studying the effect of proxemic-awareness of robots on users' first impressions*, 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), IEEE, 2019, pp. 1–7.

[145] Nicola Pezzotti, Thomas Höllt, Jan Van Gemert, Boudewijn PF Lelieveldt, Elmar Eisemann, and Anna Vilanova, *Deepeyes: Progressive visual analytics for designing deep neural networks*, IEEE transactions on visualization and computer graphics 24 (2017), no. 1, 98–108.

[146] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach, *Manipulating and measuring model interpretability*, Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–52.

[147] Tahrima Rahman, Prasanna Kothalkar, and Vibhav Gogate, *Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of chow-liu trees*, Joint European conference on machine learning and knowledge discovery in databases, Springer, 2014, pp. 630–645.

[148] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati, *Toward effective robot–child tutoring: Internal motivation, behavioral intervention, and learning outcomes*, ACM Transactions on Interactive Intelligent Systems (TiiS) 9 (2019), no. 1, 1–23.

[149] Arthur S Reber, *Implicit learning and tacit knowledge.*, Journal of experimental psychology: General 118 (1989), no. 3, 219.

[150] John K Rempel, John G Holmes, and Mark P Zanna, *Trust in close relationships.*, Journal of personality and social psychology 49 (1985), no. 1, 95.

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

[152] _____, *Anchors: High-precision model-agnostic explanations*, Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[153] Caleb Robinson, Fred Hohman, and Bistra Dilkina, *A deep learning approach for population estimation from satellite imagery*, Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities, 2017, pp. 47–54.

[154] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele, *Coherent multi-sentence video description with variable level of detail*, German conference on pattern recognition, Springer, 2014, pp. 184–195.

[155] Chiradeep Roy, Mahesh Shanbhag, Mahsan Nourani, Tahrima Rahman, Samia Kabir, Vibhav Gogate, Nicholas Ruozzi, and Eric D Ragan, *Explainable activity recognition in videos.*, IUI Workshops, 2019.

[156] Cynthia Rudin and Berk Ustun, *Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice*, Interfaces 48 (2018), no. 5, 449–466.

[157] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer, *I can do better than your ai: expertise and explanations*, Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 240–251.

[158] Astrid Schepman and Paul Rodway, *Initial validation of the general attitudes towards artificial intelligence scale*, Computers in human behavior reports 1 (2020), 100014.

[159] Norbert Schwarz, Herbert Bless, Fritz Strack, Gisela Klumpp, Helga Rittenauer-Schatka, and Annette Simons, *Ease of retrieval as information: another look at the availability heuristic.*, Journal of Personality and Social psychology 61 (1991), no. 2, 195.

[160] TB Sheridan, B Fischhoff, M Posner, and RW Pew, *Supervisory control system. research needs for human factors*, 1983.

[161] Thomas B Sheridan and Robert T Hennessy, *Research and modeling of supervisory control behavior. report of a workshop*, Tech. report, NATIONAL RESEARCH COUNCIL WASHINGTON DC COMMITTEE ON HUMAN FACTORS, 1984.

[162] Ben Shneiderman, *The eyes have it: A task by data type taxonomy for information visualizations*, The craft of information visualization, Elsevier, 2003, pp. 364–371.

[163] Cornelia Sindermann, Peng Sha, Min Zhou, Jennifer Wernicke, Helena S Schmitt, Mei Li, Rayna Sariyska, Maria Stavrou, Benjamin Becker, and Christian Montag, *Assessing the attitude towards artificial intelligence: Introduction of a short measure in german, chinese, and english language*, KI-Künstliche intelligenz 35 (2021), 109–118.

[164] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman, *Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence*, ACM Transactions on Computer-Human Interaction (2022).

[165] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady, *explainer: A visual analytics framework for interactive and explainable machine learning*, IEEE transactions on visualization and computer graphics (2019).

[166] Aaron Springer and Steve Whittaker, *Progressive disclosure: empirically motivated approaches to designing effective transparency*, Proceedings of the 24th international conference on intelligent user interfaces, 2019, pp. 107–120.

[167] Nancy Staggers and Anthony F. Norcio, *Mental models: concepts for human-computer interaction research*, International Journal of Man-machine studies 38 (1993), no. 4, 587–605.

[168] Florian Stoffel, Hannah Post, Marcus Stewen, and Daniel A Keim, *polimaps: supporting predictive policing with visual analytics*, EuroVA 2018: EuroVis Workshop on Visual Analytics, 2018, pp. 43–48.

[169] Stefan Strauß, *Deep automation bias: How to tackle a wicked problem of ai?*, Big Data and Cognitive Computing 5 (2021), no. 2, 18.

[170] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M Rush, *S eq 2s eq-v is: A visual debugging tool for sequence-to-sequence models*, IEEE transactions on visualization and computer graphics 25 (2018), no. 1, 353–363.

[171] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan, *Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs*, arXiv preprint arXiv:2101.09824 (2021).

[172] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, *Going deeper with convolutions*, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[173] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert, *Visual, textual or hybrid: the effect of user expertise on different explanations*, 26th International Conference on Intelligent User Interfaces, 2021, pp. 109–119.

[174] Dairazalia Sánchez, Monica Tentori, and Favela Jesús, *Activity recognition for the smart hospital*, IEEE Intelligent Systems 23 (2008), no. 02, 50–57.

[175] Philip E Tetlock, *Accountability and the perseverance of first impressions*, Social Psychology Quarterly (1983), 285–292.

[176] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal, *Suspicious human activity recognition: a review*, Artificial Intelligence Review 50 (2018), no. 2, 283–339.

[177] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M Carroll, *Exploring and promoting diagnostic transparency and explainability in online symptom checkers*, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–17.

[178] Preethi Vaidyanathan, Jeff Pelz, Cecilia Alm, Pengcheng Shi, and Anne Haake, *Recurrence quantification analysis reveals eye-movement behavior differences between experts and novices*, Proceedings of the symposium on eye tracking research and applications, 2014, pp. 303–306.

[179] Gilles Vandewiele, Olivier Janssens, Femke Ongenae, Filip De Turck, and Sofie Van Hoecke, *Genesim: genetic extraction of a single, interpretable model*, arXiv preprint arXiv:1611.05722 (2016).

[180] Jennifer Wortman Vaughan and Hanna Wallach, *A human-centered agenda for intelligible machine learning*, Machines We Trust: Getting Along with Artificial Intelligence (2020).

[181] Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert, *A formative study of interactive bias metrics in visual analytics using anchoring bias*, IFIP Conference on Human-Computer Interaction, Springer, 2019, pp. 555–575.

[182] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert, *Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics*, 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), IEEE, 2017, pp. 104–115.

[183] Emily Wall, Leslie M Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert, *Four perspectives on human bias in visual analytics*, Cognitive biases in visualizations, Springer, 2018, pp. 29–42.

[184] Clarice Wang, Kathryn Wang, Andrew Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan, *Do humans prefer debiased ai algorithms? a case study in career recommendation*, 27th International Conference on Intelligent User Interfaces, 2022, pp. 134–147.

[185] Dakuo Wang, Josh Andres, Justin D Weisz, Erick Oduor, and Casey Dugan, *Autods: Towards human-centered automation of data science*, Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–12.

[186] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim, *Designing theory-driven user-centric explainable ai*, Proceedings of the 2019 CHI conference on human factors in computing systems, ACM, 2019, pp. 1–15.

[187] Junpeng Wang, Liang Gou, Wei Zhang, Hao Yang, and Han-Wei Shen, *Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation*, IEEE transactions on visualization and computer graphics 25 (2019), no. 6, 2168–2180.

[188] Xiang Wang, Xiangnan He, Fuli Feng, Liqiang Nie, and Tat-Seng Chua, *Tem: Tree-enhanced embedding model for explainable recommendation*, Proceedings of the 2018 World Wide Web Conference, International World Wide Web Conferences Steering Committee, 2018, pp. 1543–1552.

[189] Xinru Wang and Ming Yin, *Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making*, 26th International Conference on Intelligent User Interfaces, 2021, pp. 318–328.

[190] Christine T Wolf, *Explainability scenarios: towards scenario-based xai design*, Proceedings of the 24th International Conference on Intelligent User Interfaces, ACM, 2019, pp. 252–257.

[191] Bingjun Xie and Jia Zhou, *The influence of mental model similarity on user performance: Comparing older and younger adults*, Human Aspects of IT for the Aged Population. Applications, Services and Contexts (Cham) (Jia Zhou and Gavriel Salvendy, eds.), Springer International Publishing, 2017, pp. 569–579.

[192] Bingjun Xie, Jia Zhou, and Huilin Wang, *How influential are mental models on interaction performance? exploring the gap between users' and designers' mental models through a new quantitative method*, Advances in Human-Computer Interaction 2017 (2017).

[193] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt, *How do visual explanations foster end users' appropriate trust in machine learning?*, Proceedings of the 25th International Conference on Intelligent User Interfaces, 2020, pp. 189–201.

[194] Serena Yeung, Francesca Rinaldo, Jeffrey Jopling, Bingbin Liu, Rishab Mehra, N. Lance Downing, Michelle Guo, Gabriel M. Bianconi, Alexandre Alahi, Julia Lee, and et al., *A computer vision system for deep learning-based detection of patient mobilization activities in the icu*, npj Digital Medicine 2 (2019), no. 11, 1–5.

[195] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach, *Understanding the effect of accuracy on trust in machine learning models*, Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–12.

[196] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen, *User trust dynamics: An investigation driven by differences in system performance*, Proceedings of the 22nd International Conference on Intelligent User Interfaces, 2017, pp. 307–317.

[197] Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor, *Graying the black box: Understanding dqns*, International Conference on Machine Learning, 2016, pp. 1899–1908.

[198] Leslie A Zebrowitz, *First impressions from faces*, Current directions in psychological science 26 (2017), no. 3, 237–242.

[199] John R Zech, Jessica Zosa Forde, and Michael L Littman, *Individual predictions matter: Assessing the effect of data ordering in training fine-tuned cnns for medical imaging*, arXiv preprint arXiv:1912.03606 (2019).

[200] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S Ebert, *Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models*, IEEE transactions on visualization and computer graphics 25 (2018), no. 1, 364–373.

[201] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick, *" an ideal human" expectations of ai teammates in human-ai teaming*, Proceedings of the ACM on Human-Computer Interaction 4 (2021), no. CSCW3, 1–25.

[202] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma, *Explicit factor models for explainable recommendation based on phrase-level sentiment analysis*, Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM, 2014, pp. 83–92.

[203] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy, *Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making*, arXiv preprint arXiv:2001.02114 (2020).

[204] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li, *Storybuddy: A human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement*, Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–21.

[205] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba, *Temporal relational reasoning in videos*, Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 803–818.

[206] Martina Ziefle and Susanne Bay, *Mental models of a cellular phone menu. comparing older and younger novice users*, International Conference on Mobile Human-Computer Interaction, Springer, 2004, pp. 25–37.

BIOGRAPHICAL SKETCH

Mahsan Nourani is an interdisciplinary researcher whose work lies at the intersection of Artificial Intelligence (AI) and Human-Computer Interaction (HCI). She received her computer science Ph.D. degree from the University of Florida in the Summer of 2023, focusing on the human-centered roles of AI systems in the society, particularly through studying how people understand, collaborate, and interact with AI-supported systems. In her doctoral research, she has studied various psychological and sociotechnical aspects of people in the context of AI technology, such as user biases, mental models, trust, and prior experiences. During this time, her research was generously supported by DARPA XAI and NSF grants, and has earned her various accolades, including a Best Paper, Honorable Mention Award from ACM Intelligent User Interfaces (IUI) conference (2021); Gartner Group Graduate Fellowships (2021, 2022); and an Outstanding Achievement Award from the Herbert Wertheim College of Engineering.

She is a co-founder and organizer of the Workshop on TRust and EXpertise (TREX) in Visualization at IEEE Visualization and Visual Analytics Conference. During her graduate school, she has collaborated with researchers, engineers, and practitioners at Microsoft Research and Apple Technology Development Group, as well as other academic institutes. She earned her Master of Science degree in computer science from the University of Florida in 2021 and her Bachelor of Engineering degree in information technology (computer engineering) from University of Tehran in 2017.