Towards Effective Human-AI Collaboration

Promoting Appropriate Reliance on AI Systems

He, G.

DOI
10.4233/uuid:b5862ede-deca-4f6a-bcc7-9674051daf58

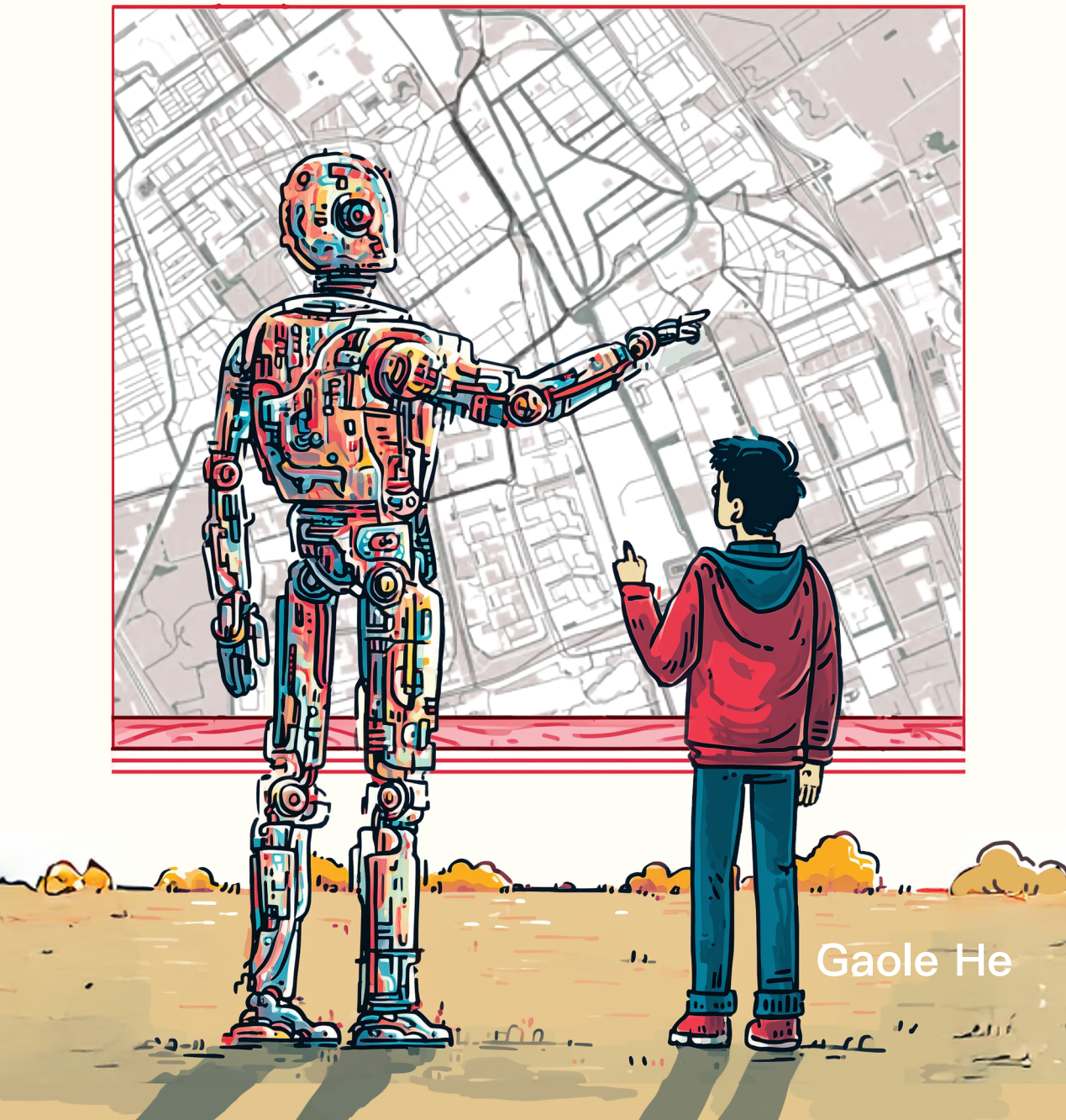
Publication date
2025

Document Version
Final published version

Citation (APA)
He, G. (2025). *Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:b5862ede-deca-4f6a-bcc7-9674051daf58

Towards Effective Human-AI Collaboration:

Promoting Appropriate Reliance on AI Systems



Gaole He

Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems

Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on Thursday 23, October 2025 at 12:30 o'clock

by

Gaole HE

Master of Engineering in Computer Application Technology,
Renmin University of China, China,
born in Chongqing, China.

This Dissertation has been approved by the promotors.

Composition of the doctoral Committee:

Rector Magnificus,	Chairperson
Prof. dr. ir. G.J.P.M. Houben,	Delft University of Technology, Promotor
Dr. ir. U.K. Gadiraju,	Delft University of Technology, Copromotor

Independent members:

Prof. dr. ir. D.N. Nas,	Delft University of Technology
Prof. dr. P.A. Lloyd,	Delft University of Technology
Prof. dr. T. Miller,	The University of Queensland, Australia
Dr. Q.V. Liao,	University of Michigan, United States of America
Dr. ir. M.C. Willemsen,	Eindhoven University of Technology
Dr. A. Anand,	Delft University of Technology, reserve member

SIKS Dissertation Series No. 2025-53
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Appropriate Reliance, Human-AI Collaboration, Human-AI Decision Making, Human-centered AI system, Empirical Studies
Cover: Cover Designed by studio shu | shu.artdesign@gmail.com
Style: TU Delft House Style, with modifications by Moritz Beller
`https://github.com/Inventitech/phd-thesis-template`

ISBN 978-94-6518-106-6

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

Contents

Summary

As AI technologies gain widespread acceptance across society, human-AI collaboration has emerged as a promising avenue to enhance the accountability and reliability of task outcomes where AI is used in task completion. Although AI systems are advancing rapidly, most people in society – particularly laypeople – still lack sufficient understanding and experience in collaborating with them. This gap becomes a barrier when interacting with deep learning-based AI systems, where users often struggle to assess the trustworthiness of AI advice. Consequently, individuals may develop uncalibrated trust or misperceptions about AI capabilities, hindering appropriate reliance and degrading overall team performance. Empirical studies have shown that human-AI teams often underperform compared to AI systems operating alone, highlighting that current human-AI collaboration remains suboptimal. These observations underscore a substantial need to advance our understanding of fostering effective human-AI collaboration.

This dissertation contributes to the growing literature on human-AI collaboration by analyzing potential approaches to promoting appropriate reliance. Specifically, to ensure effective human-AI collaboration, we aim to achieve both reliable task outcomes and a positive, engaging user experience. Through a series of empirical studies, we explored promoting appropriate reliance by calibrating user perception of competence (Part I), improving user understanding of AI systems with human-centered explainable AI (Part II), and enhancing user control with collaborative workflows (Part III). Our findings confirm that an uncalibrated perception of AI competence and self-competence can be a cause to trigger over-reliance and under-reliance, respectively. Additionally, we observed that both XAI methods (*e.g.,* analogy-based explanation) and interactive XAI interfaces (*e.g.,* conversational XAI interfaces) may induce an illusion of explanatory depth, which can trigger over-reliance. Finally, our analysis of fine-grained reliance patterns within multi-step decision workflows, as well as user involvement in plan-then-execute LLM agents, offer valuable insights for designing effective collaborations with agentic AI systems.

Taken together, the findings and implications in this dissertation advance our understanding of how to foster appropriate reliance on AI systems. By examining human and contextual factors that shape user reliance and perception, and by proposing novel methods for explanation and interaction, this work contributes both theoretical insights and quantitative evidence to the design of human-centered AI systems. We hope the key findings and implications reported in this dissertation will inspire further research on promoting appropriate reliance and facilitating effective human-AI collaboration.

Samenvatting

Naarmate AI-technologieën steeds breder worden geaccepteerd in de samenleving, is samenwerking tussen mens en AI naar voren gekomen als een veelbelovende manier om de verantwoordelijkheid en betrouwbaarheid te verbeteren van uitkomsten van taken waarbij AI gebruikt wordt. Hoewel AI-systemen zich in rap tempo ontwikkelen, ontbreekt het veel mensen – met name leken – nog steeds aan voldoende begrip en ervaring om effectief met deze systemen samen te werken. Deze kloof vormt een belemmering, vooral bij interactie met op deep learning gebaseerde AI-systemen, waarbij gebruikers vaak moeite hebben om de betrouwbaarheid van AI-adviezen in te schatten. Als gevolg hiervan kunnen mensen een verkeerd gekalibreerd vertrouwen of onjuiste opvattingen over de capaciteiten van AI ontwikkelen, wat leidt tot een ongepaste afhankelijkheid en een verminderd prestatieniveau van het team als geheel. Empirisch onderzoek toont aan dat mens-AI-teams vaak slechter presteren dan AI-systemen die zelfstandig opereren, wat benadrukt dat de huidige samenwerking tussen mens en AI nog verre van optimaal is. Deze observaties onderstrepen de noodzaak om onze kennis over effectieve mens-AI-samenwerking verder te verdiepen.

Dit proefschrift levert een bijdrage aan de groeiende literatuur over mens-AI-samenwerking door mogelijke benaderingen te analyseren die gepaste afhankelijkheid bevorderen. Concreet streven we naar zowel betrouwbare taakuitkomsten als een positieve, betrokken gebruikerservaring om effectieve samenwerking met AI te waarborgen. In een reeks empirische studies onderzochten we drie benaderingen om gepaste afhankelijkheid te stimuleren: het kalibreren van de gebruikersperceptie van competentie (Deel I), het verbeteren van het begrip van AI-systemen via mensgerichte uitlegbare AI (Deel II), en het vergroten van gebruikerscontrole door middel van collaboratieve werkstromen (Deel III). Onze bevindingen bevestigen dat een verkeerd gekalibreerde perceptie van zowel AI-competentie als eigen competentie kan leiden tot respectievelijk overmatige en onvoldoende afhankelijkheid. Daarnaast zagen we dat zowel XAI-methoden (zoals analogiegebaseerde uitleg) als interactieve XAI-interfaces (zoals conversatiegerichte XAI) een illusie van diepgang in de uitleg kunnen oproepen, wat op zijn beurt overmatige afhankelijkheid kan uitlokken. Tot slot bieden onze analyses van fijnmazige afhankelijkheidspatronen binnen meerstapsbesluitvorming en gebruikersbetrokkenheid bij plan-then-execute LLM-agenten waardevolle inzichten voor het ontwerpen van effectieve samenwerking met steeds autonomere AI-systemen.

Samenvattend dragen de bevindingen en implicaties in dit proefschrift bij aan ons begrip van hoe gepaste afhankelijkheid van AI-systemen kan worden bevorderd. Door menselijke en contextuele factoren die gebruikersperceptie en afhankelijkheid beïnvloeden te analyseren, en door nieuwe methoden voor uitleg en interactie te introduceren, levert dit werk zowel theoretische inzichten als kwantitatief bewijs voor het ontwerp van mensgerichte AI-systemen. We hopen dat de belangrijkste bevindingen en implicaties in dit

proefschrift verdere onderzoeksinspanningen zullen inspireren rond het bevorderen van gepaste afhankelijkheid en het faciliteren van effectieve mens-AI-samenwerking.

Acknowledgments

It all began with curiosity. On a September afternoon in 2019, I was given two career options: an internship at a well-known company or a research visit to Singapore Management University. Without hesitation, I chose to explore the path of academia. That decision has since led me into a challenging yet rewarding journey—the pursuit of a PhD. Looking back, it has been more demanding than I ever imagined, but it has also offered me invaluable experiences that have shaped my thinking and broadened my perspective. I feel fortunate to spend some of the most wonderful years of my life pursuing a career that excites and inspires me every day.

First, I wish to convey my deepest gratitude to my promotors, Ujwal Gadiraju and Geert-Jan Houben, for their trust, recognition, and constant support throughout my doctoral journey. They witnessed my aspirations to publish 'best paper', and together we made it a reality. Without their guidance and encouragement, I would not have gained the knowledge and experience necessary to complete my Ph.D. journey. They have not only taught me how to conduct research but also shared valuable lessons on teaching and life. Their sense of humor and wisdom have made this journey both inspiring and enjoyable, and I feel truly fortunate to have them as my promotors.

I am deeply grateful to have Deborah Nas, Peter Lloyd, Tim Miller, Q. Vera Liao, Martijn Willemsen, and Avishek Anand as my Ph.D. committee members. I deeply appreciate your generous commitment of time in reading and providing feedback on my thesis.

On my research journey, I have been incredibly fortunate to be supervised by and collaborate with many remarkable colleagues. I am especially grateful to my master's supervisors, Wayne Xin Zhao and Ji-Rong Wen. They opened the door to academia for me, introducing me to a fascinating world full of opportunities for exploration and discovery. Their mentorship, encouragement, and example have left a lasting impact on me, for which I am deeply thankful. I would also like to thank Jing Jiang and Gianluca Demartini for their guidance and supervision during my visits to Singapore and Brisbane. Collaborating with them allowed me not only to strengthen my profile through advanced research projects but also to gain new knowledge and broaden my professional perspective. The experience of working with them has profoundly influenced the way I approach and conduct research.

I would also like to express my appreciation to the students I had the privilege to mentor during my Ph.D.: Lucie, Abri, Nilay, Liang, Rohan, Varun, Xinyu, Jiacheng, and Yankun. Mentoring their master's thesis projects has been an invaluable experience, preparing me for a future career in academia. At the same time, witnessing their growth and development has given me a profound sense of accomplishment and fulfillment.

I feel truly fortunate to be part of the WIS group. From my very first day in Delft, I have been warmly welcomed and generously supported by colleagues and friends. Especially the Kappa members: Agathe, Alisa, Anne, Esra, Garrett, Lorenzo, Philip, Sara, and Shreyan. I would also like to express my gratitude to Sihang and Jie for generously sharing knowledge and experience, guiding me in both my career and life in the Nether-

lands. I am grateful for the support from colleagues: Aditya, Andra, Andrea, Alessandro, Arthur, Asterios, Claudia, Christos, Cristoph, Danning, Daphne, David, Felipe, Georgios, Gustavo, Hrishita, Jie, Jurek, Kyriakos, Lijun, Lixia, Manuel, Marcus, Nadia, Nirmal, Oto, Petros, Peide, Rihan, Robin, Sarah, Sepideh, Sihang, Shabnam, Shahin, Sole, Tim, Venktesh, Wenbeo, Yuandou, Ziyu, as well as many others, whose contributions I sincerely appreciate. During the past four years, we gathered, cooked, drank, and laughed together. The memories we created will forever be treasures in my life.

I want to express my sincere gratitude to my friends and colleagues who have accompanied me on my research journey. Special thanks to Hongjian, who, as a senior in the master's lab, not only offered guidance and support in finding opportunities but also became a valued collaborator and friend. I would also like to appreciate the friends I met during my master's program at RUC AI Box: Jin, Jinhao, Junyi, Kun, Ruiyang, Siqing, Shanlei, Shuqing, Tianyi, Xiaolei, Yingqian, Yushuo, etc. I would like to appreciate friends: Cheng, Haiyan, Heng, Shuo, Tao, Wenkai, Xiaojun, Xue, Zimeng, Yunshi, Wei, Xiaosen, Lei, Jipeng, Xuemei, Zhe, Zhihao, Chaoxia, Peng, Guocheng, etc.

Finally, I want to thank my family. Over the past four years, we have maintained a close connection through video calls. Although you were physically far away, I always felt your emotional support. I am especially grateful for my father, Fei He, and my mother, Fenglan Gao. Without their unwavering support, love, and encouragement, I could not have grown up as a happy and healthy person, nor could I have completed this journey.

Best wishes to all my friends, colleagues, and families!

Gaole
Delft, Sep 2025

1

Introduction

The use and fabrication of instruments of labour,
although existing in germ amongst certain species of animals,
is specifically characteristic of the human labour-process,
and Franklin therefore defines man as a 'tool-making animal'.

– Karl Marx

Humans show distinguished intelligence from other animals by inventing and using tools [1, 2]. Incorporating tools into daily life and work, human capabilities and work efficiency get amplified significantly [3]. For instance, humans are the only species that have learned to control fire, which makes their food more digestible and enhances nutrient absorption. Similarly, by replacing machines powered by humans or other animals (*e.g.,* horses) with machines supplied with fuels and electricity, humans obtained much higher efficiency in factories, transportation, and healthcare [4]. As the saying goes, "necessity is the mother of invention" — tools emerge to meet human needs [5]. However, to fully harness their potential, humans must learn to control the tools they create [6–8]. In other words, tool users should know when the tools can effectively address their empirical needs and how to apply tools in practice.

Artificial Intelligence (AI) systems [9], which have demonstrated promising effectiveness across various tasks [10] (*e.g.,* machine translation [11] and face recognition [12]), represent the latest advancement in human-made tools. Formally, an AI system is a computational entity that perceives its environment, processes data, and generates outputs — such as predictions, classifications, or decisions — based on learned models or predefined rules [9]. Unlike traditional tools designed primarily to extend human physical capabilities, AI is a cognitive technology that processes information, identifies correlation patterns in datasets, and makes predictions by mimicking aspects of human intelligence. Its impact spans nearly every domain of modern society, from healthcare and finance to transportation and education [9]. Moreover, AI continuously evolves through human feedback, achieving significant performance improvements over the past decade [13].

1

The promising potential of AI systems has led to widespread adoption in intelligent services like customer service [14], image generation [15], etc. However, their proliferation also raises significant ethical and safety concerns. One major issue is the potential for AI-generated content to spread misinformation [16], reinforce stereotypes [17], or even propagate hate speech [18]. For instance, AI chatbots can generate misleading narratives that quickly circulate on social media [16], influencing public opinion and exacerbating societal divisions [19]. Similarly, biased AI models used in hiring or law enforcement can perpetuate discrimination, leading to unfair outcomes [20]. Given these risks, it would be dangerous to apply AI systems without human oversight [21]. Existing studies have shown that incorporating human oversight can help mitigate AI failures [22], reduce harmful consequences [23], and ensure AI technologies are aligned with ethical and societal values [24].

Although AI systems have shown promising performance across various tasks, they are typically imperfect. They often struggle with out-of-distribution data [25] and may rely on spurious correlations rather than genuine reasoning [26]. The reasons behind this are multi-fold. First, many AI systems function as black boxes [27], making them difficult to interpret and understand. Nowadays, the most popular AI systems are based on deep neural networks trained with large volumes of task-specific data. Due to intrinsic opacity of deep neural networks, it would be challenging to trace their decision-making process or diagnose errors when they produce flawed outputs [10, 27]. Second, most AI systems are trained to simulate the collected data from the real world and are inherently probabilistic. This means there can be significant uncertainty about their outcomes, particularly when addressing out-of-distribution data [25]. Thus, due to accountability and reliability concerns, it may be undesirable to automate AI systems in high-stakes tasks (*e.g.,* medical diagnosis) [22]. Under these circumstances, there is a substantial need for **effective human-AI collaboration**, where humans can decide to adopt or override AI advice when necessary to ensure the quality of task outcomes and user experiences.

The above concerns have been substantial barriers to AI development and AI applications. To take advantage of AI systems and address potential risks, we need to further our understanding of human-AI collaboration. Specifically, this dissertation focuses on the **appropriate reliance within human-AI decision making**. With the wish of complementary team performance, one goal of effective human-AI collaboration is appropriate reliance: human decision makers rely on an AI system when it is accurate (or perhaps more precisely, when it is more accurate than humans) and do not rely on it when the system is inaccurate (or, ideally, whenever it is wrong). Understanding and fostering appropriate reliance is crucial for ensuring AI serves as a reliable and effective partner rather than a source of unchecked automation risks.

1.1 Overview: Effective Human-AI Collaboration

To ensure effective human-AI collaboration, this dissertation focuses on two key aspects: (1) performance-related outcomes, aiming for optimal teamwork and reliable task results; and (2) experience-related outcomes, emphasizing the quality of the human experience when interacting with AI systems. Figure 1.1 illustrates key elements in this section.

Performance-related Outcomes. In the last decade, we have witnessed AI systems

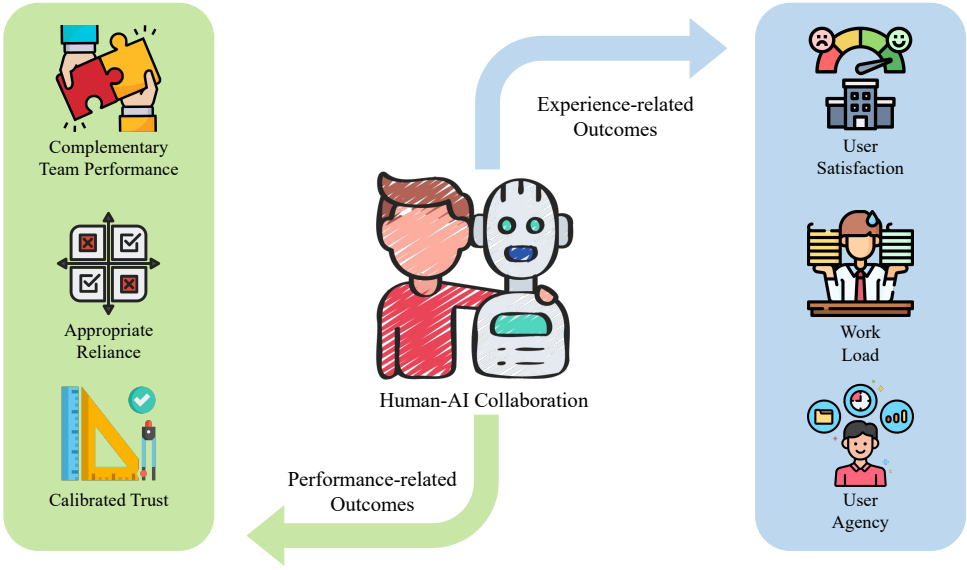


Figure 1.1: Overview of Human-AI Collaboration.

evolve at a fast pace. Their promising performance across various domains envisions a future AI-supported world. However, many current AI systems are found vulnerable without human oversights [21]. They can generate harmful content (*e.g.,* hate speech and fake news [16]) and make wrong predictions (*e.g.,* detect the moon as a yellow traffic light[1]). These shortcomings degrade user experience and erode people's trust in all AI systems [28]. This dissertation aims to leverage human oversights to ensure the trustworthiness of task outcomes from imperfect AI systems.

To achieve the goal of ***complementary team performance***, humans should recognize when the AI systems are problematic and override flawed outcomes from the AI systems [29]. At the same time, humans are also supposed to realize when the AI systems are correct (or more capable than themselves) and adopt AI assistance [30]. These behavior patterns reflect appropriate reliance on AI systems. However, many empirical user studies [22] on human-AI collaboration have indicated that humans typically rely on AI systems either too much (*i.e., over-reliance*) or too little (*i.e., under-reliance*). As a result, human-AI teams often underperform compared to AI systems operating alone [30, 31]. Accordingly, a growing body of research [22, 29, 32] aims to promote ***appropriate reliance*** on AI systems.

When interacting with AI systems that may outperform human experts, people exhibit two contradictory attitudes: *Algorithm Aversion* [33] and *Algorithm Appreciation* [34]. Algorithm aversion refers to a biased assessment of an algorithm, resulting in negative behaviors and attitudes toward it compared to a human agent [33, 35]. In contrast, algorithm appreciation describes people's preference for algorithmic advice over human advice [34]. These altitudes reflect that humans can easily develop uncalibrated trust in AI systems,

[1]https://www.autoweek.com/news/green-cars/a37114603/tesla-fsd-mistakes-moon-for-traffic-light/

which may lead to misuse or disuse [22, 36]. Meanwhile, prior work has pointed out that subjective trust may substantially affect user reliance behaviors [37]. Thus, besides the appropriate reliance behaviors, we also highlight the need to facilitate ***calibrated trust*** in the AI systems.

Experience-related Outcomes. Effective human-AI partnerships should not only enhance task performance but also provide tangible advantages to human participants [21]. Prior research [7, 38] on user interaction with tools has shown that users who engage more positively and actively with tools tend to demonstrate more effective and efficient tool use. Therefore, we highlight the importance of positive user experiences for effective human-AI collaboration.

One fundamental experience-related goal we seek is ***user satisfaction***. If users have a negative experience when collaborating with AI systems, they may be reluctant to adopt AI assistance, even if it leads to improved task outcomes [37]. Thus, to foster effective human-AI collaboration, we should avoid flawed designs that may decrease user satisfaction (*e.g.,* increased user cognitive load or time pressure). Specifically, AI systems should be designed to amplify human capabilities, like providing expert-level advice support and improving efficiency in repetitive jobs. Furthermore, AI systems should improve user experience by offering proactive suggestions, personalized support, and streamlined workflows that align with human preferences and needs. By reducing cognitive and physical ***work load***, humans can enjoy working with AI assistance and focus on tasks where they show distinct advantages (*e.g.,* tasks require human intuition or tasks need to align with human values).

Besides user satisfaction, ***user agency*** also plays a crucial role in facilitating effective human-AI collaboration. Prior work on tool use [7, 39] emphasizes the importance of user agency in promoting effective and empowering interactions with technology. In the context of human-AI collaboration, preserving user agency is not only critical for appropriate reliance and trust [40], but also for ensuring that AI augments rather than replaces human roles. While society is excited about the promising future AI systems bring, there are also concerns about unemployment and job displacement due to increased automation and efficiency [41]. These concerns are compounded when AI systems are deployed without preserving user agency, which can reduce worker autonomy and control over their roles. In a nutshell, human-AI collaboration should be mutually beneficial, where AI systems empower users to achieve their goals more effectively and efficiently while preserving user agency and satisfaction.

1.2 Our Focus: Appropriate Reliance

Trust and Reliance. With the growing interest in human-AI collaboration, researchers have increasingly focused on user trust and reliance on AI systems. Building on prior work in AI-assisted decision-making, this dissertation defines user trust as a subjective attitude and user reliance as objective behavior (*e.g.,* adoption of AI advice). Prior studies suggest that subjective trust can influence reliance behaviors; for instance, users who trust an AI system may blindly rely on it, while those who distrust it may choose to avoid using it altogether. However, user trust and reliance are not always closely coupled, and observations of one are not sufficient to infer the other. For example, even when users

rely on an AI system, it does not necessarily indicate trust. They may instead perceive the AI system as more capable than themselves in a particular context or prefer to avoid the responsibility of making decisions (*e.g.,* when facing a potential moral dilemma [42]).

Compared to subjective trust, objective reliance behaviors have a more direct connection with task outcomes. To help understand the association between reliance and task outcomes, we will explain with common setup in human-AI decision making — two-stage decision making [22, 30]. In the first stage, users make decisions independently (without checking AI advice). Then, users will check the AI advice (and relevant evidence or explanations) and make the final decisions. The two-stage decision making setup is widely adopted by prior work in human-AI decision making [22, 30]. Thus, we also adopt it in the dissertation for all one-step decision making tasks.

Initial human decision	AI advice	Relationship between initial human decision and AI advice	Human decision after receiving AI advice	Reliance
Correct	Correct	Confirmation	Correct	n/a
Correct	Correct	Confirmation	Incorrect	n/a
Incorrect	Incorrect	Confirmation	Correct	n/a
Incorrect	Incorrect	Confirmation	Incorrect	n/a
Incorrect	Correct	Positive advice (PA)	Correct	Positive AI reliance
Incorrect	Correct	Positive advice (PA)	Incorrect	Negative self-reliance
Correct	Incorrect	Negative advice (NA)	Correct	Positive self-reliance
Correct	Incorrect	Negative advice (NA)	Incorrect	Negative AI reliance

Table 1.1: Effect of AI advice on reliance. The form is taken from prior work [29].

Following prior work on appropriate reliance measurement [29], we considered user behavior when their initial decision differ from AI advice as a signal for user reliance. We enumerate all potential cases in two-stage decision making with Table 1.1. As we can see, when AI systems provide correct AI advice and the initial human decision is wrong, users are supposed to adopt AI advice. Similarly, when AI systems provide incorrect advice and humans hold a correct initial decision, humans are supposed to insist on their initial decision and disregard AI advice. Then, based on the four reliance patterns, we can calculate relative positive AI reliance (RAIR) and relative positive self-reliance (RSR) with:

$$\textbf{RAIR} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance} + \text{Negative self-reliance}},$$

$$\textbf{RSR} = \frac{\text{Positive self-reliance}}{\text{Positive self-reliance} + \text{Negative AI reliance}}.$$

The two measures can reflect appropriate reliance on two dimensions: low RAIR indicates under-reliance on AI systems, while low RSR indicates over-reliance on AI systems. The two measures can help us develop insights into user reliance patterns and are frequently used in this dissertation (Chapter 2-6).

1.3 Research Questions and Original Contributions

This dissertation is divided into three parts, each advancing the journey to promote appropriate reliance on AI systems by focusing on a distinct aspect: calibrating user perception of Competence (Part I), facilitating user understanding with human-centered XAI (Part

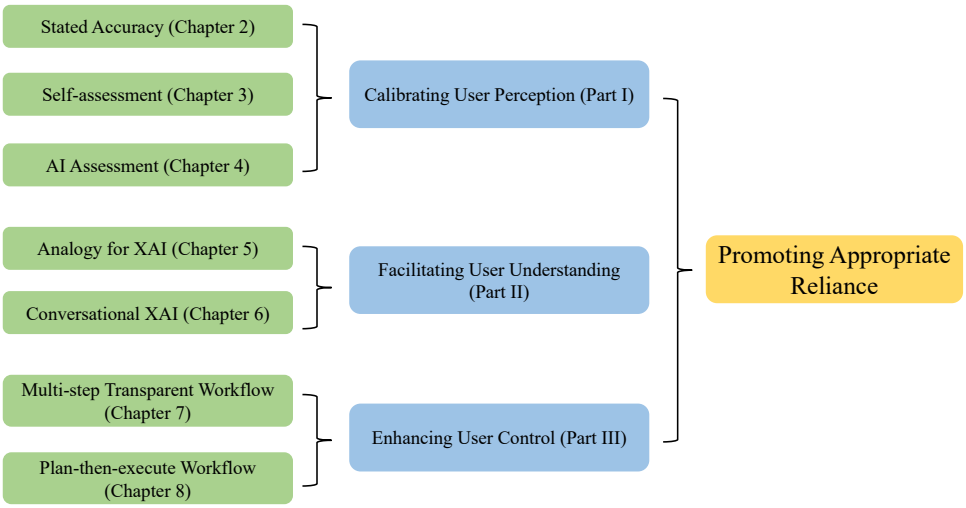


Figure 1.2: Hierarchy of the dissertation content.

II), and enhancing user control with collaborative workflows (Part III). We visualize their connections and hierarchical structure of content chapters with Figure 1.2.

Based on prior observations of miscalibrated trust in AI systems, we first look into the impact of the perception of task competence. We analyzed scenarios that contained performance feedback (*e.g.,* stated accuracy, Chapter 2) and those without performance feedback (Chapter 3 and Chapter 4). We also proposed user interventions to calibrate user perception of task competence (self-assessment and AI assessment). Based on empirical studies' findings, we realize that calibrating task competence solely may not be enough to make informed decisions. User understanding may be a bottleneck to making informed decisions and estimating the trustworthiness of AI advice. Thus, we developed and analyzed the impact of analogy-based explanations (Chapter 5) and conversational XAI decision support (Chapter 6). Furthermore, we noticed that the users may need effective workflows and have more flexible interaction with AI systems when handling complex tasks. Thus, we analyzed the impact of fine-grained transparency (Chapter 7) and user involvement with plan-then-execute LLM agents (Chapter 8). While the three parts focus on different aspects, they are also interconnected. For example, human-centered XAI and user involvement in the multi-step workflow can help calibrate user perception. Meanwhile, user perception and user understanding of the AI system may also affect how often they are involved in fixing AI systems in a multi-step workflow.

In the remainder of this section, we describe the motivation, research questions, and original contributions of each part.

Part I: Calibrating User Perception of Competence

In the last several years, decision making has been a popular scenario for analyzing user reliance on AI systems. When users' initial decision conflicts with AI advice, users need to either insist on their own decision or switch to AI advice in the final decision, which is a

clear signal of user reliance on the AI system. Prior research has shown that accuracy and other performance metrics significantly influence user trust and reliance on AI systems. Such an impact may stem from the perceived performance of different actors (*i.e.,* human, AI system, human-AI team). While prior work has extensively analyzed the effects of accuracy levels and confidence/uncertainty variations, less attention has been given to how perceived performance and associated cognitive biases (e.g., the Dunning-Kruger effect). In Part I of this dissertation, we seek to bridge this gap by investigating how the perceived performance of different actors influences user reliance on AI. This leads us to the following research questions:

RQ1-a: How does the perceived performance of humans and AI systems shape user reliance on AI systems?
RQ1-b: How to mitigate the impact of cognitive bias associated with misperception on user reliance on AI systems?

To understand the impact of the perceived competence of different actors (*i.e.,* human, AI system, human-AI team). We conducted a series of controlled empirical studies. To begin with (Chapter 2), we explore how the degree to which humans understand system accuracy influences their reliance on the AI system by investigating numeracy levels and using analogies to explain system accuracy. Based on the experimental results, we reason that the meta-cognitive bias Dunning-Kruger effect can be a potential cause for under-reliance. To confirm the impact of the Dunning-Kruger effect on user reliance, we conducted an empirical study on logical question answering (Chapter 3), which has been observed to trigger the Dunning-Kruger effect. At the same time, we also propose a tutorial intervention to help calibrate self-assessment and mitigate the impact of the Dunning-Kruger effect. Similar to the reasoning for under-reliance, we take the perceived performance of AI systems as one potential cause for over-reliance on AI systems. Inspired by existing literature on critical thinking and a critical mindset, we propose debugging an AI system as an intervention to foster appropriate reliance (Chapter 4). To make sure that laypeople (*e.g.,* crowd workers) can be able to debug AI systems, we adopted a deceptive hotel review detection task along with guidelines about deceptive patterns.

Contributions of Part I:

- We present findings of three empirical studies to advance our understanding of the impact of stated accuracy (Chapter 2), self-assessment (Chapter 3), and AI assessment (Chapter 4) in human-AI decision making.
- We reason that under-reliance on the AI system may be a result of users' overestimation of their own ability to solve the given task (Chapter 2). With further analysis, we confirm that users with the Dunning-Kruger effect will rely less on AI systems, and such under-reliance hinders them from achieving optimal team performance (Chapter 3).
- We propose tutorial intervention to mitigate the impact of Dunning-Kruger effect (Chapter 3). Based on the mixed results on participants with different self-assessment, we synthesize guidelines for better tutorial designs.

- We propose debugging intervention to develop critical mindset and foster appropriate reliance on AI systems (Chapter 4). Our results suggest that we should be careful in presenting the weakness of the AI system to users, to avoid any anchoring effect which may result in under-reliance.

Part II: Facilitating User Understanding with Human-centered XAI

As pointed out by GDPR, the users of AI systems should have the right to access meaningful explanations of model predictions. In response, a growing body of research has focused on developing human-centered explainable AI (XAI) solutions to enhance human-AI collaboration. To foster appropriate reliance, researchers have extensively explored XAI as a means of providing supporting evidence to help users make informed decisions. However, prior work has consistently shown that users can develop an illusion of explanatory depth — overestimating AI competence after viewing explanations generated with XAI methods. Inspired by research on analogy-based learning in education, we propose that analogy-based explanations can improve user comprehension of AI decision criteria and promote appropriate reliance. Specifically, we introduce concept-level analogy-based explanations to help users understand the causal relationships between key concepts and model predictions. This leads to the following research questions:

RQ2-a: How do analogies for concept-level explanations shape the understanding of an AI system among non-expert users?
RQ2-b: How do analogy-based explanations affect user reliance on AI systems?

As there are no off-the-shelf solutions for generating high-quality analogies as explanations, we proposed a crowd computing method for generating analogies with templates (Chapter 5). To help advance our understanding of the analogy quality on their usefulness in decision making, we conducted two empirical studies for evaluation. First, we synthesized a set of structured dimensions to assess analogy quality (see Chapter 5) and recruited five experts to evaluate the analogies generated with our crowd computing method. Our findings reveal that the proposed dimensions show a positive contribution to the perceived helpfulness of explanations. To further our understanding of the impact of analogy-based explanation on user reliance, we conducted an empirical study on cancer diagnosis (Chapter 5). While the analogy-based explanations do not work as expected to significantly boost user understanding of AI systems or facilitate appropriate reliance, we figure out key challenges in generating high-quality analogies and the potential for personalization.

Beyond the explanation methods, the user interfaces used to present XAI methods may also impact user understanding and reliance on AI systems. The recent advancement of large language models (LLMs) has enabled conversational interactions with AI systems, which envisions a promising future of conversational XAI support. Conversational user interfaces can provide a human-like interaction and simplify complex tasks with filtered information, which can bring better user experience and higher user engagement. To confirm the impact of conversational XAI interface on user reliance, we conducted a systematic comparison with XAI Dashboard, a widely adopted user interface in practical XAI applications. This leads to the following research questions:

RQ3-a: How does a conversational XAI interface shape user understanding of an AI system, in comparison with an XAI Dashboard?
RQ3-b: How does a conversational XAI interface influence user reliance on an AI system, in comparison with an XAI Dashboard?

To answer **RQ3-a** and **RQ3-b**, we conducted an empirical study by comparing the conversational XAI interface with the XAI dashboard (Chapter 6). We considered both rule-based agents and LLM-based agents to support the conversation. Meanwhile, we also adopted an evaluative AI perspective to adjust the conversation and nudge users to check and compare their own decision criteria with those of AI systems. Our results indicate that we should be careful in presenting XAI methods with an interactive XAI interface, which may cause over-reliance on the AI system.

Contributions of Part II:

- We present findings of two empirical studies to advance our understanding of the impact of concept-level analogy-based explanations (Chapter 5) and conversational XAI interfaces (Chapter 6) in human-AI decision making.
- We propose analogical inference as a bridge to help end-users leverage their commonsense knowledge to better understand the concept-level explanations (Chapter 5).
- We design an effective analogy-based explanation generation method and collect 600 analogy-based explanations from 100 crowd workers (Chapter 5).
- We propose a set of structured dimensions for the qualitative assessment of analogy-based explanations and conduct an empirical evaluation of the generated analogies with experts (Chapter 5).
- We find that, compared to concept-level explanations, the additional analogies do not cause a significant delay in decision making or pose a significantly higher cognitive load. Based on the qualitative analysis of participants' feedback and user reliance patterns, we summarized guidelines for future work about generating effective analogy-based explanations and on the appropriate usage of analogy-based ex-planations (Chapter 5).
- We find that, compared to XAI dashboard, the conversational XAI interface showed a slightly better understanding, and demonstrated a slightly higher trust in the AI system (Chapter 6).
- We provide empirical evidence that the XAI interfaces were persuasive and have the potential to bring about an illusion of the AI systems' capability, which in turn increased over-reliance on the AI system (Chapter 6).
- We find that boosting the conversation quality and flexibility (*i.e.,* with LLM-based conversational agent) may further reinforce over-reliance and hurt user understanding as well as user trust (Chapter 6). Our insights and observations can inform the

1

future design of conversational XAI interfaces to promote complementary human-AI collaboration.

Part III: enhancing user control with collaborative workflows

In real-world applications, AI systems are increasingly tasked with handling complex problems that require multi-step decision-making and reasoning. In such workflows, errors in earlier steps may propagate, making flawed AI outcomes harder to diagnose. To address such concerns, we followed prior work to provide transparency along with AI assistance. Specifically, we investigate how fine-grained transparency methods can be structured to improve the reliability of mutli-step decision workflows, leading to the following research questions:

RQ4-a: How does a multi-step decision workflow shape user reliance on an AI system?
RQ4-b: How do global transparency and local transparency shape user reliance in a multi-step decision workflow?

To investigate the impact of fine-grained transparency on user reliance within a multi-step decision workflow, we conducted an empirical study on composite fact-checking tasks (Chapter 7). The AI system is implemented by prompting LLMs to conduct task decomposition and leverage retrieval-augmented generation as the solution for each sub-task. By completing the task in a multi-step workflow, we analyzed appropriate reliance with fine-grained levels. Users make a final decision based on intermediate decisions in the multi-step decision workflow. Meanwhile, each intermediate decision is supported by retrieved documents, which users rely on to conduct fact-checking. Thus, we look into appropriate reliance on the intermediate steps and appropriate reliance on the retrieved documents. Our insights help deepen the understanding of the role of decision workflows in facilitating appropriate reliance. Furthermore, we synthesize important implications for designing effective means to facilitate appropriate reliance on AI systems in composite tasks, positioning opportunities for human-centered AI and broader HCI communities.

Along with the growing popularity of large language models (LLMs), LLM agents have shown promising capabilities in handling complex tasks like playing games and research assistance. The workflow of LLM agents can be abstracted into two stages: planning and execution. With the planning stage, LLM agents decompose the complex task into a sequence of sub-tasks. Then, the LLM agents can leverage external toolkits to solve the sub-tasks step-by-step. While such framing is attractive, the uncertainty associated with LLMs can lead to unintended or unexpected errors. To address such gap, we analyzed how human involvement in the LLM agent can provide more reliable task outcomes, leading to the following research questions:

RQ5-a: How does human involvement in the high-level planning and real-time execution shape their trust in an AI system powered by LLM agents?
RQ5-b: How does human involvement in the high-level planning and real-time ex-

ecution of tasks with an AI system powered by LLM agents affect the overall task performance?

To address **RQ5-a** and **RQ5-b**, we conducted an empirical study with plan-then-execute LLM agents on six daily scenarios (Chapter 8). To provide necessary user involvement within the LLM agent workflow, we allow users to check the outcomes from LLM agents in both the planning and execution stages. We analyzed how user involvement at each stage affects their trust and collaborative team performance. Our work has important implications for the future design of daily assistants and human-AI collaboration with LLM agents.

Contributions of Part III:

- We present findings of two empirical studies to advance our understanding of the impact of fine-grained transparency in multi-step decision workflow (Chapter 7) and user involvement in plan-then-execute LLM agents (Chapter 8).
- We demonstrate that fine-grained transparency within multi-step decision workflow can facilitate human-AI collaboration in specific contexts (Chapter 7).
- We synthesize two metrics to evaluate appropriate reliance at intermediate steps and appropriate reliance on evidence (*i.e.,* input to AI systems). Further analysis demonstrates that the fine-grained appropriate reliance promotes appropriate reliance on the global level (Chapter 7).
- We propose LLM agents as a daily assistant to help end-users solve daily tasks. By abstracting the workflow of LLM agents, we adopted a plan-then-execute workflow (Chapter 8).
- We examine the impact of user involvement at high-level planning and real-time execution stage (Chapter 8). Our findings demonstrate that LLM agents can be a double-edged sword — (1) they can work well when a high-quality plan and necessary user involvement in execution are available, and (2) users can easily mistrust the LLM agents with plans that seem plausible.
- We find that user involvement in both high-level planning and real-time execution fails to calibrate user trust in LLM agents (Chapter 8). At the same time, user involvement in planning may hurt plan quality. By comparison, user involvement in execution provides more stable positive contributions to task outcomes.
- We synthesize key insights for using LLM agents as daily assistants to calibrate user trust and achieve better overall task outcomes (Chapter 8).

1.4 Research Methodology

To analyze user reliance on AI systems, we conducted a series of controlled empirical studies. Figure 1.3 provides an overview of our research methodology. In the beginning, we

1

first formulate the research questions about facilitating appropriate reliance on AI systems. In this step, we conduct a systematic literature review and extensively gather ideas with open discussion. After we settle down the research questions to analyze, we conduct the study design. In this step, we decide the key factors to analyze and design a controlled study regarding key factors. Meanwhile, we synthesize hypotheses based on prior findings and logical reasoning. Before implementing the experiment, we pre-register our experiments. Then, we strictly follow the study design to implement user interfaces and provide online demos. All data collected in this dissertation are collected based on the crowdsourcing platform Prolific[2]. After the data collection, we conduct quantitative and qualitative analyses to verify our hypotheses and advance our understanding of the topic.

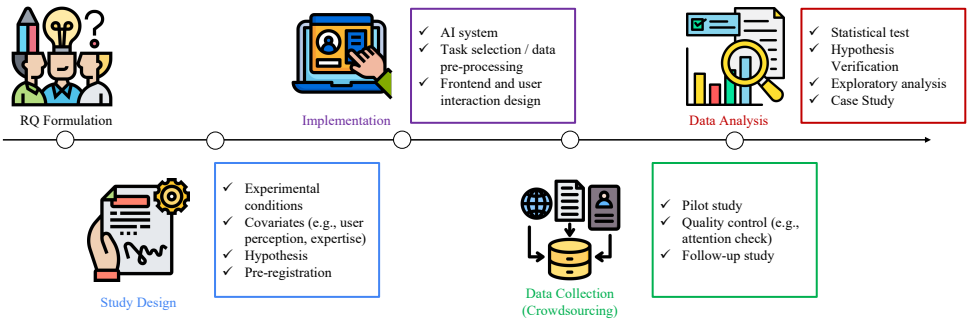


Figure 1.3: High-level overview of the empirical studies in this dissertation.

In this dissertation, we included 2,304 valid participants in data analysis. In total, we spent around 15k GBP (including bonus) on crowdsourcing, which compensates for around 1.5k human hours. We analyzed human-AI collaboration with varying task scenarios, such as question answering, cancer diagnosis, loan approval, fact-checking, planning, and knowledge collection. With these controlled empirical studies, we obtained valuable insights for human-AI collaboration.

1.5 Content Organization

The remainder of this dissertation consists of the chapters listed below. Each chapter (except the conclusions) is based on projects I led and collaborated on with co-authors. To help track the origin of chapters, we also provide associated publications along with each chapter.

Chapter 2: The Impact of User Understanding of Stated Accuracy. It is based on a peer-reviewed paper at CSCW'23:

- **Gaole He***, Stefan Buijsman*, Ujwal Gadiraju. *How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System.* Proceedings of the ACM on Human-Computer Interaction 7, no. CSCW2 (2023): 1-29. https://doi.org/10.1145/3610067

[2] https://www.prolific.com/

Chapter 3: The impact of the Dunning-Kruger effect. It is based on a peer-reviewed paper at CHI'23:

- **Gaole He**, Lucie Kuiper, and Ujwal Gadiraju. *Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems.* In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1-18. 2023. https://doi.org/10.1145/3544548.3581025

Chapter 4: Developing critical mindset with debugging AI systems. It is based on a peer-reviewed paper at HT'24:

- **Gaole He**, Abri Bharos, and Ujwal Gadiraju. *To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems.* In Proceedings of the 35th ACM Conference on Hypertext and Social Media, pp. 98-105. 2024. https://doi.org/10.1145/3648188.3675130

Chapter 5: Analogy-based concept-level explanations. It is based on two peer-reviewed papers at HCOMP'22 and JAIR'24:

- **Gaole He**, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. *It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge.* In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 10, pp. 89-101. 2022. https://doi.org/10.1609/hcomp.v10i1.21990; **Best Paper Award**

- **Gaole He**, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. *Opening the Analogical Portal to Explainability: Can Analogies Help Laypeople in AI-assisted Decision Making?* Journal of Artificial Intelligence Research 81 (2024): 117-162. https://doi.org/10.1613/jair.1.15118

Chapter 6: Conversational XAI decision support. It is based on a peer-reviewed paper at IUI'25:

- **Gaole He**, Nilay Aishwarya, Ujwal Gadiraju. *Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant.* 30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy. https://doi.org/10.1145/3708359.3712133

Chapter 7: Fine-grained appropriate reliance in multi-step decision workflow. It is based on a work-in-progress paper:

- **Gaole He**, Patrick Hemmer, Michael Vössing, Max Schemmer, Ujwal Gadiraju. *Fine-Grained Appropriate Reliance: Human-AI Collaboration with a Multi-Step Transparent Decision Workflow for Complex Task Decomposition.* Revised after reviews from CSCW'25 and CHI'25, now under review.

Chapter 8: User involvement in plan-then-execute LLM agent. It is based on a peer-reviewed paper at CHI'25:

- **Gaole He**, , Gianluca Demartini, Ujwal Gadiraju. *Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant.* CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan. https://doi.org/10.1145/3706598.3713218.

Chapter 9: Discussions and conclusions in this dissertation.

I

Calibrating User Perception of Competence

2

Using Analogies to Explain Accuracy of AI systems

AI systems are increasingly being used to support human decision making. It is important that AI advice is followed appropriately. However, according to existing literature, users typically under-rely or over-rely on AI systems, and this leads to sub-optimal team performance. In this context, we investigate the role of stated system accuracy by contrasting the lack of system information with the presence of system accuracy in a loan prediction task. We explore how the degree to which humans understand system accuracy influences their reliance on the AI system, by investigating numeracy levels and with the aid of analogies to explain system accuracy in a first of its kind between-subjects study (N = 281). We found that explaining the stated accuracy of a system using analogies failed to help users rely on the AI system appropriately (i.e., the tendency of users to rely on the system when the system is correct, or on themselves otherwise). To eliminate the impact of subjective attitudes towards analogy domains, we conducted a within-subjects study (N = 248) where each participant worked on tasks with analogy-based explanations from different domains. Results from this second study confirmed that explaining stated accuracy of the system with analogies was not sufficient to facilitate appropriate reliance on the AI system in the context of loan prediction tasks, irrespective of individual user differences. Based on our findings from the two studies, we reason that the under-reliance on the AI system may be a result of users' overestimation of their own ability to solve the given task. Thus, although familiar analogies can be effective in improving the intelligibility of stated accuracy of the system, an improved understanding of system accuracy does not necessarily lead to improved system reliance and team performance.

2.1 Introduction

It is becoming more and more common for humans to make decisions supported by machine learning algorithms. Whether it is in financial risk assessment [43, 44], medical diagnosis [45, 46] or in public employment services [47], such collaborative, socio-technical

This chapter is based on a peer-reviewed paper: **Gaole He***, Stefan Buijsman*, Ujwal Gadiraju. *How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System.* Proceedings of the ACM on Human-Computer Interaction 7, no. CSCW2 (2023): 1-29. https://doi.org/10.1145/3610067

systems (*i.e.,* a decision procedure where humans and AI are jointly involved in making the decision) are ubiquitous. And while initial hopes were that such a combination would lead to better decisions [48], it has proved tough to mitigate unexpected reliance (*i.e.,* under-reliance and over-reliance) on the AI system. *Appropriate reliance* is defined as the tendency for users to rely on the system in situations where it is accurate (or more precisely, where it is more accurate than humans) and not to rely on it when the system is inaccurate (or, ideally, whenever it is wrong). This follows the conceptualization of appropriate system reliance established in the Human-AI interaction, collaboration, and teaming fields over the last few years [29, 40, 49–51]. Users in the real world, however, find it difficult to determine their own accuracy in difficult tasks as well as the system's accuracy (in individual cases). That in turn means they have a hard time deciding when an AI system is more accurate than they are. This tension has been shown to result in both under-reliance [34, 52] and over-reliance [50] of users on AI systems, often leading to detrimental outcomes.

There are several complementary approaches to facilitating appropriate system reliance, such as research in explainable AI attempting to elucidate the reasons for model output [53, 54]. Such tools can help, especially if users are actively made to reflect on explanations using cognitive forcing interventions [55]. Another approach, and one which is explored further in this chapter, is to give users information on the confidence and overall accuracy of the system. Papenmeier *et al.* [56], Yin *et al.* [57] found that users adjust their reliance on AI systems based on the reported system accuracy. However, even after seeing the high stated accuracy, users do not rely on the system as often as the accuracy warrants (*e.g.,* adopting system advice 80% of the time while system accuracy is 95%, resulting in an inferior overall performance than the theoretical potential). We explore if this under-reliance among users is a result of their potentially limited understanding of the system accuracy measure. We do not hold the position that reliance on AI systems is universally good. On the contrary, preventing over-reliance on AI systems is just as important. However, a fundamental pre-requisite to designing and facilitating human-AI interactions that can effectively support humans in a given task, is to advance our current understanding of how users rely on AI systems. An unanswered question in this context pertains to why users tend to under-rely on AI systems despite their relatively high stated accuracy. Perhaps users do not properly calibrate their reliance on the AI system because they have trouble identifying the right accuracy level when presented only with an overall accuracy value.

We use analogies to counter such lack of understanding of global accuracy measures, which is to our knowledge the first attempt of its kind to elucidate system measures. An analogy can be interpreted as a structural mapping of a target domain which is to be clarified (in this case, overall system accuracy) onto a source domain which the recipient of the analogy is more familiar with [58, 59]. As a simple example, one might elucidate how hard a task is by saying 'it is as hard as finding a needle in a haystack'. As the recipient is likely to know that finding a needle will be difficult in this case, the inference on the target domain can be made that the relevant task will also be difficult. While such simple examples may not make a convincing case for the use of analogies, there is strong empirical evidence that more specific analogies can help people to individuate and identify risk levels, as discussed further in Section 2.2.

To address the aforementioned research gap in this chapter, we aim to find answers for the two research questions:

> **RQ1**: How does the understanding of stated system accuracy affect reliance of users on the AI system?
> **RQ2**: How does explaining stated system accuracy using analogies affect the reliance of users on the AI system?

To answer these questions, we proposed four hypotheses considering the effect of the stated accuracy level on user reliance, the effect of using analogies to explain accuracy measures on reliance, and two important user factors (numeracy level and familiarity with the analogy domain). We tested these hypotheses in an empirical study of human-AI collaborative decision making in a loan approval task.[1] In this chapter, we present a between-subjects exploration (N = 281) as the main study to verify the proposed hypotheses. To ensure that our results do not suffer from the impact of domain-specific user characteristics (trust in and familiarity with the analogy domain) caused by individual user experiences, we conducted a further within-subjects study (N = 248) to investigate the effects of seeing different analogies. We found that well-understood stated accuracy is insufficient for users to calibrate their reliance on an AI system, for a 75% accuracy level. Explaining stated system accuracy, even for users with low numeracy skills, had no significant effect on our (behavioral) reliance measure. We did find a limited effect of the successful use of analogies on subjective measures of trust in the system. However, this improvement in subjective measures did not translate to an improvement in reliance or performance. This suggests that the issue is not with users' trust in the system, but with an overestimation of their own skill at the task.

Our results highlight that a limited understanding of the system accuracy measure is not the reason why users rely on AI systems lesser than warranted by the relatively higher system accuracy. Instead, it is likely that users' overestimation of their own ability to solve the given task drives their under-reliance on the system. This interpretation is supported by various findings in prior work [30, 60–62]. We outline this as a direction for further study. Empirical studies that explore why and how humans tend to rely on AI systems play a vital role in furthering our understanding of how we can build better human-AI interactions in a variety of tasks, scenarios, and domains. It is in this context that our work makes important contributions by (a) advancing our understanding of user under-reliance on AI systems, (b) exploring the effectiveness of analogies as an instrument to explain measures like stated system accuracy, and (c) investigating whether an improved understanding of global AI system measures can lead to more appropriate reliance.

In addition, although we considered several potentially important user factors (such as numeracy level and familiarity with and trust in the analogy domain), most of them did not significantly impact user reliance behaviors. Only users' general propensity to trust automated systems emerged as an important user factor which contributes to both subjective trust and objective reliance. Based on the results from our empirical study, we synthesized and discussed favorable conditions for the use of analogies and pointed out

[1] All data and code can be found at: `https://osf.io/9jqma/?view_only=c0c0dd12fa804b028cd29fbf9fd2ef4f`

promising future directions for further research exploring user reliance on AI systems. Our findings contribute to the growing body of literature on human-AI decision making and further our understanding of under-reliance on AI systems.

2.2 Related Work

This chapter contributes to the growing literature on user reliance on AI systems by focusing on how users might be helped to calibrate their reliance by analogies that clarify stated accuracy measures. Our goal is to explore whether a limited understanding of stated accuracy is to blame for under-reliance on an AI system (within the scope of **RQ1**) and whether improving this understanding can lead to more appropriate reliance (within the scope of **RQ2**). As such, the research combines three strands of literature: the general literature on user reliance of AI systems (2.2.1). The more specific literature on how that reliance is affected by stated accuracy measures (2.2.2) and finally the literature on analogies, which have been shown to benefit risk perception (2.2.3).

On the one hand, the research focuses on the use of accuracy scores to engender (appropriate) reliance on AI systems. As merely stating the accuracy has been found to be insufficient for reaching appropriate reliance, the contribution of this chapter is to explore whether that is due to a limited grasp of the implications of the accuracy scores. Another area of research that is therefore relevant for this chapter is the literature on analogies in risk perception, where the use of analogies to elucidate percentages in a similar setting has been investigated. That gives us a basis to postulate that analogies improve this understanding.

2.2.1 Reliance on AI Systems

There is a wide range of factors that affects how users rely on AI systems. For example, Dietvorst *et al.* [33] and Dzindolet *et al.* [63] found that users stop relying on a system after seeing it make a mistake. Meanwhile, Yeomans *et al.* [64] found that people did not rely on system advice in a highly subjective domain – namely a task to predict which jokes others will find funny – even if the system performed better than they did. At the same time, Dietvorst *et al.* [35] saw that participants are more willing to rely on systems if they are able to alter the final decision somewhat, rather than having to follow the exact prediction. Such prior research has generally found that it is hard to get users to rely on a system appropriately. Inspired by the design of these studies, in our study we used a two-stage decision making process that allows users to alter their final decision after seeing the AI advice (see Section 2.3.1).

Different solutions for this challenge have been examined. We investigate the option of presenting users with accuracy measures (2.2), but the other major option is to provide users with explanations of the system output (XAI). In a risk assessment task (for a loan approval and a pretrial domain), Green *et al.* [65] looked at whether explanations or feedback per decision help users calibrate their reliance, but found mostly null effects. They show that people are unable to evaluate their own accuracy at risk assessments, do not calibrate their reliance based on observed accuracy and only had a positive effect from explanations on the loan approval task. And whereas Green *et al.* [65] found some positive effects of explanations, Zhang *et al.* [66] failed to find similar appropriate reliance

when users were given (feature importance) explanations. However, they did observe an improvement in reliance when presenting confidence scores for the system, with users switching more often to (*i.e.,* relying on) AI predictions with high confidence scores than to those with lower confidence scores or none at all. This is in line with the proposal of Bhatt *et al.* [67] to use uncertainty measures to help users rely appropriately on AI systems. Yet the addition of confidence scores in the study by Green *et al.* did not improve the accuracy of participants using the AI system.

One complicating factor here is the interplay between subjective trust and objective reliance. In this chapter, we consider that subjective trust influences objective reliance. And indeed Lu *et al.* [49] found similar patterns for both objective reliance and subjective trust when feedback on model performance is limited. Both trust and reliance are significantly affected by the level of agreement between people and a model on decision making tasks that people have high confidence in. However, other conflicting results have also been found. Through an extensive user study, Buçinca *et al.* [68] pointed out that "when using actual decision making tasks, subjective results do not predict objective performance results," which reveals a gap between the subjective trust attitude of users and their objective reliance behavior. Similarly, a gap between stated trust and actual reliance was reported by Schmitt *et al.* [69], and Bansal *et al.* [70] observed that explanations can promote blind trust rather than lead to appropriate reliance on AI systems. We thus hold that subjective trust *can* promote objective reliance, but keep in mind that subjective trust measures can give an overly optimistic image of reliance and therefore focus on objective reliance.

2.2.2 Reliance and System Accuracy

Though research specifically on stated accuracy is sparse, prior experiments do show that the stated accuracy of a system has an effect on the degree to which people rely on the system. Yin *et al.* [71] first reported a significant effect of stated accuracy on reliance and further expanded on this in [57]. Here, in a task where users had to predict if someone wanted to see his or her date a second time, they compared reliance on the system across conditions with different stated accuracies (and included a control with no stated accuracy). They observed significant differences in the fraction of cases in which users agreed with the system and in the fraction of cases in which users changed their initial decision so that their final decision agreed with the system advice. However, they found that participants struggle to calibrate their reliance. When there was no stated accuracy, users agreed in about 75% of the final decisions with the system. For decisions with an initial disagreement between users and the system, users switched to agree with the system in 30% of cases. This did not change for a stated accuracy of 60% or 70% and only increased for a stated accuracy of 90 and 95%. However, the effect of the stated accuracy is not as high as it should be: for 90% and 95%, users only agreed with the system in 80% of cases. Finally, the effect of stated accuracy was canceled out by the effects of observed accuracy when these were presented to users midway through the study.

This relevance of observed accuracy has further been underscored by Papenmeier *et al.* [56], who found that the effect of varying observed accuracy on reliance was stronger than the effect of explanations of system outputs (either no, low-fidelity, or high-fidelity explanation). So, system accuracy has been shown to be relevant for calibrating reliance, and therefore the extent to which users understand what this system accuracy means.

Recent work by Nourani *et al.* has shown that users do not rely on what they do not understand [72]. It is this lack of understanding that we hope to alleviate through the use of analogies.

2.2.3 Analogies in Risk Perception

There is a long-standing use of analogies to explain statistical concepts [73, 74] and medical risk levels [75, 76]. What emerges from this is that it can be difficult to get analogies to deliver benefits, as the meta-study by Sopory *et al.* [77] on the effect of metaphor's persuasive effects underlines. Analogies, as they intricately depend on how they are perceived by the recipient, can be hard to calibrate to the audience. If successful, however, they can have clear cognitive benefits. Sopory *et al.* [77] found that when they are novel, have a familiar source domain (*i.e.,* the 'needle in a haystack' part in 'x is as difficult as finding a needle in a haystack') and are used early in the message then they are used optimally and have a clear effect on persuasiveness. A later meta-study by Van *et al.* [78] confirms this, finding that metaphorical messages are, when using a familiar source domain, more effective than literal messages.

Such effects can be found in the existing literature on risk perception too. Barilli *et al.* [79] tested the use of analogies to improve the risk perception between a 1 in 100 chance and a 1 in 900 chance. While adding analogies does not make these risks more discriminable, they do lower the overall risk perception on a 7-point scale (from 3.5 to 2.5 for 1 in 100, from 3.1 to 2.1 for 1 in 900). The lack of effects here has, however, been hypothesized to be due to the choice of analogies: stated analogies were about the odds of drawing a red ball out of a jar, something which we do not encounter or deal with on a regular basis. More familiar analogies studied by Galesic *et al.* [76], such as 'as a flu vaccine is to flu' or 'as a car alarm is to theft', did show a clear effect of analogies. Performance on difficult medical problems was improved for people with high numeracy skills and performance on easy problems was improved for people with low numeracy skills. Numeracy here means the ease and skill with which participants work with numbers. Their interpretation of the finding, therefore, was that analogies help when problems are not too difficult and performance is not at ceiling. Interestingly for the current study, Galesic *et al.* [76] also looked at what makes analogies helpful and again ranked familiarity with the source domain highly.

The effect of numeracy level on findings has, moreover, been collaborated in other studies. Pighin *et al.* [80] found that high-numeracy participants do improve on discrimination of risk levels after seeing analogies. Participants with low numeracy showed no improvement in the discrimination between a 1 in 5390, 1 in 770 and 1 in 110 risk on a 7-point Likert scale. Similarly, with a more visual analogy in the form of a risk ladder, Keller *et al.* [81] found the visualisation to suffice for high-numeracy participants in discriminating between different risk levels. Low-numeracy participants only managed to do so after also seeing analogies with the number of cigarettes one would smoke a day. So, here too, familiarity with the source domain is likely to have been high, to support understanding of the risk levels.

To sum up, analogies have been found to be effective tools to improve risk perception and performance on related medical problems, though a number of relevant factors have emerged that interact with the effectiveness. These have informed our hypotheses

3 and 4. Numeracy level is important, as also underlined by a recent overview study [82], and especially low numeracy individuals can use help in understanding the meaning of percentages. This finding supports our motivation to look into the possibility that participants fail to calibrate reliance to accuracy scores because they might not fully understand the presented information. Aside from numeracy, familiarity with the source domain used to explain the percentages is an important factor for the success of analogies. Hence, we have used a range of analogies in our study that vary with respect to familiarity and included a question in the post-task questionnaire to measure user's (subjective) familiarity with the source domain.

2.3 Task and Hypothesis

In this section, we describe the loan prediction task and present our hypotheses, which have all been preregistered before any data collection.

2.3.1 Loan Prediction Task

The basis for our experimental setup is a task where participants have to decide whether to accept or reject a loan application using the publicly available loan prediction dataset.[2] This task was chosen as a realistic scenario for human-AI collaboration, where there is a clear risk and a benefit to the adoption of AI advice. As such, it fits in with the risk perception research where analogies were pioneered. It has also been adopted by existing research in behavioral economics [83] and human-AI collaboration [65].

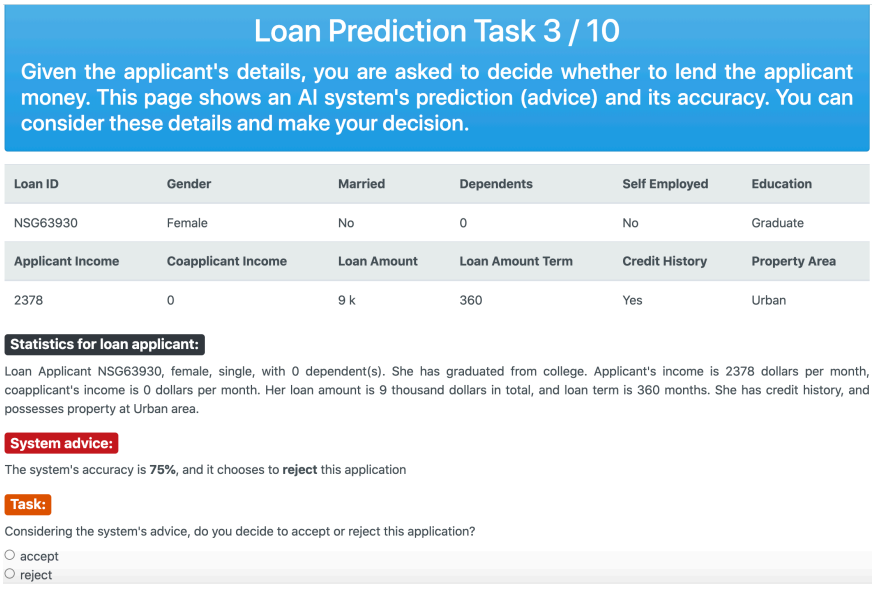


Figure 2.1: Illustration of the interface that participants used to complete the loan prediction task.

[2] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

Participants thus made decisions on whether to grant a loan or not based on twelve features such as income, the absence or presence of a credit history and the loan amount. This simulates a realistic scenario where participants interact with an AI system and may rely on it due to the complexity in simultaneously considering multiple features for successful decision making, but also due to a relatively high stated accuracy of the AI system. Furthermore, we consider this to be a suitable task to test the influence of user numeracy level, as almost all the presented information is in numerical format. The task interface is shown in Figure 2.1.

Task Selection. Participants were presented with twelve such cases, of which two were example cases and ten trial cases. These cases were selected by first training a linear regression model on the full dataset. The two example cases were the top-*1* most confident correct cases for approval and rejection (with respect to the linear regression model). The ten trial cases used in the actual experimental task were: two high confidence correct predictions, two medium confidence correct predictions, two borderline correct predictions, two borderline wrong predictions and the two least confident wrong predictions (again, with respect to the linear regression model). Cases were evenly split between those where the loan should be approved and those where the loan should be rejected and the order of the trial cases was randomized to prevent order effects [84].

Two-stage Decision Making. In trial cases, participants of all conditions were first presented with the applicant information corresponding to the case and then asked to make a decision whether to accept or reject the loan application (see screenshot in Figure 2.1). This first time, they were not presented with the systems' prediction, or with any additional information. After making an initial choice they saw the same case again, but now additionally saw the systems' prediction and (depending on the experimental condition) also the system accuracy and analogy. Participants were then asked to make a final decision. This setup of an initial unaided decision and the presentation of system advice in order to make a second and final choice is similar to the update condition in [65], and in line with findings that people first make a decision on their own and only then decide whether to incorporate system advice [85]. It also fits with the research of Dietvorst *et al.* [35] on trust in two-stage decision making.

2.3.2 Hypotheses

Our study was designed to answer questions about the effectiveness of well-understood stated accuracy on reliance, and the use of analogies to improve user understanding of the accuracy level. As stated accuracy has been found to be effective in improving (appropriate) reliance [57], we expect to observe the same effect here:

> **(H1)** The stated accuracy of a system has a significant effect on user reliance on the system.

Analogies, as we have discussed above, have the potential to make stated accuracy more intuitive to users and thus increase their sensitivity to it. Therefore, we hypothesize:

(H2) The stated accuracy of a system presented using an analogy has a significantly larger effect on user reliance on the system than the stated accuracy presented without an analogy.

In particular, we expect that this effect will depend on how familiar users are with the target (the stated accuracy) and source (*e.g.,* train punctuality) domain of the analogy, as discussed in Section 2.2. Thus, we further hypothesize that the numeracy level of users, *i.e.,* how familiar they are with quantitative measures, shapes the usefulness of analogies. Participants with a high numeracy level might understand the task and stated accuracy well enough already for analogies to offer little improvement, whereas participants with low numeracy might have a lack of understanding of these numbers that is alleviated by the analogy. As the role of analogies is to make this target domain (accuracy of the system) easier to understand by creating a structural mapping onto a source domain that the user is potentially more familiar with, we also formulate a hypothesis around the familiarity with the source domain:

(H3) The numeracy level of users has a significant effect on the extent to which analogies affect user reliance on the system.
(H4) Familiarity with the source domain of the analogy has a significant effect on the extent to which the analogy affects user reliance on the system.

In addition to these last two hypotheses we will investigate the effects on reliance for all four hypotheses in light of a measure of subjective trust. Earlier research has shown that subjective trust can have an important influence on reliance and so we consider this to better understand the observed effects on reliance. The design of the study used to test these hypotheses is laid out in the next section.

2.4 Study Design

This section describes our experimental conditions, variables, procedure, and participants related to our main study. This study was approved by the human research ethics committee of our institution.[3]

2.4.1 Experimental Conditions

The main aspects of our hypotheses concern the effect of stated (overall) system accuracy, fixed in this experiment to 75%, and the addition of analogies to explain this stated accuracy. As a consequence, there are three conditions in the experiment: {`SysPred`, `PredAcc`, `AccAnalogy`}. Participants in all these conditions saw the systems' advice, but the three conditions differed in the inclusion of additional information:

- `SysPred`: does not include any further information. Example: *The system chooses to accept/reject this application.*

[3] `https://osf.io/9jqma/?view_only=c0c0dd12fa804b028cd29fbf9fd2ef4f`

- `PredAcc`: includes system accuracy in percent. Example: *The accuracy of the system is 75%, and it chose to accept/reject this application.*
- `AccAnalogy`: includes system accuracy *and* an analogy-based explanation for system accuracy. Example: *The system is 75% accurate, which is about as accurate as the five day weather forecast, and it chose to accept/reject this application* (with the weather report analogy used as an example here).

Participants in the `AccAnalogy` conditions were presented with one of three possible analogies along with the stated accuracy, with the prompts shown (ordered by how familiar we expected participants to be with these at the time of the experiment):

1. Vaccine efficacy: 'the system is 75% accurate, which is about as reliable as the AstraZeneca vaccine is for protecting against covid' (which is about 70% effective against the then-current Delta variant and somewhat more effective against earlier variants [86]). [4]
2. Accuracy of weather predictions: 'the system is 75% accurate, which is about as reliable as the five-day weather prediction' (which is also typically around 75% accurate). [5]
3. Train punctuality: 'the system is 75% accurate, which is about as reliable as the French trains are on punctuality' (which is 75% as listed in the 7th Rail Market Monitoring Report of the European Commission).

2.4.2 Measures And Variables

As mentioned, we use analogies to investigate whether a lack of appropriate reliance is due to a lack of understanding of global accuracy measures. It is important for this investigation to note the difference between (objective) reliance, which is the focus of our study, and (subjective) trust. We follow Lee *et al.* [40] in postulating that "trust in automation guides reliance when the complexity of the automation makes a complete understanding impractical and when the situation demands adaptive behavior that procedures cannot guide." Thus, we operationalize trust as a subjective user attitude, and reliance as objective user behavior that can be influenced by trust. As such, subjective trust can help us illuminate the effects we see on objective reliance [87].

To answer **H1** and **H2** we measure the reliance of participants on the system via two metrics: the agreement fraction and the switch fraction. These look at the degree to which participants are in agreement with system advice, and how often they adopt system advice in cases of initial disagreement. They are commonly used in the literature, for example in [57, 66]. In addition, we consider the overall accuracy and the accuracy under initial disagreement (*i.e.,* accuracy-wid) to measure participants' performance and appropriate reliance respectively. Since cases without initial disagreement do not clearly signal reliance on the system we restrict the scope of the appropriate reliance measure to accurately understand how participants handle divergent system advice. Following Schemer *et al.* [29],

[4] `https://www.nature.com/articles/d41586-021-02261-8`

[5] `https://spectrumlocalnews.com/tx/austin/weather/2020/10/08/wisconsin-weather-blog-meteorologist-wrong-rudd`

we adopted the relative positive AI reliance (RAIR) and relative positive self-reliance (RSR) metrics to measure appropriate reliance. When the AI system provides correct advice and the user makes a wrong initial decision, there are two possible reliance patterns: positive AI reliance (users switch to AI advice), negative self-reliance (users do not follow correct AI advice). When the AI system provides wrong advice and the user makes a correct initial decision, there are two other possible reliance patterns: positive self-reliance (users insist on their own initial decision) and negative AI reliance (users switch to another option). These measures are computed as follows:

$$\text{Agreement Fraction} = \frac{\text{Number of decisions same as the system}}{\text{Total number of decisions}},$$

$$\text{Switch Fraction} = \frac{\text{Number of decisons where the user switched to agree with the system}}{\text{Total number of decisions with initial disagreement}},$$

$$\text{Participant Accuracy} = \frac{\text{Number of correct final decisions}}{\text{Total number of decisions with initial disagreement}},$$

$$\text{Accuracy-wid} = \frac{\text{Number of correct final decisions with initial disagreement}}{\text{Total number of decisions with initial disagreement}},$$

$$\text{RAIR} = \frac{\text{Number of positive AI reliance}}{\text{Total number of positive AI reliance and negative self-reliance}},$$

$$\text{RSR} = \frac{\text{Number of positive self-reliance}}{\text{Total number of positive self-reliance and negative AI reliance}}.$$

To answer **H3**, we measured the numeracy level of the participants in our study. To do so we used the Subjective Numeracy Scale [88, 89], which has been widely validated as a measure for numeracy level in risk perception literature. We chose this subjective scale as opposed to an objective measure (asking participants to answer a number of quantitative questions) since prior work by Zikmund-Fisher *et al.* revealed that participants find objective tests stressful and unenjoyable [88]. Furthermore, the subjective scale has also been shown to correlate with the helpfulness of analogies in increasing risk perception [81], motivating our hypotheses.

To answer **H4**, perceived familiarity and helpfulness of the analogies is measured using 5-point Likert scale questions in the post-task questionnaire for those participants who were in the `AccAnalogy` condition. In addition to perceived familiarity and helpfulness, we gathered feedback from participants on their perception of the analogy-based explanations. To this end, we used the questions: "*Why did you find the analogy to be helpful or not helpful?*" and "*Please share any comments, remarks or suggestions regarding the use of analogies to explain the accuracy of the system.*"

For a deeper analysis of our results, a number of additional measures were taken:

- The Trust in Automation (TiA) (post-task) questionnaire [90], a validated instrument to measure (subjective) trust [87] consisting of 6 subscales: *Reliability /Competence* (TiA-R/C), *Understanding/Predictability* (TiA-U/P), *Propensity to Trust* (TiA-PtT), *Familiarity* (TiA-Familiarity), *Intention of Developers* (TiA-IoD), and *Trust in Automation* (TiA-Trust). Thus, we consider possible effects of trust on reliance, in accordance with Lee *et al.* [40].

- The Affinity for Technology Interaction Scale (ATI) [91], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [87].

Table 2.1 presents an overview of all the variables considered in our study.

Table 2.1: The different variables considered in our experimental study. "DV" represents a dependent variable.

Variable Type	Variable Name	Value Type	Value Scale
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
	Accuracy-wid	Continuous	[0.0, 1.0]
	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
Performance (DV)	Participant Accuracy	Continuous, Interval	[0.0, 1.0]
Trust (DV)	TiA-Reliability/Competence	Likert	5-point, 1: poor, 5: very good
	TiA-Understanding/Predictability	Likert	5-point, 1: poor, 5: very good
	TiA-Intention of Developers	Likert	5-point, 1: poor, 5: very good
	TiA-Trust in Automation	Likert	5-point, 1:strong distrust, 5: strong trust
Perception (DV)	Usefulness of Explanation	Likert	5-point, 1: useless, 5: very useful
Covariate	Analogy Domain	Categorical	{train, weather, vaccine}
	Numeracy Level	Likert	6-point, 1: low, 6: high
	Familiarity with Analogy Domain	Likert	5-point, 1: unfamiliar, 5: very familiar
	ATI	Likert	5-point, 1: low, 5: high
	TiA-Familiarity	Likert	1: unfamiliar, 5: very familiar
	TiA-Propensity to Trust	Likert	5-point, 1: tend to distrust, 5: tend to trust

2.4.3 Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [92]. To correct for testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.9$, and that we will investigate 3 different experimental conditions/groups. This resulted in a required sample size of 273 participants. We thereby recruited 316 participants from the crowdsourcing platform Prolific[6], in order to accommodate potential exclusion.

Compensation. All participants were rewarded with £1.5, amounting to an hourly wage of £7.5 deemed to be "*good*" payment by the platform (estimated completion time was 12 minutes). We rewarded participants with extra bonuses of £0.1 for every correct decision in the 10 trial cases. By incentivizing participants to reach a correct decision, we operationalize the concomitant "vulnerability" discussed by Lee and See[40] as a contextual requirement to encourage appropriate system reliance.

Filter Criteria. All participants were proficient English-speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one attention check (2 participants), or represented an outlier in terms of the amount of time they spent on our study. Outliers were participants (33 in total) who spent less than 7 minutes on the entire study. The resulting sample

[6] https://www.prolific.co

of 281 participants had an average age of 27 (SD = 8.64) and a gender distribution (70.1% female, 28.5% male, 1.4% other).

2.4.4 Procedure

The full procedure that participants followed in our study is illustrated in Figure 2.2. All participants first read the same basic instructions on the loan prediction task. Next, participants were asked to complete a pre-task questionnaire to measure their numeracy level and affinity for technology interaction.

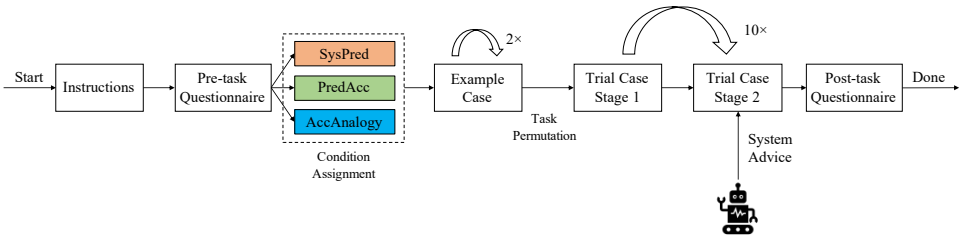


Figure 2.2: Illustration of the procedure that participants followed within our study.

Participants were then randomly assigned to one of three different experimental conditions, that differed in whether or not the system's prediction was supplemented with its accuracy and an analogy to explain the accuracy. After assignment, the participants were trained with two example cases before 10 trial cases. Selection of these cases is described in section 2.3.1. Finally, a post-task questionnaire was administered, using the 6 subscales of the TiA questionnaire discussed in section 2.4.2. Participants in the `AccAnalogy` condition were additionally asked for their familiarity with the source domain and the perceived helpfulness of the analogy they were presented with. To further ensure reliability of responses gathered in the questionnaires and the loan decisions, we added five attention check questions spread out at random through the different stages of the procedure [93].

2.4.5 Pilot Study

To determine the accuracy of the system (which was set to 75%) and verify the experimental procedure, a pilot study was conducted with 20 participants. They followed the same procedure as for the main experiment, except that no system advice was presented and so the ten trial tasks were only displayed once. In addition to the basic reward of £0.88 (equivalent to an hourly wage of £7.5), we set up a bonus of £0.1 for every correct decision to incentivize and encourage participants to concentrate on their individual decisions. On average, the pilot study was completed in 8.5 minutes, with an average accuracy of 0.43 (SD = 0.13). Moreover, participants performed better (M = 0.68, SD = 0.47) on the tasks that were estimated to be easy (based on linear regression) and relatively poorly on the tasks that we estimated to be difficult (M = 0.20, SD = 0.41).

This validated our task selection strategy, and suggested that the task is relatively difficult for humans to complete accurately, and decision support from an AI system would be realistic and meaningful. A 75% accuracy of the system is, then, a level which is helpful if the system is relied on, but still involves some risks and so calls for *appropriate* reliance,

as opposed to blindly following the system advice. Note that this design choice is motivated by Lee and See's work which emphasizes the role of uncertainty in dictating the need to facilitate appropriate reliance [40]. Had we set the accuracy at 90 or 95%, the situation would have been less clearly one of uncertainty for participants following the system advice.

2.5 Results

In this section, we present the results of our study. We discuss descriptive statistics, the outcomes of the hypothesis tests we conducted, and our exploratory findings pertaining to user perception of the analogy-based explanations.

2.5.1 Descriptive Statistics

Participants were distributed over the three experimental conditions: 87 (`SysPred`), 92 (`PredAcc`), 102 (`AccAnalogy`). The number of participants in the `AccAnalogy` condition was balanced between three analogy domains: there were 36, 35, and 31 participants in the *train punctuality*, *vaccine efficacy*, and *weather prediction* domains respectively.

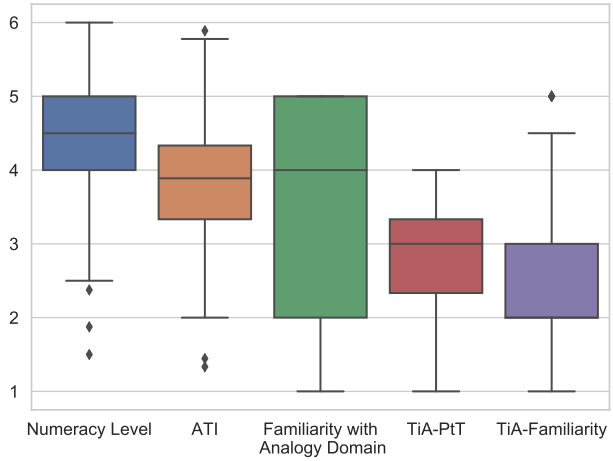


Figure 2.3: Box plot illustrating the distribution of the different covariates considered in our study. Among these covariates, *numeracy level* and *ATI* were measured on a 6-point scale, while others were measured on a 5-point scale.

Distribution of Covariates and Reliance Behavior. The covariates' distribution is as follows: *numeracy level* (M = 4.48, SD = 0.78, 6-point Likert scale, *1: low*, *6: high*), *ATI* (M = 3.82, SD = 0.78, 6-point Likert scale, *1: low, 6: high*), *familiarity with analogy domain* (M = 3.36, SD = 1.52, 5-point Likert scale, *1: unfamiliar*, *5: very familiar*), *TiA-Propensity to Trust* (M = 2.79, SD = 0.60, 5-point Likert scale, *1: tend to distrust*, *5: tend to trust*), and *TiA-Familiarity* (M = 2.38, SD = 0.98, 5-point Likert scale, *1: unfamiliar*, *5: very familiar*). This is illustrated in the boxplots in Figure 2.3.

Overall, all participants had at least one initial disagreement with system advice and 83.6% participants switched at least one decision after viewing the system's advice. On

average, the initial decision was the same as the final decision in 77.6% of all decisions. A small portion of participants (0.5% across all conditions) changed their mind despite an initial agreement with the system, to reach a final decision different from both their initial decision and the system advice.

Performance Overview. Recall that, informed by the pilot study, system accuracy was fixed to 75%. This meant that the system was in fact correct in 7 out of the 10 cases (which, though 70% accurate, is consistent with the reported 75% accuracy). The accuracy of the 281 participants in our main study was found to be 0.52 on average (SD = 0.14), rather worse than the overall system accuracy.

Table 2.2 shows the accuracy and error analysis for each of the 10 loan prediction tasks. In all tasks, we observe that the average accuracy of task and participants' error cause is highly correlated to its difficulty level (determined as described in Section 2.4.4). On relatively easy tasks, participants achieved high accuracy, and the errors in such cases are mainly caused by adopting incorrect system advice. In contrast, participants achieved a low accuracy on hard tasks, and demonstrated a reluctance to rely on the AI system which achieved superior performance. On average, however, we see that the mistakes made by participants are evenly split between cases where they should have relied on the system (49.3%) and cases where they should have disagreed with the system (50.7%).

Table 2.2: Participant performance on loan prediction tasks. Observed errors are split into two cases: 'Error-reliance' refers to the fraction of errors that were a result of participants agreeing with the system when it was wrong. 'Error-non-reliance' refers to the fraction of errors that were a result of participants disagreeing with the system when it was in fact correct. The difficulty levels are from 1 (*very easy*) to 5 (*very hard*), obtained by leveraging the predictions from a linear regression model. 'Accuracy', 'Error-reliance' and 'Error-non-reliance' are reported in percent (%).

Task-ID	Difficulty Level	Correct Answer	Accuracy	Error-reliance	Error-non-reliance
LP001030	1	accept	82.9	79.2	20.8
LP001849	1	reject	68.7	55.7	44.3
LP001806	2	accept	61.2	67.0	33.0
LP002142	2	reject	68.3	48.3	51.7
LP002534	3	accept	59.8	46.0	54.0
LP001451	3	reject	35.2	44.5	55.5
LP001882	4	accept	50.9	52.2	47.8
LP002181	4	reject	37.7	48.0	52.0
LP002068	5	accept	40.2	54.2	45.8
LP002840	5	reject	16.4	34.0	66.0

2.5.2 Hypothesis Tests

H1 and H2: the effect of accuracy and analogies on reliance and trust

Effect on Objective Reliance. To analyze the main effect of system accuracy (**H1**) and analogies (**H2**) on reliance, we conducted a Kruskal-Wallis H-test by considering the *experimental condition* as independent variable. The results showed no significant effects of *experimental condition* on reliance measures. The only effect that was significant was one of *experimental condition* on *participant accuracy*; $H(2) = 11.42$, $p = 0.003$. Participants in the `AccAnalogy` condition perform worse on *participant accuracy* (M = 0.48, SD = 0.14)

than those in the SysPred condition (M = 0.54, SD = 0.15) and the PredAcc condition (M = 0.55, SD = 0.14). Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) were used to compare all pairs of conditions. The difference in *participant accuracy* between SysPred condition and PredAcc condition was not significant; $U(N_{\texttt{SysPred}} = 87, N_{\texttt{PredAcc}} = 92) = 3682, p = 0.345$.

Thus, **H1** is not supported, as there is no change in reliance when system accuracy is given. **H2** is not supported either, as also providing analogies did not improve reliance on the system. Instead, we observed reduced participant accuracy, although this was not reflected in significantly lower agreement or switch fraction. To look for an explanation of these findings, we turn first to subjective trust, to see if this can explain the lack of effect of system accuracy information, as well as the counter-productiveness of analogies (more reliance would, after all, have been beneficial, given the accuracy scores reported earlier).

Effect on Subjective Trust. The impact of subjective trust was analyzed using an *Analysis of Covariance* (ANCOVA) with the *experimental condition* as between-subjects factor and *numeracy level*, *ATI*, *TiA-Familiarity* and *TiA-Propensity to Trust* as covariates. This allows us to explore the main effects of system accuracy (**H1**) and analogy-based explanation (**H2**) on subjective trust as measured by the relevant four subscales of the TiA. We decided to conduct AN(C)OVAs despite the anticipation that our data may not be normally distributed because these analyses have been shown to be robust to Likert-type ordinal data [94]. Table 2.3 shows the ANCOVA results pertaining to the four trust-related dependent variables.

Table 2.3: ANCOVA test results for **H1** and **H2** on trust-related dependent variables. "†" indicates the effect of the variable is significant at the level of 0.0125.

Dependent Variables	**TiA-R/C**			**TiA-U/P**			**TiA-IoD**			**TiA-Trust**		
Variables	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
Experimental Condition	0.00	0.997	0.00	1.18	0.309	0.01	0.78	0.459	0.00	0.02	0.979	0.00
Numeracy Level	0.608	0.436	0.00	1.47	0.227	0.00	4.97	0.027	0.01	0.89	0.346	0.00
ATI	5.17	0.024	0.01	6.66	**0.010**†	0.02	6.71	**0.010**†	0.02	2.40	0.123	0.01
TiA-Familiarity	1.55	0.214	0.00	2.51	0.114	0.01	11.57	**0.000**†	0.03	3.14	0.077	0.01
TiA-Propensity to Trust	158.92	**0.000**†	0.361	15.72	**0.000**†	0.05	62.92	**0.000**†	0.17	169.1	**0.000**†	0.38

As can be seen, there is no effect on any of the four subjective trust subscales by experimental condition. This suggests that the reduced accuracy in the analogy group (considered broadly) is not due to a lack of subjective trust in the system. Subjective trust in the particular system participants was presented with did correlate significantly with their familiarity with similar systems (*TiA-Familiarity*) and their general propensity to trust automated systems (*TiA-PtT*), as one would expect. Likewise, general affinity to technology (*ATI*) had a significant effect on subjective feeling of understanding the system (*TiA-U/P*) and trusting the intentions of the designers (*TiA-IoD*). This strengthens our confidence that we did succeed in measuring subjective trust in the system, as it depends on other subjective measures in the way one would expect. In a further Spearman rank-order test we observed that *TiA-PtT* significantly affects reliance and accuracy. Namely, there is a significant positive correlation between *TiA-PtT* and the reliance-based measures: *agreement fraction*, $r(279) = 0.277$, $p = 0.000$; *switch fraction*, $r(279) = 0.271$, $p = 0.000$; *accuracy-wid*, $r(279) = 0.191$, $p = 0.001$; *participant accuracy*, $r(279) = 0.203$, $p = 0.001$; *RAIR*, $r(279) = 0.266, p = 0.000$; *RSR*, $r(279) = -0.177, p = 0.003$. This confirms our postulated link

between subjective trust and objective reliance and so our null findings on objective reliance *w.r.t.* the experimental conditions can be partially explained by the observed lack of improvement in subjective trust. However, this fails to explain why the accuracy decreased in the analogy condition. We discuss this further while assessing the results for **H4**, where we examine the different analogy domains in detail.

H3: Numeracy level

To verify **H3**, we calculated Spearman rank-order correlation coefficients for *numeracy level* and dependent variables on the different experimental conditions and the sub-groups of the `AccAnalogy` condition. As can be seen in Table 2.4, we found that *numeracy level* does not significantly correlate with reliance measures when considering all participants in the `AccAnalogy` condition. Nor does it significantly correlate with reliance measures when focusing on participants in any of the three subgroups. We thus find no evidence in support of **H3**.

Table 2.4: Spearman rank-order correlation coefficient for numeracy level on reliance.

Dependent Variables	Agreement Fraction		Switch Fraction		Accuracy-wid		Participant Accuracy		RAIR		RSR	
Group	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
`AccAnalogy`	-0.019	0.852	0.066	0.510	-0.011	0.912	-0.083	0.408	0.025	0.804	-0.080	0.425
`AccAnalogy-train`	0.028	0.870	0.181	0.291	0.083	0.631	0.004	0.980	0.120	0.484	-0.180	0.292
`AccAnalogy-weather`	0.082	0.661	-0.009	0.963	-0.100	0.592	-0.010	0.957	-0.069	0.714	0.051	0.787
`AccAnalogy-vaccine`	-0.122	0.484	0.031	0.861	-0.073	0.676	-0.219	0.206	-0.006	0.971	-0.146	0.402

We carried out an exploratory analysis to examine the overall effect of numeracy level on reliance. To do so, we split the participants in all conditions into three groups: those with high (top 25%), medium (25-75%) and low (bottom 25%) numeracy. We conducted Kruskal-Wallis H-test with *numeracy group* and all dependent variables. The results indicate that there is no statistically significant difference between the three groups with different numeracy levels in terms of either reliance or subjective trust measures (see Table 2.5).

Table 2.5: Mean of dependent variables on different numeracy groups. "p" refers to the p-value for Kruskal-Wallis H-test results between three groups.

Dependent Variables	High Numeracy	Medium Numeracy	Low Numeracy	p
Agreement Fraction	0.69	0.69	0.71	0.578
Switch Fraction	0.39	0.44	0.41	0.509
Accuracy-wid	0.37	0.45	0.42	0.101
Participant Accuracy	0.50	0.52	0.55	0.248
RAIR	0.35	0.41	0.39	0.329
RSR	0.35	0.43	0.44	0.392
TiA-R/C	3.02	2.96	2.93	0.894
TiA-U/P	3.14	3.15	3.17	0.988
TiA-IoD	3.31	3.12	2.94	0.016
TiA-Trust	3.25	2.93	2.84	0.022

However, as shown in Table 2.5, participants in the low numeracy group did exhibit a higher agreement fraction and as a result had a higher accuracy in the task. Meanwhile, in

cases with an initial disagreement between user decision and system advice, participants in the medium numeracy group achieved higher appropriate reliance and switch fraction than other two groups. Oddly enough, low numeracy participants report virtually the same subjective understanding of the system as high numeracy participants, but lower subjective trust on the other measures. Though these results were not statistically significant, they potentially suggest that participants with lower numeracy might have felt the need to rely more on the system as they were less comfortable with the numerical task.

H4: Familiarity with analogy domains

Impact of Familiarity on Trust and Reliance. Finally, we investigated the role of analogy domains in detail. In line with **H4** we analyzed the main effect of *familiarity with analogy domain* on reliance. The results are: *agreement fraction*, $H(4) = 2.691$, $p = 0.611$; *switch fraction*, $H(4) = 8.165$, $p = 0.086$; *accuracy-wid*, $H(4) = 6.169$, $p = 0.187$; *participant accuracy*, $H(4) = 5.598$, $p = 0.231$; *RAIR*, $H(4) = 5.262, p = 0.261$; *RSR*, $H(4) = 5.233, p = 0.520$. There was no significant effect of familiarity on these objective measures. We, therefore, did not find support for **H4**, presumably because analogies generally speaking failed to improve user reliance.

To better understand the lack of effectiveness of analogies in shaping the reliance of users, we conducted a number of analyses. First, we considered the effect of familiarity with the analogy domain (which is a proxy for its effectiveness in clarifying a given measure, such as the stated system accuracy) on the subjective measures of trust. We found a significant effect of familiarity on the (subjective) *TiA Understanding/Predictability* measure with a Kruskal-Wallis H-test; $H(4) = 15.05$, $p = 0.005$. Participants who reported familiarity levels of '4' ($M = 3.30$, $SD = 0.52$) and '5' ($M = 3.39$, $SD = 0.47$) perform better than those who reported levels of '1' ($M = 2.88$, $SD = 0.51$) and '2' ($M = 3.01$, $SD = 0.51$). Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) were used to compare all pairs of conditions. The results suggest that participants with a higher *familiarity with analogy domain* tend to achieve higher *TiA-Understanding/Predictability*.

Familiarity and Usefulness (domain-agnostic). In the `AccAnalogy` condition, 56 participants reported a familiarity score greater than 3, and we considered them as the `familiar` group, while the remaining 46 participants were considered as being `unfamiliar` with the presented analogy domain. We conducted a Kruskal-Wallis H-test with *familiarity with analogy domain* and the self-reported *usefulness of analogy*. This analysis only considered participants in the `AccAnalogy` condition who were exposed to analogy-based explanations. The results showed that *familiarity with analogy domain* significantly affected the perceived *usefulness of analogy*; $H(4) = 41.46$, $p = 0.000$. Participants who reported familiarity scores of '*4*' ($M = 3.52$, $SD = 1.03$) and '*5*' ($M = 4.00$, $SD = 1.00$) also performed better than those who reported '*1*' ($M = 2.06$, $SD = 1.00$), '*2*' ($M = 2.45$, $SD = 0.74$) and '*3*' ($M = 2.38$, $SD = 1.19$). Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) were used to compare performance across all pairs of conditions. The difference in performance between both the `familiar` group and `unfamiliar` group was not significant.

Familiarity and Usefulness (domain-specific). To further confirm the effect of *familiarity with analogy domain*, we conducted a Kruskal-Wallis H-test with *analogy domain*

and *usefulness of analogy*. This effect was significant; $H(2) = 20.74$, $p = 0.000$. Participants in the `AccAnalogy-train` condition ($M = 2.42$, $SD = 1.08$) indicated a lower subjective usefulness of the analogy than those in the `AccAnalogy-weather` condition ($M = 3.74$, $SD = 1.09$) and the `AccAnalogy-vaccine` condition ($M = 3.34$, $SD = 1.16$). The results are in line with our expectations about how familiar participants were with the chosen analogy domains, given the global pandemic situation at the time of the experiment. This shows that choosing the right analogy makes a difference for these subjective measures, and that a well-chosen analogy can improve subjective measures of usefulness and understanding. As we did not have objective measures of understanding we cannot say whether this translates to objective understanding. However, we can draw further insights into the role of analogies by analyzing the participant perception of analogy-based explanations.

2.5.3 Participant Perception of Analogy-based Explanations

Finally, we analyzed the written responses of participants to the prompts "*Why did you find the analogy to be helpful or not helpful?*", and "*Please share any comments, remarks or suggestions regarding the use of analogies to explain the accuracy of the system.*" Authors of this chapter manually coded all participants' responses about the analogy-based explanations into the mutually exclusive categories of — `positive` ($N = 32$), `negative` ($N = 57$), `neutral` ($N = 4$), or `not reported` ($N = 9$). Using a random sample of the responses from participants, authors agreed on the categories for coding. We do not report inter-rater reliability, as disagreement between the authors was resolved through detailed discussions and critical reflection [95]. Example excerpts of the feedback received from participants are presented in Table 2.6. Using the thematic analysis software, ATLAS.ti,[7] we conducted a thematic analysis and selected the top-3 topics mentioned by users across three analogy domains (shown in Table 2.7).

Table 2.6: Excerpts from participants' responses to open questions regarding the analogy-based explanations.

Participant Feedback	Sentiment	Reason
I found the analogy to be helpful, because the weather forecast is something I am familiar with, and it gave me a pretty good idea of the accuracy of the system. I think the analogy was a perfect way to explain the accuracy of the system because it is something most people are very familiar with.	Positive	helpful with familiar reference
The weather can be unpredictable, and so even the experts cannot be 100% sure at all times. The analogy helped to determine whether I should take the system's advice 100% or not.	Positive	helpful with risk perception
I've never experienced the punctuality of a French train to know how reliable it is. I like the idea of using an analogy to explain the accuracy of the system.	Negative	unfamiliar with analogy domain
I usually don't trust the weather forecast 7 days out so I thought the same of the system. I find the weather forecast to be wrong most of the time so I thought it was ironic that it was compared to be 75% accurate.	Negative	distrusts or dislikes analogy domain

[7] https://atlasti.com

Table 2.7: Resulting main themes from the thematic analysis of participants' responses to the open questions pertaining to analogy-based explanations across domains.

Topic	**Participant Feedback**		
	Train	Weather	Vaccine
Familiarity	(1) Not helpful because it requires an understanding of the French train system, I would use an analogy that is easier for more people to relate to. (2) I don't know the punctuality of French trains. Analogies only work if they are commonly known. (3) I've never experienced the punctuality of a French train to know how reliable it is. I like the idea of using an analogy to explain the accuracy of the system.	I found the analogy to be helpful, because the weather forecast is something I am familiar with, and it gave me a pretty good idea of the accuracy of the system. I think the analogy was a perfect way to explain the accuracy of the system because it is something most people are very familiar with.	(1) It is a useful comparison that everyone is familiar with in today's world. I would get a vaccine with 75% efficacy. This was a strong explanation. (2) I am familiar with the vaccine analogy and it is something that is very relevant today.
Risk Perception	— no responses —	The weather can be unpredictable, and so even the experts cannot be 100% sure at all times. The analogy helped to determine whether I should take the system's advice 100% or not.	Just like a vaccine will not work effectively 100% of the time due to variations in human biology, a system to determine creditworthiness cannot take into consideration certain aspects of human behavior and therefore will not always be 100% correct.
Personal Experience	From experience I perceive the French train system to be highly efficient, therefore I did not trust the analogy and it did not collate with my experience. As we are working in facts and figures I prefer to not use an analogy that corresponds to something that is open to such a variation of circumstances that could arise as a train being delayed or on time.	I usually don't trust the weather forecast 7 days out so I thought the same of the system. I find the weather forecast to be wrong most of the time so I thought it was ironic that it was compared to be 75% accurate.	(1) I just found it kind of funny to be honest, I figure people will take it differently based on how they perceive the vaccine. For me it was just something funny and interesting. (2) I guess it let me know it only had about a 25% failure rate, but it also wasn't helpful because computer systems and vaccines are very different.

By analyzing the responses of participants who were satisfied with the analogy-based explanations for system accuracy (N = 32), we found the following main causes:

- 12 participants (37.5%) found it helpful to provide a reference frame that they are familiar with.
- 10 participants (31.3%) thought the analogy-based explanation made it easier to understand the system's accuracy.
- 3 participants (9.4%) felt the analogy-based explanation improved their risk perception.

By analyzing the responses of participants who were not satisfied with the analogy-based explanations for system accuracy (N = 57), we found the following main causes:

- 14 participants (24.6%) believed that the stated system accuracy itself, expressed in a percentage was sufficient for them to understand and inform their decisions.
- 14 participants (24.6%) reported that they were unfamiliar with the analogy domain and were therefore unable to use it in their decision making.
- 9 participants (15.8%) found that the explanations were not specific enough to be helpful in informing their decisions in the task.
- 8 participants (14.0%) reported that they did not trust the corresponding analogy domain and therefore found the analogies to be less helpful.
- 5 participants (8.8%) found that the analogy was irrelevant to the task at hand and therefore less helpful.

31.4% of the participants expressed positive opinions about the analogy-based explanation in our experiment, and 10 participants who expressed negative opinions (17.5%) also thought that a better analogy may be helpful. Overall, we observe that analogies can be (perceived as) useful if the target domain is not well-understood and the analogy is familiar. A third of the participants in the analogy domain found the analogies helpful, another 25% considered the accuracy measure as already well-understood. Even so, familiarity and the subjective helpfulness and understanding with which it correlates, did not lead to improvements in appropriate reliance or accuracy. On the contrary, participant accuracy was significantly lower in the `AccAnalogy` condition than in the other conditions.

We believe that this is due to the explanation that well-understood accuracy highlighted the fact that the system can be wrong, thereby making users more aware of the risk (for example, the second comment in Table 2.6), and leading to a slight change in decision making that led to lower accuracy. As discussed in Section 2.5.2, we found that accuracy decreased in the `AccAnalogy` condition, but subjective trust did not. If analogies indeed improved risk perception, as prior work [33, 63] have shown in other contexts, then participants may have viewed relying on the system as riskier than making their own decisions. We discuss this further in the next section, in light of the earlier findings on reliance when users are presented with information on system accuracy.

2.6 Follow-up Study: The Influence of Differing User Trust in Analogy Domains

To further understand the impact of users' trust in the analogy domains on their appropriate reliance, we conducted a within-subjects study in which each participant worked with AI systems where their stated accuracy was explained using analogies from three different

analogy domains. This study was approved by the human research ethics committee of our institution.[8]

2.6.1 Experimental Setup

Task Selection. To assess the impact of user factors on each analogy domain, we balanced the difficulty of the tasks for each analogy. We selected 4 tasks for each analogy domain in the same way as in the main study, using a regression model. Tasks were all predictions where the model had borderline confidence (*i.e.,* difficult tasks for the model) and were evenly split between two tasks where the model predicts approval and two tasks where the model predicts rejection.

We thus obtained three groups of 4 tasks each, where each group was explained by a different analogy domain. To maintain an accuracy level of 75%, we manually provide one incorrect prediction among the four tasks in each group. To prevent any bias caused by ordering, we kept the relative order of 3 groups, but shuffled the order of analogy domains provided to each participant and the task order within each group.

Procedure. We followed a similar procedure as in the main study (see Section 2.4.4). The main difference is that we did not separate participants into different experimental conditions. Instead, we separately assessed the user factors in each analogy domain before participants worked on one group of tasks explained with a single analogy domain.

Measures. We consider all covariates and reliance-based measures in the main study (see Section 2.4.2). However, we calculated the reliance-based measures according to each analogy domain. In addition, we assessed familiarity, trust, and confidence with the relevant analogy domain before each block of 4 tasks using that analogy domain. This was done using the following questions on a 6-point Likert scale:

- How familiar are you with **[analogy domain]** (punctuality of French trains / five-day weather forecasts / AstraZeneca vaccine for COVID-19)?
- To what extent do you trust the **[analogy domain]** (French train punctuality / five-day weather forecast / effectiveness of AstraZeneca vaccine for COVID-19) ?
- How confident are you with estimating the **[analogy domain]** (punctuality of French trains / accuracy of five-day weather forecasts / effectiveness of AstraZeneca vaccine for COVID-19) numerically?

As 4 tasks may be inadequate to assess the trust related measures for AI systems on each analogy domain, we did not consider the trust-related measures (*i.e., TiA-R/C*, *TiA-U/P*, *TiA-IoD*, and *TiA-Trust*) in this follow-up study.

Participants. Before recruiting participants, we computed the required sample size in a power analysis for a Within-Subjects ANOVA using G*Power [92]. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.025$ (*i.e.,* due to testing multiple hypotheses, **H3** and **H4**), a statistical power of $(1 - \beta) = 0.95$. This resulted in a required sample size of 245 participants.

[8] https://osf.io/9jqma/?view_only=c0c0dd12fa804b028cd29fbf9fd2ef4f

We therefore recruited 261 participants from the crowdsourcing platform Prolific, in order to accommodate potential exclusion. All participants were rewarded with £1.5, amounting to an hourly wage of £9 deemed to be "*good*" payment by the platform (estimated completion time was 10 minutes). Similar to the main study, we rewarded participants with extra bonuses of £0.1 for every correct decision in the 12 trial cases. All participants were proficient English speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform. Meanwhile, we pre-screened all participants in the main study from this study to prevent any learning effect. After data collection, we excluded participants from our analysis if they failed at least one attention check (2 participants), or represented an outlier in terms of the amount of time they spent on our study. Outliers were participants (11 in total) who spent less than 6 minutes on the entire study. The resulting sample of 248 participants had an average age of 38 (SD = 12.98) and a gender distribution (50% female, 50% male).

2.6.2 Results and Analysis

Domain-specific User Factor Distribution. The distribution of analogy-specific user factors is visualized in Figure 2.4. Most participants reported a low *Familiarity* with the punctuality of French trains (M = 1.70, SD = 1.14). In comparison, most participants were familiar with the five-day weather forecast (M = 5.08, SD = 0.94) and AstraZeneca vaccine (M = 4.65, SD = 1.25). *Trust* was similar for all analogy domains, with the punctuality of French trains scoring lowest (M = 3.57, SD = 0.99), the weather report scoring slightly higher (M = 3.85, SD = 1.04) and the AstraZeneca vaccine getting the highest trust scores (M = 4.36, SD = 1.33). As for *Confidence*, this too was lowest for the French train punctuality (M = 2.77, SD = 1.48). Both the weather report (M = 3.79, SD = 1.03) and AstraZeneca vaccine (M = 4.00, SD = 1.26) scored higher on *Confidence*. As can be seen, standard deviations indicate that there were individual differences in how participants perceived these different analogies, while the aggregate results also show that the choice of analogy has an overall impact. Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.025 ($\frac{0.05}{2}$) were used to compare all pairs of analogy domains. Our results indicate that: (1) participants showed a significantly higher *Familiarity*, *Trust*, and *Confidence* in the five-day weather report accuracy and the AstraZeneca vaccine effectiveness than the French train punctuality; (2) comparing the weather report and the AstraZeneca vaccine domains, we found that although participants reported a significantly higher *Familiarity* with the five-day weather report accuracy, they showed a significantly higher *Trust* and *Confidence* in the AstraZeneca vaccine effectiveness.

Main Effect of Domain-specific User Factors. To analyze whether these differences had an effect on performance, we conducted Friedman tests for reliance-based measures across the different analogy domains. The results show that no significant difference exists between the reliance-based measures across the three analogy domains: *Agreement Fraction*, χ^2 = 0.19, p = 0.91; *Switch Fraction*, χ^2 = 0.41, p = 0.81; *Accuracy-wid*, χ^2 = 1.28, p = 0.53; *Participant Accuracy*, χ^2 = 1.37, p = 0.50; *RAIR*, χ^2 = 0.62, p = 0.73; *RSR*, χ^2 = 2.89, p = 0.24. While participants show relatively lower *Familiarity*, *Trust*, and *Confidence* on French train punctuality, no significant difference exists in the reliance-based measures. This indicates that, although participants perceive the three analogy domains differently,

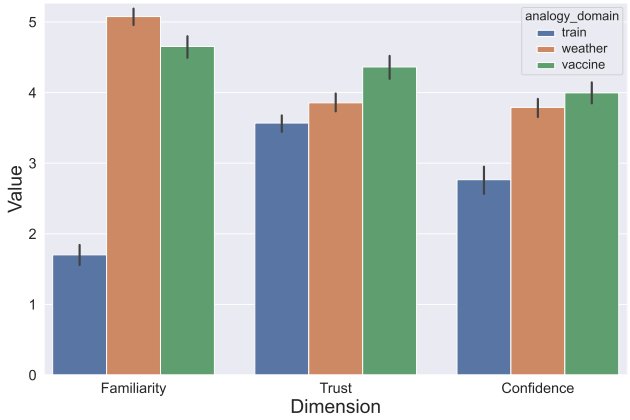


Figure 2.4: Bar plot illustrating the distribution of the different user factors considered in our study. All user factors were measured on a 6-point scale.

their reliance on the system is not affected by these differences in perception. Thus, we are reassured that our findings in the first study were not biased due to individual differences.

Table 2.8: Spearman rank-order correlation coefficient for user characteristics on reliance. "†" indicates the effect of variable is significant at the level of 0.025.

Dependent Variables	**Agreement Fraction**		**Switch Fraction**		**Accuracy-wid**		**Participant Accuracy**		**RAIR**		**RSR**	
User Factor	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Trust	0.039	0.286	0.077	0.036	0.053	0.151	-0.009	0.811	0.068	0.065	-0.041	0.266
Familiarity	-0.012	0.751	0.020	0.578	0.050	0.174	0.017	0.638	0.043	0.245	0.035	0.342
Confidence	-0.025	0.504	0.034	0.359	0.027	0.469	-0.054	0.139	**0.087**	**0.017**†	-0.018	0.619
Numeracy Level	-0.044	0.228	-0.048	0.189	-0.016	0.661	-0.041	0.262	-0.008	0.833	0.019	0.598
ATI	-0.061	0.097	-0.106	0.004	-0.035	0.334	-0.020	0.578	-0.082	0.026	0.050	0.173
TiA-Familiarity	-0.012	0.753	0.002	0.957	0.016	0.667	0.024	0.522	0.016	0.659	0.041	0.266
TiA-Propensity to Trust	**0.151**	**0.000**†	**0.102**	**0.005**†	0.075	0.040	**0.096**	**0.009**†	0.067	0.068	-0.060	0.103

Correlation Analysis for User Factors on Reliance. For further insights about all user factors on user reliance behaviors, we calculated Spearman rank-order correlation coefficients for reliance-based dependent variables across all groups of tasks. As can be seen in Table 2.8, we found that participants' trust, familiarity, and confidence with the analogies do not significantly affect reliance on the system. This further confirms our finding that differences in the perception of analogies do not affect reliance. Only participants' general *Propensity to Trust* shows a significant positive correlation with *Agreement Fraction*, *Swith Fraction*, and *Participant Accuracy*. This also aligns with our findings in main study (see Table 2.3) where the subjective trust in the AI system correlated significantly with their general *Propensity to Trust*. We also observed a positive correlation between users' Confidence and the RAIR they demonstrated, which indicates that users who have more confidence in the AI system, tend to more appropriately rely on the AI system.

2.7 Discussion

2.7.1 Key Findings

Our analysis of the responses to the analogies suggests that the problem is not one of a lack of understanding of what the stated accuracy measure means. Nor was the decline in reliance observed in the analogy case the result of a reduction in subjective trust. As discussed, there were no significant effects on the various TiA subscales, even though these subscales correlated as expected with other subjective measures. In fact, the cases where participants were familiar with the analogies led to a significantly higher subjective understanding of the system, though here too there was no translation into higher reliance. We thus see a significant decline in accuracy that does not seem to be explainable in terms of a decline in subjective trust. According to the results discussed in Section 2.5.2, participants who reported a higher numeracy level tended to rely less on the AI system and achieved worse appropriate reliance and team performance (*i.e.,* accuracy). Therefore, we argue it is likely that participants overestimated their skills to deal with numeracy and loan prediction task, and did so more in the `AccAnalogy` condition. Combined with existing findings that analogies help improve risk perception in dealing with numeracy, the reduced reliance on AI system may be caused by the risk perception brought by analogies. The only unexpected effect is that it improved risk perception to their detriment: making users think that relying on the relatively accurate AI system was riskier than trusting their own answer. User comments such as the second and fourth in Table 2.6 match this interpretation of the results. For example, "The weather can be unpredictable, and so even the experts cannot be 100% sure at all times. The analogy helped to determine whether I should take the system's advice 100% or not".

Positioning in Existing Work. Our findings may seem at first to contrast with the findings of Yin *et al.* [57], where the authors found a significant effect of stated accuracy on reliance. We did not find this to be the case in our study using the loan prediction task. When aiming to better explain the stated accuracy measure through the aid of analogies, we even saw a reduction in reliance. How do these contrasting findings fit together? We consider the crucial difference to their study [57] to be that the observed effect of stated accuracy on reliance was only found for very high stated accuracy levels (90 and 95%) and even then users only agreed with the system in 80% of cases (up from 75% with no/lower stated accuracy). Our study intentionally did not consider these high accuracy levels, to avoid inducing system reliance simply due to the near certain promise of making the right decision when relying on the system (and thus acquiring the monetary reward). At 75% accuracy, though significantly better than human performance, users (especially those with high self-reported numeracy level) were reluctant to rely on the AI system. And indeed, for stated accuracies around 75% Yin *et al.* also did not find an improvement in reliance. In fact, even for a stated accuracy of 50% the observed agreement fraction was around 80% – they did not find effective calibration of reliance, especially for lower levels of stated accuracy.

This explanation of the findings is also in line with the findings of Yin *et al.*, where participants started to rely more on the system after they were given an overview of their own performance and that of the system midway through the task (where generally the system performed better) [57]. This also aligns with the observed effect of *Propensity to*

Trust and *Numeracy Level* in our study where the AI system shows superior performance than human performance. Participants who reported higher numeracy levels tended to rely less on the AI system — potentially due to thinking they can do better than the AI system with a 75% accuracy. Their reduced reliance and accuracy can be caused by the illusion of their own competence with numeracy and this task [30]. In contrast, participants who showed a higher propensity to trust tended to treat the AI system advice as more trustworthy, and relied more on the AI system.

Potential Cause — Dunning-Kruger Effect. Prior work in human behavior and psychology that have studied poor task performance have observed participants' overestimation of their own performance as an important reason. These studies attribute the overestimation to a cognitive bias called the Dunning-Kruger effect [60, 61]. The Dunning–Kruger effect describes a tendency for incompetent individuals to overestimate their ability, and has been replicated across several tasks in different domains including crowd work [96]. While we cannot entirely attribute the under-reliance of participants on the AI system in our study to the overestimation of their skills on the loan prediction task, there is a substantial amount of support for this plausible explanation in existing literature [30, 97].

Numeracy Levels Did Not Play a Role. Following on from overestimation of one's skills as the potential cause for under-reliance on the AI system, our results suggest that this occurs regardless of the numeracy level of participants. Having said that, we did observe that participants with low numeracy levels exhibited a higher reliance, *i.e.,* agree with and switch towards system advice more often (see Table 2.5), though this effect is not significant. Furthermore, participants with lower numeracy levels tend to have lower Trust in Automation scores, which is significant for the Intention of Developers measure (cf. Tables 2.5). As these findings are statistically insignificant, we refrain from drawing conclusions from them. At most, we think that should it turn out that findings regarding numeracy are significant in later studies then they make intuitive sense. Low-numeracy participants might rely more on a system not because of higher subjective trust, but rather due to a struggle with the range of numerical information they have to deal with. Hence, they report lower subjective trust but display higher objective reliance.

2.7.2 Caveats and Limitations

Observations on Single Accuracy Level. While it is informative to observe a lack of calibration to the stated accuracy level of 75%, our study is limited due to the restriction to a single accuracy level. As discussed above, the research of [57] only found an effect for higher accuracy levels when participants were not given feedback on their own performance, so perhaps the lack of findings regarding analogies is partly a result of our chosen accuracy level. That being said, participants would have been significantly better off relying more on the AI system, so even with a single accuracy level the question of how to get users to rely appropriately on such a system remains a valuable and important one. Thus, the findings of our study are important even though a single accuracy level was used.

Limitations of Analogy Domains. Furthermore, while the analogies we chose differed on the main feature of familiarity (with participants generally being unfamiliar with French trains and familiar with weather reports and covid vaccines), and all had a relevant structural mapping from accuracy in the AI domain and reliability in the various analogy

domains, none were very close to the AI domain. Thus, it may be that participants' knowledge of the analogy domains was hard to apply in the AI domain. Alternatively, they might have preferred analogies closer to the task domain (loan predictions), to clarify the meaning of accuracy in that context. That being said, participants who were familiar with the presented analogy domains did rate their understanding of the system higher and found the analogies to be helpful. According to the results in the follow-up study, we also found that the differences in perception of analogies (on *Familiarity*, *Trust*, and *Confidence*) did not show a significant impact on reliance-based measures. We, therefore, do not consider the choice of analogies to be the reason behind the significant decrease in user reliance on the AI system in the `AccAnalogy` condition.

Framing of Analogies. The presentation of the analogies might also have been a limiting factor in our experimental study. In our study design participants saw the same analogy-based explanation in each task where they made a choice that was possibly informed by the system. While it seems realistic that the overall system accuracy would remain the same for the duration of the study, participants may have come to ignore the information after the first few tasks. That being said, we did observe a significant effect when analogies were added, suggesting that they were not completely ignored despite a static application to the system accuracy measure.

Analogies can benefit users in understanding something that is not easy to digest [51, 98]. So in tasks with input data which is easy to comprehend (*e.g.,* visual input), our findings may not apply. Furthermore, as reported by Nourani *et al.* [99], the domain knowledge (expertise) plays an important role in facilitating reliance. In the presence of such potentially dominant factors, which appear to have a significant impact on trust formation and reliance behavior of users, our findings may not hold. In short, if users do not lack in their understanding (e.g., of measures like the AI system accuracy) analogies may be of little help, and explanations may not be needed in the first place.

Consideration of Task Type. The loan prediction task has been widely used to study human-AI decision making where there is a clear risk associated with the decision and a potential benefit in adopting AI advice [62, 65, 100, 101]. This task also follows the scenario-based exploration of end-user interpretability of AI systems championed by prior work [102]. However, the external validity beyond this scenario and domain (i.e., in other human-AI decision making tasks) and type of data (i.e., other than numerical data) cannot be ascertained. Future work could explore the effectiveness of analogy-based explanations, and consider alternative XAI methods altogether, in different scenarios [103].

2.7.3 Implications and Future Work

Based on our findings, we reason that an overestimation of users' skills in the task may explain their under-reliance on the AI system. Future work should further explore the effects of providing feedback to users on their performance. For whereas Green *et al.* [65] found that feedback on single decisions was of little use, Yin *et al.* [57] found feedback of average user accuracy to be a good motivator for increased reliance on system advice (though note, again, that reliance in their study was not optimal either). The question is whether and how this increased reliance can be calibrated properly to the system accuracy. Note that it is not the aim of our work to treat reliance on AI systems as universally

desirable. However, to design and facilitate optimal team performance in human-AI decision making, it is pivotal to understand why users fail to achieve the theoretically possible higher accuracy — particularly when aided by a relatively more accurate AI system — and why users tend to demonstrate under-reliance. This is the spirit in which we explored the RQs in our work.

Regarding the use of analogy-based explanations, a complementary direction would be to consider the use of analogies to elucidate other general features of algorithms (*e.g.,* their decreased reliability when applied on outlier data, as such explanations have helped for appropriate reliance [50]), or to use analogies to explain more technical measures such as confidence scores and Shapley values. These instance-level measures may be harder to interpret than the global accuracy measure explored in our work, and allow for a more dynamic presentation of analogies. If users lack enough expertise to comprehend these instance-level measures, then we believe that analogies can be helpful. Analogies may fit how humans actually reason, as Wang *et al.* note in their discussion of analogical reasoning [104] and we have observed some subjective effects from the use of analogies for stated accuracy. For that reason, they might be useful in explaining other parts of AI systems. An interesting finding from our work in this context, is that an improved risk perception can lead to under-reliance on AI systems and perhaps result in sub-optimal final decisions. Thus, more work is required to understand how to balance these two — promote criticality with which users rely on AI systems to prevent over-reliance on the one hand, and encourage reliance on AI systems when the advice is accurate to decrease under-reliance on the other hand. The ultimate aim should be to support users in their decision making, while fostering a better understanding of the AI system and promoting appropriate reliance of users on the system.

In the pursuit of this goal, analogy-based explanations can be an option if the measures in question are not clearly understood by users. However, there are several questions that need to be explored. First, not all users may need the help of analogies. Second, the familiarity of the analogy is crucial to it being helpful. Third, analogies in some domains (such as vaccines, or indeed the five-day weather report which many consider less reliable than it actually is) may carry with them undesirable connotations that impact their usefulness or even increase distrust. At the same time, these findings also provide guidelines to generate and apply high-quality analogies for explainability. For example, when users explicitly indicate that they find it difficult to interpret an explanation, we can provide an analogy as an alternative. This gives laypeople a better chance to understand challenging explanations. Here, user's beliefs and experiences may play an important role in the adoption of analogy-based experience and so we need to understand these users previous knowledge better in order to ensure the effectiveness of provided analogy-based explanations. In line with that, future work should consider exploring the potential of adaptive and personalized analogy-based explanations.

2.8 Conclusions

The two main research questions for this chapter were: 'How does the understanding of stated system accuracy affect reliance of users on the AI system?' and 'How does explaining stated system accuracy using analogies affect the reliance of users on the AI system?'. As we have discussed, the conclusion to draw from our experiment is that users are no

better at calibrating their reliance on the system when they better understand system accuracy. In fact, analogies made users less accurate, presumably because they became more aware of the risk that the system makes mistakes. A lack of understanding of the accuracy level is not the reason users fail to rely on the system appropriately. Thus, the limited understanding of stated accuracy is not to blame for under-reliance. This tallies with our finding that numeracy level, a factor one would expect to be relevant for a task filled with numerical information, had no significant effects on system reliance or accuracy.

Although our findings do not directly inform how we can facilitate appropriate reliance, we have identified important research directions that can further our understanding of system reliance in the complex and timely area of *Human-AI interaction.* Based on what is understood in the HCI community, we consider it likely that users' overestimation of their own skills is the main reason that explains why participants failed to rely on the AI system's advice as much as would be appropriate given the system accuracy, and their own lower performance. It seems that they considered 75% accuracy to be on the low side, and estimated their own performance to be better than that. This would fit in with the significant results observed for higher accuracies and the effect of *Propensity to Trust* on reliance. Further research is needed here, but it is striking that the level of understanding of the presented numerical information has little bearing on user reliance.

We also found that explaining the stated accuracy of the AI system with analogies was not the helpful tool we hypothesized it to be. However, our findings revealed that analogy-based explanations can be experienced as helpful by users when adjusted to their needs. In particular, we observed a set of guidelines for the use of analogies in line with that of earlier research on analogies in risk perception, which will help in the implementation of analogies in cases where a problematic lack of understanding is observed. If analogies are chosen to alleviate such a problem, one should pay attention to: (1) users' familiarity with the source domain, (2) their sentiments and expectations about the source domain, and (3) users' risk perception. We hope our findings and implications may help researchers have more insights about facilitating appropriate reliance and leveraging analogies to explain numerical attributes.

3

The impact of Dunning-Kruger Effect

The dazzling promises of AI systems to augment humans in various tasks hinge on whether humans can appropriately rely on them. Recent research has shown that appropriate reliance is the key to achieving complementary team performance in AI-assisted decision making. This chapter addresses an under-explored problem of whether the Dunning-Kruger Effect (DKE) among people can hinder their appropriate reliance on AI systems. DKE is a metacognitive bias due to which less-competent individuals overestimate their own skill and performance. Through an empirical study (N = 249), we explored the impact of DKE on human reliance on an AI system, and whether such effects can be mitigated using a tutorial intervention that reveals the fallibility of AI advice, and exploiting logic units-based explanations to improve user understanding of AI advice. We found that participants who overestimate their performance tend to exhibit under-reliance on AI systems, which hinders optimal team performance. Logic units-based explanations did not help users in either improving the calibration of their competence or facilitating appropriate reliance. While the tutorial intervention was highly effective in helping users calibrate their self-assessment and facilitating appropriate reliance among participants with overestimated self-assessment, we found that it can potentially hurt the appropriate reliance of participants with underestimated self-assessment. Our work has broad implications on the design of methods to tackle user cognitive biases while facilitating appropriate reliance on AI systems. Our findings advance the current understanding of the role of self-assessment in shaping trust and reliance in human-AI decision making. This lays out promising future directions for relevant HCI research in this community.

3.1 Introduction

In the last decade, powerful AI systems (especially deep learning systems) have shown better performance than human experts on many tasks, sometimes outperforming humans

This chapter is based on a peer-reviewed paper: **Gaole He**, Lucie Kuiper, and Ujwal Gadiraju. *Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems.* In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1-18. 2023. https://doi.org/10.1145/3544548.3581025.

by a large margin [66, 105]. Attracted by the predictive capability of such AI systems, researchers and practitioners have started to adopt such systems to support human decision makers in critical domains (*e.g.,* financial [65], medical domains [106]). With the wish of complementary team performance, one goal of such human-AI collaboration is *appropriate reliance*: human decision makers rely on an AI system when it is accurate (or perhaps more precisely, when it is more accurate than humans) and do not rely on it when the system is inaccurate (or, ideally, whenever it is wrong). In such a collaborative decision process, human factors (*e.g.,* knowledge, mindset, cognitive bias) and the explanations for AI advice are important for trust in the AI system and for human reliance on the system. Several prior works have carried out empirical studies within this context of human-AI decision making, to explore the effectiveness of different kinds of explanations and the role of human factors in shaping such collaboration [25, 52, 65, 66, 70, 107–109].

In recent literature exploring human-AI interaction, researchers have shown a great interest in understanding what shapes user trust and reliance on AI systems. They found that factors like first impression [87], AI literacy [108], risk perception [65, 110], and performance feedback [49, 111] among others, play important roles in shaping human trust and reliance on AI systems. Explanations (*e.g.,* feature attribution of input) have been found to be useful in promoting human understanding and adoption of AI advice [25, 66, 70, 107] and He *et al.* [98] recently proposed analogies as an instrument to increase the intelligibility of explanations. However, prior studies observed improvements in performance in the presence of explanations only when the AI system outperformed both the human and the best team [70]. One reason for such phenomenon is under-reliance, which indicates humans do not rely on accurate AI predictions as often as it is ideal to [107, 112, 113]. In this chapter, we explore whether Dunning-Kruger effect (DKE) [60] – a metacognitive bias due to which individuals overestimate their competence and performance – affects user reliance on AI systems. This a particularly important metacognitive bias to understand in the context of human-AI decision making, since one can intuitively understand how inflated self-assessments and illusory superiority over an AI system can result in overly relying on oneself or exhibiting under-reliance on AI advice. This can cloud human behavior in their interaction with AI systems. However, to the best of our knowledge no prior work has addressed this. In addition, DKE is closely related to user confidence in decision making, which has been identified as an important user factor and has been recently explored in the context of human-AI decision making [65, 114]. To achieve the goal of appropriate reliance, users are expected to adequately calibrate their self-confidence and their confidence in the AI system. Our work can lead to fundamental HCI insights that can help facilitate appropriate reliance of humans on AI systems.

To explore the impact of DKE on user reliance, we need to first identify participants who demonstrate the DKE (*i.e.,* participants who perform relatively poorly but overestimate their performance). According to existing research on the DKE [115, 116], the participants representing the bottom performance quartile tend to overestimate their skill and depict an illusory superiority, while those in the top performance quartile do not exhibit such a trend. Researchers have also operationalized self-assessments to serve as indicators of competence in different online tasks [96]. Informed by such prior work, we consider overestimated self-assessments in the context of human-AI decision making as an indicator of the DKE and explore it further. Through an explicit analysis of partici-

pants' performance in the bottom quartile, we verified that the overestimation in their performance is highly indicative of DKE in our study. In this scope, we explore whether we can design interventions to help users improve their own calibration of their skills in the task at hand.

Inspired by existing work in mitigating cognitive biases such as the DKE [60] and promoting appropriate reliance [107, 108, 117], we propose to leverage tutorials to calibrate their self-assessment through revealing the actual performance level of participants with performance feedback. In such a tutorial, after the initial decision making, participants are provided with correct answers and explanations to contrast with their final choice (if they make a wrong choice). As pointed out by existing research [118], one cause of DKE can be that people place too much confidence in the insightfulness of their judgments. When the correct answer differs from their own choice, they may refrain from trusting such ground truth in the absence of additional rationale. To ensure the effectiveness of revealing users' shortcomings, we provide them with contrastive explanations which point out not only the reason for correct answers, but also why their choice was incorrect. Based on prior work, we expect such a training session to help users realize their errors and calibrate their self-assessment. Furthermore, they become more skillful at the task, which is also highlighted by Kruger *et al.* [60] in mitigating DKE.

When AI advice disagrees with human decisions, the lack of rationales may be a reason not to adopt AI advice. To help participants interpret the AI advice, we leverage logic units-based explanations which reveal the AI system's internal states. When users recognize that an explanation provides reasonable evidence for supporting AI advice, it is much easier for them to resolve disagreement in their decision making. As a result, participants have a better opportunity to know and understand when they "should" in fact rely on AI systems. From this standpoint, effective explanations alongside the tutorial may help mitigate the impact of the Dunning-Kruger Effect on user reliance. To analyze the impact of DKE on user reliance on AI systems in this chapter, we aim to find answers for the following two research questions:

RQ1: How does the Dunning-Kruger Effect shape reliance on AI systems?
RQ2: How can the Dunning-Kruger Effect be mitigated in human-AI decision making tasks?

To answer these questions, and based on existing literature, we proposed four hypotheses considering the effect of the overestimation of performance on (appropriate) reliance, the effect of the tutorial intervention on self-assessment calibration and reliance for participants with miscalibrated self-assessment, the effect of logic units-based explanations and tutorial intervention on reliance and team performance. We tested these hypotheses in an empirical study (N = 249) of human-AI collaborative decision making in a logical reasoning task (*i.e.,* multi-choice logical question answering based on a context paragraph). We found a negative impact of the DKE on human reliance behavior, where participants with DKE relied significantly less on the AI system than their counterparts without DKE. To mitigate such effects, we designed a tutorial intervention for making users aware of their miscalibrated self-assessment and provided logic units-based explanations to help explain AI advice. Although we found that the intervention tutorial was highly effective

in improving participants' self-assessments, their improvement in appropriate reliance and performance is limited (statistically non-significant). Moreover, no obvious benefits were found with introducing logic units-based explanations in the logical reasoning task.

Our results highlight that the overestimation of performance will result in under-reliance, and such miscalibrated self-assessment can be improved with our proposed tutorial intervention. We also found that participants who overestimated their performance demonstrated an increased appropriate reliance, which the calibration of self-assessment can partially explain. However, this was in contrast to participants who initially underestimated their performance – while they calibrated their self-assessment, they achieved significantly worse appropriate reliance and performance. One potential cause is that such tutorials help them recognize their actual performance but also cause the illusion of superiority to AI systems. Such finding is also in line with algorithm aversion [33], where users are less tolerant of the mistakes made by AI systems. In addition, we found that the users' propensity to trust goes a long way in shaping trust in AI systems, despite our tutorial not having an effect in reshaping subjective trust. Based on the results from our empirical study, we provide guidelines for designing more comprehensive user tutorials and point out promising future directions for further research around self-assessments in the context of human-AI decision making. Although we found that miscalibrated self-assessments may hinder appropriate reliance (*i.e.,* participants with DKE relied less on AI systems), the participants with accurate self-assessment did not necessarily show optimal appropriate reliance (*e.g.,* we found that participants with underestimation showed better appropriate reliance and performance). This interplay between self-assessment and reliance on AI systems is potentially more complex than what can be explained by a linear relationship and, therefore, deserves further research.

In summary, we explored the effectiveness of a tutorial intervention to mitigate the DKE and, in turn, facilitate appropriate reliance. We found evidence suggesting its effectiveness through an empirical study in a logical reasoning task. Our work has important implications for HCI research in the realm of human-AI interaction. Our findings indicate that incorrect self-assessments and a prevalent meta-cognitive bias can affect user objective reliance on the AI system. Thus, while designing for optimal human-AI interaction, it is important to consider the extent to which users are aware of their own abilities and that of the AI system. Our work is an important first step towards furthering our understanding of how cognitive biases shape human reliance on AI systems, an understudied aspect in this quickly evolving realm of research. Considering the unique and evolving landscape of AI systems, the associated metaphors, and end-user expectations that are mediated through abstractions and their own experiences, we believe that studying the role of the DKE in the human-AI decision making context is a timely and unique contribution. We hope that our work can inform future research on designing human-AI interactions that can facilitate appropriate reliance on AI systems.

3.2 Background and Related Work

This chapter contributes to the growing literature on human-AI interaction, collaboration, and teaming, by exploring **how the Dunning-Kruger Effect shapes user reliance on AI systems** and **whether such effect can be mitigated with a user tutorial that highlights the fallibility of AI advice and logic units-based explanation**. Thus, we

position our work in different strands of related literature: the general literature on AI-assisted decision making and what roles explanations play in such collaboration (3.2.1), more specific literature on promoting appropriate reliance (3.2.2), the contradicting literature on algorithm aversion and algorithm appreciation (3.2.3), and finally the literature on self-assessments, which has been explored in psychology and other HCI studies (3.2.4).

3.2.1 Human-AI Collaborative Decision Making

In recent years, AI-assisted decision making has received more and more attention. In such collaboration, user factors and interaction with AI systems are observed to be of much impact on final user behaviors. Among these work, most researchers are interested in how users shape their trust in AI systems and how user behaviors will be affected by AI systems. Topics like performance feedback [49, 119], risk perception [110, 120], uncertainty [121] and confidence [66, 107, 114] of machine learning models, impact of explanations [70, 122] have been extensively studied in human-AI decision making. Meanwhile, fairness, accountability, and transparency of incorporating AI systems for collaborative decision making received more and more attention from a wide range of stakeholders [123, 124]. For a more comprehensive survey of existing work on Human-AI decision making, readers can refer to [22].

According to GDPR, the users of AI systems should have the right to access meaningful explanations of model predictions [125]. Under this perspective, more and more researchers have started to provide human-centered explainable AI (XAI) solutions to promote human-AI collaboration [27, 104, 124, 126, 127]. Up to now, the benefits of incorporating XAI methods in human-AI collaboration are still limited [22, 70]. As reported by most existing work, though XAI methods can aid understanding of AI advice, such effect does not necessarily lead to clear performance improvement [25, 70]. For instance, Liu *et al.* [25] observed that interactive explanations may "reinforce human biases and lead to limited performance improvement". Based on a comprehensive literature review, Wang *et al.* [107] proposed three desiderata of AI explanations to promote appropriate reliance: (1) critical for people to understand the AI, (2) recognize the uncertainty underlying the AI, and (3) calibrate their trust in the AI in AI-assisted decision making. With such ideal properties, effective explanations may also potentially help participant realize their weakness and mistake when they disagree with AI advice. Under this perspective, we also explored whether logic units-based explanations can help participants calibrate their self-assessment and promote appropriate reliance.

3.2.2 Empirical Studies on Appropriate Reliance

AI systems and human decision makers are supposed to achieve complementary team performance through taking advantage of both powerful predictive capability of AI systems and flexibility of human users to handle complex decision tasks. However, existing literature still struggles to find such complementary team performance — in most empirical studies, AI alone performs much better than human-AI team [22, 25]. With further analysis, researchers point out two main causes: (1) under-reliance, users fail to fully take advantage of powerful AI systems, and (2) over-reliance, users fail to rely on themselves when they actually outperform AI systems.

To promote appropriate reliance, existing research mainly focused on mitigating

under-reliance and over-reliance. Different interventions like cognitive forcing functions [55], user tutorial [50, 108] and explanations [107] are proved to be highly effective in mitigating such unexpected reliance patterns. Buçinca *et al.* [55] introduced three types of cognitive functions to mitigate over-reliance: show AI advice on demand, update decision with AI advice after the initial decision, and keep participants waiting for a while before providing advice. Their experimental results indicate that such cognitive forcing functions are even more effective than simple XAI methods in mitigating over-reliance. With a comparative study of four types of different explanations, Wang *et al.* [107] reported that feature importance and feature contribution explanations can promote appropriate reliance with mitigating under-reliance.

"User tutorials, when presented in appropriate forms, can help some people rely on ML models more appropriately" [108]. Another important branch is educating users with user tutorials, which stands out in recent years. On one hand, such user tutorials make users aware of the weakness of AI systems, which further calibrate user trust and reliance on AI systems. For example, Chiang *et al.* [50] found that a brief education session (to increase people's awareness of the machine learning model's possible performance disparity on different data) can effectively reduce over-reliance on out-of-distribution data. On the other hand, such a system can educate participants with domain-specific knowledge extracted from an AI system, which further improves users' capability. As a typical example, Lai *et al.* [117] proposed model-driven tutorials to help humans understand patterns learned by models in a training phase. Inspired by this series of research, we also explored whether DKE can be mitigated with user tutorial. For the purpose of calibrating self-assessment, we include performance feedback and explanations to contrast wrong user choice with correct answers.

3.2.3 Algorithm Aversion and Algorithm Appreciation

In the face of intelligent predictive agents, which may outperform human experts, people show two contradicting altitudes: *Algorithm Aversion* and *Algorithm Appreciation*. Compared to human forecasters, people more quickly lose confidence in AI systems after seeing them make the same mistakes [33]. Thus, some users are reluctant to use superior but imperfect algorithms [128]. Such a phenomenon is called "Algorithm Aversion," which has been observed across multiple domains, like moral decision making [129], economic bargains [113], medical diagnosis [130], and autonomous driving [131]. Burton *et al.* [128] summarized the cause and solution of algorithm aversion with five aspects: expectations and expertise, decision autonomy, incentivization, cognitive compatibility, and divergent rationalities. Meanwhile, Dietvorst *et al.* [35] found that such algorithm aversion can be overcome with the chance to modify algorithm advice. Readers can refer to two recent survey papers [128, 132] for a comprehensive literature review. In contrast, Logg *et al.* [34] found that users were influenced more by the algorithmic decision instead of human decision, and they first coined the notion of "Algorithm Appreciation" to describe such a phenomenon. Others revealed similar findings in contexts where tasks are perceived as being more objective [133], machines share rationale with humans [134] or with prior exposure to similar systems [135].

Besides contradicting attitudes towards the use of AI systems, prior work has shown how different human factors such as algorithmic literacy [136], expertise [34], and cogni-

tive load [137] can affect users' final adoption of algorithmic advice. For example, users' algorithmic literacy [136] about fairness, accountability, transparency, and explainability is found to greatly affect their trust and privacy concern in adopting the advice from AI systems. Logg *et al.* [34] found that experts may even show more tendency to discount algorithmic advice when compared to laypeople. Furthermore, these factors can also affect the extent to which users show algorithm aversion or algorithm appreciation. For instance, You *et al.* [137] argue that algorithm appreciation declines when the transparency of the advice source's prediction performance further increases. In their study, they used a series of numbers instead of aggregated average performance, which increases the transparency of prediction performance. But they observed a decrease in algorithm appreciation, which was explained by the greater cognitive load imposed by the elaborated format. A recent work [138] found that the choice of framings of human agents and algorithmic agents may affect user perception of agent competence (*i.e.,* expert power), which further affects user behavior and cause inconsistent observations of algorithm aversion and algorithm appreciation. In this chapter, since we explore means to facilitate appropriate reliance of humans on AI systems, we position our findings in the context of the research breaching algorithm aversion and appreciation. Future work can further explore the role of algorithmic aversion and appreciation in the context of interventions to facilitate appropriate reliance on AI systems.

3.2.4 Self-assessment in HCI Studies

Evaluating one's own performance on a task, typically known as "self-assessment", is perceived as a fundamental skill, but people appear to calibrate their abilities [139] poorly. In general, most people tend to overestimate their own abilities. The cause of such an effect is multi-fold, like people tend to think they are above average and people place too much confidence in the insightfulness of their judgments [118]. With self-assessment, existing HCI research has explored using it as a measure for different purposes: Gadiraju *et al.* [96] used self-assessment for competence-based pre-selection, Green *et al.* [65] measured users' risk assessment by comparing self-reported confidence with their actual performance, and Chromik *et al.* [62] compared perceived understanding of XAI methods with their actual understanding to reveal users' illusion of explanatory depth.

Dunning-Kruger effect (DKE) [60] described the dual burden the unskilled suffer from, besides the low performance, the unskilled will also lack the skill to estimate their own ability. Kruger *et al.* also found that a training session to increase the skills of participants is highly successful in mitigating such effect [60]. It had some positive effects and showed that by increasing knowledge, the overestimation could also be reduced. Further work also proved the effectiveness of such training in different domains like medicine [140] and economics [141].

Besides the popularity in psychology research, Dunning-Kruger effect was also studied in human-computer interaction field. In a recent study, Schaffer *et al.* [97] conducted a user study based on Diner's Dilemma game. They found that participants who considered themselves very familiar with the task domain showed more trust in an intelligent assistant but relied less on it. Presenting explanations was not as effective as expected, and sometimes even resulted in automation bias. Using logical reasoning tasks with varying difficulty levels, Gadiraju *et al.* [96] showed that online crowd workers also fall prey

to the DKE. The authors proposed the use of self-assessments in a pre-selection strategy to improve quality-related outcomes. Informed by prior literature, we selected logical reasoning tasks as the exploratory lens to address our research questions since the tasks themselves are straightforward to understand for laypeople, but with increasing difficulty, they also create room for inviting AI advice. This serves suitably to study the DKE in the context of human-AI decision making.

3.3 Method and Hypothesis

In this section, we describe the logical reasoning task (*i.e.,* multi-choice logical question answering based on a context paragraph) and present our hypotheses.

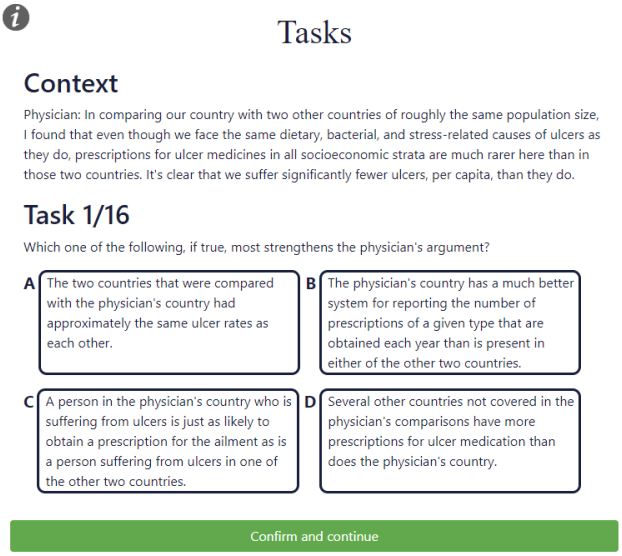


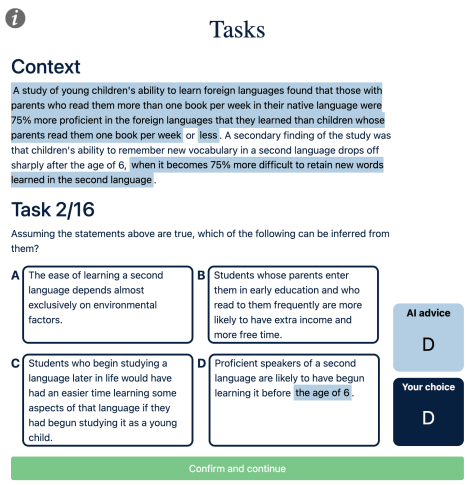
Figure 3.1: An example of a logical reasoning task used to obtain an initial human decision in the two-stage decision making process.

3.3.1 Logical Reasoning Task

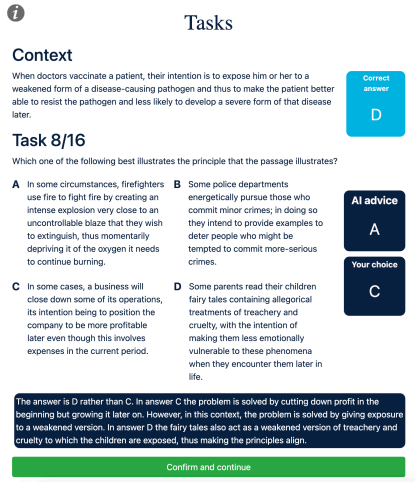
Prior work in the human-AI decision making context has explored how one can reliably study human behavior in proxy tasks. These work has established the importance of designing tasks, where users can find that there is a need to rely on AI (*e.g.,* owing to the task difficulty or a perceivable benefit) and where there is a risk associated with such reliance (*e.g.,* dealing with an imperfect AI system) [55, 87]. This follows from the work of lee *et al.* [40] who defined trust in the Human-AI interaction context as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability.*" The basis for our experimental setup is a task where participants are asked to choose an option in a multi-choice setting based on a paragraph of context presented to them (an example of the interface page is shown in Figure 3.1). We use the publicly avail-

able Reclor[1] [142] dataset to this end. The dataset corresponds to characteristically high difficulty of logical reasoning tasks and has been used in prior work exploring Human-AI team performance [70]. This task was chosen as a realistic scenario for human-AI collaboration, where humans incentivized to complete the task accurately, may have the capability to reason accurately and find the right answer, but may also evidently perceive a benefit in adopting AI advice. In addition, the Dunning-Kruger Effect which has been widely replicated in a variety of contexts has been shown to be prevalent in the domain of logical reasoning as well [60, 116].

In the basic setting of the task, participants are presented with three snippets of information: (1) a context paragraph, (2) a question related to this context, and (3) four different options corresponding to the question. Among the four options, a single option is deemed to be the best match to the question (*i.e.,* ground truth). Participants are asked to first go through the context paragraph, and then make a choice based on the question. This simulates a realistic scenario where participants make decisions in a reading comprehension setting. While humans are capable of handling such tasks, AI systems may outperform them by extracting useful information and dealing with complex reasoning structures which require a larger working memory capacity. The task interface is shown in Figure 3.2a.



(a) Logical question answering page with AI advice.



(b) Tutorial page with manual explanation.

Figure 3.2: Screenshots of the task interface. In panel (a), logic unit-based explanations are highlighted. In panel (b), the rationale of correct answers is shown.

Two-stage Decision Making. To analyze human reliance on AI systems, all participants in our study worked on tasks with a two-stage decision making process. In the first stage, only task information was provided, and participants were asked to make decisions themselves (example shown in Figure 3.1). After that, we showed the same task with AI advice

[1] https://whyu.me/reclor/

(and *explanations* depending on the experimental condition) and provided an opportunity for the participants to alter their initial choice. An example of second stage is shown in Figure 3.2a, where "Your choice" shows the initial decision participants made in the first stage. This setup of an initial unaided decision and the presentation of advice from an AI system in order to make a second and final choice is similar to the update condition in [65], and in line with findings that people first make a decision on their own and only then decide whether to incorporate system advice [85]. It also fits with the research of Dietvorst *et al.* [35] on trust in two-stage decision making.

Quality Control. To ensure participant reliability and that participants worked on the logical reasoning tasks genuinely (*i.e.,* read the context paragraph and question carefully), we employed three attention check questions during the study process [143]. For this purpose, we embedded explicit instructions asking participants to select a specific option either in the context paragraph (once) or the question (twice). For example, we embedded the instruction, "*Confirm that you have read the context by selecting answer B.*" into a context paragraph on the task interface (which looks nearly identical to other tasks). A conservative estimate through trial runs reflected that participants would take at least 1 minute to complete each task. As a further quality control measure, we deactivated the submit button corresponding to each task page (including tasks in tutorial phase) for 30 seconds. Since attention check pages do not require deliberation, we reduced that time to 5 seconds.

3.3.2 Logic Units-based Explanations

In natural language processing tasks, feature attribution methods (*e.g.,* text highlights on input) are the most popular in existing literature. However, multiple pieces of research work point out that such token-level highlights are still hard to interpret [144–146]. Meanwhile, since logical reasoning tasks highlight the potential for logical reasoning congruent to human understanding, explanations based on logic units (*i.e.,* text spans) may be a better choice to reveal how AI systems reach their final decision. With this perspective, we drew inspiration from LogiFormer, proposed by Xu *et al.* [147], who conducted logical reasoning with logic units based on pre-trained language models to generate such explanations. LogiFormer adopted a graph transformer network for logical reasoning of logic units, where the logic units are text spans connected with causal relations. Following this interpretability design, we also relied on the self-attention matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ (n indicates the number of logic units) from the last layer of the graph transformer network and identified the important logic units with the following formula:

$$E = Argmax_k(\sum_{j=1}^{j=n} \mathbf{A}_{ij}), \tag{3.1}$$

where E is the top-k logic units which receive most attention from other logic units (*i.e.,* we calculated it with the sum along each column of the self-attention matrix). One example of such explanation is shown in figure 3.2a.

Our implementation and extracted logic units-based explanations can be found in

Github repo.[2] To generate the explanations described above, we first trained the LogiFormer model on the Reclor dataset. With the trained model, we generated logic units-based explanations according to Equation 3.1. In this study, we specify $k = 5$ to highlight the most important logic units for each task. Notice that, such explanations are generated for each option, and the spans are only extracted from the context paragraph and each option. For more details about the LogiFormer model, we refer readers to the original paper [147] and the corresponding implementation.[3]

3.3.3 Proposing a Tutorial Intervention to Help Users Calibrate Their Skills

To answer RQ2, we need to verify whether our proposed intervention can help mitigate the DKE among the same participants who demonstrated it in the absence of the intervention. This requires two batches of tasks that can facilitate comparative performance assessment and on which participants can be asked to self-assess their performance. Based on the effectiveness of tutorials as interventions in previous HCI literature [50, 108, 117], we designed a tutorial as a means to shed light on the fallibility of AI advice. In our paper, we, therefore, considered the tutorial as an intervention and analyzed its effectiveness by comparing participants' reliance and self-assessment before and after the tutorial was delivered. Inspired by existing work to mitigate different kinds of cognitive biases through revealing such biases to users [148, 149], we decided to adopt a tutorial to help users calibrate their skills through self-assessment on logical reasoning tasks. To this end, we designed a tutorial with the aim of revealing to users that they may not be as capable in such tasks as they may believe. Furthermore, to ensure the effectiveness of revealing their mistakes, we designed persuasive explanations for users. To achieve that goal, we chose to provide contrastive explanations which point out not only the reason for correct answers but also the reason to reject users' wrong choices. As none of the existing off-the-shelf toolkits can be used to obtain such strongly persuasive explanations, we manually created explanations for each option in the four tasks considered in the tutorial phase. These explanations corresponding to each task have also been made available on the Open Science Framework companion page. An example of such performance feedback and contrastive explanation can be found in Figure 3.2b. On this page, we showed the correct answer in a box with light blue background color. The final decision of the participant after receiving AI advice, and the AI advice itself are shown in boxes with a dark blue background color. The contrastive explanation is shown at the bottom of this page. Through such a performance feedback intervention, we hope that users with inflated self-assessments can realize their true capability with respect to the tasks and recalibrate their self-assessment. Such an intervention can potentially help users improve their reliance on AI systems [150].

3.3.4 Pilot Study for Task Selection

To answer our research questions, we need to analyze the impact of the Dunning-Kruger effect on reliance measures and the effectiveness of the proposed intervention to mitigate such an effect. Note that the Dunning-Kruger effect corresponds to one's skills in a given task [116]. To operationalize this, we need two batches of tasks with similar difficulty

[2] https://github.com/RichardHGL/CHI2023_DKE
[3] https://github.com/xufangzhi/Logiformer

levels, through which we can verify the effectiveness of the intervention by comparing performance before and after the intervention. Meanwhile, for the tutorial tasks, we need tasks that may trigger the Dunning-Kruger effect. In other words, tasks that participants may make mistakes on with high confidence. For these purposes, we conducted a pilot study with 10 participants from the Prolific crowdsourcing platform.[4] In the pilot study, each participant worked on 30 questions randomly sampled from the validation set of the Reclor dataset. We collected their choice and confidence level for each task. With six participants who passed all the attention checks, we assessed the difficulty of each task based on the number of participants who answered the task correctly. Considering that most participants spend around 1 minute to fully understand a task and make a decision, we considered the batch size to be six. We collected two batches of tasks which are of similar difficulty (informed by the average accuracy on the tasks in the pilot study). To make the tutorial effective but not cumbersome, we selected four tasks for the tutorial. The tasks for the tutorial were selected in a similar fashion as the other batches, as the tutorial only has four questions instead of six, the tasks with the lowest and highest accuracy were removed. Such selection strategy creates a batch similar in difficulty to the other batches. Among the four tasks, we configured the AI advice to be correct on two of them and misleading on the other two. All participants were rewarded with hourly wage of £7.5 (estimated completion time was 33 minutes), and extra bonus of £0.05 for each correct decision.

3.3.5 Hypotheses

Our experiment was designed to answer questions surrounding the impact of Dunning-Kruger effect on user reliance on AI systems, and how to mitigate such potentially undesirable impact. People who are less competent in a task struggle more with estimating their own performance in the task, compared to the more competent counterparts [60]. Impacted by DKE, users with the option to rely on AI advice may overestimate their own performance in a task and tend to rely on themselves when they are actually less capable than the AI systems. Apart from them, some users can exhibit accurate self-assessment. Such accurate self-assessments can be indicative of a good understanding of the task difficulty and personal skills, which may help these users rely on AI systems more appropriately. Meanwhile, effective explanations may amplify such an effect. Thus, we hypothesize that:

(H1) Users overestimating their own performance will demonstrate relatively less reliance on AI systems than users demonstrating accurate self-assessment.

According to previous work [151, 152], interventions that provide users with feedback on their performance may help improve their self-assessment. By providing users with an opportunity to reflect on their skills and recalibrate their skills on the given task, we argue that the impact of the DKE can be mitigated. As a result of an improved calibration of oneself, such users are better suited to rely on AI systems appropriately when making decisions. Therefore, we hypothesize that:

[4] https://www.prolific.co/

(H2) Making users aware of their miscalibrated self-assessment, will help them improve their self-assessment.
(H3) Making users aware of their miscalibrated self-assessment will result in relatively more appropriate reliance on AI systems.

Performance feedback can potentially help participants improve their self-assessment, which may facilitate appropriate reliance. At the same time, explanations have been shown to improve the human understanding and interpretation of AI advice [25, 70, 107], which can also potentially contribute to appropriate reliance. Thus, we hypothesize to observe the following in a human-AI decision making context:

(H4) Providing performance feedback and meaningful explanations can facilitate appropriate reliance on the AI system.

3.4 Study Design

This section describes our experimental conditions, variables, statistical analysis, procedure, and participants in our main study.

3.4.1 Experimental Conditions

In our study, all participants worked on logical reasoning tasks with two-stage decision making process (described in Sec. 3.3.1). The only difference is whether tutorial is presented and whether explanations are provided along with AI advice. To comprehensively study the effect of each factor and their interaction effect, we considered a 2×2 factorial design with four experimental conditions: (1) no tutorial, no XAI (represented as × `Tutorial`,× `XAI`), (2) with tutorial, no XAI (represented as ✓ `Tutorial`,× `XAI`), (3) no tutorial, with XAI (represented as × `Tutorial`,✓ `XAI`), (4) with tutorial, with XAI (represented as ✓ `Tutorial`,✓ `XAI`). In conditions with tutorial, participants were presented with four selected tasks with performance feedback and contrastive explanation for correct answers against wrong choice (when participants missed the wrong answer). While in conditions without tutorial, the four tasks selected are presented as normal tasks without any performance feedback or explanation for correct answers, to prevent any learning effect. In conditions with XAI, the top-5 most important logic units are highlighted as an explanation for AI advice.

For each batch of six tasks, the AI system was configured to provide correct advice on four of them and misleading advice on two tasks. So the accuracy of AI systems is around 66.7%. To avoid any ordering effect, we randomly assign one batch of tasks as first batch of tasks for each participant and further shuffled the order of tasks within each batch.

3.4.2 Measures and Variables

We measure the reliance of participants on the AI system via two metrics: the **Agreement Fraction** and the **Switch Fraction**. These look at the degree to which participants are in

Table 3.1: The different appropriate reliance patterns considered in [29]. d_i is initial human decision, while d_f is the final decision after AI advice.

d_i	**AI advice**	d_f	**Reliance**
Incorrect	Correct	Correct	Positive AI reliance
Incorrect	Correct	Incorrect	Negative self-reliance
Correct	Incorrect	Correct	Positive self-reliance
Correct	Incorrect	Incorrect	Negative AI reliance

agreement with AI advice, and how often they adopt AI advice in cases of initial disagreement. They are commonly used in the literature, for example in [57, 66]. In addition, we consider the accuracy in batches to measure participants' performance with AI assistance. Since cases without initial disagreement do not clearly signal reliance on the system we restrict the scope of the appropriate reliance measure to accurately understand how participants handle divergent system advice. Schemer *et al.* [29] presented four conditions of appropriate reliance patterns (see Table 3.1) when the disagreement exists and correct answer exists in human initial decision or AI advice. We followed them to adopt *Relative positive AI reliance* (**RAIR**) and *Relative positive self-reliance* (**RSR**) as appropriate reliance measures. The two measures assessed users' appropriate reliance from two dimensions, which can help analyze the dynamics of reliance. To provide an overview of participants' appropriate reliance under initial disagreement, we considered **Accuracy-wid** (*i.e.,* accuracy with initial disagreement). These measures are computed as follows:

Table 3.2: The different variables considered in our experimental study. "DV" refers to the dependent variable. **RAIR**, **RSR**, and **Accuracy-wid** are indicators of appropriate reliance.

Variable Type	**Variable Name**	**Value Type**	**Value Scale**
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
	Accuracy-wid	Continuous	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
Assessment (DV)	Degree of Miscalibration	Continuous, Interval	[0,6]
	Self-assessment	Continuous, Interval	[-6,6]
Trust (DV)	TiA-Trust	Likert	5-point, 1:strong distrust, 5: strong trust
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-PtT	Likert	5-point, 1: tend to distrust, 5: tend to trust
Other	Helpfulness of Explanation	Likert	5-point, 1: not helpful, 5: very helpful

$$\textbf{Agreement Fraction} = \frac{\text{Number of decisions same as the system}}{\text{Total number of decisions}},$$

$$\textbf{Switch Fraction} = \frac{\text{Number of decisions user switched to agree with the system}}{\text{Total number of decisions with initial disagreement}},$$

$$\textbf{Accuracy} = \frac{\text{Number of correct final decisions}}{\text{Total number of decisions}},$$

$$\textbf{Accuracy-wid} = \frac{\text{Number of correct final decisions with initial disagreement}}{\text{Total number of decisions with initial disagreement}},$$

$$\textbf{RAIR} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance + Negative self-reliance}},$$
$$\textbf{RSR} = \frac{\text{Positive self-reliance}}{\text{Positive self-reliance + Negative AI reliance}}.$$

To measure the self-assessment of users, we gathered responses on the following question after each batch of tasks – "From the previous 6 questions, how many questions do you estimate to have been answered correctly? (after receiving AI advice)". Comparing that estimation with the actual correct number, we can calculate the degree of miscalibration and self-assessment as: **Degree of Miscalibration** = |Estimated correct number - Actual correct number|, **Self-assessment** = Estimated correct number - Actual correct number. Meanwhile, for conditions with explanations, we also assessed the helpfulness of explanations with the question, "To what extent was the explanation (*i.e.,* the highlighted words/phrases) helpful in making your final decision?" Responses were gathered on a 5-point Likert scale from *1* to *5* corresponding to the labels *not helpful*, *very slightly helpful*, *slightly helpful*, *helpful*, *very helpful*.

For a deeper analysis of our results, a number of additional measures were considered based on observations from existing literature [87, 153, 154]:

- Trust in Automation (TiA) questionnaire [90], a validated instrument to measure (subjective) trust [87]. In this study we adopted two subscales: *Propensity to Trust* (TiA-PtT), *Trust in Automation* (TiA-Trust). Thus, we consider possible effects of trust on reliance, in accordance with Lee *et al.* [40].
- Affinity for Technology Interaction Scale (ATI) [91], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [87].

Table 3.2 presents an overview of all the variables considered in our study.

3.4.3 Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for the 2 × 2 factorial design using G*Power [92]. To correct for error-inflation as a result of testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and the consideration of 4 different experimental conditions. This resulted in a required sample size of 244 participants. We thereby recruited 314 participants from the crowdsourcing platform Prolific[5], in order to accommodate potential exclusion.

Compensation. All participants were rewarded with £2.5, amounting to an hourly wage of £7.5 (estimated completion time was 20 minutes). We rewarded participants with extra bonuses of £0.1 for every correct decision in the 16 trial cases. By incentivizing participants to reach a correct decision, we operationalize the concomitant "vulnerability" discussed by Lee and See [40] as a contextual requirement to encourage appropriate system reliance.

[5] https://www.prolific.co

Filter Criteria. All participants were proficient English speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one attention check (65 participants). The resulting sample of 249 participants had an average age of 38 (SD = 12.8) and a gender distribution (48.6% female, 51.4% male).

3.4.4 Procedure

The full procedure that participants followed in our study is illustrated in Figure 3.3. All participants first read the same basic instructions on the logical reasoning task. Next, participants were asked to complete a pre-task questionnaire to measure their propensity to trust and affinity for technology interaction.

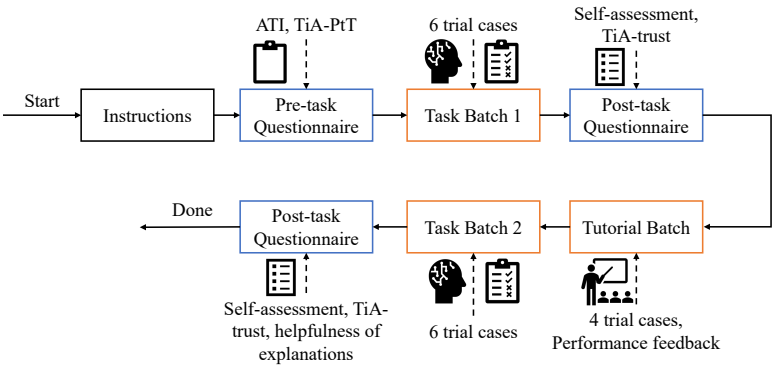


Figure 3.3: Illustration of the procedure participants followed within our study. This flow chart describes the experimental condition ✓ `Tutorial`,✓ `XAI`. Blue boxes represent the questionnaire phase, orange boxes represent the task phase.

Participants were then assigned to one experimental condition, which differed in whether or not tutorial feedback is provided and the system's prediction is supplemented with explanation. In × `Tutorial`,× `XAI` and × `Tutorial`,✓ `XAI` conditions, participants worked on the four trial cases without any difference with the task batch, no extra information was provided. After that, participants will work on 16 tasks (two task phases with six tasks, and one tutorial phase with four tasks). Selection of these cases is described in section 3.3.4. After each task phase, post-task questionnaires were adopted to assess their self-assessment and trust in AI systems (TiA-trust). Participants in the × `Tutorial`,✓ `XAI` and ✓ `Tutorial`,✓ `XAI` conditions were additionally asked for their perceived helpfulness of the explanations they were presented with. To further ensure the reliability of responses gathered in the questionnaire and the task phases, we added four attention check questions spread out at random through the different stages of the procedure [93].

3.5 Results

In this section, we present the results of our study. We discuss descriptive statistics, the outcomes of the hypothesis tests we conducted, and our exploratory findings. Our code

and data can be found on Github.[6]

3.5.1 Descriptive Statistics

In our analysis, we only kept participants who passed all attention checks, which deemed to be more reliable. Participants were distributed in a balanced fashion over the four experimental conditions as follows: 63 (× `Tutorial`, × `XAI`), 62 (✓ `Tutorial`, × `XAI`), 62 (× `Tutorial`, ✓ `XAI`), 62 (✓ `Tutorial`, ✓ `XAI`). On average, participants spend around 32 minutes (*SD* = 11 minutes) in our study. We found no significant difference in the time spent across the four experimental conditions.

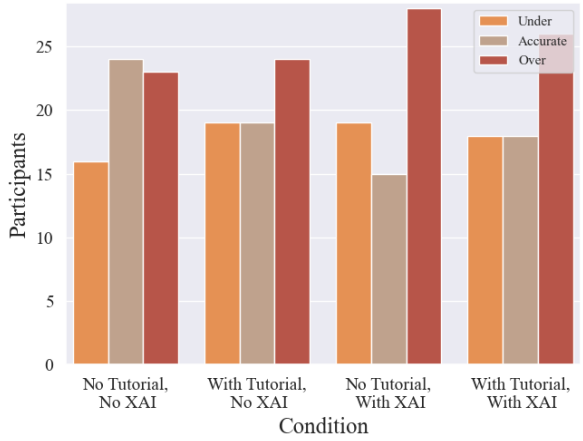


Figure 3.4: Distribution of participants with underestimated, accurate, and overestimated self-assessment across all experimental conditions in the first batch of tasks.

Distribution of Covariates. The covariates' distribution is as follows: *ATI* (*M* = 3.73, *SD* = 0.99, 6-point Likert scale, and *1: low, 6: high*), *TiA-Propensity to Trust* (*M* = 2.95, *SD* = 0.60, 5-point Likert scale, *1: tend to distrust, 5: tend to trust*).

Distribution of Participants. Among 249 participants, we identified the participants who underestimated their performance (*i.e.,* Self-assessment < 0), those with an accurate self-assessment (*i.e.,* Self-assessment = 0), and those with overestimation of their performance (*i.e.,* Self-assessment > 0) according to their performance in the first batch of tasks (shown in Figure 3.4). In general, participants showed relatively balanced distribution into the three types of self-assessment across conditions: (1) the number of participants with underestimated self-assessment lies in the range of 15 ∼ 20, (2) the number of participants with accurate self-assessment lies in the range of 15 ∼ 25, (3) the number of participants with overestimated self-assessment was in the range of 20 ∼ 30. We also compared the time spent by participants with different self-assessment and participants with different experimental conditions, and found no statistically significant difference with Kruskal-Wallis H-tests.

[6] `https://github.com/RichardHGL/CHI2023_DKE`

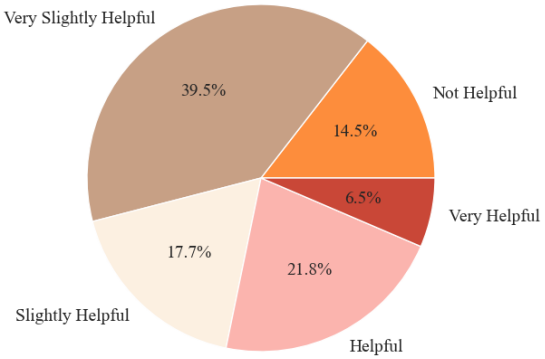


Figure 3.5: Distribution of participants with perceived helpfulness of logic units-based explanations.

For participants in conditions with explanation (*i.e.,* [× `Tutorial`,✓ `XAI`] and [✓ `Tutorial`,✓ `XAI`]), we assessed the helpfulness of logic units-based explanations. The ratios of perceived helpfulness are illustrated with Figure 3.5. As we can see, most people (57.2%) think it slightly or very slightly helpful, while only 28.3% participants show positive feedback to the logic units-based explanations.

Performance Overview. On average across all conditions, participants achieved an accuracy of 56.9% ($SD = 0.16$) over the two batches of tasks, still lower than the aforementioned AI accuracy of 66.7%. The agreement fraction is 0.665 ($SD = 0.17$) while the switching fraction is 0.453 ($SD = 0.27$). With these measures, we confirm that when disagreement appears participants in our study did not always switch to AI advice or blindly rely on the AI system. As all dependent variables are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

3.5.2 Hypothesis Tests

H1: effect of inflated self-assessments on AI system reliance

To analyze the main effect of participants' inflated self-assessment (*i.e.,* overestimation of performance) on their reliance on the AI system, we conducted Kruskal-Wallis H-tests by considering how participants varied in their self-assessment. We categorize all participants into three groups according to the self-assessment: (1) participants who underestimated their performance (*i.e.,* Self-assessment < 0), (2) participants with accurate performance self-assessment (*i.e.,* Self-assessment = 0), and (3) participants who overestimated their performance (*i.e.,* Self-assessment > 0). For this analysis, we considered all participants across the four experimental conditions, and the performance metrics are calculated based on the first batch of tasks. The results are shown in Table 3.3.

Effect of Overestimated Self-Assessments on Objective Reliance. For all reliance-based measures, we found a statistically significant difference between the performance

Table 3.3: Kruskal-Wallis H-test results for inflated self-assessments (**H1**) on reliance-based dependent variables. "††" indicates the effect of variable is significant at the level of 0.0125. "Under", "Accurate", abd "Over" refers to participants who underestimated , accurately estimated, and overestimated their performance on the first batch of tasks, respectively.

Dependent Variables	H	p	$M \pm SD$(Under)	$M \pm SD$(Accurate)	$M \pm SD$(Over)	Post-hoc results
Accuracy	74.06	**<.001**††	0.72 ± 0.16	0.61 ± 0.15	0.45 ± 0.19	Under > Accurate > Over
Agreement Fraction	10.87	**.004**††	0.70 ± 0.18	0.69 ± 0.21	0.59 ± 0.24	Under, Accurate > Over
Switch Fraction	23.31	**<.001**††	0.50 ± 0.28	0.53 ± 0.31	0.32 ± 0.32	Under, Accurate > Over
Accuracy-wid	87.94	**<.001**††	0.65 ± 0.21	0.53 ± 0.27	0.28 ± 0.22	Under > Accurate > Over
RAIR	46.91	**<.001**††	0.65 ± 0.36	0.58 ± 0.37	0.27 ± 0.33	Under, Accurate > Over
RSR	30.23	**<.001**††	0.67 ± 0.44	0.41 ± 0.47	0.27 ± 0.43	Under > Accurate, Over

of the participants who overestimated their performance and those with accurate self-assessment. Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) were used to make pairwise comparisons of performance, revealing that participants who did not overestimate their performance in fact performed significantly better than those who did (The only exception is on metric **RSR**). Overall, participants with accurate self-assessment and underestimation of their own performance performed much better than participants who overestimated their own performance. The main reason is that they showed more reliance on the AI system and achieved better appropriate reliance when their initial decision disagreed with AI advice. The results indicate that participants who overestimate their own performance rely significantly less on AI systems compared to those who do not, which indicates more severe under-reliance. As a result, they achieved a significantly lower accuracy on average. Thus, we find support for hypothesis **H1**.

We also found that participants who underestimated their performance achieved significantly higher **Accuracy**, **Accuracy-wid**, and **RSR** than participants demonstrating accurate self-assessment. Since they showed similar degrees of reliance (**Agreement Fraction** and **Switch Fraction**) on the AI system, the improvement of overall accuracy is mainly due to appropriate reliance. In general, they showed significantly better **RSR**, which indicates that they have a better chance to rely on themselves to make correct decisions when they initially disagree with misleading AI advice.

In the first batch of tasks, we found no difference (with Kruskal-Wallis H-tests) in reliance and accuracy metrics when comparing participants in XAI conditions (*i.e.,* `× Tutorial,✓ XAI` and `✓ Tutorial,✓ XAI`) with participants in non-XAI (*i.e.,* `× Tutorial,× XAI` and `✓ Tutorial,× XAI`). To verify how the provided logic units-based explanations affect participants with different self-assessments, we compared the performance and reliance measures of participants with XAI and without XAI in underestimation, accurate self-assessment, and overestimation. No significant effects were found from the logic units-based explanation on performance and reliance for participants with overestimated self-assessment.

H2: effect of the tutorial on self-assessment

To verify **H2**, we used Wilcoxon signed rank tests to compare the performance of participants before and after the tutorial. We considered participants who are provided with the tutorial for self-assessment calibration (*i.e.,* `✓ Tutorial,× XAI` and `✓ Tutorial,✓ XAI`). Meanwhile, we exclude participants who have accurate assessment on the first batch of

tasks from this analysis. Finally, we have 87 participants reserved for analysis of **H2**. On average, the participants' self-assessment get improved after receiving the tutorial (*i.e.,* decreased **Degree of Miscalibration**, $M \pm SD$(first) = 1.67 ± 0.91, $M \pm SD$(second) = 1.14 ± 1.04; a smaller value indicates more accurate self-assessment). A Wilcoxon signed rank test indicated that the difference was statistically significant, T=1175.0, p<0.001, which supports **H2**. To further check how the tutorial intervention has an impact on participants with different types of miscalibration, we separately conducted Wilcoxon signed rank tests on participants underestimating their own performance and overestimating their own performance separately. The results indicate that: (1) participants underestimating their own performance calibrated their self-assessment, the difference is significant (T=229.0, p=0.002); (2) participants overestimating their own performance calibrated their self-assessment, the difference is significant (T = 381.5, p = 0.012). The detailed analysis of participants with different types of miscalibration also supports **H2**.

To further explore the effect of logic units-based explanation on calibrating self-assessment, we conducted a Kruskal-Wallis H-test (among these participants) by considering whether the explanation is provided. We found no significant results, which indicates that logic units-based explanations cannot amplify the effect of the tutorial intervention (*i.e.,* calibrating self-assessment).

Table 3.4: Wilcoxon signed ranks test results for **H3** on reliance-based dependent variables. For participants with initial underestimation, we report results with one-sided hypothesis that the performance / reliance decrease after tutorial. For participants with initial overestimation, we report results with one-sided hypothesis that the performance / reliance increase after tutorial. "†" and "††" indicates the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Participants	**Underestimation**				**Overestimation**			
Dependent Variables	p	$M \pm SD$(first)	$M \pm SD$(second)	Trend	p	$M \pm SD$(first)	$M \pm SD$(second)	Trend
Accuracy	**.000**††	0.73 ± 0.17	0.55 ± 0.21	↓	.075	0.46 ± 0.18	0.51 ± 0.22	-
Agreement Fraction	.543	0.68 ± 0.20	0.70 ± 0.23	-	.605	0.60 ± 0.23	0.57 ± 0.23	-
Switch Fraction	.592	0.47 ± 0.29	0.48 ± 0.36	-	.147	0.31 ± 0.33	0.36 ± 0.31	-
Accuracy-wid	**.000**††	0.68 ± 0.22	0.44 ± 0.29	↓	**.013**†	0.27 ± 0.20	0.41 ± 0.28	↑
RAIR	**.006**††	0.68 ± 0.37	0.45 ± 0.38	↓	**.038**†	0.24 ± 0.32	0.36 ± 0.36	↑
RSR	**.000**††	0.72 ± 0.43	0.30 ± 0.44	↓	**.020**†	0.29 ± 0.45	0.52 ± 0.48	↑

H3: effect of the tutorial on appropriate reliance

Similar to the analysis for **H2**, we only considered the participants who showed miscalibration in the first batch of tasks. Overall, there is no significant difference in reliance and performance measures when we compare the participants' performance before and after receiving the tutorial. To further check how our tutorial intervention will affect participants with different miscalibration of self-assessment, we conducted analysis for participants with underestimation and overestimation separately. The results of Wilcoxon signed rank tests corresponding to each of the reliance measures are shown in Table 3.4. Both participants with underestimation and overestimation did not show any significant difference in reliance measures (*i.e.,* **Agreement Fraction** and **Switch Fraction**). For participants who underestimated their performance in the first batch of tasks, they showed significantly worse performance and appropriate reliance after receiving the tutorial. In contrast, we found some improvement of **Accuracy** and appropriate reliance measures (*i.e.,* **Accuracy-wid**, **RAIR**, **RSR**) for participants who overestimated their performance

in the first batch of tasks. However, the improvement is non-significant at the level of 0.0125. Thus, on the whole, we find partial support for **H3**.

Meanwhile, to check how the tutorial intervention affects the participants with initial accurate self-assessment, we also conducted Wilcoxon signed rank tests for their performance before and after the tutorial intervention. No significant difference is found. Combined with the findings from participants with initial miscalibration, we found that: (1) the designed tutorial intervention does not show much impact on participants with accurate self-assessment, (2) the designed tutorial intervention has positive impact on appropriate reliance for participants who initially overestimate themselves, while negative impact on participants with initial underestimation of their performance.

Relation Between Self-assessment Calibration and the Change in Reliance. To further explore the relationship between the change in self-assessment and change with (appropriate) reliance, we conducted the Spearman rank-order test separately for participants with overestimation and underestimation in the first batch of tasks. As the impact of tutorial intervention on **Agreement Fraction** and **Switch Fraction** is insignificant, we ignore the two metrics in calculating the correlation. The results are shown in Table 3.5. We found a strong negative monotonic relationship between the two variables in participants with overestimation. Thus, in logical reasoning tasks, the calibration effect in self-assessment accounted for 59.3% of the improved **Accuracy** ($\rho^2 = 0.593, p < 0.001$), 55.5% of the improved **Accuracy-wid** ($\rho^2 = 0.555, p < 0.001$), 32.0% of the improved **RAIR** ($\rho^2 = 0.320, p < 0.001$), and 12.9% of the improved **RSR** ($\rho^2 = 0.129, p = 0.005$). Similarly, the calibration of self-assessment also accounted for 26.2% of the decreased **Accuracy** ($\rho^2 = 0.262, p = 0.001$), 14.8% of the decreased **Accuracy-wid** ($\rho^2 = 0.148, p = 0.009$) for participants with underestimation.

Table 3.5: Correlation of self-assessment change and reliance change. "††" indicates the effect of variable is significant at the level of 0.0125. "†" indicates the effect of variable is significant at the level of 0.05.

Participants	**Underestimation**		**Overestimation**	
Dependent Variables	ρ	p	ρ	p
Accuracy	-0.512	**.001**††	-0.770	**.000**††
Accuracy-wid	-0.385	**.009**††	-0.745	**.000**††
RAIR	-0.293	**.039**†	-0.566	**.000**††
RSR	-0.349	.068	-0.359	**.005**††

In general, for all participants with miscalibrated self-assessment, the difference in self-assessment shows strong negative correlation with the difference in performance and appropriate reliance. In other words, the increase in self-assessment (trend to overestimation) will lead to decrease in performance and appropriate reliance, which is consistent with our findings in **H1**. While the significant negative correlation exists for performance measures in all participants with miscalibrated self-assessment, only participants with overestimation showed significant correlation (in the level of 0.0125) with **RAIR** and **RSR**. The difference indicates that the change of self-assessment can hardly explain why participants with underestimation showed worse appropriate reliance.

To further explore the impact of logic units-based explanations on performance improvement (the difference between performance metrics from the second batch of tasks

Table 3.6: Kruskal-Wallis H-test results for logic units-based explanations on performance improvement of reliance-based dependent variables.

Participants	Underestimation				Overestimation			
Dependent Variables	H	p	$M \pm SD$(Exp)	$M \pm SD$(No Exp)	H	p	$M \pm SD$(Exp)	$M \pm SD$(No Exp)
Accuracy	0.00	.963	-0.19 ± 0.15	-0.18 ± 0.24	1.38	.241	0.10 ± 0.27	0.00 ± 0.30
Agreement Fraction	0.00	.963	0.01 ± 0.25	0.04 ± 0.32	0.88	.349	0.01 ± 0.38	-0.06 ± 0.28
Switch Fraction	0.04	.843	-0.03 ± 0.39	0.04 ± 0.41	0.02	.884	0.06 ± 0.47	0.05 ± 0.33
Accuracy-wid	0.00	.951	-0.25 ± 0.30	-0.22 ± 0.31	0.50	.478	0.16 ± 0.36	0.11 ± 0.39
RAIR	0.02	.878	-0.23 ± 0.48	-0.24 ± 0.57	0.00	.968	0.11 ± 0.46	0.14 ± 0.46
RSR	0.96	.327	-0.33 ± 0.50	-0.50 ± 0.51	1.84	.175	0.35 ± 0.72	0.10 ± 0.66

and those from the first batch of tasks), we conducted a Kruskal-Wallis H-test (among these participants) by considering whether explanations are provided. Overall, no significant difference is found for all behavior-based dependent variables considering all 87 participants who showed miscalibration in the first batch and then received the tutorial intervention. We further check the logic units-based explanation impact according to participants with underestimation (37 participants) and overestimation (50 participants) respectively (cf. Table 3.6). No significant difference is found for all behavior-based dependent variables. Although participants with explanations show better performance improvement in **RSR**, such difference is not significant.

H4: Two-factor analysis for final performance

To verify H4, we conducted a two-way ANOVA to compare the performance and (appropriate) reliance measures of participants under the effect of providing tutorial intervention and logic units-based explanations. In this analysis, only the second batch of tasks are taken into consideration, as the performance of the first batch of tasks is not affected by the tutorial intervention. According to the test results shown in Table 3.7, no significant impact (in the significance level of 0.0125) is found for tutorial intervention, logic units-based explanations and their interaction effect. Thus, **H4** is not supported.

Table 3.7: ANOVA test results for **H4** on behavior-based dependent variables in the second batch of tasks.

Dependent Variables	Accuracy		Agreement Fraction		Switch Fraction		Accuracy-wid		RAIR		RSR	
Variables	F	p	F	p	F	p	F	p	F	p	F	p
Tutorial	2.41	.122	3.74	.054	3.87	.050	1.63	.203	4.70	.031	0.20	.652
XAI	2.10	.148	0.30	.587	1.00	.319	3.35	.068	2.05	.153	0.23	.632
Tutorial × XAI	0.05	.824	0.00	.990	0.00	.956	0.10	.746	0.00	.923	0.05	.832

According to the results of **H3**, the tutorial intervention shows positive impact on participants with initial overestimation, no significant effect on participants with accurate self-assessment, and negative impact on participants with initial underestimation. As indicated by Figure 3.4, the participants show compatible distribution in the three groups with different initial self-assessment. The contradicting effects on the participants with miscalibrated self-assessment get canceled. That may explain why the tutorial intervention does not show significant impact across experimental conditions. On the other hand, we did not find any support for effectiveness of logic units-based explanations in reliving DKE or facilitating appropriate reliance in analysis of **H1** - **H3**.

3.5.3 Further Analysis On the DKE

According to Dunning and Kruger [60], participants demonstrating the DKE are less competent and overestimate their performance. For further analysis of DKE in our study, we follow the method in the original study as well as consequent replications [60, 96], to split the participants in all conditions into performance-based quartiles. The top-quartile corresponds to those demonstrating high performance (top 25%), the bottom quartile corresponds to those with low performance (bottom 25%), and we combine the two quartiles in the middle comprising of participants with a medium level of performance in the first batch of tasks. As our tutorial is demonstrated to be effective in calibrating self-assessment, we do not take the second batch of tasks into consideration. In total, 101 participants among 249 participants showed an overestimation of performance in the first batch of tasks. In high accuracy group (63 participants), 35 participants showed underestimation of their own performance, and 21 participants demonstrated accurate self-assessment, while only 7 participants (11.1%) show overestimation of performance in the first batch of tasks. In comparison, 46 participants (73.0%) in low accuracy group (63 participants) show an overestimation of performance in the first batch of tasks, while only 6 participants and 11 participants showed underestimation of their performance and demonstrated accurate self-assessment, respectively. This aligns with the observation of Dunning and Kruger [115, 116]: top-performance group shows the tendency to underestimate their performance, while low-performance group shows tendency to overestimate their performance. With this observation, we can take low accuracy group as a representative group of participants with DKE, and take high accuracy group as a representative group of participants without DKE. This aligns with and validates our motivation to design a tutorial intervention to mitigate DKE, and improve self-assessment and appropriate reliance on AI systems.

Table 3.8: Kruskal-Wallis H-test results for reliance-based measures on high accuracy group and low accuracy group. "††" indicates the effect of variable is significant at the level of 0.0125.

Dependent Variables	H	p	$M \pm SD$(High)	$M \pm SD$(Low)
Agreement Fraction	54.68	**<.001**††	0.75 ± 0.15	0.46 ± 0.18
Switch Fraction	13.09	**<.001**††	0.46 ± 0.32	0.27 ± 0.21
Accuracy-wid	81.00	**<.001**††	0.74 ± 0.24	0.21 ± 0.15
RAIR	25.71	**<.001**††	0.64 ± 0.45	0.21 ± 0.21
RSR	46.41	**<.001**††	0.76 ± 0.39	0.18 ± 0.37

The impact of DKE on Reliance. To further analyze how the DKE affects user reliance on AI systems, we compared the reliance-based measures of high accuracy group and low accuracy group using a Kruskal-Wallis H-test. The results are shown in Table 3.8. Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) also confirmed the significant difference. As we can see, participants in the low accuracy group (representative for participants with DKE) achieve a relatively poorer appropriate reliance than participants in the high accuracy group. Participants in the low accuracy group demonstrate significantly less reliance and appropriate reliance on AI systems, which also reflects that under-reliance is to blame for their low performance. We also compared the time spent by participants in the high accuracy group with participants in low accuracy

group through a Kruskal-Wallis H-test. The difference of time spent on tasks between the two groups is non-significant (p = 0.018, borderline significance in Kruskal-Wallis H-test). On average, the high accuracy group spent around 30 minutes (SD=12 minutes), while the low accuracy group spent around 34 minutes (SD=13 minutes). Interestingly, despite the fact that participants in the low accuracy group spent longer time on the task they still relied poorly on the AI system. This is consistent with what has been widely understood as an impact of the DKE metacognitive bias.

3.5.4 Further Analysis of Trust

In addition to the behavior-based reliance measures, we also assessed the subjective trust of participants in AI systems. In this subsection, we explore the impact of our tutorial intervention and logic units-based explanation on user trust in the AI system.

The effect of tutorial intervention on trust. To explore whether our tutorial intervention had any effect on user trust in AI system, we conducted Wilcoxon signed ranks test comparing the trust before and after the tutorial. On average, participants' trust in the AI system does not show significant difference after the tutorial intervention (increased from 2.996 to 3.016; T = 1063.5, p = 0.952). This suggests that the main impact of the tutorial was on helping users calibrate their competence (*i.e.,* their self-assessment) without directly shaping their trust in the AI system.

Table 3.9: ANCOVA test results on trust-related dependent variables. With different self-assessmnet patterns, we divide all participants into three groups. "††" indicates the effect of variable is significant at the level of 0.0125.

Variables	F	p	η^2
Group	1.15	.318	.009
ATI	1.22	.271	.004
TiA-PtT	10.21	**.002**††	.040

To further analyze how other covariates shape user trust in AI system, we decided to conduct AN(C)OVAs despite the anticipation that our data may not be normally distributed because these analyses have been shown to be robust to Likert-type ordinal data [94]. As no significant difference is found between the trust before and after the tutorial, we aggregated the trust across the two batches of tasks as users' trust in the AI system. Considering our main hypothesis, we aimed to explore whether overestimation of performance and accurate self-assessment shape user trust in the AI system. For that purpose, we consider the three groups of participants (based on self-assessment, the same criteria in **H1**) with different self-assessment patterns. The results are shown in Table 3.9. As we can see, propensity to trust was the only user factor which corresponded to a significant impact on **TiA-Trust**. In a further Spearman rank-order test, we observed that there is a significant positive correlation between **TiA-PtT** and **TiA-Trust**, $\rho(249) = 0.22$, $p < .001$; suggesting a weak linear relationship between users' propensity to trust an AI system and the subjective trust measured with respect to the AI system in our study. We also conducted the Spearman rank-order tests with **TiA-PtT** and other reliance-based variables. No significant correlation was found between **TiA-PtT** and reliance measures.

3.6 Discussion

3.6.1 Key Findings

Our analysis of the impact of miscalibrated self-assessment on reliance suggests that participants with DKE tend to overestimate their own competence and rely less on AI systems, which results in under-reliance and much worse performance. To mitigate such cognitive bias, we introduced a tutorial intervention including performance feedback on tasks, alongside manually crafted explanations to contrast the correct answer with the users' mistakes. Experimental results indicate that such an intervention is highly effective in calibrating self-assessment (significant improvement), and has some positive effect on mitigating under-reliance and promoting appropriate reliance (non-significant results). We also note that after making participants who overestimated their performance aware of their miscalibrated self-assessment, participants tend to rely more (appropriately) on the AI system (*i.e.,* increased **Switch Fraction** and appropriate reliance measures, non-significant results, from Table 3.4) and achieve a higher performance improvement when logic units-based explanations are provided (insignificant results from Table 3.6). However, we did not find any significant evidence to support that the logic units-based explanations can amplify the effect of the tutorial intervention in calibrating self-assessment, or relieving the impact of DKE.

The tutorial and calibrated self-assessment demonstrate a positive impact in facilitating appropriate reliance for participants who overestimated themselves, but an opposite trend was observed on participants who underestimated themselves. We found such difference can be explained partially by the change of self-assessment. The calibration of overestimation can bring positive impact, while the calibration of underestimation may also turn into overestimation or algorithm aversion, which may explain the decrease in performance and appropriate reliance. The tutorial was initially designed to reveal the shortcomings of participants with DKE. While for participants without DKE, there is a risk that some participants did not get exposed to their shortcomings in this tutorial and only found the AI system also made mistakes, which in turn even caused overestimation of themselves. An alternative explanation is that the performance feedback in tutorial intervention showed one mistake from the AI system, which led to algorithm aversion. As pointed out by [33]: "people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake." These findings advance our current understanding of human-AI decision making, and provide useful insights that can drive guidelines for designing interventions to promote appropriate reliance.

Positioning in Existing Literature. In our study, we found that DKE can have a negative impact on user reliance on the AI system and our proposed tutorial intervention can mitigate such an impact. In the context of human-AI decision making, DKE is closely relevant to a popular stream of research around user confidence[65, 114]. For the participants who overestimated their performance, the designed tutorial intervention calibrated their self-confidence (as reflected in their self-assessment) and facilitated appropriate reliance. In contrast, the negative impact on participants who underestimated their performance can be explained by: (1) the calibrated self-assessment which can also bring overconfidence, or (2) their confidence/trust in the AI system being eroded by the observed mistake(s) of the AI system [70, 87]. The latter is consistent with findings in the literature

on algorithm aversion [33]. More empirical studies are required to confirm and explain these observations, breeding promising grounds for future research.

The participants with DKE show under-reliance on AI systems, which also aligns with the finding from Schaffer *et al.* [97]. Authors found that participants who reported higher familiarity with the task domain relied less on the intelligent assistant. The effectiveness of our tutorial intervention to calibrate self-assessment and mitigate under-reliance is also consistent with existing work using user tutorial / education interventions to mitigate unexpected and undesirable reliance patterns. All these tutorial interventions share a common objective of changing the mindset of users. For example, Chiang *et al.* [108] reported that user tutorials such as machine learning literacy interventions can effectively help high-performance individuals to reduce over-reliance without affecting the reliance of low-performance individuals. Similarly, Chiang *et al.* [50] showed that a brief education session about the possible performance disparity of an ML model (on data with different distribution) can effectively reduce over-reliance on such cases. While their work focused more on changing human understanding of AI systems (performance, uncertainty, etc.), our work aims to help users calibrate their competence (*i.e.,* their self-assessment) on specific tasks. As a result, their main objective was to realize when AI systems are not reliable to reduce over-reliance, while we attempt to mitigate under-reliance for participants who overestimate themselves.

Logic Units-based Explanations Do Not Have the Expected Effect. In our study, the logic units-based explanations did not aid in further amplifying the calibration effect of the tutorial intervention. This is in line with the findings of Wang *et al.* [107] and Schaffer *et al.* [97]. With a comparative study about four types of different explanations, authors found that "on decision making tasks that people are more knowledgeable, explanation that is considered to resemble how humans explain decisions (*i.e.,* counterfactual explanation) does not seem to improve calibrated trust." One potential explanation is that such explanations do not fulfill the three desiderata of AI explanations [107] (refer to section 3.2.1): the logic units-based explanations may help participants understand the AI, but fail to help them recognize the uncertainty underlying the AI or calibrate their trust in the AI in AI-assisted decision making. Another potential cause is such explanations may introduce automation bias [97], which will cause over-reliance. Our results suggest that logic units-based explanations may still be hard to follow, because participants still need to connect and interpret the logic units by themselves. A limitation of our current work is that we did not gather explicit input from participants on their perceived understanding of the explanations. One further step to ground such logic units into readable logical claims may work better for users. However, we do not deny the prospect that some XAI methods may have the potential to help mitigate DKE and calibrate user confidence in human-AI decision making. For example, contrastive explanations may work in the context of human-AI decision making [155, 156].

3.6.2 Implications

As our findings suggest that participants with DKE tend to rely less on AI systems, it implies that future work should look more closely at the effects of self-assessment in human-AI collaboration. Although our tutorial intervention shows significant improvement in calibrating self-assessment, the improvement in appropriate reliance is still limited (with

borderline significance). Meanwhile, such calibration of self-assessment may even hurt the team performance for participants with initial underestimation of their performance. For these participants, the tutorial calibrated their underestimation, which may also lead to illusion of superior performance (overestimation of themselves). In order to further promote appropriate reliance in human-AI collaboration, we need to develop more effective human-centered tutorials. Meanwhile, participants who show lower performance in our scenario have significantly higher probability to overestimate their performance, which aligns with DKE properties. Thus, we can leverage overestimation of individual performance as an indicator of such a meta-cognitive bias and further mitigate it with personalized or appropriate interventions.

Guidelines for Tutorial Designs to Promote Appropriate Reliance. While our tutorial intervention proved to be effective in helping users calibrate their self-assessment, accurate self-assessment does not necessarily translate to optimal appropriate reliance. Compared with participants with accurate self-assessment, the participants with underestimation showed a significantly better performance in **RSR** (see Table 3.3), and calibrating such underestimation may even lead to decreased appropriate reliance (see Table 3.4), which indicates accurate self-assessment does not necessarily lead to optimal appropriate reliance. One possible cause is that while the tutorial makes such users aware that they underestimated themselves and they can make correct decisions when the AI system is wrong in the task, users may have an illusion of superior capability than the AI system. As a result, on tasks where AI systems are more capable, users make mistakes by exhibiting under-reliance on the AI system due to recalibrated overestimation of their own competence. Our findings suggest that we should pay attention to avoiding such side effects of making users overestimate themselves in comparison to the AI system. To avoid such side effects, tutorials designed to mitigate a specific kind of bias should be carefully checked before subjecting them to broad participant pools. This also implies that tutorials designed for promoting appropriate reliance should not only reveal the shortcomings of users or AI systems (*i.e.,* when they are less capable of making the right decision), but also their strengths (*i.e.,* when they are capable or more capable). This has useful implications for the future design of interventions to mitigate cognitive biases in human-AI collaboration.

In previous work on mitigating over-reliance with a tutorial intervention, researchers focused on revealing the AI systems' brittleness [50, 108]. Combined with their findings, we argue that a more effective tutorial to promote appropriate reliance can be one that helps users understand both themselves and AI systems, and not only revealing the weakness but also showing the strengths of each. With such a comprehensive understanding, human decision makers can potentially have a better chance to understand when they should rely on AI systems, and when they should rely on themselves, ultimately leading to (more) appropriate reliance. More work is required to understand whether and how explanations can mediate this process of creating a better understanding among users of AI system capabilities in comparison to their own. This resonates with recent work exploring human-AI complementarity [22, 25, 70].

3.6.3 Caveats and Limitations

Potential Biases. Our research questions focused on DKE and reliance and how to mitigate such impact. As we cannot pre-identify which participants have DKE, we recruit the participants and determine it with performance assessment. However, such assessment may be affected by other factors, which can lead to biased results. For example, although we relied on a pilot study to inform our task selection while creating two batches of tasks with comparable difficulty levels, we cannot be certain that they would be perceived the same way on average across the participants.

As pointed out by Draws *et al.* [157], cognitive biases introduced by task design and workflow may have a negative impact on crowdsourcing experiments. With the help of Cognitive Biases Checklist introduced [157], we analyzed potential bias in our study. *Self-interest bias* is possible, because crowd workers we recruited from the Prolific platform are motivated by monetary compensation. To alleviate any participants with low effort results, we put attention checks to remove ineligible participants from our study. As the question and context in Reclor dataset may be something participants familiar with, *familiarity bias* and *availability bias* can also affect our results.

Transferability Concern. In our study, all analyses are based on the logical reasoning task, which most laypeople are capable of dealing with. However, in practice, the application scenarios may be affected by more factors (like user expertise, familiarity, and input modality). This gap can be a potential threat to the transferability of our findings and implications. However, Dunning and Kruger [60] showed that participants suffer from DKE across multiple scenarios: "participants scoring in the bottom quartile on tests of humor, grammar, and logic grossly overestimated their test performance and ability." These effects were replicated in a number of other tasks, like human-AI collaboration [97] and crowdsourcing [87, 158]. Our findings are therefore highly relevant and can play an important role in informing the design for appropriate reliance in the context of human-AI interaction, collaboration, and teaming.

3.7 Conclusions and Future Work

In this chapter, we present a quantitative study to understand the impact of the Dunning-Kruger effect (DKE) on reliance behavior of participants in a human-AI decision making context. We propose a tutorial intervention and explore its effectiveness in mitigating such an effect. Our results suggest that participants who overestimate their own performance tend to rely less on the AI system. Combined with the findings that participants with DKE show a much higher probability of overestimating their performance, we conclude that participants with DKE rely less on AI systems, and such under-reliance hinders them in achieving better performance on average (RQ1). Through a rigorous experimental setup and statistical analysis, we found the effectiveness of our tutorial intervention in mitigating DKE (RQ2). However, we found that the tutorial may mislead some participants (*i.e.,* participants who underestimated themselves) to overestimate their performance or exhibit algorithm aversion, which in turn harms their appropriate reliance on the AI system. Our findings suggest that, to fully mitigate the negative impact of the Dunning-Kruger effect and achieve appropriate reliance, more comprehensive, insightful, and personalized user tutorials are required. We reflected on guidelines for better tutorial designs based on

our key findings.

We found that our tutorial intervention failed to make a difference in participants' subjective trust in the AI systems. Instead, we found that users' general propensity to trust has a significant impact on shaping their subjective trust in the AI system. Future work can further look into how user trust can be reshaped with different interventions or by using more effective explanations (*e.g.*, contrastive explanations or logical explanations in natural language). We hope the key findings and implications reported in this chapter will inspire further research on promoting appropriate reliance.

4

Developing Critical Mindset with Debugging AI systems

Powerful predictive AI systems have demonstrated great potential in augmenting human decision making. Recent empirical work has argued that the vision for optimal human-AI collaboration requires 'appropriate reliance' of humans on AI systems. However, accurately estimating the trustworthiness of AI advice at the instance level is quite challenging, especially in the absence of performance feedback pertaining to the AI system. In practice, the performance disparity of machine learning models on out-of-distribution data makes the dataset-specific performance feedback unreliable in human-AI collaboration. Inspired by existing literature on critical thinking and a critical mindset, we propose the use of debugging an AI system as an intervention to foster appropriate reliance. In this chapter, we explore whether a critical evaluation of AI performance within a debugging setting can better calibrate users' assessment of an AI system and lead to more appropriate reliance. Through a quantitative empirical study (N = 234), we found that our proposed debugging intervention does not work as expected in facilitating appropriate reliance. Instead, we observe a decrease in reliance on the AI system after the intervention — potentially resulting from an early exposure to the AI system's weakness. We explore the dynamics of user confidence and user estimation of AI trustworthiness across groups with different performance levels to help explain how inappropriate reliance patterns occur. Our findings have important implications for designing effective interventions to facilitate appropriate reliance and better human-AI collaboration.

4.1 Introduction

With the rise of deep learning systems over the last decade, there has been a widespread adoption of AI systems in supporting human decision makers [22], albeit without always fully understanding the societal impact or downstream consequences of relying on such systems [52, 113]. Due to the opaqueness of some AI systems, users have struggled to

This chapter is based on a peer-reviewed paper: **Gaole He**, Abri Bharos, and Ujwal Gadiraju. *To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems.* In Proceedings of the 35th ACM Conference on Hypertext and Social Media, pp. 98-105. 2024. https://doi.org/10.1145/3648188.3675130

determine when exactly they are trustworthy and have failed to achieve a complementary team performance. As a result, several previous studies that have explored human-AI collaboration and teaming across different contexts have reported improvements over human performance stemming from AI assistance, although this often falls short of AI performance [25, 70]. To realize the full potential of complementary team performance, human decision makers need to identify when they should rely on AI systems (*i.e.,* identifying instances where AI systems are capable or more capable than humans) and when they are better off relying on themselves (*i.e.,* identifying instances where AI systems are less capable than humans). Such a reliance pattern has been defined as *appropriate reliance* [22, 49].

In practice, it is common that users need to deal with data from unknown distributions and unseen contexts, meaning that AI systems in the real-world need to provide users with advice on out-of-distribution data [25, 50]. Under such circumstances, the estimated performance of an AI system or the so-called 'stated accuracy' of the system (*i.e.,* accuracy on pre-defined test sets) cannot faithfully reflect the trustworthiness of the AI system. Only a few works [49] have explored how humans rely on AI systems when performance feedback is limited or scarce. Previous work has found that user agreement with AI advice in tasks where they have high confidence significantly affects their reliance on the system, in the absence of the stated accuracy or performance of the system [49]. To help users assess the trustworthiness of AI systems, a practical solution that has been proposed, is to provide meaningful explanations along with AI advice [159, 160]. Post-hoc explanations have been found to improve user understanding of AI advice in empirical studies exploring human-AI decision making [22, 107]. However, most existing XAI methods have remained ineffective in helping users assess the trustworthiness of AI advice at the instance level, adversely affecting the degree of appropriate reliance of users on AI systems [107, 161].

To realize the goal of appropriate reliance, human decision makers need to be capable of evaluating AI advice and the trustworthiness of the AI system critically. We argue that such a critical mindset can help users avoid blindly following AI advice (*i.e.,* avoiding *over-reliance*), and also prevent them from distrusting AI advice when it can be productive (*i.e.,* avoiding *under-reliance*). Inspired by recent works on explanation-based human debugging of AI systems [162, 163], we propose explanation-based debugging as a training intervention to increase appropriate reliance on AI systems. We posit that such a debugging intervention has the potential to help users understand the limitations of AI systems — ***that neither explanations of the AI advice nor the advice itself are always reliable***. Recognizing these limitations can help users better understand when an AI system is trustworthy and thereby increase appropriate reliance on the system. In this chapter, we aim to empirically evaluate the effectiveness of using a debugging intervention as a means to increase appropriate reliance and address the following research questions.

RQ1: How can a debugging intervention help users to estimate the performance of an AI system, both at the instance and at the global level?
RQ2: How does a debugging intervention affect the reliance of users on an AI system?

To answer the above questions, we propose three hypotheses considering the effect of the debugging intervention on AI performance assessment as well as reliance, and the task ordering effect of debugging intervention on appropriate reliance. We tested these

hypotheses in an empirical study (N = 234) of human-AI collaborative decision making in a deceptive review detection task (*i.e.,* identifying whether one piece of review is written based on real experience). Interestingly, however, we found that the proposed debugging intervention fails to calibrate user estimation of AI performance and further promote appropriate reliance.

Our results highlight that when presented with the weakness of the AI system in an early stage of the debugging intervention, users underestimate AI performance and rely less on the AI system. Users' overestimation of their own competence may further amplify such an effect. We analyzed user confidence evolution across the different reliance patterns exhibited, which helps explain why inappropriate reliance occurs. Through an analysis exploring relatively less-competent individuals, we found that the underestimation of AI trustworthiness may also play a role in shaping under-reliance, which is potentially relevant to the metacognitive bias called the Dunning-Kruger effect [60]. Our work has important implications for designing effective interventions to promote appropriate reliance in the context of human-AI decision making.

4.2 Related Work

Our work is closely related to the studies on human-AI decision making, appropriate reliance on AI systems, and explanation-based debugging of machine learning systems.

Human-AI Decision making. With the technical advances of deep learning methods in the recent decade, researchers have shown much interest in putting such methods for a wide arrange of applications (like medical image analysis [164], autonomous driving [165]). However, due to the intrinsic uncertainty and opaqueness of such AI systems, it would be undesirable to make AI systems automate the decision making, especially in high-stakes scenarios (*e.g.,* legal judgment, medical diagnosis). Under such circumstances, AI systems are expected to play a supporting role for human decision makers. According to GDPR, users have the right to obtain meaningful explanations to work with such AI systems [166]. Motivated by this, a series of work has proposed to construct human-centered explainable AI systems [27, 124, 126] for better human-AI collaboration. Existing work has widely explored how different user factors (*e.g.,* expertise [167, 168], risk perception [110], machine learning literacy [108]) and interaction designs (*e.g.,* performance feedback [57, 111, 119], explanation [107], user tutorial [117, 169]) will affect user trust in and reliance on AI systems.

Appropriate Reliance on AI Systems. One important goal of human-AI decision making is complementary team performance [25, 70], which requires appropriate reliance [40]. In practice, however, humans always misuse (*i.e.,* over-reliance [170], relying on automation when it performs poorly) or disuse (*i.e.,* under-reliance [107, 112, 171], rejecting automated predictions when it is correct) AI systems. Such inappropriate reliance results in sub-optimal team performance, which is always worse than AI alone [25, 70]. To mitigate such issues, existing work has proposed different interventions including user tutorials [108, 117], cognitive force functions [55], and improving AI literacy of the use case [50]. Another stream of work proposed to improve the transparency of AI systems with effective explanations [107, 122], performance feedback [49], and global model properties [172]. In summary, these works presented users with extra information about AI systems (more

than advice) or changed users' mindset and knowledge of AI systems.

Explanation-based Debugging. Explanation-based debugging was found to be helpful for improving human understanding of machine learning system [173]. Recent works in both natural language processing tasks [162] and computer vision tasks [163] have explored how to leverage explanations for model debugging. The core idea of such explanation-based debugging is to check whether the explanations from AI systems misalign with human (expert) knowledge. From human feedback, it would be possible to improve machine learning models' robustness, *e.g.,* with reducing spurious reasoning patterns [174, 175] and bias in dataset [176]. Debugging in programming is the process by which programmers can determine the potential errors in the source code and resolve these errors [177]. Inspired by such an idea, we proposed debugging as an intervention to help participants understand the limitations of both explanations and advice of AI systems. In such an error finding and resolution process, users may learn when the AI system is trustworthy.

Compared with these studies, our focus is to promote appropriate reliance on AI systems by improving users' capability to critically evaluate AI performance at the instance level. For that purpose, we design an elaborate debugging intervention to help users realize the limitations of both AI advice and AI explanation, which may result in calibrated trust in and appropriate reliance on the AI system.

4.3 Task, Hypotheses, and Intervention

In this section, we describe the deceptive review detection task and present how we designed the debugging intervention. Based on the explanation-based debugging setting, we further proposed our hypotheses to verify.

4.3.1 Deceptive Review Detection Task

In the context of AI-assisted decision making, the decision tasks are typically challenging for humans, while the AI system may achieve superior performance. In this chapter, we base our experiment within such a challenging task – deceptive review detection – where AI advice can be a realistic need. In each task, based on a hotel review, participants are asked to identify whether it is genuine (*i.e.,* written by real customers) or deceptive (*i.e.,* reviews written by people who did not stay at the hotel). An example of this task is shown in Figure 4.1. This task has been used in prior work exploring Human-AI decision making [117, 122]. We also used the same public dataset [122].[1]

Using Text Highlights as Explanations. In our study, we consider a real-world scenario where the performance of an AI system is not provided or available. To help participants assess the trustworthiness of advice from the AI system in each instance of decision making, we provide local explanations for each prediction. Following Lai *et al.* [117], we adopted BERT-LIME (a popular explanation method in text classification tasks) to generate text highlights as local explanations for each AI advice. We first finetuned the BERT [178] (bert-base-uncased) on the deceptive review detection dataset, and then generated the top-10 highlighted features from post-hoc XAI method LIME [159] as explanations.

[1] https://github.com/vivlai/deception-machine-in-the-loop

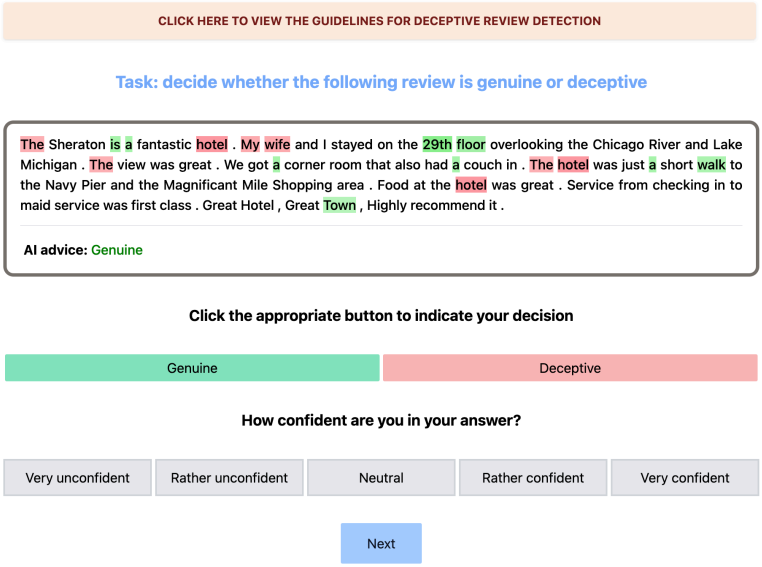


Figure 4.1: Task interface and an example of the deceptive review detection task.

Selection of Tasks. To measure the effect of the debugging intervention in our study, two batches of tasks with compatible difficulty levels are required. For that purpose, we conducted a pilot study on human performance over 20 tasks randomly sampled from evaluation and test set of the deceptive review detection dataset. We divided the trial cases into two sets of 10 tasks with equal human performance in a pilot study (10 participants).

Two-stage Decision Making. Following existing empirical study design of human-AI decision making [65, 85], all participants in our study work on each trial case with two stages of decision making. In the first stage, only task input (*i.e.,* one paragraph of hotel review) is provided; participants need to make an initial decision on themselves. After that, the same task input along with a local explanation (*e.g.,* text highlights in review, one example shown in Figure 4.1) and AI advice are provided. They will make the final decision based on all information. To help participants work on this challenging task, we provide a button to access the guidelines in each stage. In addition to making a decision for each task, we also collected participants' confidence in each decision with a 5-point Likert scale: *Very Unconfident*, *Rather Uconfident*, *Neutral*, *Rather Confident*, *Very Confident*.

4.3.2 Hypotheses

Our experiment was designed to answer questions surrounding the impact of the proposed explanation-based debugging intervention on user estimation of AI performance, and user reliance on AI systems. Putting users into a debugging setting, they will try to challenge the AI advice and explanations. Along with the real-time feedback about the debugging results, they can have a better understanding of how the AI system works and when the explanation and advice are reliable. Thus, they can more accurately estimate the perfor-

mance of the AI system when no performance of the AI system is provided, and rely on the AI system more appropriately. Based on this, we expect to observe:

> **(H1)** Encouraging users to critically evaluate the trustworthiness of AI advice at the instance level in a debugging intervention, will improve their assessment of the AI system's performance at the instance and global levels.
> **(H2)** Encouraging users to critically evaluate the trustworthiness of AI advice at the instance level in a debugging intervention, will improve the extent to which users appropriately rely on the system.

Within a debugging intervention, to present a balanced view of AI systems, we considered showing both the strength and weakness of an AI system (by providing accurate or inaccurate advice). Thus, multiple tasks of different characteristics will be presented in the debugging intervention. When these tasks are presented in different orders, users may show different learning effects, which further affects the reliance on AI systems. Thus, we hypothesize that:

> **(H3)** The trustworthiness of AI advice at the instance level in a debugging intervention corresponds to an ordering effect with respect to appropriate reliance.

4.3.3 Debugging Intervention

To help participants accurately assess the trustworthiness of AI advice at the instance level and calibrate their reliance on the AI system, we designed a debugging intervention with explanations generated with post-hoc explanation methods LIME [159]. Our data and code is available with anonymous companion page.[2]

Explanation-Based Human Debugging. Through the debugging phase, all participants are supposed to learn two important facts about the AI system: (1) the AI advice is not always correct, and (2) explanations are not always informative and helpful in identifying the trustworthiness of AI advice. Thus, we considered two main factors for each task: (1) the correctness of AI advice, and (2) whether an explanation is informative (*i.e.,* combined with guidelines, whether or not such explanations can help participants easily identify the correct answer). Participants subjected to training were presented with a hotel review with explanatory elements consisting of a model prediction and color-coded highlights showing 10 predominant features. Each token highlight shows the contribution of the token to the model prediction on a 5-point Likert scale: *deceptive*, *somwhat deceptive*, *neutral*, *somewhat genuine*, *genuine*. This difference in the contribution is distinguished by the color and intensity of the highlight shown in the interface. An example of the debugging phase is shown in Figure 4.2. They are instructed to read the text, and, when deemed necessary, refine the explanations by adjusting the highlights and indicating whether the AI advice is correct. After each task, the correctness of AI advice and missed adjustments will be shown to the participant as real-time feedback. In practice, the explanations obtained

[2] https://osf.io/dh34y/?view_only=6a6833eafdbd4d5daa8c036579247159

Figure 4.2: Screenshot of debugging interface.

from XAI methods may not always align with human understanding [179]. Besides realizing the explanations are not always helpful, we hope participants can learn patterns they can rely on to make the decision given the guidelines. With that wish, the authors manually adjusted the highlights generated with BERT-LIME according to the task guidelines (from [117]) and take the adjusted highlights as ground truth for debugging phase.

Real-time Feedback in Debugging Intervention. We provide the real-time feedback in each debugging task, to show whether AI advice is correct and which highlights participants missed to adjust according to guidelines. One example of the feedback is shown in Figure 4.3.

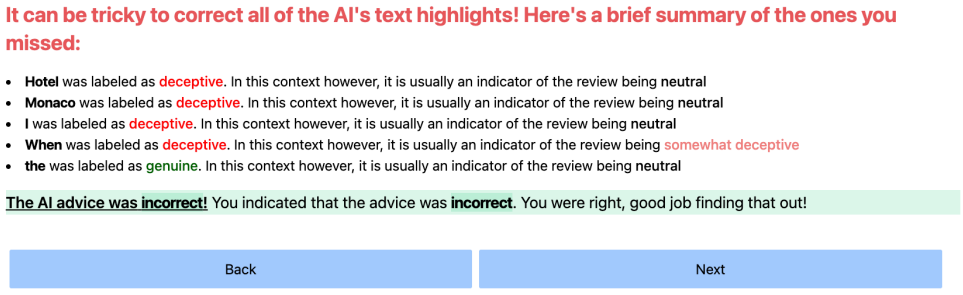


Figure 4.3: Screenshot of debugging feedback.

Selection of the Debugging Tasks. To create a balance between the strength and weakness of the AI system, we manually selected four tasks with informative explanations (where explanations and guidelines can help participants easily identify the correct answer) and four tasks with uninformative explanations. The eight tasks presented in our debugging phase are: (1) two tasks with correct AI advice and informative explanations, (2) two tasks with correct AI advice and uninformative explanations, (3) two tasks with incorrect AI advice and informative explanations, (4) two tasks with incorrect AI advice and uninformative explanations. The tasks are balanced in whether explanations are informative and whether AI advice is correct. While the informative explanations are manually selected, the correctness of AI advice is determined randomly.

Ordering Effect. When presenting the debugging phase to participants, the order of tasks may have an impact on their estimation of AI performance and reliance on the AI system. According to existing work [84, 87], first impressions (either good or bad) greatly affect user estimation of AI performance and user trust in AI systems. Overall, both correct AI advice and informative explanations tend to leave positive impression on users. As pointed out by a recent study [180], the public would prioritize the accuracy of AI systems over interpretability. Thus, compared with "wrong AI advice, informative explanation" case, we would consider "correct AI advice, uninformative explanation" will leave participants a better impression. With these concerns, we designed three orders of tasks:

- Random order.
- Decreasing impression order (*i.e.,* from good to bad): correct AI advice, informative explanation → correct AI advice, uninformative explanation → wrong AI advice, informative explanation → wrong AI advice, uninformative explanation.
- Increasing impression order (*i.e.,* from bad to good): wrong AI advice, uninformative explanation → wrong AI advice, informative explanation → correct AI advice, uninformative explanation → correct AI advice, informative explanation.

4.4 Study Design

This section describes our experimental conditions, variables, participants, and procedure in our study. This study was approved by the human research ethics committee of our institution. More implementation details can be found in the appendix (4.8.1).

4.4.1 Experimental Conditions

In our study, all participants worked on deceptive review detection tasks with a two-stage decision making process (described in Sec. 4.3.1). In all conditions, the top-10 most important features obtained from BERT-LIME are highlighted as an explanation for AI advice to help participants identify the trustworthiness of AI advice.

The differences between conditions are whether debugging intervention is adopted and the order of debugging tasks. To comprehensively study the effect of debugging intervention, we considered four experimental conditions in our study: (1) no debugging intervention (represented as `Control`), (2) with debugging intervention, debugging tasks in random order (represented as `Debugging-R`), (3) with debugging intervention, debugging tasks in decreasing impression order (represented as `Debugging-D`), (4) with debugging intervention, debugging tasks in increasing impression order (represented as `Debugging-I`). In

Table 4.1: The different variables considered in our experimental study. "DV" refers to the dependent variable.

Variable Type	Variable Name	Value Type	Value Scale
Assessment (DV)	EAP	Continuous, Interval	[0, 10]
	ETP	Continuous, Interval	[0, 10]
	MAP	Continuous, Interval	[0, 10]
	MTP	Continuous, Interval	[0, 10]
	CCD	Continuous, Interval	[0, 10]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous, Interval	[0.0, 1.0]
	RAIR	Continuous, Interval	[0.0, 1.0]
	RSR	Continuous, Interval	[0.0, 1.0]
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
Trust (DV)	TiA-R/C	Likert	5-point, 1: poor, 5: very good
	TiA-U/P	Likert	5-point, 1: poor, 5: very good
	TiA-IoD	Likert	5-point, 1: poor, 5: very good
	TiA-Trust	Likert	5-point, 1:strong distrust, 5: strong trust
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-PtT	Likert	5-point, 1: tend to distrust, 5: tend to trust
	TiA-Familiarity	Likert	5-point, 1: not familiar, 5: very familiar

conditions with debugging intervention, participants were presented with eight selected tasks with performance feedback and manually adjusted contribution of tokens. While in `Control` condition, the eight tasks selected are presented as normal tasks without any feedback of AI advice correctness or adjusted explanation feedback. Such a control setting is designed to compare with debugging intervention and eliminate the learning effect brought by the eight tasks.

For each batch of ten tasks, the AI system was configured to provide correct advice on eight of them and incorrect advice on two tasks. To eliminate the potential ordering effect of trial cases, we randomly assigned one batch of selected tasks (see section 4.3.1) as the first batch and further shuffled the task order within each batch.

4.4.2 Measures And Variables

To have a more comprehensive view of variables used in our experimental analysis, we listed the main variables in Table 4.1. Notice that we do not add the confidence and dimensions from the NASA-TLX questionnaire [181] into it.

To verify **H1**, we assessed participants' global estimation of AI system's performance with two questions: "From the previous 10 tasks, on how many tasks do you estimate the AI advice to be correct?" and "From the previous 10 tasks, how many questions do you estimate to have been answered correctly? (after receiving AI advice)". The answers to the two questions correspond to participants' estimation of AI performance and team performance respectively. We can refer to the estimated trustworthiness as estimated AI performance (**EAP**) and estimated team performance (**ETP**). Comparing that performance estimation with actual performance in abstract difference, we can calculate the degree of miscalibration of AI performance (**MAP**) and team performance (**MTP**). If participants can accurately estimate the performance of AI system at instance level, they may make the final decision with high confidence. Thus, for the AI performance estimation at instance level, we calculated the number of tasks they made the correct final decision with indication of "Very Confident" (**CCD**).

To verify **H2** and **H3**, we measured both reliance and appropriate reliance of partici-

Table 4.2: The different appropriate reliance patterns considered in [29]. d_i and d_f refer to initial human decision and final human decision respectively. ✓ and × refer to correct and incorrect respectively.

d_i	AI advice	d_f	Reliance
×	✓	✓	Positive AI reliance
×	✓	×	Negative self-reliance
✓	×	✓	Positive self-reliance
✓	×	×	Negative AI reliance

pants on the AI system. The reliance is measured with two widely adopted metrics: the **Agreement Fraction** and the **Switch Fraction**. These look at the degree to which participants are in agreement with AI advice, and how often they adopt AI advice in cases of initial disagreement. They are commonly used in the literature, for example in [49, 57, 66]. As for the appropriate reliance, we followed Schemer *et al.* [29] to calculate the appropriate reliance based on four reliance patterns (shown in Table 4.2). According to the four reliance patterns where human initial decision disagree with AI advice and the correct answer occurs in one of them, we can assess the appropriate reliance from two dimensions:

$$\textbf{RAIR} = \frac{\text{Positive AI reliance}}{\text{Positive AI reliance} + \text{Negative self-reliance}},$$

$$\textbf{RSR} = \frac{\text{Positive self-reliance}}{\text{Positive self-reliance} + \text{Negative AI reliance}}.$$

They stand for whether users switch to AI advice when AI outperforms them, and whether users can insist on correct decisions made by themselves when AI advice is incorrect. In addition, we consider the accuracy in batches to measure participants' performance with AI assistance.

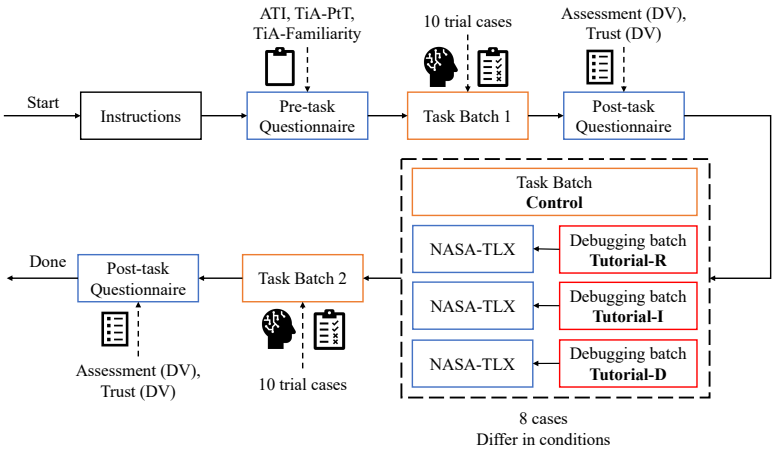


Figure 4.4: Illustration of the procedure that participants followed within our study. Blue boxes represent questionnaire phase, orange boxes represent task phase, and the red box represents the debugging intervention.

For a deeper analysis of our results, a number of additional measures were considered based on observations from existing literature [87, 153, 154]:

- Trust in Automation (TiA) questionnaire [90], a validated instrument to measure trust [87] consisting of 6 subscales: *Reliability/Competence* (TiA-R/C), *Understanding/Predictability* (TiA-U/P), *Propensity to Trust* (TiA-PtT), *Familiarity* (TiA-Familiarity), *Intention of Developers* (TiA-IoD), and *Trust in Automation* (TiA-Trust).
- Affinity for Technology Interaction Scale (ATI) [91], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [87].
- NASA-TLX questionnaire [181] for the working load assessment of the debugging intervention.

4.4.3 Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [92]. To correct for testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{3} = 0.017$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.017$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and that we will investigate 4 different experimental conditions. This resulted in a required sample size of 230 participants. We thereby recruited 324 participants from the crowdsourcing platform Prolific[3], in order to accommodate potential exclusion.

Compensation. All participants were rewarded with £3.8, amounting to an hourly wage of £7.6 (estimated completion time was 30 minutes). We rewarded participants with extra bonuses of £0.05 for every correct decision in the 20 trial cases. Such extra bonus for correct decisions provides a monetary motivation for crowd workers to try their best on each task, which is also widely adopted by existing work [50, 117].

Filter Criteria. All participants were proficient English speakers above the age of 18. For a high-quality study, we require participants to have an approval rate of at least 90% and more than 80 successful submissions on the Prolific platform. After reading the basic introduction and guidelines about the deceptive review detection task, participants who failed any qualification test (about understanding the task) were removed from our study. After data collection, we excluded participants from our analysis if they failed any attention check (90 participants). The resulting sample of 234 participants had an average age of 39 (SD = 13) and a balanced gender distribution (48.7% female, 49.6% male, 1.7% other).

4.4.4 Procedure

The full procedure of our study can be visualized in Figure 4.4. In the beginning, all participants will be presented with a basic introduction of the deceptive review detection task. According to Lai *et al.* [117], guidelines about how to identify deceptive reviews are highly useful in improving user performance on this task. Thus, we also follow them to provide the guidelines in the introduction. Then, participants will be checked with two qualifi-

[3] https://www.prolific.co

cation questions to ensure they carefully read the instruction and understand this task. Any failure at the qualification test will result in removal from our study. All reserved participants will then be asked to answer a pre-task questionnaire consisting of affinity for technology interaction, TiA-PtT, and TiA-Familiarity.

As described in section 4.3.1, we selected two batches of tasks (10 for each batch) as trial cases and 8 tasks for debugging intervention. For all conditions, participants will first work on the first batch of tasks and go through a post-task questionnaire for assessment of AI performance and subjective trust in AI system (*i.e.,* with TiA subscales). The main difference between conditions (shown in the dashed box of Figure 4.4) is the 8 tasks presented after the post-task questionnaire. In condition `Control`, participants will work on the 8 tasks as normal trial cases. No debugging intervention and result feedback will be provided. In comparison, we show debugging intervention and result feedback in conditions `Debugging-R`, `Debugging-I`, and `Debugging-D`. In conditions with debugging intervention, the participants will go through the debugging tasks with different task orders and be asked about the task working load resulting from the debugging intervention, using the NASA-TLX [181] questionnaire. Then, participants in all conditions will continue to work on another batch of tasks and answer the same post-task questionnaire as the one after the first task batch.

4.5 Results and Analysis

In this section, we present the main results of our study (*i.e.,* hypothesis tests) and further exploration about reliance shaping with confidence dynamics.

4.5.1 Descriptive Statistics

In our analysis, we only consider participants who passed all attention checks, as a measure of participant reliability [93]. Participants were distributed in a balanced fashion across conditions: 57 (`Control`), 59 (`Debugging-R`), 60 (`Debugging-D`), 58 (`Debugging-I`). On average, participants spend around 51 minutes (SD = 14) in our study.

Variable Distribution. The covariates' distribution is as follows: *ATI* (M = 3.91, SD = 0.94, 6-point Likert scale, *1: low, 6: high*), *TiA-PtT* (M = 2.89, SD = 0.61, 5-point Likert scale, *1: tend to distrust, 5: tend to trust*), *TiA-Familiarity* (M = 2.29, SD = 1.09, 5-point Likert scale, *1: unfamiliar with AI system used in study, 5: familiar with AI system used in study*).

The working load of debugging intervention is measured with NASA-TLX questionnaire (scale in [-7, 7]). For all dimensions except "Performance", a higher value indicates a higher working load. In the dimension "Performance", a smaller value indicates a higher estimated performance on tasks. We visualized the dimensions in Figure 4.5. In general, participants think the debugging intervention requires high "Mental Demand" and "Effort", but low "Physical Demand" and "Temporal Demand". While most participants do not show high expectations in achieved "Performance", they also do not get troubled with "Frustration".

Performance Overview. On average across all conditions, participants achieved an accuracy of 0.64 (SD = 0.11) over the two batches of tasks, still lower than the aforementioned AI accuracy of 0.8. The agreement fraction is 0.66 (SD = 0.13) while the switching fraction is 0.31 (SD = 0.22). With these measures, we confirm that when disagreement appears par-

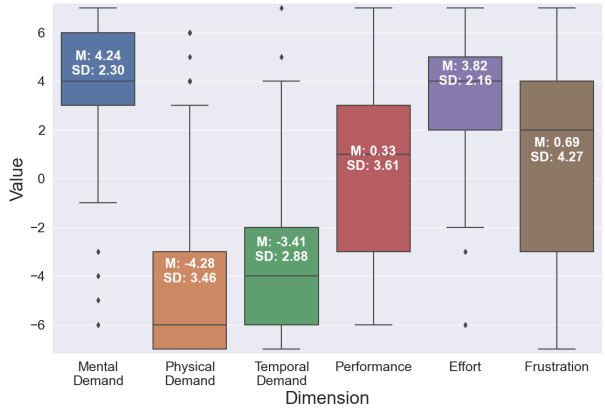


Figure 4.5: Box plot illustrating the distribution of the different dimensions in NASA-TLX questionnaire. *M* and *SD* represent mean and standard deviation respectively.

ticipants in our study did not always switch to AI advice and participants did not blindly rely on the AI system. In the two batches of tasks (10 for each batch), the average estimated AI performance are 5.81 (*SD* = 1.91) and 5.79 (*SD* = 1.71) respectively; the average estimated team performance is 6.64 (*SD* = 1.74) and 6.44 (*SD* = 1.87) respectively. Overall, participants underestimated the performance of the AI system and believed they could outperform the AI system on this task after receiving AI advice.

4.5.2 Hypothesis Tests

H1: the effect of critical evaluation setting on AI performance estimation

To verify H1, we used Wilcoxon signed rank tests to compare all assessment-based dependent variables of participants before and after the debugging intervention (only participants in condition `Debugging-R`, `Debugging-D`, `Debugging-I` are considered). The results are shown in Table 4.3. Although no significant results were found to support **H1**, we found that participants in `Debugging-D` condition showed a worse **MTP** after the debugging intervention, in contrast to our expectations. Thus, **H1** is not supported.

Table 4.3: Wilcoxon signed ranks test results for **H1** on AI performance estimation. "†" indicates the effect of variable is significant at the level of 0.017 (adjusted alpha).

Condition	**Debugging**		`Debugging-R`		`Debugging-D`		`Debugging-I`	
DV	T	p	T	p	T	p	T	p
MAP	3833	.662	363	.742	463	.238	457	.826
MTP	4006	.957	512	.892	324	**.992**†	528	.160
CCD	3761	.753	474	.717	379	.660	429	.603

To have a closer look at how participants' assessment of the AI performance and Team performance change after debugging intervention. We compared the assessment of AI performance and team performance across conditions. With Kruskal-Wallis H-test, we found no significant difference in the estimation across conditions with debugging inter-

Table 4.4: Participants' estimation of AI performance and Team Performance.

Condition	Debugging-R		Debugging-D		Debugging-I	
Estimation	Before	After	Before	After	Before	After
EAP	6.05 ± 1.63	5.97 ± 1.65	5.92 ± 1.89	6.07 ± 1.67	6.00 ± 2.20	5.64 ± 1.90
ETP	6.81 ± 1.55	6.36 ± 1.85	6.68 ± 1.48	6.57 ± 1.65	6.81 ± 1.90	6.60 ± 1.92

Table 4.5: Wilcoxon signed ranks test results for **H2** on reliance-based dependent variables. "†" indicates the effect of variable is significant at the level of 0.017 (adjusted alpha).

Condition	**Debugging**		Debugging-R		Debugging-D		Debugging-I	
Dependent Variables	T	p	T	p	T	p	T	p
Accuracy	6494	**.998**†	840	.952	679	.935	684	.970
Agreement Fraction	6332	.896	756	.639	512	.475	897	.971
Switch Fraction	6812	.953	703	.735	762	.565	817	.979
RAIR	6340	.981	628	.829	722	.618	807	**.995**†
RSR	2494	.736	241	.325	311	.953	292	.461

vention. We show their estimation with mean value and standard deviation ($M \pm SD$) in Table 4.4. We found that (1) generally, participants showed a worse estimation of AI performance and team performance after the debugging intervention; (2) only participants in the Debugging-D condition showed a slight increase in the estimation of AI performance.

H2: the effect of critical evaluation setting on appropriate reliance

Similarly, to analyze the effect of the debugging intervention on user reliance on the AI system (**H2**), we used Wilcoxon signed rank tests to compare all reliance-based dependent variables of participants before and after the debugging intervention. The results are shown in Table 4.5. Overall in all conditions with the debugging intervention, the improvement in reliance caused by debugging intervention was not statistically significant. With a post-hoc Mann-Whitney test on **Accuracy**, we found that: after the debugging intervention, the accuracy drops significantly. For a fine-grained analysis, we further conducted Wilcoxon signed rank tests on each condition with the debugging intervention. We found that participants in the Debugging-I condition show a significant difference in **RAIR**, while no significant difference is found with post-hoc Mann-Whitney test. The observed results do not support the **H2**.

Although no significant improvement was found in the performance and reliance measures due to debugging intervention, we did witness a drop in reliance measures generally: **Accuracy** (0.67 → 0.63), **Agreement Fraction** (0.68 → 0.66), **Switch Fraction** (0.34 → 0.28), **RAIR** (0.38 → 0.30), **RSR** (0.64 → 0.61). This is evident in the condition Debugging-I: **Accuracy** (0.68 → 0.63), **Agreement Fraction** (0.71 → 0.66), **Switch Fraction** (0.39 → 0.29), **RAIR** (0.43 → 0.29), **RSR** (0.59 → 0.61). When AI advice is in disagreement with users' initial decision, users tend to rely on themselves more than they should. This results in decreased (appropriate) reliance and accuracy. In the deceptive review detection tasks, the AI system performs generally better than participants. The reduced reliance may help explain why we found a decreased accuracy on average.

H3: ordering effect of debugging tasks
For the analysis of the ordering effect, meanwhile mitigating the individual differences and learning effect brought by the eight tasks used in debugging phase, we compared the difference of reliance-based dependent variables (calculated with the difference between the second batch and the first batch) and user reliance on the second batch with participants of all conditions with Kruskal Wallis test. No significant difference is found with such comparisons. To compare the task working load brought by debugging intervention of different ordering, we conducted Kruskal-Wallis H-test on the six measures in the NASA-TLX questionnaire. No significant difference is found. Thus, **H3** is also not supported.

To further look at how the ordering effect of debugging tasks affects the final performance of participants. We counted the participants who achieved an accuracy level above 80% (*i.e.,* compatible with or better than provided AI system) in the second task batch. After filtering out the participants who blindly rely on the AI system (*i.e.,* **Agreement Fraction** is 1.0), we found the number of participants in condition `Debugging-D` (14) is clearly more than in condition `Debugging-R` (9) and `Debugging-I` (9). In comparison, the number of participants who achieved an accuracy level above 80% in condition `Control` is 11. Although the ordering effect does not show a significant statistical difference, such an observation lends partial support to **H3**.

4.5.3 Exploratory Analyses

Trust Analysis
To explore whether our debugging intervention had any effect on user trust in AI system, we conducted Wilcoxon signed ranks test comparing the trust before and after the debugging intervention. On average, there is a slight drop in assessed trust in automation subscales (*i.e.,* **TiA-R/C**, **TiA-U/P**, **TiA-IoD**, **TiA-Trust**) after the debugging intervention, but no statistically significant difference are found in test results. This suggests that the designed debugging intervention can calibrate user reliance and estimation of AI performance without directly shaping their trust in the AI system.

Table 4.6: Kruskal-Wallis H-test results for user estimated trustworthiness and miscalibration of estimated performance based on performance quartiles. "††" indicates the effect of the variable is significant at the level of 0.017. "Top", "Middle", and "Bottom" refer to participants in the top quartile, middle quartiles, and bottom quartile based on the performance of the first batch of tasks, respectively.

Variables	H	p	$M \pm SD$(Top)	$M \pm SD$(Middle)	$M \pm SD$(Bottom)	Post-hoc results
EAP	41.54	**<.001**††	6.85 ± 1.48	5.84 ± 1.91	4.69 ± 1.68	Top > Middle > Bottom
ETP	15.85	**<.001**††	7.29 ± 1.40	6.65 ± 1.66	5.98 ± 1.96	Top > Middle, Bottom
MAP	40.89	**<.001**††	1.32 ± 1.33	2.29 ± 1.74	3.31 ± 1.68	Top < Middle < Bottom
MTP	22.67	**<.001**††	1.46 ± 1.24	1.34 ± 1.12	2.24 ± 1.28	Top, Middle < Bottom
CCD	23.17	**<.001**††	2.08 ± 1.70	2.06 ± 1.78	0.86 ± 1.00	Top, Middle > Bottom

Covariates Impact on Trust and Reliance
To analyze the impact of covariates on user trust and reliance, we conducted the Spearman rank-order tests with covariates and the average trust and reliance-based dependent variables on two batches of tasks. The results show that, propensity to trust (*i.e.,* **TiA-**

PtT) is the only factor which shows sinificant positive correlations with trust-based measures: **TiA-R/C** ($r(234) = 0.270, p = .000$), **TiA-U/P** ($r(234) = 0.165, p = .011$), **TiA-IoD** ($r(234) = 0.234, p = .000$), **TiA-Trust** ($r(234) = 0.303, p = .000$).

Users' Estimation of AI Trustworthiness

To further understand how users' estimation of AI trustworthiness affects their reliance and performance, we split the participants in all conditions into performance-based quartiles. To avoid the impact of debugging intervention, we only considered user performance in the first batch of tasks. The top quartile corresponds to those demonstrating high accuracy (top 25%), the bottom quartile corresponds to those with low accuracy (bottom 25%), and we combine the two quartiles in the middle comprising of participants with a medium level of performance in the first batch of tasks. To show how these participants differ in their appropriate reliance and estimation of AI trustworthiness, we adopted the Kruskal-Wallis H test to compare the estimated performance and their assessment of the AI system's performance at the instance and global levels. Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.017 ($\frac{0.05}{3}$) were used to make pairwise comparisons of performance. Generally, participants in the high accuracy group showed more appropriate reliance (*i.e.,* **RAIR** and **RSR**) than the low accuracy group (with statistical significance). The results of user estimation of performance, AI trustworthiness, and miscalibration of performance are shown in Table 4.6. Overall, participants in the high accuracy group showed significantly higher AI performance and team performance in comparison with the low accuracy group. Meanwhile, the high accuracy group also has a more precise estimation of AI performance and team performance (*i.e.,* significantly lower **MAP** and **MTP**) and makes more correct decisions confidently (significantly higher **CCD**). It also indicates that the underestimation of AI trustworthiness can be the main cause of the under-reliance, which results in lower accuracy.

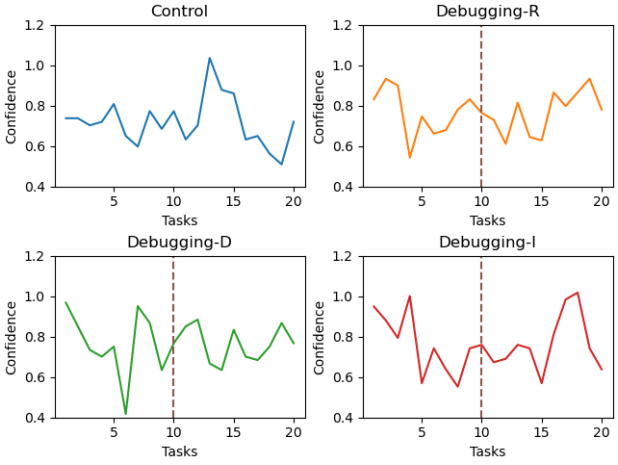


Figure 4.6: Illustration of dynamics of confidence change in the 20 tasks of each condition. The brown dashed line represents the debugging intervention.

Table 4.7: Reliance and confidence correlation.

Pattern	Dependent Variables	*M*	*SD*
Reliance	**Initial agreement**	0.38	0.75
	Initial disagreement	-0.42	0.99
	Final agreement	0.23	0.90
	Final disagreement	-0.44	0.90
	Switch behavior	-0.32	1.17
Appropriate Reliance	**Positive AI reliance**	-0.34	1.17
	Negative AI reliance	-0.23	1.18
	Positive self-reliance	-0.41	0.88
	Negative self-reliance	-0.48	0.89

Confidence Analysis

We show the difference in confidence dynamics of four conditions in Figure 4.6. On average, participants show positive confidence (above neutral) in their final decisions. After receiving the debugging intervention, both `Debugging-I` and `Debugging-R` conditions showed decreased confidence, but it comes back to the average level soon and keeps vibrating around it. By contrast, participants in condition `Debugging-D` showed increased confidence after the debugging intervention and keeps relatively stable compared with all other conditions.

We calculated the confidence change after receiving AI advice based on nine different reliance patterns: whether initial decision agrees with AI advice, whether final decision agrees with AI advice, switch behavior, and four reliance patterns considered in calculating appropriate reliance (see Table 4.2). The results are shown in Table 4.7. In general, participants indicated increased confidence when AI advice agreed with their initial decision, and showed decreased confidence when AI advice disagreed with their initial decision. And even if participants choose to switch to AI advice given initial disagreement, they tend to show decreased confidence in the final decision. Considering the four patterns in calculating appropriate reliance, users' confidence drop seems to be more severe when insisting on their own decision, compared with adopting AI advice.

4.6 Discussion

4.6.1 Key Findings

In order to promote appropriate reliance on AI systems by calibrating user estimation of AI performance, we proposed a debugging intervention to educate participants that AI systems are not always reliable and that the explanations may also not always be informative. We hypothesized that the proposed debugging intervention could improve critical thinking about the AI system, which can facilitate appropriate reliance on the AI system. As opposed to our hypotheses, such a debugging intervention fails to calibrate participants' estimation of AI performance at both the global and local levels. Participants tended to rely less on the AI system after receiving the debugging intervention. Through an exploratory analysis based on different performance quartiles, we found that participants who performed worse in our study tended to underestimate AI performance. Thus, they

achieved suboptimal team performance, which is largely impacted by the under-reliance on the AI system. These findings can also be explained using the lens of plausibility of the XAI intervention. According to Jin *et al.* [182], plausibility can substantially affect user perceived trustworthiness of the AI system. The debugging intervention may make the XAI (*i.e.,* text highlights in our study) less plausible to users, which results in more tendency to underestimate AI performance.

In our study, no significant difference was found between the different ordering of debugging tasks across experimental conditions. However, participants who were exposed to the weakness of the AI system at the beginning of the debugging intervention, showed a more obvious tendency to disuse the AI system. Such under-reliance was found to result in sub-optimal team performance. This finding is in line with recent work that has uncovered similar ordering effects and cognitive biases influencing outcomes in human interaction with intelligent systems [84, 87]: a bad first impression of an AI system can lead to an underestimation of AI competence and reduced reliance on the system.

Confidence Analysis. We calculated the confidence change after receiving AI advice based on nine different reliance patterns: whether initial decision agrees with AI advice, whether final decision agrees with AI advice, switch behavior, and four reliance patterns considered in calculating appropriate reliance (see appendix). In general, participants indicated increased confidence when AI advice agreed with their initial decision (+0.38 on average), and showed decreased confidence when AI advice disagreed with their initial decision (−0.42 on average). And even if participants choose to switch to AI advice given initial disagreement, they tend to show decreased confidence in the final decision (−0.32 on average). Considering the four patterns in calculating appropriate reliance, users' confidence drop seems to be more severe when insisting on their own decision, compared with adopting AI advice.

In further analysis of covariates (cf. Sec 4.5.3), we found that general propensity to trust shows a positive correlation with all trust subscales. However, no significant correlations were found between the propensity to trust and reliance, which indicates that the increased trust due to the propensity to trust does not translate to reliance behaviors. Meanwhile, the confidence dynamics in different reliance patterns showed that AI advice may amplify the confidence of user decisions when in agreement and decrease user confidence when in disagreement. Under disagreement, users appear to rely more on themselves (*i.e.,* indicated by confidence decrease), as opposed to adopting AI advice.

4.6.2 Implications

Our findings suggest that the debugging intervention and similar interventions with training purposes (*e.g.,* user tutorial) may suffer from the cognitive bias brought by the ordering effect within such interventions. If we want to use such interventions to show users both strength and weakness of AI systems, we should avoid leaving users with a bad first impression of the weakness of the AI system. Meanwhile, in our study, participants tend to be optimistic about the team performance while underestimating the AI performance. It is possibly caused by meta cognitive bias — Dunning-Kruger effect [30, 60]. According to previous work [60], Dunning-Kruger Effect is mainly triggered among less-competent individuals overestimating their own competence/performance in a task. In our study, we found that less-competent individuals showed a greater tendency to underestimate the AI

performance and make fewer correct decisions with confidence (see Table 4.4). This indicates that the underestimation of AI systems can also contribute to under-reliance in the context of human-AI decision making. According to He *et al.* [30], an overestimation of self-competence can result in under-reliance on the AI system. Both the overestimation of self-competence and the underestimation of AI competence can contribute to an illusion of superior competence over the AI system. As a result, users with such an illusion tend to disuse the AI system. To conclude whether the underestimation of AI performance plays a role in triggering the Dunning-Kruger effect in the context of human-AI decision making, more work is required in the future.

Through our study, we also found that the reliance patterns (*e.g.,* agreement, disagreement) have a clear correlation with user confidence change. When the AI system disagrees with human initial decision, decision makers' confidence shows a clear decrease. And compared with insisting on their own decision, they may have higher confidence when giving agency to AI advice. Such observation may be a dangerous signal for appropriate reliance. Further research is required to explore how to keep user confidence on themselves when exposed to a disagreement from an AI system.

4.6.3 Caveats and Limitations

Our debugging intervention may have left participants with a negative impression of the AI system, which could irreversibly harm the trust and reliance on the system (as shown by prior literature exploring first impressions of AI systems [87]). To make the debugging intervention more effective in building up critical mindsets and facilitating appropriate reliance, future research can explore how to avoid such side-effects. The high difficulty of the task and the debugging intervention may have influenced our findings. In a highly complex task, crowd workers may not be patient and engaged enough to fully understand the AI system at both the global and local level. Although we only focus on one specific task to verify the effectiveness of the proposed debugging intervention, such an application-grounded evaluation is still highly valuable [183]. In this chapter, we used a rigorous setup to explore the effectiveness of a debugging intervention, which can inform the future design of effective interventions for better human-AI collaboration.

Potential Bias. As pointed out by Draws *et al.* [157], cognitive biases introduced by task design and workflow may have negative impact on crowdsourcing experiments. With the help of Cognitive Biases Checklist introduced [157], we analyzed potential bias in our study. *Self-interest bias* is possible, because crowd workers we recruited from the Prolific platform are motivated the monetary compensation. Thus, it would be challenging to keep participants engaged in the debugging intervention and highly motivated to learn from the weakness of AI system. That could be also potential reason why the debugging intervention does not work as expected. To alleviate any participants with low effort results, we put attention checks to remove ineligible participants from our study. The observation of reduced reliance brought by bad first impression also happens with *Anchoring Effect*. Meanwhile, the participants generally under-estimate the AI performance and believe they can outperform AI system, which also may fall into *Overconfidence or Optimism Bias*.

4.7 Conclusion

In this chapter, we present an empirical study to understand the impact of the debugging intervention on the estimation of AI performance and user reliance on the AI system. Our results suggest that we should be careful in presenting the weakness of the AI system to users, to avoid any anchoring effect which may result in under-reliance. While our experimental results do not provide support to our original hypotheses, we can not fully reach a conclusion that debugging intervention does not help with facilitating appropriate reliance on the AI system. Future work may explore how to mitigate potential bias brought by the users' overestimation of themselves along with the underestimation of AI performance. Meanwhile, our observations of confidence dynamics in different reliance patterns also provide insights for future study of human-AI decision making.

4.8 Appendix

4.8.1 Experimental Details

Guidelines. Following Lai *et al.* [117], we provided the following guidelines in the user study:

- Deceptive reviews tend to focus on aspects that are external to the hotel being reviewed, *e.g.,* husband, business, vacation.
- Deceptive reviews tend to contain more emotional terms; positive deceptive reviews are generally more positive and negative deceptive reviews are more negative than genuine reviews.
- Genuine reviews tend to include more sensorial and concrete language, in particular, genuine reviews are more specific about spatial configurations, *e.g.,* small, bathroom, on, location.
- Deceptive reviews tend to contain more verbs, *e.g.,* eat, sleep, stay.
- Deceptive reviews tend to contain more superlatives, *e.g.,* cleanest, worst, best.
- Deceptive reviews tend to contain more pre-determiners, which are normally placed before an indefinite article + adjective + noun, *e.g.,* what a lovely day!

Timer. Besides attention checks, we also add a timer to ensure each participant spends enough time on the questionnaire, task instruction, and decision tasks. A conservative estimate through trial runs reflected that participants would take at least 25 seconds to complete each decision task and 30 seconds to complete each debugging task. We reduced the time for the decision making in the second stage to 15 seconds.

Qualification Test. To ensure participants carefully read the task instruction and understand the task, we used two questions for the qualification test.

- In this study, the deceptive reviews written by? Option 1: An AI system, option 2: People without actual experience.

- Indicate whether the following statement is true or false: ”Guidelines are provided for finding deceptive reviews”. Option 1: True, option 2: False.

Attention Checks. To prevent participants from providing low-effort results in questionnaires and decision tasks, we add attention check tasks that are similar to normal ones. For example, we asked participants to select a specified option in the questionnaire. One example of attention check in decision tasks is shown in Figure 4.7. To ensure participants read the hotel review with attention, we put the instruction in the last sentence to select a specific option (*e.g.,* “In order to confirm you have read this paragraph, please select Genuine and indicate that you are Very confident in this answer.”). We have such attention checks in the middle of each task batch, long questionnaires, and the debugging intervention.

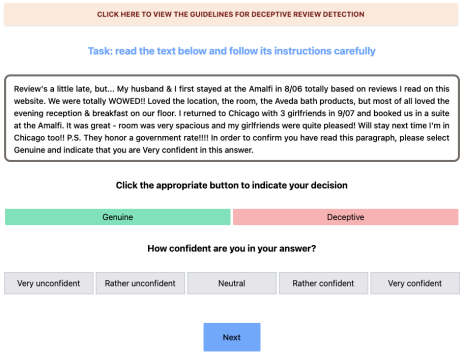


Figure 4.7: Screenshot of attention check in decision tasks.

II

Facilitating User Understanding with Human-centered XAI

4

5

Analogy-based Concept-level Explanations

Concepts are an important construct in semantics, based on which humans understand the world with various levels of abstraction. With the recent advances in explainable artificial intelligence (XAI), concept-level explanations are receiving an increasing amount of attention from the broad research community. However, laypeople may find such explanations difficult to digest due to the potential knowledge gap and the concomitant cognitive load. Inspired by prior work that has explored analogies and sensemaking, we argue that augmenting concept-level explanations with analogical inference information from commonsense knowledge can be a potential solution to tackle this issue. To investigate the validity of our proposition, we first designed an effective analogy-based explanation generation method and collected 600 analogy-based explanations from 100 crowd workers. Next, we proposed a set of structured dimensions for the qualitative assessment of such explanations, and conducted an empirical evaluation of the generated analogies with experts. Our findings revealed significant positive correlations between the qualitative dimensions of analogies and the perceived helpfulness of analogy-based explanations, suggesting the effectiveness of the dimensions. To understand the practical utility and the effectiveness of analogy-based explanations in assisting human decision-making, we conducted a follow-up empirical study (N = 280) on a skin cancer detection task with non-expert humans and an imperfect AI system. Thus, we designed a between-subjects study spanning five different experimental conditions with varying types of explanations. The results of our study confirmed that a knowledge gap can prevent participants from understanding concept-level explanations. Consequently, when only the target domain of our designed analogy-based explanation was provided (in a specific experimental con-

This chapter is based on two peer-reviewed papers: **Gaole He**, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. *Opening the Analogical Portal to Explainability: Can Analogies Help Laypeople in AI-assisted Decision Making?* Journal of Artificial Intelligence Research 81 (2024): 117-162. https://doi.org/10.1613/ jair.1.15118. and **Gaole He**, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. *It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge.* In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 10, pp. 89-101. 2022. https://doi.org/10.1609/hcomp.v10i1.21 990; **Best Paper Award**

dition), participants demonstrated relatively more appropriate reliance on the AI system. In contrast to our expectations, we found that analogies were not effective in fostering appropriate reliance. We carried out a qualitative analysis of the open-ended responses from participants in the study regarding their perceived usefulness of explanations and analogies. Our findings suggest that human intuition and the perceived plausibility of analogies may have played a role in affecting user reliance on the AI system. We also found that the understanding of commonsense explanations varied with the varying experience of the recipient user, which points out the need for further work on personalization when leveraging commonsense explanations. In summary, although we did not find quantitative support for our hypotheses around the benefits of using analogies, we found considerable qualitative evidence suggesting the potential of high-quality analogies in aiding non-expert users in their decision making with AI-assistance. These insights can inform the design of future methods for the generation and use of effective analogy-based explanations.

5.1 Introduction

5

In recent years, we have witnessed the rise of machine learning (ML) methods for various applications (*e.g.,* machine translation and object detection). Despite their high accuracy, more and more researchers recognize the necessity to obtain meaningful explanations of these ML methods for real-world scenarios, especially in high-stakes scenarios like medical diagnosis. Machine learning models may provide unreliable predictions based on spurious patterns (*e.g.,* Tesla's self-driving system mistook the moon for a yellow traffic light[1]), which may cause catastrophic consequences [184]. With meaningful explanations, humans can better understand the internal working mechanisms and exercise control over powerful machine learning models. With this perspective, a growing number of explainable artificial intelligence (XAI) methods are being proposed to provide explanations for ML model behaviors [159, 183, 185].

Identifying and communicating the salient parts of the input (*e.g.,* through pixels in image, or highlighted tokens in text) as explanations is a typical and model-agnostic XAI method [159, 163, 186], called feature attribution. While such salient parts of the input may be helpful for AI practitioners who have the relevant knowledge, it is still challenging for laypeople to interpret them. To provide more human-friendly explanations, Kim *et al.* [187] proposed to derive high-level concepts to describe the internal state of models. Compared with low-level salient features, high-level concepts have been shown to be more understandable for laypeople. However, in many real-world tasks, these high-level concepts (*e.g.,* chemicals, cells in medical diagnosis) are still not comprehensible for laypeople due to the gap of domain knowledge and expertise. At the same time, it is unnecessary for users or stakeholders (*e.g.,* patients or loan applicants taking medical or financial advice) to fully understand the explanation technically. Their information need is often satisfied by understanding explanations adequately enough to achieve better decision making for their own benefit.

The challenge, therefore, is to provide the right kind of explanations. Transparency about systems, and the provision of explanations, is likely to be a requirement in the AI

[1] https://www.autoweek.com/news/green-cars/a37114603/tesla-fsd-mistakes-moon-for-traffic-light/

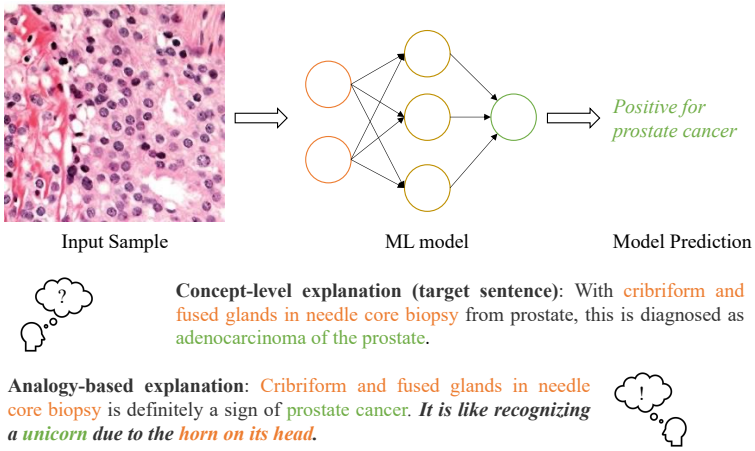


Figure 5.1: Example of analogy-based explanation in prostate cancer detection. The medical image and the concept-level explanation are sourced from [188].

Act [189] for a wide range of systems. Likewise, according to General Data Protection Regulation (GDPR),[2] the users of AI systems should have the right to access meaningful explanations of model predictions [166]. This implies that intelligible explanations which can facilitate such an understanding for laypeople are required. We argue that analogy-based explanations can be a potential solution to fill in this gap in understanding. We illustrate our motivation through an example in Figure 5.1. Given a concept-based explanation extracted from an ML model, laypeople may still have difficulties connecting the concepts (*i.e.,* cribriform and fused glands in needle core biopsy) with specific model predictions (*i.e.,* positive for prostate cancer). Such explanations can be difficult to understand due to the lack of domain knowledge and expertise, and they can be a heavy burden when figuring out the causality or relevance of observing these concepts to make the prediction [51, 190, 191].

An analogy can be interpreted as a structural mapping from a target domain to be clarified, onto a source domain which the recipient of the analogy is more familiar with [58, 59]. For example, in Figure 5.1, the target domain, *medical diagnosis*, is clarified based on a source domain: *fantasy*. Through everyday experiences, laypeople master commonsense knowledge of the world and build up sophisticated mental models to deal with regular tasks; *e.g.,* a single horn on the head of a beast is an important pattern for recognizing a unicorn. With analogy-based explanations, high-level concepts and model predictions can be translated into everyday concepts that laypeople are familiar with, by triggering their capabilities of analogical inference. From this standpoint, we argue that laypeople can leverage the sophisticated mental models of their worldly experiences to interpret the behavior of ML models and generate meaningful analogy-based explanations. Thus, users can understand that the complex concepts in "*cribriform and fused glands in needle core biopsy*" are also a strong pattern which indicates the model prediction "*positive*

[2] https://gdpr-info.eu/

for prostate cancer." Laypeople (or non-expert users) can thereby use the explanation adequately enough to inform their decisions, without having to understand the concepts from a technical standpoint, addressing the knowledge gap while reducing their cognitive load.

Despite the intuitive promise and potential of analogy-based explanations, how to generate such analogy-based explanations remains an open question. In addition, we also lack a framework to qualitatively characterize and evaluate the generated analogies. Hence, in this chapter, we first address the following research questions:

> (**RQ1**) *How can we generate high-quality analogy-based explanations using non-experts*?
> (**RQ2**) *How can we systematically assess the quality of analogy-based explanations?*

To the best of our knowledge, no work has yet investigated whether conceptually high-quality, analogy-based, explanations can be helpful for human-AI collaborative decision making. Inspired by recent literature on human-centered explainable AI [27, 126], a human-grounded evaluation [183] can further our understanding of the impact of analogy-based explanations in decision support. Hence, as a second step of our work beyond generating analogy-based explanations and evaluating their *conceptual quality*, it is also important to validate their effectiveness in assisting human decision making *in practice*. To this end, we aim to address the following questions:

> (**RQ3**): *How do analogies for concept-level explanations shape the understanding of an AI system among non-expert users*?
> (**RQ4**): *How do analogy-based explanations affect user reliance on AI systems*?

To answer **RQ1**, we designed a novel analogy generation method that leverages templates and crowd computing to obtain high-quality analogy-based explanations. To answer **RQ2**, we first defined a structured set of dimensions through which one can conceptually assess the quality of analogy-based explanations. Then we recruited 100 crowd workers as non-experts to generate analogy-based explanations using our method. After that, we carried out an expert evaluation of the quality of the collected explanations across the different dimensions. To answer **RQ3** and **RQ4**, we formulated four hypotheses about the effect of the analogy-based explanations on user understanding, appropriate reliance, cognitive load, and decision making efficiency. We tested these hypotheses in an empirical study with crowd workers (N = 280), asked to perform a skin cancer detection task, in four different human-AI collaborative decision making settings.

In our empirical study, we found that the mere presence of the target domain information within the analogy-based explanations was most effective in mitigating under-reliance but also gave rise to over-reliance. However, we did not find an improved understanding of the AI system or a statistically significant increase in appropriate reliance when all the information contained in the analogy-based explanations was presented. This was particularly the case when analogies were provided on demand. Surprisingly, such analogy-based explanations could even have some negative impact on the appropriate re-

liance. Analyzing the participants' qualitative feedback about the analogy-based explanations helped us understand the unexpected reliance patterns (*i.e.,* over-reliance and under-reliance) and the potential role of human intuition and plausibility in shaping our findings. Introducing analogies did not pose a significantly higher cognitive load on users, or cause a significant delay in decision making efficiency. Collectively, our findings suggest that although analogies may not be universally effective in fostering appropriate reliance in the context of human-AI decision making, there is some potential for analogy-based explanations in assisting laypeople for efficient decision making if they can be personalized. Our main contributions can be summarized as follows:

- A novel analogy-based explanation generation method with non-expert crowds and a dataset of analogies generated using this method.
- An elaborate set of qualitative dimensions to assess the quality of analogy-based explanations.
- An extensive evaluation of the quality of the analogy-based explanations collected from two distinct AI tasks.
- A rigorous empirical study in the context of human-AI decision making to understand the effectiveness of analogy-based explanations in a skin cancer detection task.
- Guidelines for the generation of effective analogy-based explanations and for the appropriate use of such analogy-based explanations.

Note that this manuscript is an extended version of the paper [98], extended in the following ways: To validate the effectiveness of analogy-based explanations, (1) we proposed new research questions and hypotheses about the impact of analogy-based explanations on a user's understanding of an AI system and their appropriate reliance on the system; and (2) we conducted an empirical study of human-AI decision making on a skin cancer detection task to test these hypotheses; (3) based on the results from our empirical study, we synthesized guidelines for future work on the generation and use of analogy-based explanations in the context of human-AI decision making.

If not used appropriately, analogy-based explanations may not work as expected to improve human-AI collaborative decision making. To the best of our knowledge, this is the first work that combines analogy-based explanations with commonsense knowledge in the context of human-centered explainable AI. Based on the results from our empirical study, we synthesize promising future directions for further XAI research.

5.2 Background and Related Work

We position our work in the following realms of related literature: *commonsense knowledge*, *analogy-based explanation*, *human-AI decision making* and the context of *human-centered explainable AI.*

5.2.1 Commonsense Knowledge

Commonsense knowledge is "information that humans typically have that helps them make sense of everyday situations" [192]. It has been proved to be highly useful in various AI applications, like question answering [193], dialogue systems [194] and visual reasoning [195]. However, due to the intrinsic implicitness, commonsense knowledge is usually omitted in oral or written communication [192]. To collect such implicit knowledge, re-

searchers have proposed to make use of the wisdom of crowds, through text mining of corpora [196, 197], and via games with a purpose [198, 199].

In recent years, commonsense knowledge has been used to also improve the explainability of AI models. In commonsense reasoning tasks, explanations from humans which contain rich commonsense knowledge, have been shown to be highly useful both to boost performance and to aid understanding [200]. In addition to generating commonsense explanations with humans, some studies have also demonstrated that commonsense knowledge can help build connections between multiple statements [201] and enhance natural language explanation generation with extractive rationales [202].

To facilitate the understanding of concept-level explanations, we propose to generate commonsense explanations for laypeople. The commonsense knowledge contained within such explanations forms the source domain over which laypeople can exercise their analogical reasoning, to improve their understanding of the concept-level explanations.

5.2.2 Analogy-based Explanations

Analogy-based explanations have been extensively studied in many research domains such as logic, linguistics, and philosophy. "An analogy is created when some aspects of an unknown target are compared with those of a source about which more is known" [203]. Due to such intrinsic property for elucidating new knowledge with existing knowledge, analogies have been adopted as explanation in education, and supported by multiple research work [204–206].

In the context of artificial intelligence, the importance of analogies has been recognized by multiple AI applications such as representation learning [207], preference learning [208], and image processing [209]. Readers can refer to [210] for a more comprehensive survey of analogical inference in the context of AI, which is beyond the scope of this chapter. However, only a few works [51, 211] explored the potential of analogy-based explanations in the context of XAI. While such works show and argue that analogy-based explanations have great potential in XAI, it is still unclear how we can measure the quality of analogy-based explanations and how we can efficiently generate such analogy-based explanations for machine learning applications.

As for analogy generation, in addition to previous methods that relied on human intelligence for drawing out analogies in instructional, teaching and educational contexts [212, 213], some research has also explored the automatic generation of analogies. Veale *et al.* [214] explored how lexical resource HowNet [215] can support analogy generation with two approaches: (1) abstraction via a taxonomic backbone, (2) selective projection via structure-mapping on propositional content. Chiu *et al.* [216] propose to generate lexical analogies with the help of dependency relations from unstructured text data. However, such methods do not incorporate commonsense knowledge, making it inappropriate for explaining to laypeople the complex concept-level explanations. That is why we adopt a crowd computing-based method to generate analogy-based explanations.

In this chapter, we propose structured dimensions for the qualitative assessment of analogy-based explanations. We also design a crowd computing method to generate such explanations, and empirically evaluate its effectiveness.

5.2.3 Human-Centered XAI and the Human-AI Decision Making

Explainability is a concern for AI systems, especially for black box deep learning models. To provide meaningful explanations for AI predictions, a wide range of explainable artificial intelligence (XAI) tools have been proposed [217]. However, due to the inherent human-centric property of explainability (*i.e.,* explanations are only successful if they match the specific needs of the person receiving them), there is no one-size-fits-all solution in the growing collection of XAI techniques [27]. Consequently, researchers have increasingly begun to explore the area of human-centered explainable artificial intelligence (HCXAI) [27, 104, 126, 127], by putting the human at the center of technology design [126].

Human-AI decision making has emerged as an important paradigm to augment human capabilities with the computational prowess of AI systems, leading to complementary teamwork and effective decision making [22]. In the collaborative decision making process, human factors (*e.g.,* AI literacy [108] and cognitive bias [218]) and interaction with AI systems (tutorial intervention [30, 117] and performance feedback [49]) are observed to affect subjective trust and reliance behaviors greatly. In recent works with human-AI decision making, researchers have shown great interest in achieving complementary team performance with appropriate reliance on the AI system by exploring a multitude of factors including human and task factors [29, 30, 219, 220].

To help users better understand AI advice and inform the trustworthiness, XAI methods are widely analyzed in human-AI decision making. Based on a comprehensive literature review, Wang *et al.* [107] summarized three desiderata of AI explanations to facilitate complementary teamwork: (1) Explanations of an AI should improve people's understanding of it, (2) Explanations of an AI should help people recognize the uncertainty underlying the AI, and rely on the high-confidence predictions when model confidence is calibrated, (3) Explanations of an AI should empower people to trust the AI appropriately. However, most XAI methods are rarely found helpful in achieving a complementary performance in human-AI decision making [25, 70, 221]. Sometimes, XAI methods can even make users suffer from automation bias [222], which will cause over-reliance on the AI system.

AI systems have become ubiquitous in intelligent applications around our daily life, and involve nearly everyone as stakeholder rather than experts only. Different communities of stakeholders [223] have different goals and explainability needs. For example, system developers require explainability to debug the system, while system users may place more emphasis on the explainability of outputs in order to aid their own decision making [223, 224]. As a result, explanations should be tailored to different stakeholders.

Inspired by previous studies about analogy-based explanations [51, 211], we focus on explainability for laypeople using such explanations:

- Laypeople lack technical expertise and domain knowledge to interpret AI systems. Analogy-based explanations fill in such knowledge gap with concepts they are familiar with.
- Analogy-based explanations provide familiar information for laypeople, which reduces the cognitive load for comprehension compared to concept-level explanations which contain uncommon terminologies.

5.3 Quality of Analogy-based Explanations

We first conducted a systematic review of existing works in the area of analogy-based explanations, in order to understand how the quality of analogy-based explanations has

been empirically investigated in prior literature.

5.3.1 Effective Analogies

Properties of analogical argument. Analogies have been widely used as explanations for educational and learning purposes [204, 206]. With analogical inference, humans can compare one new topic that is being introduced with another topic they are already familiar with, which leads to a better understanding of the new topic by relating back to previous knowledge [225]. However, to make the analogy-based explanations work as an aid to understand new knowledge or events, several properties need to be satisfied by the analogical arguments. Aristotle's theory provides us with four important and influential criteria for the evaluation of analogical arguments [226]:

- The strength of an analogy depends upon the number of similarities.
- Similarity reduces to identical properties and relations.
- Good analogies derive from underlying common causes or general laws.
- A good analogical argument need not pre-suppose acquaintance with the underlying universal (generalization).

In previous studies, researchers also emphasized the importance of the quality of structural mapping. According to [58, 203], an analogy needs to fullfill certain constraints to work as expected – (i) there should only be a single one-to-one correspondence between each pair of elements; (ii) it must involve common relationships across the source domain and target domain (iii) an analogy must describe systems of connected relations, which permits the generation of inferences. According to the multiconstraint theory [227], people use analogies guided by a series of constraints that favour coherence in analogical reasoning [206]. The constraints are semantic similarity, structural correspondence, and purpose. Specifically, the similarity in concept level contributes to analogical reasoning, while the structural constraint helps to establish an isomorphism between source domain and target domain. Furthermore, the analogical reasoning is guided by the purpose. In addition to ensuring the analogical properties of the structural mapping, Thalheim *et al.* [228] further considered the "degree of structural adjustment" (*i.e.,* the extent to which the structure is considered independent on the later use). This dimension evaluates the *transferability* of the generated source artifact.

Factors shaping the effectiveness of analogies. Apart from the properties of analogical argument, there are other factors which affect the effectiveness of analogy-based explanations. To guarantee the usefulness of analogy-based explanations, explanation consumers should be familiar with the source domain (*e.g.,* the generated commonsense explanations in our case). According to Galesic *et al.* [76], the most helpful analogies boast a high relational similarity between the source and target domain and a high familiarity with the source domain. Thalheim *et al.* [228] also argued that the source domain of effective analogies should be "easily interpretable and understandable".

5.3.2 Synthesizing a Structured Set of Dimensions

Analogical Properties. According to the above, the quality of generated analogy-based explanations is largely reflected by the quality of the analogical properties, that rely on comparing the source domain (*i.e.,* generated commonsense explanation) to the target

sentence. In this chapter, we base the quality of analogical properties on four aspects: (1) **structural correspondence** between the target domain (*i.e.,* observed concepts and model prediction) and source domain (*i.e.,* concepts used in the explanation), (2) **relational similarity** between the target domain (*i.e.,* relation between observed concepts and model prediction) and source domain (*i.e.,* relation between concepts in explanation), (3) **transferability**, *i.e.,* the extent to which the structure is considered independent of its later use, and (4) **helpfulness**, *i.e.,* the extent to which the generated commonsense explanation is considered helpful to understand the target sentence.

Among these dimensions, "relational similarity" and "structural correspondence" have been highlighted by existing works with phrases like "semantic similarity" [227] and "structural alignment" [229]. "Helpfulness" corresponds to the "purpose" mentioned in Holyoak and Thagard's multiconstraint theory [227], while "transferability" corresponds to the "degree of structural adjustment" [228]. To assess the "helpfulness" of explanations, we need to ground them within specific tasks. In this chapter, we conduct human-based evaluation to assess the extent to which the analogy-based explanations can be helpful to explain the original concept-level explanations. In practice, the generated analogy-based explanation may also be fit to explain other concept-level explanations which show similar information. To serve that purpose, one can argue that high-quality analogy-based explanations should be capable of generalizing to more tasks. Thus, we also consider the "transferability" of generated analogy-based explanations.

As mentioned above, the generated analogy-based explanations can be used to explain other tasks than the one used for generation. In such cases, it is also necessary to evaluate the quality of the explanations. All the dimensions we propose can be used to assess such quality for these new tasks.

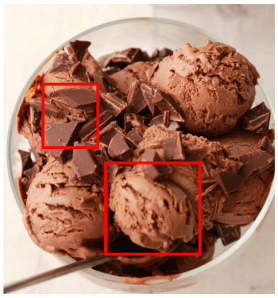
Utility. In addition to the above dimensions, we identified dimensions specifically related to the generated commonsense explanations. These dimensions are independent of the target sentence, but may also affect the effectiveness of analogy-based explanations.

Some dimensions are identified from the factors shaping the effectiveness of analogies mentioned previously. They are: (5) explainee's **familiarity** with the concepts mentioned in generated explanation; (6) **simplicity** of the analogy-based explanation, which describes how easily laypeople can interpret and understand the explanation would be [228]. We also identify other dimensions based on intuitively desirable expectations from effective explanations. Reducing the scope for misunderstanding can aid the overall comprehension of analogy-based explanations. Thus, we also consider the dimension of (7) **misunderstanding**, which occurs when different interpretations exist for a single analogy-based explanations. For example, the phrase "*subway definitely contains seats*" can be interpreted as referring to *e.g.,* either the restaurant, "Subway", or an underground railway. To ensure the utility of generated explanations, it is vital to ensure that they are (8) **syntactically correct**, and (9) **factually correct**. That means the explanations are comprehensible according to syntactic grammar, and describe the truth about the world. Further details including our annotation of these dimensions are provided in section 5.5.

5.4 Analogy Generation

We propose a crowd computing method to generate analogy-based explanations using image classification tasks as an empirical lens, and verify the effectiveness of our proposed set of dimensions in determining the quality of the analogy-based explanations.

Tasks for Analogy Generation. To collect useful analogy-based explanations from crowd workers, we need to adopt task contexts which non-experts are capable of interpreting and explaining. We also consider the relationship explicitness in the task domain. In some domains, it is difficult to elucidate relationships between concepts and labels other than ascribing correlation (*e.g.,* food to calorie level). In others (such as furniture to places), most concepts and the labels have a clear indication of relationships like "PartOf", "SignOf", and "FoundAt", which also appear in commonsense knowledge bases like ConceptNet [197]. Hence, we select two image classification tasks: calorie level classification (CLC) and scene classification (SC).



(a) Calorie dataset.

(b) Places dataset.

Figure 5.2: Examples of tasks used to generate analogies.

For the calorie level classification task, we used the dataset provided by Buçinca *et al.* [68], where two possible labels are attached to images: (1) *high calorie level*, fat more than 30%, (2) *low calorie level*, otherwise. In this task, participants are given an image (see Figure 5.2a) along with concepts highlighted with bounding boxes (*i.e.,* chocolate and ice cream) and the predicted calorie level. For the scene classification task, we used a subset of the Places dataset [230], which covers six place labels: *living room*, *bathroom*, *hospital room*, *conference room*, *bedroom*, *dining room* (Figure 5.2b is an example of a *conference room*). In both tasks, we ask participants to describe the relevance of given concept(s) and labels, *e.g.,* the relevance of food concept(s) and calorie levels, with explanations constructed using everyday concepts and given templates.

Templates for Analogy-based Explanations. To help crowd workers associate the concepts with model predictions, we provide templates for generating analogy-based explanations. Machine learning models may learn both useful concepts and spurious concepts to make predictions [187]. Some of the useful concepts can directly lead to the correct conclusion, while others are highly relevant and helpful to predict the label but not definite. In comparison, the spurious concepts are irrelevant or insufficient (like predicting a *dog* in image by focusing on *grass field*) to make the prediction, and sometimes even contradict

Table 5.1: Templates used in *analogy generation* with placeholders presented to the users (bold text in square brackets).

Relevance	Template	Example
Positive Evidence	Definite Sign Of	Mayonnaise is definitely a sign of high calorie food. This is like a **[trunk]** is a definitely sign of **[an animal being an elephant]**.
	Typically Associated with	Chocolate is typically associated with high calorie food, while rarely associated with low calorie food. This is like **[printers]** can typically be associated with **[offices]**, but it's also possible to associate **[printers]** with **[homes]**.
Inconclusive Evidence	Insufficient	Bread is not sufficient to indicate high calorie, as both high calorie food and low calorie food may contain it. This is similar to how we can find **[chair]** in both **[a living room]** and **[a bedroom]**, you can't determine which room it is by seeing a **[chair]**.
	Irrelevant	A plate is irrelevant to indicate high calorie food. This is similar to to how **[an arbitrary stone]** is irrelevant for **[recognising a continent]**.
Negative Evidence	Seldom Found At	Carrots are seldom found in high calorie food. This is like **[cats]** can seldom be found in **[water]**.
	Contradict With	A vegetable salad contradicts with high calorie food. This is similar to how one cannot find **[water]** in **[electrical appliances]**.

with our commonsense knowledge, leading to an incorrect prediction. Hence, we decide to use six templates based on three different relevance levels (*i.e.,* positive evidence, inconclusive evidence, and negative evidence). For each relevance level, we have one template to indicate the type of relationship and another one to indicate relevance. The templates along with examples can be found in Table 5.1.

Task Selection. To balance the generated analogies in each relevance category, we manually selected two tasks for each category according to the authors' interpretation of their relevance levels. Thus, we use 12 tasks for analogy generation: 6 for calorie level (CLC) and 6 for scene classification (SC).

Hints for Analogy Generation. Through a pilot study, we learned that although non-expert crowd workers can generate analogies based on their own experience, it becomes challenging to generate new analogies after a handful of tasks. To help crowd workers in generating high-quality analogies, we provide a list of hint domains with a clickable button in the interface. The list contains: weather, animals and plants, place, transportation, food, art, education, sports, finance, clothes, electronics, games and toys, health.

Analogy Generation Procedure. To generate high-quality analogies, we provide the six templates shown in Table 5.1 to each participant. Participants are first asked to select one template, comprising one sentence with placeholders for concepts. They can then refer to our example analogies and everyday domains provided as hints. Next, based on the template, they are asked to fill in one word or phrase (up to five words) as a concept in each placeholder. All participants are forbidden to fill in concepts belonging to the task domain (such as places and furniture in the Places task). An example of the analogy generation interface is shown in Figure 5.3.

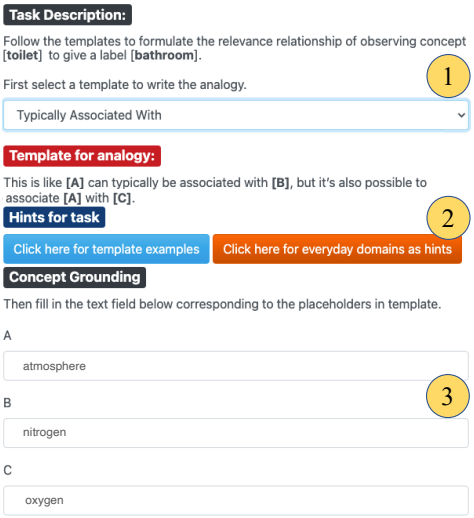


Figure 5.3: Analogy generation main interface and workflow. (1) Participants select a template to describe the relevance level; (2) refer to examples and everyday domains as hints; and (3) fill in concepts in placeholders to generate analogy.

5.5 Study I: Analogy Generation and Evaluation

In the first study, our experiment mainly consists of two stages: (1) analogy generation with crowd workers, (2) evaluation of generated analogies with third-party experts.

5.5.1 Analogy Generation Based on Non-experts

Pilot Study. We conducted a pilot study with 7 participants hired from Prolific[3] crowdsourcing platform. All participants were asked to complete 12 tasks (6 for CLC, 6 for SC). Through the pilot study, we gained the following insights:

- After generating several analogies, participants found it difficult to generate new analogies (*i.e.,* required more time for analogy generation and repeated concepts used). To help with this issue, we provided a list of daily domains as hints. As a consequence, we also reduced the number of tasks that each participant was required to complete in the analogy generation phase of the main study.
- Some participants used the examples or concepts shown in one task (*e.g.,* calorie) as answers for another one (*e.g.,* places). To counter such behavior, we decided to limit each participant to a single generation task.

Informed by these observations, we asked each participant in the main study to work on 6 analogy generation tasks from one task domain (either CLC or SC).

Participants. In the main study, we recruited 50 crowd workers for the calorie task, and 50 crowd workers for the places task. In total, 600 analogy-based explanations were generated. We compensated each worker with £1.35 (*i.e.,* 9 min × hourly salary £9). All

[3] https://www.prolific.com/

participants were proficient English-speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform.

Quality Control. To discourage unreliable behavior (*e.g.,* copy-pasting concepts from the task description and examples provided), we enforce all concepts mentioned in the task description and possible labels in each task as taboo phrases (words). We also prevent participants from generating the same analogy-based explanations twice.

Table 5.2: Structured dimensions used in qualitative assessment of analogy-based explanations.

Category	Dimension	Questionnaire	Scale
Analogical Properties	Structural Correspondence	How well can you align the properties of the explanation concepts to the properties of the concepts in the target sentence?	5-point Likert
	Relational Similarity	How similar do you perceive the relationship between concepts in the explanation and the relationship between concepts in the target sentence?	5-point Likert
	Transferability	How well can the explanation be used in other contexts?	5-point Likert
	Helpfulness	How helpful is this explanation for you to understand the target sentence?	5-point Likert
Utility	Familiarity	How familiar are you with the concepts in the explanation?	5-point Likert
	Simplicity	Do you think the explanation is simple enough for others to understand?	5-point Likert
	Misunderstanding	Do you think this explanation lead to more than single interpretation?	{Yes, No}
	Syntactic Correctness	Whether the analogy sentence is syntactically correct?	{Yes, No}
	Factual Correctness	Whether it describes a fact about real world? Can we switch it to make it factual? (switch concept A and concept B in template)	{Yes w/o switch, Yes & switch, No}

5.5.2 Analogy Evaluation with Experts

Experts. To ensure a fair evaluation of the quality of generated analogies, we recruited 5 external AI experts from the department of the authors' institute using a purposeful sampling strategy [231]. All experts had at least a basic knowledge of machine learning and explainable AI.

For the purpose of this evaluation, we considered a subset of the analogies generated from 23 participants in the calorie task and 26 participants in the place task (we randomly sampled around half of the participants in our study). In total, we consider 294 analogy-based explanations for evaluation. We ensured a 10% (*i.e.,* 29 analogy-based explanations) overlap across experts. Thus, each expert evaluated 82 different analogy-based explanations. On average, each expert spent 2.5 hours on this qualitative evaluation.

Qualitative Assessment. Based on our synthesis of the dimensions for quality of analogies (cf. Section 5.3.2), the quality of analogy-based explanations was mainly assessed across two categories: (1) analogical properties and (2) utility. We followed an iterative coding process [232] to characterize the quality of the analogy-based explanations across dimensions informed by our synthesis from literature. While different terminologies (*e.g.,* degree of structural parallelism [226], degree of structural analogy [228], semantic similarity [227]) were adopted to assess the quality of analogies and their quality as explanations, we aimed to address the redundant definitions and integrate a structured

set of dimensions for the qualitative assessment (see dimension and questionnaire in Table 5.2).

Annotation Rubrics. Through iterative coding interspersed with discussions, the authors finally constructed the following annotation rules to guide the qualitative assessment:

- If the concepts of commonsense explanation are of the same domain as the target sentence (regarded as invalid due to non-compliance with analogy generation instruction), annotators can skip that annotation.
- For *Factual Correctness*, take the generated explanation "*The pink feather is definitely a sign of flamingo*" as an example. This explanation can be factually correct after we switch the order of "pink feather" and "flamingo".
- When *Misunderstanding* exists, we consider one analogy as factually correct when a single interpretation can be true. For example, "subway is definitely a sign of seat". When interpreting the "subway" as the one in transportation, we can consider it as being factually correct.
- For *Transferability* and *Helpfulness*, assign '1' when *Factual Correctness* = No
- We devised additional, concrete rubrics for each of the other dimensions. While we do not present them here for space consideration, they can be found online.[4]

Procedure. In the beginning, we provided an annotation manual for each expert. They spent around 10 minutes on reading the annotation manual which contains both dimensions and annotation rules we mentioned above. In this process, we also answered their questions to clarify any issues related to quality evaluation. After that, each expert independently worked on the 82 samples provided according to the rubric we provided.

Annotation Agreement. We calculated the annotation agreement based on 29 samples (overlap for experts) in evaluation experiment. As 7 analogy-based explanations are recognized as invalid (crowd workers generate the explanation with concepts via the same domain as target sentence), we calculated the Krippendorff's α scores based on the valid 22 analogy-based explanations. Due to the subjectivity in evaluating the dimensions in the 5-point Likert scales, we merge the 5 items into three levels of attitude (*i.e.,* Negative={1, 2}; Neutral={3}; Positive={4,5}) when calculating the Krippendorff's α scores. The results are respectively 0.15 for *Structural Correspondence*, 0.17 for *Relational Similarity*, 0.22 for *Factual Correctness*, 0.64 for *Syntactic Correctness*, 0.35 for *Misunderstanding*, 0.03 for *Familiarity*, 0.14 for *Helpfulness*, 0.11 for *Transferability*, and 0.14 for *Simplicity*. Naturally, the experts show relatively higher agreement on *Factual Correctness*, *Syntactic Correctness*, and *Misunderstanding*, which are more objective than the other dimensions. The disagreement on other dimensions is due to the subjectivity of the task [233]: knowledge and the quality of an analogy-based explanation vary depending on one's own experience of the world.

For further illustrative analysis, let us consider an example analogy-based explanation which received disagreement among experts on most dimensions — "*Lemon is seldom found in high calorie food. This is similar to how having hair is irrelevant for recognising a human*". All experts see this analogy-based explanation as factually correct and syntactically

[4] https://github.com/delftcrowd/HCOMP2022_ARCHIE/blob/main/annotation_manual/annotation_manual.pdf

Table 5.3: Evaluation of the following analogy by 5 experts illustrating disagreement – "*Lemon is seldom found in high calorie food. This is similar to how having hair is irrelevant for recognising a human.*"

Dimension	E_1	E_2	E_3	E_4	E_5
Structural Correspondence	4	3	5	1	2
Relational Similarity	1	1	5	1	3
Familiarity	4	5	5	5	2
Helpfulness	1	5	5	1	2
Transferability	4	5	5	1	2
Simplicity	3	5	5	2	3

correct without any misunderstanding. As the experts assessment reveals in Table 5.3, the experts diverge on most dimensions of the Likert scale.

For further insights in the disagreement, we ask the experts to explain their scoring. We find multiple user factors can lead to disagreement. For instance, we observed that: (i) The overall negative attitude of E_4 ("*I just gave it a low number because I didn't really understand what it was trying to tell me*") towards this explanation, and the severity of E_5 make them rate most dimensions lower. (ii) As the relationship between "lemon" and "high calorie" is not explicit, experts seem to have different interpretation of the relationship, leading to disagreement on *Relational Similarity*. While E_1, E_2, E_5 would rate it low, E_3 judge it high, because *"calorie is a common property of food, which is not unique to Lemon. having hair is also a common (mostly) property of humans, which is not unique to a specific person"*. (iii) Some experts have more abstract thinking on the properties and relations, again causing disagreement. E_1 gives a 4 to *Structural Correspondence* because they think "human" and "high calorie" have some connections. And E_2 would rate *Relational Similarity* as 1 because "*people have hair, lemon are not high calorie food*". Besides, we also notice that both E_1 and E_5 take this explanation as unhelpful due to poor *Relational Similarity*.

5.5.3 Results and Analysis

In this subsection, we present the quality assessment results for the generated analogies with the proposed approach.

Descriptive Statistics

In the analogy generation experiment, crowd workers are asked to generate explanations with concepts in a different domain from the target sentence. The generated analogies that violate this requirement are then regarded as being invalid. Among the 294 generated analogy-based explanations, 255 (nearly 87%) were recognized as valid by all five experts. As the annotation rubric described, experts only provide qualitative evaluation for valid analogy-based explanations. Finally, we gathered 358 valid evaluation results for 410 samples (82×5, with 29 samples overlap for each).

When generating the analogy-based explanations, crowd workers used everyday concepts in domains "Animals", "Scene/Place", and "Weather" most frequently, which are also in the hint list we provide. For the identified relationship between concepts in generated analogy, crowd workers prefer to use "FoundAt" (175 times), "SignOf" (158 times), and "PartOf" (24 times).

Dimension	Label	Example
Structural Correspondence	1	Chocolate and cream contradict with low calorie food. This is similar to how one cannot find tsumanis in uk.
	3	Nuts is insufficient to indicate high calorie. This is similar to how we can find hairdryer in both hotel and hairdresser, you can't determine where it is if you see hairdryer.
	5	A medical monitor is a definite sign of hospital room. This is like an echocardiogram is definitely a sign of pulse oximeter.
Relational Similarity	1	Nuts are seldom found in high calorie food. This is similar to how one cannot find fire hydrants in boats.
	3	Fireplace is not sufficient to indicate bedroom. This is similar to how we can find wig in both pantomime and courtroom, you can't determine where it is if you see wig.
	5	A medical monitor is a definite sign of hospital room. This is like doctor is definitely a sign of surgery.
Transferability	1	A fireplace is a definite sign of bedroom. This is like art is definitely a sign of human expression.
	3	Beet and apple contradict with high calorie food. This is similar to how one cannot find toys in a clothes store.
	5	Chocolate and ice cream is a definite sign of being high-calorie. This is like keyboard is definitely a sign of having a computer.
Helpfulness	1	Toothbrush and towel are insufficient to recognize a bathroom. This is similar to how we can find reading in both education and hobby.
	3	Chocolate and cream are definitely a sign of high calorie food. This is like udders are definitely a sign of cow.
	5	A fireplace can seldom be found in a bedroom. This is like dogs can seldom be found in a fishtank.
Familiarity	1	Chocolate and cream contradict with low calorie food. This is similar to how one cannot find bargains in harrods.
	3	Chocolate and cream are seldom found in low calorie food. This is like roar can seldom be found in big animal.
	5	Nuts is not sufficient to indicate high calorie food. This is similar to how we can find books in both libraries and schools, you can't determine where it is if you see books.
Simplicity	1	Carrot is not sufficient to indicate high calorie. This is like diets can typically be associated with field of hay, but it's also possible to associate diets with gemstones in a gold mine.
	3	Table and chair is insufficient to indicate a conference room. This is like atmosphere can typically be associated with nitrogen, but it's also possible to associate atmosphere with oxygen.
	5	Chocolate and ice-cream are a definite sign of high-calorie. This is like duvet is definitely a sign of bed.

Table 5.4: Examples of analogies generated for the different scale items of each dimension of the qualitative analysis.

Analogy quality. Among 358 valid evaluation results, 310 cases were found to be syntactically correct, 198 cases were factually correct without switching placeholder A and B, 49 cases are factually correct with switching (in total, 79.7% of explanations could be generated as factually correct). Meanwhile, only 53 cases were found to potentially lead to multiple interpretations. We compare the quality of analogy-based explanations based on the category of *Factual Correctness*. As shown in Figure 5.4, the factually correct analogy-based explanations show better quality in nearly all dimensions in 5 point Likert scale than factually incorrect counterparts. As factually incorrect analogies would not be taken as effective explanations for humans, we only report qualitative results on the factually correct ones in the following analysis.

The distribution of dimensions in 5-point Likert scale can be visualized with the boxplots in Figure 5.5. Overall, the generated analogies show good quality in most qualitative dimensions except *Structural Correspondence* and *Relational Similarity*. The experts consider that the analogies are easy to understand and involve familiar everyday concepts, which indicates these explanations are of relatively low cognitive load. To be concrete about how the explanations differ in quality, we show examples of scoring 1, 3, 5 for dimensions in 5 point Likert scale in Table 5.4. Note that we do not expand on examples for *Factual Correctness*, *Syntactic Correctness*, and *Misunderstanding*, which are trivial.

To further investigate how qualitative dimensions affect the perceived helpfulness of

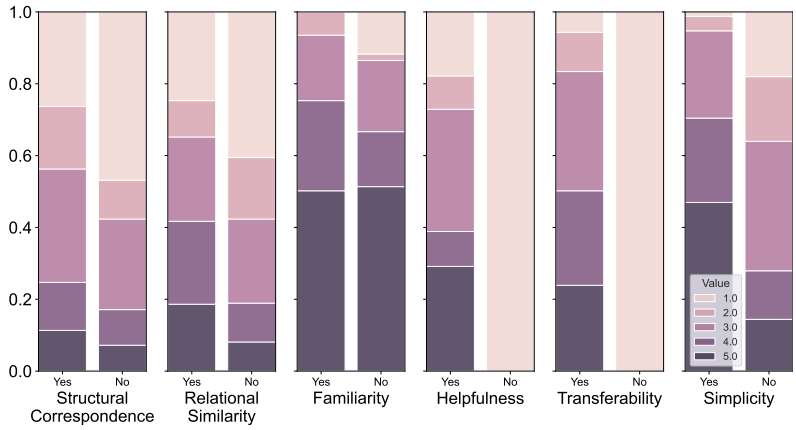


Figure 5.4: Stacked histogram illustrating the difference across the qualitative dimensions based on Factual Correctness. All dimensions were measured on a 5-point Likert scale.

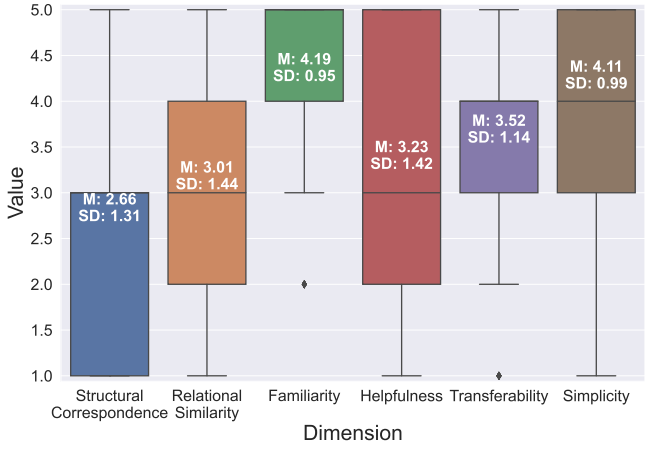


Figure 5.5: Box plot illustrating the distribution of the different dimensions considered in our study. All dimensions were measured on a 5-point Likert scale. For all dimensions, 1 indicates a poor quality while 5 indicates a good quality. *M* and *SD* represent mean and standard deviation respectively.

analogy-based explanations, we calculated Spearman rank-order correlation coefficients between *Helpfulness* and the other Likert-based dimensions. We found a significant positive correlation between all dimensions and *Helpfulness*: *Structural Correspondence*, $r(247) = 0.191$, $p = 0.003$; *Relational Similarity*, $r(247) = 0.374$, $p = 0.000$; *Familiarity*, $r(247) = 0.312$, $p = 0.000$; *Transferability*, $r(247) = 0.445$, $p = 0.000$; *Simplicity*, $r(247) = 0.467$, $p = 0.000$. This confirms that our qualitative dimensions are substantially indicative of their perceived helpfulness. Our findings suggest that if we ensure the generated explanations are of high quality across these dimensions, they have a higher likelihood of being helpful in understanding the target sentence.

Comparison between Different Tasks

Among 410 annotations, 174 cases are generated from calorie level classification (CLC) task, while 236 cases are generated from scene classification (SC) task. According to the results, 109 and 138 cases are identified as both valid and factually correct for CLC and SC tasks, respectively. We compared the difference between the quality of analogies generated with the calorie task and places task. We found a significant difference ($\alpha = 0.05$) on the assessed *Relational Similarity* ($H(1) = 7.54$, $p = 0.006$) with a Kruskal-Wallis H-test. Post-hoc Mann-Whitney tests further show that the *Relational Similarity* of analogy-based explanations generated from SC task is significantly better than the counterparts from CLC task. However, no significant difference exists in the other qualitative dimensions.

The reason for such a phenomenon may be that the relationship between "concept" and "label" in the SC task is more explicit than in the CLC task. This may make it easier for participants to generate analogy-based explanations while keeping similar relationships. However, such good analogical properties do not translate to higher perceived *Helpfulness*. This indicates that the interplay between qualitative dimensions and perceived helpfulness may be complex. Better quality on a single dimension (*Relational Similarity* here) may not necessarily lead to a better understanding.

5.6 Study II: Effectiveness of Analogy-based Explanations in Medical Diagnosis

Our first study showed that our proposed method can generate conceptually high-quality analogy-based explanations when non-expert workers are involved in the collection process. Besides evaluating analogy-based explanations with qualitative dimensions, it is also important to check how effective they are when assisting users in decision making in practice. Thus, we conducted an empirical study of human-AI decision making in medical analysis. In this section, we first present our hypotheses and experimental setup, which had all been preregistered before any data collection.[5] Then, we show the experimental results. Finally, we discuss the findings and implications of this study. This study was approved by the human research ethics committee of our institution.

5.6.1 Hypotheses

It is still unknown how analogies will affect user understanding of concept-level explanations and how analogy-based explanations affect user reliance on AI systems. Based on our

[5] https://osf.io/jm3ap

findings from Study I and findings from existing work [204–206], analogies have proven effective in aiding users in understanding new knowledge. Little has been done to build an empirical understanding of the effectiveness of analogies in real-world decision-making tasks where concept-level explanations are employed [234]. Addressing this research gap, we hypothesize that analogies can help users better understand AI systems, and that such an improved understanding will further help users rely on AI systems more appropriately.

H1: Using analogy-based explanations can help users better understand AI systems, compared to conventional concept-based explanations.
H2: Using analogy-based explanations can facilitate appropriate reliance on AI systems, compared to conventional concept-based explanations.

Analogies have proven to be effective in helping humans understand new knowledge and reduce the cognitive load for learning new knowledge [235]. While analogies can help improve users' understanding, the additional analogical inference requires more effort, which may be time-consuming. Therefore, we hypothesize that users can maintain a similar team performance and be more efficient in their decision making when engaging with analogy-based explanations when they deem it to be necessary (*i.e.,* on demand).

H3: Analogy-based explanations can reduce the perceived cognitive load of users in their decision making process.
H4: Providing analogy-based explanations on demand can improve users' efficiency in their decision making process.

5.6.2 Task

In our study, we selected a real-world medical diagnosis scenario — skin cancer detection based on skin lesions as a test bed to verify the effectiveness of analogy-based explanations in human-AI decision making. All task data are selected from the HAM10000 [236] dataset. In this task, given an image of a pigmented skin lesion, users are asked to decide whether the shown image depicts a '`malignant`' or '`benign`' skin lesion. The rationale for selecting the skin cancer detection task is three-fold: (1) This is a realistic scenario for human-AI collaboration, where humans are designated to make final decisions due to accountability concerns. (2) Medical concepts in this task are relatively challenging for laypeople to digest, which fits our motivation of providing analogy-based commonsense explanations that can be leveraged and used to communicate the explanations to laypeople. (3) There is a substantial need for AI assistance to help doctors and medical experts check increasingly large volumes of images. Thus, the setting we chose is realistic and aligned with real-world needs.

Medical Concepts. In our study, we followed Yuksekgonul *et al.* to adopt eight medical concepts to help users diagnose skin cancer based on their assessment of `malignant` versus `benign` skin lesions [237]. The eight concepts are: Blue-Whitish Veil, Regular Dots & Globules, Irregular Dots & Globules, Regression Structures, Irregular Streaks, Regular

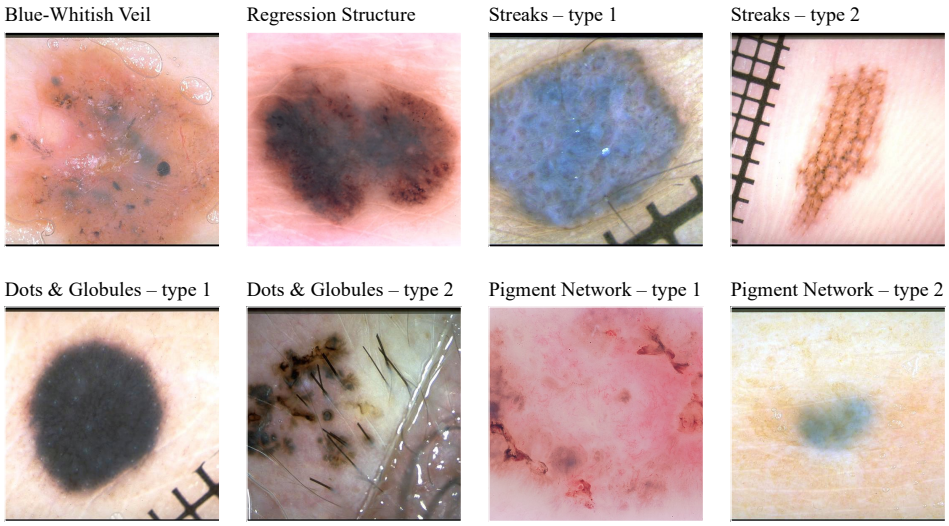


Figure 5.6: The overview of medical concepts shown to participants in Study II.

Streaks, Atypical Pigment Network, and Typical Pigment Network. Note that these concept names contain words like "Irregular" and "Atypical", which can clearly indicate their correlation to the model's prediction (*i.e.,* benign and malignant) — simplifying an otherwise complex decision making task. To test the learning effect potentially stemming from concept-based explanations, we replaced such hints with the abstractions of "type 1" and "type 2". In our study, we provided participants with an overview figure illustrating the eight different medical concepts to aid their decision making (shown in Fig 5.6). For each concept, we provided an image of an example skin lesion to swiftly illustrate the concept and help user understanding. To help participants remember and rely on these concepts along with concept-level explanations in their decision making, we provided a button (cf. Figure 5.9) below the concept-level explanations, that triggers a pop-up window containing the overview of medical concepts.

Selection of Tasks. To ensure diversity in the selected tasks and to cover the use of different medical concepts, we selected 14 tasks based on seven fine-grained categories in the HAM10000 dataset. To faithfully reflect the performance of the AI system used, we selected tasks based on performance of the post-hoc concept bottleneck model [237] on the HAM10000 dataset.

First, we generate model predictions on the validation set of the HAM10000 dataset (same split as [237]). Then, based on the performance of each category (shown in Table 5.5) and the sample size of each category, we selected 14 tasks (10 with correct predictions, 4 with wrong predictions). In our study, the accuracy of the AI system is 71.4% (10 / 14).

Pilot Study. To understand how capable non-expert crowd workers are in this task, we recruited 20 participants from Prolific. The Prolific platform has been shown to be a reliable source for participant recruitment in similar XAI studies over the last few years [62, 109, 234] and has a growing reputation as a suitable platform for human subjects research

Table 5.5: Descriptive statistics of the HAM10000 dataset and AI performance across the seven categories in the dataset.

Category	Label	#Tasks	Error Rate	Selected Task
Benign keratosis-like lesions	benign	220	9.1%	1 correct, 1 wrong
Dermatofibroma	benign	23	4.3%	2 correct
Melanoma	malignant	223	35.4%	1 correct, 1 wrong
Vascular lesions	benign	28	10.7%	2 correct
Basal cell carcinoma	malignant	103	28.2%	1 correct, 1 wrong
Melanocytic nevi	benign	1,341	2.9%	1 correct, 1 wrong
Actinic keratoses	benign	65	10.8%	2 correct

across different scientific domains [238, 239]. Each participant in our study received 2 GBP (8 GBP per hour [6]) for working on the 14 trial tasks independently. We filtered out three outliers who spent less than 5 mins on the tasks. On average, the remaining 17 participants achieved an accuracy of 59.2% on 14 tasks, which is worse than the AI performance (71.4 %). Thus, the introduction of the AI system in the decision making process within our study can be beneficial to achieve better team performance.

5.6.3 Experimental Setup

Experimental conditions

To answer the above research questions, we designed a between-subjects study consisting of four experimental conditions. Example explanations in different conditions are shown in Table 5.6. Participants in all these conditions saw the systems' advice, but the five conditions differed in the inclusion of additional explanations.

- *Control*: no additional explanation.
- *Concept*: concept-based explanation from post-hoc Concept Bottleneck Models [237], similar to ExAID [240] (see Table 5.6).
- *Concept-Imp*: we provide more details about how important each concept is, which is the target domain in our proposed analogy-based explanations (see Table 5.6).
- *Analogy*: analogy-based explanation for each concept (see Table 5.6).
- *Analogy-OD*: We show the same explanations as the *Concept-Imp* condition. When users require further clarification and indicate this by clicking the Clarify button, we provide an analogy on demand.

Explanation Generation. The AI system in our study is based on a post-hoc concept bottleneck model [237]. We trained the post-hoc concept bottleneck model following its official implementation.[7] As tested by Yuksekgonul *et al.* [237], it can provide concept-based explanations aligned with medical knowledge. The post-hoc concept bottleneck model first learned concept activation vector for skin lesions based on concept banks from the Derm7pt [241] dataset. Then a linear classifier is trained to make binary predictions.

[6] This was rated as a 'good' hourly rate by the platform at the time of running the study.
[7] https://github.com/mertyg/post-hoc-cbm

Table 5.6: Example of explanations in different conditions. In condition *Analogy-OD*, when the "clarify" button is clicked, the analogy is shown on another line for the sake of clarity.

Condition	Explanation Type
Concept	absence of Streaks - type 1: strong evidence
Concept-Imp	Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign.
Analogy	Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign. This is like how a beak is a definite sign of a bird.
Analogy-OD	Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign. Clarify
Concept	observation of Dots & Globules - type 1: moderate evidence
Concept-Imp	Dots & Globules - type 1 can typically be associated with benign.
Analogy	Dots & Globules - type 1 can typically be associated with benign. This is like fish can typically be associated with oceans, but it's also possible to associate fish with rivers.
Analogy-OD	Dots & Globules - type 1 can typically be associated with benign. Clarify

Based on the linear layer weight $\mathbf{w} \in \mathbb{R}^k$ and concept activation vector $\mathbf{c} \in \mathbb{R}^k$ for each image, we generate concept-level explanations based on the contribution of each concept. For concept $c_i, i \in [1, k]$, the contribution to final prediction is $s_i = w_i * c_i$. To generate simple heuristics-based concept-level explanations (*Concept* condition), we use two thresholds to identify the importance of each concept:

$$evidence\ strength = \begin{cases} strong, & |s_i| >= \epsilon_1 \\ moderate, & \epsilon_2 <= |s_i| < \epsilon_1 \\ ignore, & otherwise. \end{cases} \tag{5.1}$$

In our study, we set $\epsilon_1 = 0.5, \epsilon_2 = 0.1$. A positive value for contribution s_i indicates that the absence/presence of concept c_i helps predict that the lesion is `malignant`, while a negative value indicates the tendency to predict `benign`. Following the templates used in Table 5.1 for *Concept-Imp* condition, we generate the target domain of analogy-based explanations. To account for errors caused by the absence of concepts, we further clarify the target domain with relation to the alternative class prediction. Instead of claiming "absence of [`concept`] is definitely a sign of [model prediction]", we use "[`concept`] is definitely a sign of [alternative option]. Thus, absence of `concept` helps make prediction of [model prediction]." For example, *Streaks - type 1 is definitely a sign of malignant. Thus, absence of Streaks - type 1 helps make prediction of benign.* To increase clarity and reduce scope for misinterpretations caused by using double negative expressions (*e.g., Absence of [*`concept`*] seldom found at benign*), we do not provide any explanations in the form of double negative expressions.

To provide high-quality analogies, we generate analogy-based explanation with two stages. In the first stage, based on the evaluation results of Section 5.5.3, we only consider analogies which are syntactically correct, factually correct, and easy to understand (Simplicity > 3). In the second stage, we manually curated and selected the analogies reserved,

which resulted in 37 valid analogies: "Definite Sign Of" (11), "Typically Associated With" (9), "Seldom Found At" (9), "Contradict With" (8). Based on the contribution of each concept s_i and the sign of predictions, we map each concept to a template. Then we generate the analogies by randomly sampling valid candidates in each template.

Measures and Variables

Table 5.7: The different variables considered in our experimental study. "DV" refers to the dependent variable. **RAIR**, **RSR**, and **Accuracy-wid** are indicators of appropriate reliance.

Variable Type	Variable Name	Value Type	Value Scale
Learning Effect (DV)	F1 of malignant concepts	Continuous	[0.0, 1.0]
	F1 of benign concepts	Continuous	[0.0, 1.0]
Performance (DV)	Accuracy	Continuous, Interval	[0.0, 1.0]
	Accuracy-wid	Continuous	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
Trust (DV)	TiA-Reliability/Competence	Likert	5-point, 1: poor, 5: very good
	TiA-Understanding/Predictability	Likert	5-point, 1: poor, 5: very good
	TiA-Intention of Developers	Likert	5-point, 1: poor, 5: very good
	TiA-Trust in Automation	Likert	5-point, 1:strong distrust, 5: strong trust
Cognitive Load (DV)	Mental Demand	Likert	-7: very low, 7: very high
	Physical Demand	Likert	-7: very low, 7: very high
	Temporal Demand	Likert	-7: very low, 7: very high
	Performance	Likert	-7: Perfect, 7: Failure
	Effort	Likert	-7: very low, 7: very high
	Frustration	Likert	-7: very low, 7: very high
Efficiency (DV)	Time of decision making	Continuous	[0.0, +∞] (s)
Covariates	ATI	Likert	6-point, 1: low, 6: high
	TiA-Propensity to Trust	Likert	5-point, 1: tend to distrust, 5: tend to trust
	TiA-familiarity	Likert	5-point, 1: unfamiliar, 5: familar
	Medical diagnosis expertise	Likert	5-point, 1: no expertise, 5: extensive expertise
	Skin cancer expertise	Likert	5-point, 1: no expertise, 5: extensive expertise
Other	Helpfulness of Explanation	Likert	5-point, 1: unhelpful, 5: helpful
	Helpfulness of Analogy	Likert	5-point, 1: unhelpful, 5: helpful
	Experience	Category	{Yes, No}
	Confidence	Likert	5-point, -2: unconfident, 2: confident

All variables analyzed in this chapter are summarized in Table 5.7.

Dependent Variables. To assess the learning effect for participants (**H1**), we calculated the F1 measures with respect to benign and malignant cases, respectively. In the post-task questionnaire, we asked participants to select the concepts positively associated with benign and malignant labels. To analyze the impact of analogy-based explanations on user reliance, we adopted the **Switch Fraction** metrics as reliance measures [57, 66]. To assess the appropriate reliance (**H2**), we followed Schemer *et al.* [29] to adopt *Relative positive AI reliance* (**RAIR**) and *Relative positive self-reliance* (**RSR**) metrics. The two measures assessed users' appropriate reliance from two dimensions (*i.e.,* appropriate adoption of AI advice and insistence on their own decision), which can help analyze the dynamics of reliance. To provide an overview of participants' performance under initial disagreement, we considered **Accuracy-wid** (*i.e.,* accuracy with initial disagreement). To analyze the impact of analogy-based explanations on cognitive load (**H3**), we adopted NASA-TLX questionnaire [242]. For the analysis of decision making efficiency (**H4**), we measured the average time spent on each decision task, which is measured in seconds.

Covariates and Trust. As pointed out by prior studies [243], user domain expertise also affects their trust and reliance on the AI system. Thus, we assessed participants' general

medical expertise by gathering responses on a 5-point Likert-scale ranging from *1:* to *5:* ("*To what extent are you knowledgeable about medical diagnosis?*"), and specific expertise on skin cancer detection task ("Do you have any experience or knowledge about skin cancer?") on a 5-point Likert-scale ranging from *1:* to *5:*. We accounted for the effect of participants' affinity with technology through the Affinity for Technology Interaction Scale (ATI) [91]. To assess participants' subjective trust in the AI system, we adapted the Trust in Automation (TiA) questionnaire [90] to the context of the "AI system". We included six subscales from the TiA questionnaire: Reliability/Competence (TiA-R/C), Understanding/Predictability (TiA-U/P), Propensity to Trust (TiA-PtT), Familiarity (TiA-Familiarity), Intention of Developers (TiA-IoD), and Trust in Automation (TiA-Trust).

Other Variables. Meanwhile, for conditions with explanations (analogies), we also assessed the helpfulness of explanations (analogies) with the question, "*To what extent did you find the explanations (analogies) helpful to make decisions?*" Responses were gathered on a 5-point Likert scale from *1* to *5* corresponding to the labels *unhelpful*, *somewhat unhelpful*, *neutral*, *somewhat helpful*, *helpful*. We further collected the reasons (open text) for perceived helpfulness with "*Why did you find the explanation (analogies) to be helpful or not helpful?*" For participants in *Analogy* and *Analogy-OD* conditions, we collected their comments and feedback (open text) to the analogies with: "*Please share any comments, remarks or suggestions regarding the use of analogies to explain the medical concepts.*" For a deeper analysis of our results, we collected responses from participants regarding their perceived user experience ("*Have you ever had this or seen it on others?*") and confidence ("*How confident are you with your decision?*") on 5-point Likert-scales along with each trial task.

5

Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for a between-subjects study using G*Power [92]. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* $\frac{0.05}{4}$, due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and the consideration of 5 different experimental conditions. This resulted in a required sample size of 265 participants. We thereby recruited 486 participants from the crowdsourcing platform Prolific[8], in order to accommodate potential exclusion.

Compensation. All participants were rewarded with £2, amounting to an hourly wage of £8 (estimated completion time was 15 minutes). In addition to this, we rewarded participants with extra bonuses of £0.1 for every correct decision in the 14 trial cases. Such monetary bonuses have been shown to motivate and encourage participants to exert genuine effort in decision making tasks, which is also a contextual requirement to encourage appropriate system reliance [40].

Filter Criteria. All participants were proficient English speakers above the age of 18, and they had finished more than 40 tasks and maintained an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one

[8] https://www.prolific.co

attention check or any missing response. The resulting sample of 280 participants had an average age of 37 (*SD* = 13.0) and a gender distribution (51.4% female, 48.6% male).

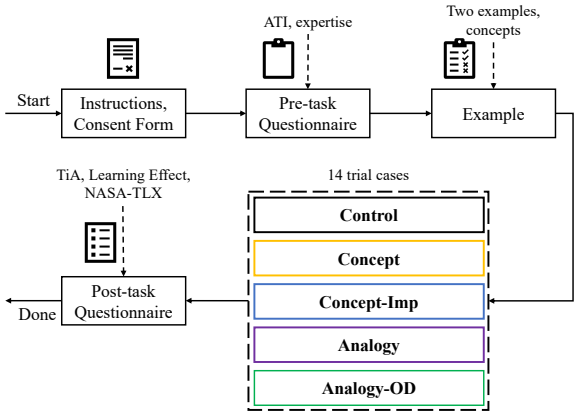


Figure 5.7: Illustration for the decision making setup.

Procedure

The entire procedure of our study is illustrated in Figure 5.7. All participants first read the same basic instructions and consent forms. Next, participants were asked to complete a pre-task questionnaire to measure their affinity for technology interaction and expertise in medical diagnosis and skin cancer. To onboard participants on the skin cancer detection task, and help them understand the labels `malignant` and `benign`, we provided them with two examples of benign and malignant skin lesions before they began working with the tasks. After the examples, all participants excluding the *Control* condition obtain an overview of the medical concepts relevant to our study (cf. Figure 5.6).

Next, participants across all conditions worked on 14 trial tasks. In each trial task, we followed a two-stage decision making process [35, 65, 85]. In the first stage, participants worked on the task without any extra information (one example shown in Figure 5.8). In the second stage, AI advice and explanations were provided, and participants had a chance to alter their decision (one example shown in Figure 5.9). After the task phase, post-task questionnaires were adopted to assess their cognitive load, their trust in the AI system, and criteria of making final decisions (open text). For all participants excluding the *Control* condition, we assessed their learning effect through a specific question ("*Please select the concepts positively associated with malignant/benign skin lesions.*"), their perceived helpfulness of explanations, and open text reasons for the perceived helpfulness. Participants in condition *Analogy* and *Analogy-OD* were additionally asked to report their perceived helpfulness of the analogies and to provide rationales/feedback in open text fields.

Attention Checks. To ensure the reliability of participants' responses, three attention check questions were placed at the pre-task questionnaire (ATI), task phase, and post-task questionnaire (Trust in automation). Each attention check asked participants to select a specific option [93, 143].

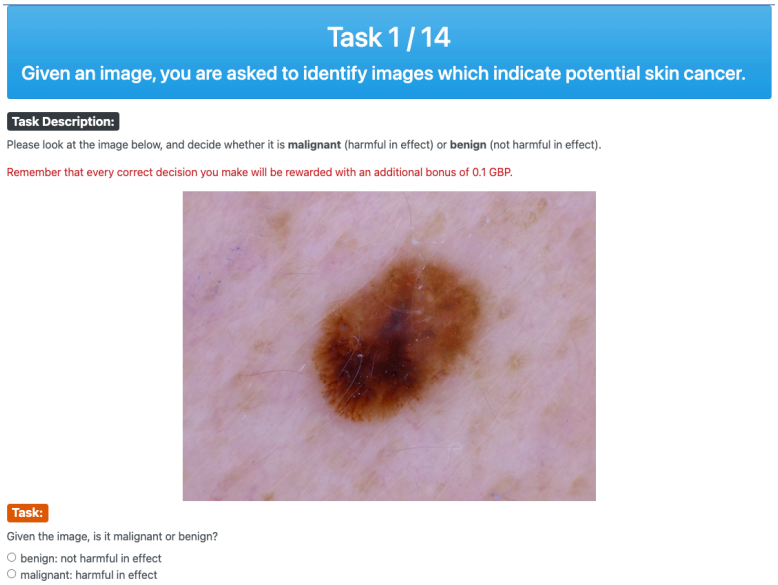


Figure 5.8: Screenshot of the task interface in the first stage of decision making.

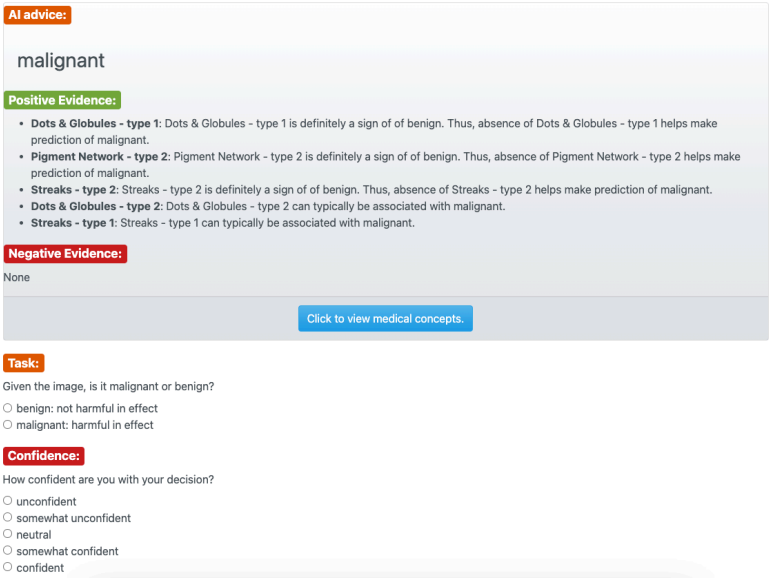


Figure 5.9: Screenshot of the task interface in the second stage of decision making for the *Concept-Imp* condition.

5.6.4 Experimental Results

Descriptive Statistics

In our analysis, we only consider participants who passed all attention checks. Participants were distributed in a balanced fashion over the four experimental conditions as follows: 55 (*Control*), 55 (*Concept*), 55 (*Concept-Imp*), 53 (*Analogy*), 62 (*Analogy-OD*).

Distribution of Covariates. The covariates' distribution is as follows: *ATI* (M = 3.87, SD = 0.87, 6-point Likert scale, and *1: low, 6: high*), *Medical Diagnosis Expertise* (M = 1.47, SD = 0.81, 5-point Likert scale, and *1: no expertise, 5: extensive expertise*), *Skin Cancer Expertise* (M = 1.59, SD = 0.81, 5-point Likert scale, and *1: no expertise, 5: extensive expertise*), *TiA-Propensity to Trust* (M = 2.76, SD = 0.57, 5-point Likert scale, *1: tend to distrust, 5: tend to trust*), *TiA-Familiarity* (M = 2.31, SD = 1.05, 5-point Likert scale, *1: unfamiliar, 5: familiar*).

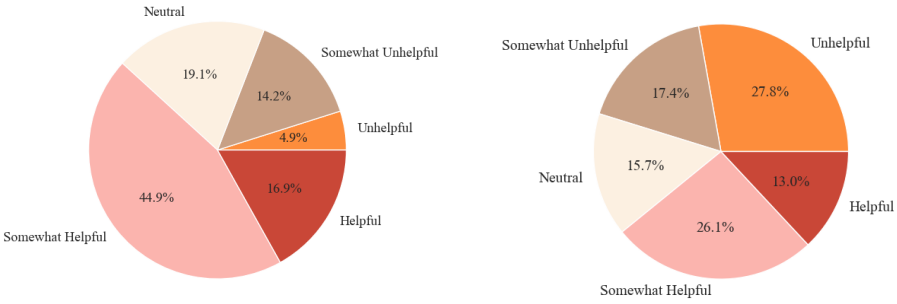
Performance Overview. On average across all conditions, participants achieved an accuracy of 63.3% (SD = 0.11), which is worse than the AI accuracy (71.4%). The agreement fraction was found to be 0.79 (SD = 0.16) while the switch fraction was 0.57 (SD = 0.30). With these measures, we confirm that in the face of disagreement with AI advice, participants in our study did not always switch to AI advice or blindly rely on the AI system. As all dependent variables are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

Table 5.8: Accuracy, experience, and confidence for the 14 tasks used in our study. "Acc" and "Con" refer to accuracy and confidence. The subscript $_i$ and $_f$ refer to the initial and final decisions, respectively. "Experience ratio" refers to the ratio of participants who reported seeing similar skin lesions in their life.

Task ID	Acc_i	Acc_f	Con_i	Con_f	Experience ratio	Ground Truth	AI correctness
ISIC-0033051	0.864	0.954	0.52	1.07	0.05	malignant	✓
ISIC-0032013	0.857	0.950	0.21	0.91	0.14	benign	✓
ISIC-0027107	0.657	0.889	0.00	0.60	0.07	benign	✓
ISIC-0028763	0.632	0.864	-0.01	0.57	0.09	benign	✓
ISIC-0034271	0.557	0.832	0.01	0.57	0.09	benign	✓
ISIC-0027665	0.554	0.818	-0.06	0.34	0.10	benign	✓
ISIC-0034155	0.443	0.793	-0.04	0.48	0.04	malignant	✓
ISIC-0033790	0.539	0.771	0.00	0.29	0.05	benign	✓
ISIC-0028076	0.457	0.750	-0.06	0.24	0.05	benign	✓
ISIC-0032557	0.043	0.368	0.93	0.30	0.05	benign	✓
ISIC-0029323	0.525	0.304	-0.05	0.29	0.05	malignant	×
ISIC-0032269	0.386	0.282	0.00	0.38	0.06	malignant	×
ISIC-0024924	0.379	0.186	0.26	0.61	0.14	benign	×
ISIC-0029260	0.311	0.100	-0.03	0.71	0.04	benign	×

Performance Per Task. Considering the 14 tasks in our study, we calculated the accuracy and confidence based on all valid participants. The results are shown in Table 5.8. Generally, the accuracy of participants increased after being exposed to correct AI advice and decreased after being exposed to wrong AI advice. Overall, participants showed higher confidence after being exposed to AI advice. The only exception is task ISIC-0032557, where participants showed less confidence in their final decision. Among all tasks, most participants indicated that they never saw the skin lesion image on themselves or on someone they know. This is illustrated by the low experience ratios observed across all tasks (cf. Table 5.8).

Helpfulness of Explanations and Analogies. In the post-task questionnaire, participants were asked to report their perceived helpfulness of explanations (for conditions with explanations) and perceived helpfulness of analogies (for condition *Analogy* and *Analogy-OD*). The distributions of perceived helpfulness are shown in Figure 5.10. Overall, 61.8% participants reported positive attitudes towards the provided concept-based explanations. Meanwhile, 39.1% participants in condition *Analogy* and *Analogy-OD* found that the provided analogies were helpful to some extent.



(a) Helpfulness of explanations.

(b) Helpfulness of analogies.

Figure 5.10: Distribution of perceived helpfulness of explanations and analogies.

H1: The impact of analogy-based explanations on learning effect

To analyze **H1**, we compared the F1 of learned concepts for the benign and malignant skin lesions. Considering that five concepts are positively correlated with label "malignant" and three concepts are positively correlated with label "benign", we adopted the weighted average F1 measures ($\text{F1}_{avg} = \frac{5}{8}\text{F1}_{malignant} + \frac{3}{8}\text{F1}_{benign}$) to assess user understanding of the AI system. The Kruskal-Wallis H-test results are: $H(279) = 1.79, p = 0.616$. The mean and std are: $M \pm SD(Concept) = 0.55 \pm 0.20$; $M \pm SD(Concept\text{-}Imp) = 0.58 \pm 0.19$; $M \pm SD(Analogy) = 0.56 \pm 0.21$; $M \pm SD(Analogy\text{-}OD) = 0.52 \pm 0.22$. No significant difference was found to suggest a learning effect. Thus, we did not find empirical support for **H1** in our study.

Table 5.9: Kruskal-Wallis H-test results for performance-based and reliance-based dependent variables across five conditions. † and †† indicate the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Dependent Variables	Accuracy	Agreement Fraction	Switch Fraction	Accuracy-wid	RAIR	RSR
H	2.18	11.03	8.42	15.81	12.77	6.16
p	.703	**.026**†	.078	**.003**††	**.012**††	.187
M(*Control*)	0.63	0.83	0.55	0.50	0.53	0.32
M(*Concept*)	0.64	0.76	0.51	0.49	0.49	**0.48**
M(*Concept-Imp*)	**0.65**	**0.83**	**0.67**	**0.65**	**0.70**	0.35
M(*Analogy*)	0.62	0.77	0.55	0.51	0.54	0.39
M(*Analogy-OD*)	0.63	0.78	0.58	0.55	0.58	0.43

To verify **H2**, we used Kruskal-Wallis H-tests to compare participants' performance across all conditions. The results are shown in Table 5.9. Among the dependent variables we analyzed across the conditions, we found that participants exhibited significant

differences in their appropriate reliance. Through post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125, we found that: (1) participants in condition *Concept-Imp* showed significantly higher **Accuracy-wid** than participants in conditions *Control*, *Concept*, *Analogy*; (2) participants in condition *Concept-Imp* showed a significantly higher **RAIR** than participants in conditions *Control*, *Concept*, *Analogy*. The results indicate that the target domain of our analogy-based explanation can help users appropriately rely on AI systems, which is mainly by addressing the under-reliance. However, this may also trigger over-reliance on the AI system, which is reflected by the relatively low **RSR** in comparison with other conditions. At the same time, we found that the analogies did not have the expected effect in facilitating appropriate reliance. However, our results suggest that providing analogies on demand can have a better impact on appropriate reliance (non-significant). Thus, we did not find empirical support for **H2** in our study.

H3: The impact of analogy-based explanations on cognitive load

Table 5.10: ANOVA test results for user cognitive load across five conditions. "Avg" refers to the average cognitive load among six dimensions. † and †† indicate the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Cognitive Load	Avg	Mental	Physical	Temporal	Performance	Effort	Frustration
F	5.81	7.01	0.65	0.67	1.08	3.03	1.98
p	**.000**††	**.000**††	.625	.616	.368	**.018**†	.098
M(*Control*)	-2.25	-0.02	-5.25	-4.35	-1.04	0.87	-3.69
M(*Concept*)	-1.42	2.05	-4.45	-4.33	-1.11	1.85	-2.53
M(*Concept-Imp*)	-0.88	2.62	-4.89	-3.85	-0.13	2.82	-1.85
M(*Analogy*)	-1.07	2.13	-4.53	-3.85	-0.32	2.04	-1.91
M(*Analogy-OD*)	-0.98	2.95	-4.68	-4.32	0.11	2.35	-2.31

To analyze **H3** for the impact of experimental conditions on cognitive load, we conducted a one-way ANOVA. Our findings are shown in Table 5.10. Overall, participants who received explanations reported a higher perceived cognitive load. Through post-hoc Turkey HSD tests using a Bonferroni-adjusted alpha level of 0.0125, we found a significant difference: For both average cognitive load and mental demand, *Control*< *Concept*, *Analogy*, *Concept-Imp*, *Analogy-OD*. Thus, we did not find support for **H3**.

H4: The impact of analogy-based explanations on decision making efficiency

To analyze **H4**, we compared participants' task completion time (units: seconds) in the 14 tasks with Kruskal-Wallis H-test. The results show a significant difference: $H(279) = 23.73, p = .000$. Post-hoc Mann-Whitney test results showed that participants who received explanations spent significantly more time making decisions: *Control* < *Concept*, *Analogy*, *Concept-Imp*, *Analogy-OD*. $M \pm SD$(*Control*) = 462 ± 309; $M \pm SD$(*Concept*) = 548 ± 210; $M \pm SD$(*Concept-Imp*) = 575 ± 209; $M \pm SD$(*Analogy*) = 574 ± 242; $M \pm SD$(*Analogy-OD*) = 658 ± 341.

For a more fine-grained analysis, we calculated the average time spent on each correct/wrong decision per person for each user group (shown in Table 5.11). With Kruskal-Wallis H-test, we compared the average time per correct/wrong decision. The post-hoc Mann-Whitney test results are still consistent with the overall decision making efficiency:

participants who received explanations spent significantly more time making decisions. Thus **H4** is not supported by our experimental results.

Table 5.11: Time per decision (in seconds). The "Decision-level" is calculated by average on all decisions in each condition. The "Human-level" is calculated by the average of all humans in each condition.

Granularity	**Decision-level**		**Human-level**	
Correctness	$M \pm SD$ (Correct)	$M \pm SD$ (Wrong)	$M \pm SD$ (Correct)	$M \pm SD$ (Wrong)
Control	34.19 ± 39.85	30.92 ± 31.06	34.17 ± 23.71	30.16 ± 21.31
Concept	37.90 ± 29.74	41.35 ± 32.17	37.83 ± 14.82	41.67 ± 19.00
Concept-Imp	39.86 ± 32.34	43.25 ± 39.69	40.78 ± 16.18	41.84 ± 20.46
Analogy	40.71 ± 35.37	41.51 ± 31.49	40.84 ± 19.79	42.78 ± 19.92
Analogy-OD	47.21 ± 59.91	46.66 ± 51.51	46.99 ± 27.13	46.63 ± 32.16

5.6.5 Exploratory Analysis

The Impact of First Impression

Prior work has demonstrated the significant impact of first impressions of AI systems in shaping user trust and reliance [87, 99, 244]. We thereby analyzed the potential impact of task ordering and the accuracy of AI advice. To this end, we grouped participants according to the AI accuracy within the first five tasks. Participants who either never encountered wrong AI advice or did so only once are grouped within "*Good First Impression*", and others are grouped within "*Bad First Impression.*" We compared participants' performance and reliance on AI systems with Kruskal-Wallis H-test. We found no significant difference, suggesting that first impressions of the AI system did not have an effect within our study.

Analysis of Trust and Covariates

An ANCOVA analysis across the experimental conditions revealed no significant difference in the perceived trust of the participants in the AI system (TiA). For all covariates, we conducted Spearman rank-order tests with dependent variables.

The impact of propensity to trust. As shown in Table 5.12, TiA-Propensity to Trust significantly affected user trust in the AI system. With Spearman rank-order test, we found that TiA-Propensity to Trust positively correlated with all trust measures: TiA-R/C, $r(278) = .650$, $p = .000$; TiA-U/P, $r(278) = .344$, $p = .000$; TiA-IoD, $r(278) = .283$, $p = .000$; TiA-Trust, $r(278) = .677$, $p = .000$. Meanwhile, TiA-Propensity to Trust also showed significant positive correlation with performance and appropriate reliance measures: Agreement Fraction, $r(278) = .227$, $p = .000$; Switch Fraction, $r(278) = .220$, $p = .000$; RAIR, $r(278) = .183$, $p = .002$; RSR, $r(278) = -.216$, $p = .000$. It is worth noting that the general propensity to trust positively correlated with all trust dimensions, and **Agreement Fraction**, **Switch Fraction**, **RAIR**, but negatively correlated with **RSR**. Thus, participants with a higher propensity to trust tend to rely more on the AI system after the XAI is provided. However, this addresses under-reliance to some extent but also causes over-reliance.

Other covariates. For TiA-Familiarity, we found a strong positive correlation with some trust measures: TiA-R/C, $r(278) = .232$, $p = .000$; TiA-Trust, $r(278) = .286$, $p = .000$. For

Table 5.12: ANCOVA test results corresponding to user trust across experimental conditions. † and †† indicate the effect of variable is significant at the level of 0.05 and 0.0125, respectively.

Dependent Variables	TiA-R/C			TiA-U/P			TiA-IoD			TiA-Trust		
Variables	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
Experimental Condition	1.02	.397	0.01	0.45	.769	0.01	4.47	**.002**††	0.05	3.06	**.017**†	0.02
Medical Expertise	0.47	.493	0.00	3.05	.082	0.01	0.03	.868	0.00	0.22	.639	0.00
Skin Cancer Expertise	2.97	.086	0.01	0.09	.766	0.00	1.64	.201	0.01	0.01	.927	0.00
ATI	0.58	.448	0.00	2.10	.149	0.01	3.68	.056	0.01	2.03	.155	0.00
TiA-Propensity to Trust	182.14	**.000**††	0.39	31.72	**.000**††	0.10	35.53	**.000**††	0.11	223.51	**.000**††	0.44
TiA-Familiarity	1.58	.210	0.00	0.22	.641	0.00	0.52	.471	0.00	3.35	.068	0.01

ATI, we found a strong positive correction with TiA-Trust, $r(278) = .149$, $p = .012$. However, according to the results of ANCOVA analysis of trust (Table 5.12), the impact of ATI and TiA-Familiarity is insignificant. No strong correlation was found for the covariates of expertise in medical diagnosis expertise. We found a strong negative correlation with the skin cancer expertise and Switch Fraction: $r(278) = -.175$, $p = .003$. Among 280 participants, 166 reported zero skin cancer experience or expertise.

Impact of user opinions towards explanations and analogies

Opinion towards explanations. To understand how users' perceived helpfulness of explanations affects user trust and reliance on the AI system, we conducted the Spearman rank-order test for participants in the condition *Concept*, *Concept-Imp*, *Analogy*, and *Analogy-OD*. The results show that, the perceived helpfulness of explanations is positively correlated with user trust: **TiA-R/C**, $r(223) = .400, p = .000$; **TiA-U/P**, $r(223) = .397, p = .000$; **TiA-IoD**, $r(223) = .249, p = .000$; **TiA-Trust**, $r(223) = .407, p = .000$. However, there is no significant correlation between the perceived helpfulness of explanations and reliance-based dependent variables.

Opinion towards analogies. Similarly, to understand how users' perceived helpfulness of analogies affects user trust and reliance on the AI system, we conducted the Spearman rank-order test for participants in condition *Analogy* and *Analogy-OD*. The results show that the perceived helpfulness of analogies is positively correlated with user trust: **TiA-R/C**, $r(113) = .303, p = .001$; **TiA-U/P**, $r(113) = .290, p = .002$; **TiA-IoD**, $r(113) = .368, p = .000$; **TiA-Trust**, $r(113) = .297, p = .001$. Meanwhile, there is no significant correlation between the perceived helpfulness of analogies and reliance-based dependent variables.

24.5% participants in the *Analogy* condition found the analogies to be helpful (perceived helpfulness > 0), while 51.6% participants in the *Analogy-OD* condition thought the analogies are helpful. This may also help explain why participants in the *Analogy* condition showed slightly lower **Switch Fraction**, **Accuracy-wid**, **RAIR** and **RSR** in comparison with the *Analogy-OD* condition. Combined with the strong positive correlation between perceived helpfulness and user trust in the AI system, we can infer that participants in the *Analogy* condition showed less trust and reliance on the AI system (*i.e.,* they exhibited under-reliance on the AI system). Meanwhile, participants in the *Concept-Imp* condition showed very low **RSR**, which indicates over-reliance on the AI system.

Qualitative Analysis of Feedback

We asked all participants in our study for their rationales in their decision making using an open-ended question ("*Please describe how you made your decisions in these tasks.*"). Using

the thematic analysis software, ATLAS.ti,[9] we conducted a thematic analysis and selected the top-5 topics mentioned by users (shown in Table 5.13).

Table 5.13: Resulting main themes from the thematic analysis of participants' responses to the open questions pertaining to the decision criteria.

Topic	Frequency	Participant Feedback
Picture	91	(1) I looked at the pictures and tried to match them with the descriptions for either malignant or benign. - *Analogy-OD* (2) based on the image content and my understanding of malignant features. - *Control* (3) by judging the photos. - *Analogy*
Examples	77	(1) Based on the examples shared and severity of the colours and depth of the shape. - *Analogy* (2) I looked at the image and referred back to the malignant and benign images and tried to think which it resembled. - *Analogy*
Explanations	77	Started off by remembering the concepts and applying them to the initial image. Then refining that based on the AI. Generally trusted the AI's decisions more than my own. I weighed up the Positive and Negative evidence. - *Analogy*
Intuition	68	(1) I went entirely on instinct. If the image made me feel uncomfortable I labelled it malignant. Funnily enough most of the time my instincts were in agreement with the AI. - *Control* (2) how i thought it maybe should look if it was something bad. - *Analogy*
AI advice	62	(1) Applied the knowledge that I previously had and the information taught in this task; used AI to help if I was a bit confused and it was labeling the image. - *Analogy-OD* (2) Based on my intuition and recommendations from the AI system. - *Analogy*

For participants who received explanations along with the AI advice, we asked for their feedback regarding the usefulness of explanations. According to the 139 participants who reported the explanations to be "somewhat helpful" or "helpful", the main reasons are summarized below:

1. the explanations enrich the context of decision making or help make the decision - 32.4%;
2. the explanations help improve the understanding of the AI system - 18.7%;
3. the explanations help confirm or validate their decision - 7.2%

The main reasons due to which workers found explanations to be either "somewhat unhelpful" or "unhelpful" are summarized below:

1. participants lack knowledge or expertise to interpret explanations - 41.9%;
2. participants failed to understand the explanations - 16.3%;
3. explanations are difficult to apply - 11.6%.

In case of analogies, the major reason that workers reported perceiving analogies as helpful were that analogies aided their understanding and reasoning (15 participants, 33.3 %). The top-5 reasons for perceiving analogies as unhelpful are as follows:

1. participants failed to connect the source domain with the target domain - 22.9%;
2. participants think the analogies do not make sense - 18.6%;
3. participants think the concepts are not relevant - 14.3%;

[9] https://atlasti.com

4. participants fail to understand the analogies - 12.9%;
5. participants think the analogies are not necessary - 10%.

We asked participants in conditions *Analogy* and *Analogy-OD* for their feedback and comments on the provided analogies. Overall, we found conflicting attitudes toward the provided analogies. While some users found merit in their use, others found them to be distracting. This is reflected in the sample quotes from two participants below.

> "*It's definitely useful and helpful for getting the point across to laymen like myself*".

> "*I don't get the relevance of using analogies to explain medical concepts. I also don't think they were explaining the concepts. It was essentially saying water is wet...*".

Insights from users to improve the effectiveness of analogy-based explanations. Based on the feedback from participants in the relevant experimental conditions in our study, we summarized the following potential directions to further improve the effectiveness of analogy-based explanations:

- *Enhancing the relation between the target domain and the source domain (analogies).* Among participants who found analogies to be "unhelpful," many of them claimed that they failed to understand the analogies or make immediate connections or associations with the target domain.
- *Providing analogies in a more relevant domain.* Some participants complained that they failed to connect the concepts used in the analogies with the context of medical analysis. Analogies in a relevant domain can potentially help improve the plausibility and trustworthiness of analogy-based explanations.
- *Providing analogies selectively or on demand.* When the original explanation is clear enough, some participants would take the analogies as unnecessary or even distracting. Some others reported feeling annoyed: "*However, when the concept is straightforward or otherwise readily met in normal daily life, the use of an analogy can easily be perceived as condescending or even irritating and thus antagonize, rather than assist, the person concerced.*" However, if a lot of analogies are used, users may feel overwhelmed, which may hurt their trust and satisfaction with the analogy-based explanations.

5.7 Discussion

In summary of the experimental results, Table 5.14 provides an overview of the findings. Based on the findings in Study I and Study II, we elaborately discussed the potential effect of analogy properties. We also identify and synthesize the limitations of our studies.

5.7.1 Key Findings and Implications

Subjectivity of Analogies. The results of the study I especially highlight the subjective nature of the qualitative dimensions that characterize analogies. According to Krippendorff's α, we find that experts show clear disagreement on most qualitative dimensions.

Table 5.14: Summary of key findings in two studies.

Study	Findings
Study I	The proposed qualitative dimensions were found to positively correlate with the perceived helpfulness of analogy-based explanations.
	The expert evaluation results show that experts do not always agree on some qualitative dimensions (*e.g., Structural Correspondence*).
Study II	The analogy-based explanations fail to bring improved user understanding, which is assessed by the learning effect of the concepts.
	Participants showed similar levels of performance across all conditions, participants showed better appropriate reliance in condition *Concept*.
	Participants who received explanations indicated higher cognitive load.
	Participants who received explanations spent significantly more time making decisions.

This is possibly because of the different experiences of the world each expert has, leading to different interpretations and familiarity of the commonsense facts in the analogies. Prior work on inter-rater disagreement suggested that disagreement is not always noise but can also be a signal [245]. With disagreement from multiple explainees, we can address the ambiguity and vagueness of analogy-based explanations and seek further improvement [246, 247]. When evaluators find that one commonsense explanation falls short in specific dimensions, we can involve another crowd worker to improve it according to the feedback.

The comparison between the quality of explanations generated from the two tasks shows that better quality on a single dimension (like *Relational Similarity*) does not necessarily translate to better helpfulness in understanding the target sentence. However, if an explainee (*e.g.,* E_1 and E_5) thinks the explanation is of poor *Relational Similarity*, they may tend to judge it unhelpful. Meanwhile other user factors (like abstract thinking, personal interpretation, and general attitude in disagreement analysis) may also affect the perceived helpfulness and other qualitative dimensions. This points out the need for further studies about the impact of user factors (*e.g.,* experience, belief) and qualitative dimensions on the helpfulness of analogy-based explanations.

Contradicting with the assumption that commonsense knowledge should be accepted and understood by all humans [192], the disagreement from experts also reveals that commonsense explanations are not one-size-fits-all solutions for laypeople. This is in line with findings for explainable AI [27, 179]. In the future, one should adjust the commonsense explanations according to the explainee's belief about the world to ensure the effectiveness of such analogical inference from commonsense knowledge. This also suggests that the role of personalization should be carefully considered when generating commonsense explanations.

Automatic Analogy Generation and Evaluation. In study I, we observed that around one-third of generated analogies are not factually correct, and that it can be difficult for workers to generate analogies that demonstrate a high *Structural Correspondence* and *Relational Similarity*. This highlights the need for strategies to support workers in generating effective analogies. Especially, we envision the development of machine-in-the-loop crowdsourcing tasks, *e.g.,* by using relational knowledge bases and machine learning methods as an auxiliary toolkit to facilitate automation [214, 216]. Knowledge bases store real world facts in a pre-defined format, typically a triplet ⟨ subject, predicate, object ⟩. Hence,

once the relationship between the concept and label in a target sentence is identified, it would be straightforward to find correct everyday facts sharing the same relationship along with high *Structural Correspondence*. This would provide high-quality candidate concepts to the crowd workers, reducing their work load.

Our results of study I highlight that most qualitative dimensions show a significant positive correlation to perceived helpfulness. Yet, it would be expensive to always obtain a human evaluation for quality control. Future work should hence investigate the (semi-)automatic assessment of the different quality dimensions (or at least of *helpfulness*). For *Syntactic Correctness*, one could involve automation toolkits (like syntactic error detection provided by Grammarly[10]) to provide suggestions for fixing syntactic errors when participants generate analogies on the fly. For *Simplicity* and *Misunderstanding*, one could maintain a list of everyday concepts and a list of concepts with multiple interpretations for ease of automatic check. Recent work on jury learning [248] proposed a method to conduct automatic pseudo-human value judgement with machine learning models, which can be an alternative to expert-based quality evaluation, while accounting for the subjectivity of each dimension.

The Role of Human Intuition. In study II, many participants reported that they relied on their intuition to make their final decisions. This indicates that human intuitions play a critical role in shaping user understanding and reliance behaviors. Our findings suggest that human intuition can be a potential factor to achieve the goal of appropriate reliance on AI systems. This is in line with prior findings about human intuition in the human-AI decision making context [249, 250].

On the one hand, human intuition may facilitate complementary collaboration with the AI system. On the other hand, human intuition can also cause bias when making decisions. In our study, we found that the **Agreement Fraction** is relatively high (on average, around 0.80 across all conditions), while **RSR** is low for most conditions. In other words, when AI advice is wrong and users disagree, they tend to rely on AI advice instead of their initial decision (which is correct). This indicates a clear over-reliance on the AI system. This is also found in prior studies about the pitfalls of XAI interventions [70, 107]. Such over-reliance can be associated with confirmation bias and the illusion of explanatory depth [218]. Meanwhile, participants also showed clear under-reliance in condition *Control*, *Concept*, *Analogy* (significantly worse than condition *Concept-Imp*). A potential cause for such under-reliance can be the Dunning-Kruger effect [60]. As reported by He *et al.* [30], "users who overestimated their capability on the task tend to exhibit under-reliance." In our study, several participants reported that they did not find explanations and analogies helpful. However, we found a strong positive correlation between the perceived helpfulness of explanations (analogies) and the subjective trust in the AI system. We can infer that participants' trust was negatively affected by the perceived unhelpfulness of analogies, which may have further impacted user reliance on the AI system. In the broader context of human-AI decision making, it would be arguably impossible for most laypeople to comprehensively understand complex AI systems. According to Lee *et al.* [40], "trust guides reliance when complexity and unanticipated situations make a complete understanding of the automation impractical." Thus, participants in our study may

[10] https://www.grammarly.com/

have exhibited under-reliance due to uncalibrated trust.

The Role of Plausibility. Through the results of the empirical study, we found that many participants thought (1) the target domain of proposed analogy-based explanations was clear enough; and (2) extra analogies are not always helpful, especially when participants fail to connect them with the target domain. Such findings can be partially explained by the plausibility of explanations. Participants implicitly hold the belief that "plausible explanations typically imply correct decisions, and vice versa" [182]. Those participants who may have found the analogies to be implausible may have perceived certain AI advice as untrustworthy and thereby relied less on the AI system. Such under-reliance could result in sub-optimal team performance. This may help explain the finding that participants in the *Analogy* condition showed worse **RAIR** than participants in the *Concept-Imp* condition. Compared to the *Analogy* condition, more participants in *Analogy-OD* took the analogies as plausible (perceived helpfulness > 0). Meanwhile, participants in *Analogy-OD* condition showed higher **Switch Fraction**, **Accuracy-wid**, and **RAIR** and **RSR**. This indicates that providing analogies on demand may be a good design to facilitate human-AI collaboration. When analogies are not used appropriately, both under-reliance and over-reliance can be triggered due to implausibility.

5.7.2 Caveats and Limitations

Bias in Templates. We used 6 pre-defined templates to help participants generate analogy-based explanations. While crowd workers can generate syntactically correct explanations to elucidate the relevance level in concept-based explanations, these templates may lead to biases in the analogy generation [149, 157]. These templates show an initial bias to relationships which may limit the participants' creativity in generating useful analogies. However, as we found through our study, participants benefit from domain cues that can help them anchor their creativity and generate high-quality analogies.

Restricted Usage. Meanwhile, analogy-based explanations may not be the ideal solution for all application scenarios. According to results from our study, we summarize several scenarios inappropriate to adopt analogy-based explanations. First, when the original task is simple enough and only involves everyday concepts, analogy-based explanations may not work as expected. In such scenarios, analogy-based explanations turn out to pose more cognitive load and make it confusing to users. Second, when no explicit properties and relationship are associated with the task domain (like CLC in our study), analogy-based explanations may not be as effective for laypeople. In these tasks, it would be very hard to generate effective analogies due to a lack of explicit structural correspondence and relational similarity.

As the analogy-based explanations are generated based on concept-level explanations, cascading effects are also a limitation for analogy-based explanations. If the concept-level explanations do not faithfully reflect the internal state of AI systems, there is no chance for analogy-based explanations to do so. Furthermore, as analogy-based explanations are more familiar to most users, they have the potential to be more persuasive than original concept-based explanations. In other words, when the concept-level explanations mislead AI system users, effective analogy-based explanations generated from them may amplify such impact.

Potential Human Biases. Draws *et al.* have demonstrated that cognitive biases introduced by task design and workflow can negatively impact crowdsourcing experiments [157]. Using the Cognitive Biases Checklist [157], we analyzed the potential biases in our study and reported our findings here. On the task ISIC-0032557 most participants thought that they made correct decisions and reported a high confidence in their decisions. However, that may have been a result of an illusion of their competence on the task. They achieved only 4.3% accuracy on this task. This suggests that **Overconfidence or Optimism Bias** bias (*i.e.,* Dunning-Kruger effect [30, 60]) may have played a role in shaping these outcomes. Meanwhile, some participants also reported that the explanations helped confirm and validate their initial decision, suggesting a potential role of **Confirmation Bias** in shaping our findings. In our study, we provide 4-7 concept-level explanations / analogy-based explanations along with each task. From the open text feedback, two participants reported an information overload. This may have some negative impact on user trust and reliance. Due to the **Self-interest Bias**, crowd workers may not have thoroughly checked explanations in each task.

Threats to generalizability. In study I, we generated and evaluated analogy-based explanations on two relatively simple and low-stake tasks. The perceived quality of analogy-based explanations should be further evaluated with more realistic decision scenarios which require AI support. Although the generated analogy-based explanations are thought to be highly transferable, it is unknown how our findings and insights can generalize to complex and high-stake tasks. If the generated analogies are not always transferable, it would be valuable to investigate how to generate effective analogy-based explanations for specific high-stake tasks, *e.g.,* with experts.

Since human intuition may have heavily affected decision making in this task, some findings in study II may not generalize to tasks where human intuition does not have a dominant role. In our studies, only the relevance level between concepts and model predictions is highlighted and explained with analogies. However, analogies can be used to express more complex structural corresponding and relationally similar events in real-world problems. Our findings may not carry forward to more complex concept-level explanations (*e.g.,* in case of a greater number of concepts or more complex relational structures between concepts).

5.8 Conclusions and Future Work

In this chapter, we propose to elucidate concept-level AI explanations with analogical inference from commonsense knowledge in order to facilitate meaningful collaborations between an AI system and non-expert humans receiving advice from the AI system. To this end, we first designed a template-based analogy generation method, and we instantiated our method by recruiting crowd workers to generate analogy-based explanations using two image classification tasks – calorie level classification and scene classification (**RQ1**). To assess the quality of the generated explanations, we then synthesized a structured set of quality dimensions and applied it to our explanations (**RQ2**). An expert-led evaluation showed that our proposed method can generate high-quality analogy-based explanations with non-expert workers.

To comprehensively explore how analogy-based explanations affect user understand-

ing of and reliance on the AI system, we then conducted a follow-up empirical study on a skin cancer detection task (**RQ3** and **RQ4**). Results from this second study showed that (1) the lack of domain expertise hinders user understanding of concept-level explanations; (2) compared to traditional concept-level explanations, the improved concept-level explanations (*i.e.,* target domain of our analogy-based explanations) can promote appropriate reliance on the AI system by mitigating under-reliance, but may also trigger over-reliance; (3) providing analogies on demand can be a good design for adoption of analogy-based explanations; (4) yet analogy-based explanations should be carefully designed and used in order to effectively elucidate concept-level explanations. Experimental results provide limited support that analogy-based explanations can facilitate user understanding of the AI system or appropriate reliance on the AI system. However, we cannot deny the potential of analogy-based explanation in assisting laypeople for effective decision making. Compared to concept-level explanations, the additional analogies do not cause a significant delay in decision making or pose a significantly higher cognitive load. Our findings suggest that the key challenge is in generating high-quality analogies and the potential for personalization. Based on the qualitative analysis of participants' feedback and user reliance patterns, we summarized guidelines for future work about generating effective analogy-based explanations and on the appropriate usage of analogy-based explanations.

In this chapter, we focused on generating high-quality analogy-based explanations using non-expert crowd workers, and evaluating their effectiveness. With the results from the first study (N = 100), it is evident that both generation and evaluation of analogy-based explanations are challenging and time-consuming. In the imminent future, we will consider including machine learning algorithms and leverage knowledge bases to automate this task while achieving scalability and efficiency. In our second study (N = 280), we found that analogy-based explanations do not work as expected in facilitating appropriate reliance. However, we found enough evidence that highlights their potential for aiding laypeople in understanding AI systems. Hence, further research about the generation of effective analogy-based explanations and their appropriate use is required. Particularly, we also found that the understanding of commonsense explanations varies with the experience of the recipient user, which points out the need for further work on the personalization of commonsense explanations.

6

Conversational XAI Decision Support

Explainable artificial intelligence (XAI) methods are being proposed to help interpret and understand how AI systems reach specific predictions. Inspired by prior work on conversational user interfaces, we argue that augmenting existing XAI methods with conversational user interfaces can increase user engagement and boost user understanding of the AI system. In this chapter, we explored the impact of a conversational XAI interface on users' understanding of the AI system, their trust, and reliance on the AI system. In comparison to an XAI dashboard, we found that the conversational XAI interface can bring about a better understanding of the AI system among users and higher user trust. However, users of both the XAI dashboard and conversational XAI interfaces showed clear over-reliance on the AI system. Enhanced conversations powered by large language model (LLM) agents amplified over-reliance. Based on our findings, we reason that the potential cause of such over-reliance is the illusion of explanatory depth that is concomitant with both XAI interfaces. Our findings have important implications for designing effective conversational XAI interfaces to facilitate appropriate reliance and improve human-AI collaboration.

6.1 Introduction

In recent years, deep learning-based AI systems have brought about tremendous possibilities to change and affect our daily life [251, 252]. Due to the intrinsic opaqueness of such systems, automating critical decision making by using AI systems is far from reliable [253]. However, leveraging such powerful AI systems to *assist* and *empower* human decision makers is an alternative that has gained prominence [22]. In such a collaborative decision making process, explanations are incorporated to increase intelligibility and ensure that decision makers can make informed decisions [254]. Post-hoc explainable AI

This chapter is based on a peer-reviewed paper: **Gaole He**, Nilay Aishwarya, Ujwal Gadiraju. *Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant.* 30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy. https://doi.org/10.1145/3708359.3712133.

(XAI) methods are typically used to help explain AI predictions from deep learning-based AI systems.

To realize the goal of complementary team performance, users of an AI system are expected to rely appropriately on AI advice [29]. Such appropriate reliance requires a comprehensive understanding of the AI system and its underlying rationale alongside the AI advice [40, 171, 255], which play important roles in calibrating user trust and reliance behaviors [66, 256]. According to several empirical studies in human-AI collaboration [22, 66, 107], most XAI methods are not as helpful as expected and are even harmful at times (*e.g.,* causing over-reliance). The reasons behind this are multi-fold: (1) Most existing XAI methods can only provide specific types of information [257] (*e.g.,* feature importance [186], counterfactual reasoning [258]). (2) In practice, there are diverse stakeholders of AI systems [223, 224] (*e.g.,* developers, experts, and laypeople) having different levels of domain expertise and AI literacy. (3) The information needs of diverse stakeholders can vary greatly. Thus, a specific type of XAI method can seldom address varying information needs, resulting in a lack of understanding of the AI system.

Based on folk concepts in the theory of mind literature, Jacovi *et al.* [259] argue that successful explanations can provide users with the necessary components to build a coherent mental model. We extrapolate that to make critical decisions with AI assistance, users need to build a relatively more complete and coherent mental model by exploring different explanations provided by XAI methods. However, such a process can be complex—it requires processing information based on a variety of aspects, depending on the XAI methods. When presenting tailored explanations for specific audiences, designers need to trade off the simplicity and completeness of the explanations [260]. Instead of selecting a single specific explanation, an XAI dashboard enables users to explore their information needs by providing them access to their desired explanations on demand. Such an interactive interface can bring forth the advantages of both simplicity and completeness and has been increasingly recognized as an effective design [261, 262]. However, not all users have the necessary AI knowledge and experience to understand or benefit from such explanations [257]. Nor can all users articulate their information needs and find suitable XAI methods to address their concerns [263]. Therefore, we need a more flexible, dynamic, and personalized approach to resolving users' explanation needs.

Conversational user interfaces can provide a human-like interaction [264] and simplify complex tasks with filtered information [265], which can bring better user experience and higher user engagement. Inspired by prior work on conversational user interfaces for XAI [263], we argue that augmenting existing XAI methods with conversational interfaces can potentially boost users' understanding of the AI system through an improved exploration of their explanation needs. Such interaction may benefit humans by fostering increased engagement and helping build a relatively more coherent and complete mental model that aids their information needs. Thus far, only a few studies [266–270] have explored how conversational interfaces can be combined with XAI methods. However, existing work has not systematically explored the impact of conversational XAI interfaces on user trust and reliance in the context of critical decision making. Our work presents a study that addresses this under-explored research and empirical gap.

In this chapter, we explored how conversational XAI interfaces shape user understanding of an AI system. To this end, we aim to address the following research questions:

RQ1: *How does a conversational XAI interface shape user understanding of an AI system, in comparison with an XAI Dashboard?*
RQ2: *How does a conversational XAI interface influence user trust and reliance on an AI system, in comparison with an XAI Dashboard?*

To answer these questions, we conducted an empirical study (N = 306), exploring human-AI collaborative decision making in a loan approval task (*i.e.,* making a binary decision based on a loan applicant profile). To further our understanding of the impact of enhanced conversation with flexible user input and high-quality text responses based on XAI outcomes, we considered large language model (LLM) agents to power the conversational XAI interface. Overall, we found that users with conversational XAI interfaces tended to rely more on the AI system. However, such increased reliance did not always translate into appropriate reliance. Instead, it was characterized by clear patterns of over-reliance. Compared to an XAI dashboard, we observed limited improvements in user understanding and trust brought forth by the conversational XAI interface. We found a strong correlation between most measures of user understanding and user trust with users' reliance behaviors.

Our results collectively suggest that both the XAI dashboard and the conversational XAI interface worked as persuasive technology. Leveraging LLM agents to power the conversational interface can increase the perceived plausibility of explanations, potentially amplifying such impact. These observations highlight that supporting specific AI advice with interactive XAI interfaces can lead to creating an illusion of explanatory depth. To this end, users may overestimate the capability of the AI system. Our findings suggest that apart from improving user experiences with conversational interfaces, addressing the illusion brought about by such persuasive technologies can be pivotal in facilitating appropriate reliance on AI systems. Systematic empirical explorations are fundamentally important to understand how conversational interfaces can be leveraged effectively to foster optimal human-AI collaboration. In the absence of such efforts, designers and practitioners are often left to make less-informed choices that can lead to unintended consequences. In this spirit, our work has important theoretical implications for promoting appropriate reliance using XAI methods, and in equal part, design implications for effective conversational interfaces to support human-AI collaboration.

6.2 Related Work

This chapter focuses on exploring the impact of an XAI dashboard and a conversational XAI interface on user understanding of an AI system (**RQ1**), which may further affect user trust and appropriate reliance (**RQ2**). Thus, we position our work in the following realms of related literature: human-AI decision making (§6.2.1), explainable AI (§6.2.2), and conversational user interfaces (§6.2.3).

6.2.1 Human-AI Decision Making

While predictive AI systems are powerful, they are seldom perfect [271]. Transparency and accountability issues prevent deep learning-based AI systems from automation in

high-stakes applications like medical diagnosis [272]. In comparison, human workers (*e.g.,* medical doctors) show strong reliability and accountability for their work outcomes and decisions, which serve as the foundation for customers to trust their services. With these concerns, human-AI collaborative decision making is regarded as a promising approach to taking advantage of both humans and AI to achieve more accurate and reliable decision outcomes.

Complementary team performance is an important goal for human-AI decision making [70, 219], and will continue to be vital in the age of LLMs [273–275]. To achieve complementary team performance, users of AI systems are expected to rely on AI advice appropriately [29]. To this end, users are expected to follow AI advice when the AI system is more capable than them, and not rely on AI advice when the AI system is less capable. When users fail to calibrate their trust in the AI system, they may misuse or disuse the AI advice, resulting in over-reliance and under-reliance, respectively. The causes for unexpected reliance behaviors are complex. For example, algorithm aversion [33, 113] and algorithm appreciation [137] can cause under-reliance and over-reliance, respectively. Existing work has extensively explored how confidence [66, 114], risk perception [85, 110], performance feedback [49, 111], and explanations [52, 107, 109] can affect human-AI decision making.

Prior studies found that human factors like expertise and domain knowledge [99, 108] and cognitive bias [30, 218] can greatly affect user trust [276] and appropriate reliance [29] on the AI system. To mitigate the negative impact of some human factors, researchers have proposed tutorial interventions [30, 108, 117, 172], cognitive forcing functions [31, 55, 277], and improving transparency of the AI system [49, 107, 122]. Chiang *et al.* [108] found that a tutorial intervention to reveal the limitations of the AI system can effectively reduce over-reliance. Others have explored the role of task factors such as task complexity and uncertainty in shaping trust and reliance in human-AI decision-making [220, 278]. Buccina *et al.* [55] proposed cognitive forcing functions to compel people to engage more thoughtfully with explanations along with AI advice. They found that such interventions can effectively mitigate over-reliance.

In previous work, researchers [107, 279–281] explored how different XAI methods may affect user understanding of an AI system, trust, and reliance. It is still unclear how the interaction interfaces to present XAI methods will substantially affect user understanding of an AI system, trust, and reliance. In this chapter, we propose to fill in such research gap and explore whether conversational XAI interface can facilitate user understanding of the AI system, which further contributes to increased trust and appropriate reliance.

6.2.2 Explainable AI

While deep learning-based AI systems have been recognized as powerful predictive toolkits, explainability has been a primary concern that prevents them from becoming widespread practice. According to GDPR, users of AI systems have the right to obtain meaningful explanations along with AI predictions [166]. Under such circumstances, researchers have proposed a diverse set of XAI methods like feature attribution explanations [159, 186], counterfactual explanations [282], and contrastive explanations [283, 284]. For a more comprehensive review of existing XAI methods and criteria to evaluate XAI methods, we encourage readers to refer to recent work [217, 285].

As humans have diverse information needs, there is no one-size-fits-all solution [27]. With a proposal of putting users/humans at the center of technology design [104, 126], more and more researchers have started to explore human-centered XAI [27, 127]. In such line of literature, researchers focus on the function of explanation — how explanations affect user understanding and what characteristics make explanations effective [190, 286]. The mental model [287] denotes how one person build an internal representation of the external reality,[1] and plays an important role for analyzing human-centered XAI [119, 288–290]. Through empirical user studies, researchers found that many properties of explanations like simplicity [190], completeness [289] will substantially affect user mental model and the effectiveness of explanations.

According to Jacovi *et al.* [259], effective explanations should produce **coherent** mental models (*i.e.,* communicate information which generalizes to contrast cases), be **complete** to avoid misunderstanding and be **interactive** to address contradictions. We recognize that conversational XAI interfaces can satisfy all the above key properties for providing effective explanations. Thus, we argue that a conversational XAI interface may benefit users with a better understanding of the AI system, which can further facilitate user trust and appropriate reliance. Existing work has explored conversational XAI interfaces in the contexts of collaborative scientific writing [269] and decision support with a focus on team performance [266]. None of the existing works, however, have systematically explored the impact of conversational XAI interfaces on trust and appropriate reliance. To fill this knowledge and empirical gap while complementing existing efforts, we designed a controlled study with loan approval tasks to analyze the impact of a conversational XAI interface on human-AI decision making.

6.2.3 Conversational User Interfaces

A conversational user interface (CUI) is a user interface for computers that emulates a conversation with a real human [291]. CUIs have been studied widely across multiple disciplines, such as natural language processing, human-computer interaction, and artificial intelligence. Since the famous *Turing Test* [292], the capability to conduct human-like conversation has for long been recognized as an important property of artificial intelligence. Researchers have shown great enthusiasm for developing intelligent conversational user interfaces. CUIs have been widely adopted in crowdsourcing [293], dialogue systems [294], search engines [295], and recommender systems [296, 297]. Nowadays, conversational assistants like Apple Siri, Amazon Alexa, and ChatGPT have shown promising potential in assisting users in their daily life and work.

The main benefits of conversational user interfaces are the natural interaction experience that they facilitate [298], improved user engagement [293], better understandability [270] and accessibility. Compared with traditional graphical user interfaces (GUIs), CUIs have the advantages of more human-like interaction [264], simplifying complex tasks with filtered information [265], and leading to a higher subjective trust in the system [299]. Informed by these prior works, we infer that a conversational XAI interface can have similar advantages over a conventional XAI Dashboard (*i.e.,* a GUI to access current XAI methods). With conversational XAI interfaces, users may better understand the AI system and develop higher trust and more appropriate reliance on the AI system.

[1] https://en.wikipedia.org/wiki/Mental_model

Compared with these studies, our focus is to analyze the impact of the XAI interfaces (*i.e.,* an XAI dashboard and a conversational XAI interface) on human-AI decision making. While several works [266, 267, 269, 300] have positioned the conversational XAI interface as a promising direction to support human-AI collaboration, this is still an under-explored research topic that requires more empirical studies.

6.3 Task, Method, and Hypotheses

In this section, we describe the loan approval task and present our hypotheses, which have been preregistered before data collection.

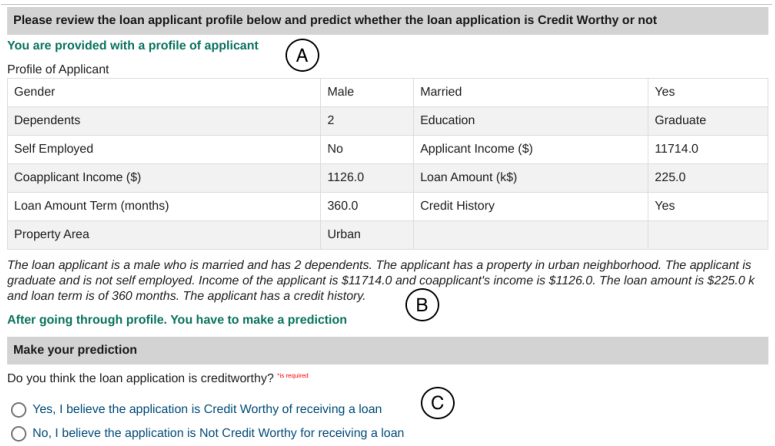


Figure 6.1: Screenshot of the loan approval task interface. This is the first stage of decision making. (A) Loan Applicant profile is shown in the table with 11 features. (B) To help understand the tabular data, we also provided a textual description below. (C) After going through the profile, participants are asked to decide whether this loan application is '**Credit Worthy**' or '**Not Credit Worthy**.'

6.3.1 Loan Approval Task

The basis for our experimental setup is a task where participants have to decide whether a loan application is **Credit Worthy** or **Not Credit Worthy** using the publicly available loan prediction dataset.[2] The rationale for selecting the loan approval task as a test bed is three-fold. Firstly, this task was chosen as a critical decision making scenario for human-AI collaboration, where there is a clear risk and a benefit when adopting AI advice. Secondly, most laypeople are familiar with this context and can make informed decisions based on their knowledge. Thirdly, It has also been adopted by existing research in behavioral economics [83] and human-AI collaboration [65, 234].

In the loan approval task, participants are presented with eleven features (including loan amount, income, and the absence or presence of credit history) in both table format and text description (as shown in Figure 6.1). Based on the application profile (composed of

[2] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

Table 6.1: Conversation setup to trigger different XAI responses. Different XAI methods can correspond to different information needs identified in the XAI question bank [257]. Queries correspond to the options provided in the conversational XAI interface.

XAI method	Information needs	Queries	User Input	XAI Response
PDP	How	How does [a given feature] influence credit worthiness in general?	Feature Dropdown Selection	Figures illustrating probability distribution when varying specific features and description messages
SHAP	Why	What are the most important features influencing the current prediction?	N/A	Figures illustrating the relative importance of all the features and description messages
MACE	Why, Why not, How to be that	What is the minimum change in the applicant's profile needed to switch the current prediction?	N/A	Text Description of minimum change in the profile
WhatIf	What if, How to be that, How to still be this	What would happen to the credit worthiness for [a different input]?	Feature Values	Model prediction on a new profile
Decision Tree	Why, How to still be this	Which sequence of steps led to the current prediction?	N/A	Figures illustrating the decision path and description message

the eleven features), participants are asked to decide whether the loan applicant is credit worthy to get the loan approved. This simulates a realistic scenario where participants interact with an AI system and may rely on AI advice and XAI methods due to the inherent complexity in decision-making [301]. As the selected loan approval task is one where decision making is fully based on the eleven features, it would be easier to assess users' decision criteria based on the top-ranked features explicitly specified by the users themselves.

Task Selection. All participants in our study were presented with ten loan approval tasks in the main task phase. All such cases are selected from the test set of a random split of the full dataset (training / test ratio 4:1). All tasks were evenly split between those where the loan applicant should be **Credit Worthy** (**CW**) for the loan being approved and those where the applicant profile should be **Not Credit Worthy** (**NCW**). As shown in Table 6.2, we selected the ten tasks according to prediction correctness and model confidence. We first trained an XGBoost Classifier [302] based on the training set. For both **CW** cases and **NCW** cases, we selected one high-confidence correct prediction, one random-confidence correct prediction, one low-confidence correct prediction, and one high-confidence wrong prediction. While we adopted another random-confidence correct prediction for class **NCW**, we selected another low-confidence wrong prediction for class **CW** to control the accuracy of the AI system to be 70%. This experimental design was also informed by a pilot study without AI advice. We recruited 20 participants from the Prolific platform to work on the selected loan approval tasks, and found that they achieved an accuracy level around 60%. To ensure the AI system is helpful to improve human decision making accuracy and maintain the risk of accepting wrong advice, we manually controlled the accuracy of the AI system to be 70%. During the study, we randomly shuffled the task order for each participant to prevent ordering effects [84].

Table 6.2: Task selection criteria for our study. 'CW' and 'NCW' refer to **Credit Worthy** and **Not Credit Worthy**, respectively.

Task ID	Groud Truth	Correctness	Model Confidence
1	CW	✓	High
2	CW	✓	Low
3	CW	✓	Random
4	CW	×	Low
5	CW	×	High
6	NCW	✓	High
7	NCW	✓	Low
8	NCW	✓	Random
9	NCW	✓	Random
10	NCW	×	High

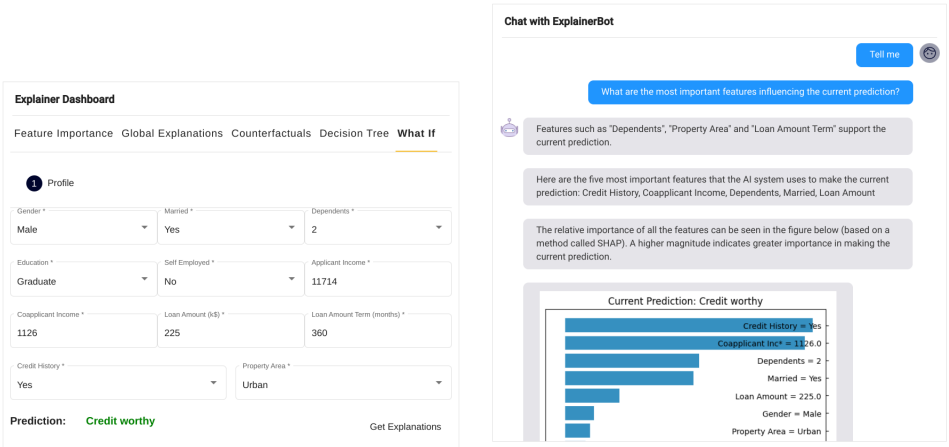
Two-stage Decision Making. In our study, we adopted a two-stage decision making process for each loan approval task. Every participant in our study is first asked to work on the loan approval task without any assistance from the AI system. After that, they were given a second chance to alter their initial choice according to the AI advice (*i.e.,* AI prediction) and AI explanations (*e.g.,* XAI dashboard, according to different experimental conditions). This setup is similar to the update condition in work by Green *et al.* [65]. This setup is apt for analyzing user incorporation of system advice and user trust in the AI system [35, 85]. It is a widely adopted setup in empirical studies exploring human-AI decision making [30, 49, 107, 108]. To assess user decision criteria, we ask users to indicate the three most important features influencing their decision at each stage along with their confidence in each decision.

6.3.2 Design of XAI Interfaces

XAI methods. Our selection of XAI methods is informed by the taxonomy of XAI methods regarding user information needs [107, 257, 285]. Following the XAI question bank [257], we selected six user information needs associated with the rationale of AI advice: *how* (global model-wide explanation), *why, why not, how to be that* (a different prediction), *how to still be this* (current prediction), and *what if*. These user information needs can be addressed with five widely-used XAI methods (correspondence summarized in Table 6.1). These are (1) A global explanation method – PDP (*i.e.,* partial dependency plot) [303], which visualizes how one feature globally impacts the model prediction, (2) Feature importance attribution method – SHAP [186]. Based on Shapley values, the SHAP method provides feature importance to indicate how each feature supports or opposes the current model prediction. (3) Counterfactual explanation method – MACE [258]. MACE will inform users of the minimum changes in the applicant profile required to flip model prediction. (4) Widely adopted interactive XAI toolkit – WhatIf.[3] Based on the WhatIf toolkit, users can modify the applicant profile and obtain the model prediction for the

[3] https://pair-code.github.io/what-if-tool/

new profile. (5) Decision tree-based explanation.[4] This is one popular XAI method, which makes decisions based on a tree-structure decision criteria. In our implementation, we provide the decision path to reach the AI advice. We implemented all these XAI methods by using the OmniXAI library.[5] More details can be found in supplementary materials.



(a) XAI Dashboard with WhatIf Response.

(b) Conversational XAI interface with SHAP Response.

Figure 6.2: Screenshots illustrating the XAI interfaces we designed. Additional screenshots demonstrating all XAI methods across both XAI interfaces are available in the supplementary materials.

XAI Dashboard. Following existing standards, the XAI dashboard is an interactive interface that provides users with XAI responses on demand when accessed through the navigation tab (see Figure 6.2a). Users can explore all XAI methods by focusing on one at a time, which ensures both simplicity and complete coverage of the available five XAI methods.

Conversational XAI Interface. Templating conversational interactions via a rule-based agent [304] can be an effective method to guide users in exploring their information needs and understanding the model decisions. Thus, we adopted a rule-based conversational agent to power the conversational XAI interface. By referring to the XAI question bank [257], we first set up five user intents (see Table 6.1), which can be answered with the corresponding XAI responses.

To provide a smooth conversational experience, we curated the five user intents into three categories: about AI advice (SHAP, MACE, Decision Tree — XAI responses required no user input), AI advice for modified applicant profile (WhatIf, where users need to revise the applicant profile), and the global impact of a specific feature (PDP, where users need to specify a feature of interest). At the beginning of the conversation, users are guided to select one category among the three and then specify one query to check or specify user input. After users receive one XAI response, we repeat the aforementioned process.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[5] https://github.com/salesforce/OmniXAI

All user intents are wrapped into an iterative loop, and users can stop the conversation after receiving at least two different XAI responses. All the conversations are guided by empowering participants to select options using custom buttons and commands (*i.e.,* dropdown selection for PDP or feature input for WhatIf, shown in Figure 6.2a). Such designs have been widely adopted in domains such as conversational crowdsourcing [293, 305], or customer service chatbots and proven to be effective in addressing user information needs and are easy to use for laypeople [306].

Evaluative Conversational XAI Interface for Decision Support. Based on the collected user decision criteria in the initial decision, we further adapted the conversation to guide users to check such features (*i.e.,* top-3 features selected in the initial decision making). This is inspired by the evaluative AI for explainable decision support [307], which argues for 'providing evidence for and against decisions made by people.' Such evaluative conversational XAI interfaces nudge users to think about their initial decision criteria further by comparing them with explanations from AI systems. To this end, it is similar to cognitive forcing functions [55], which has been adopted to calibrate user trust and reliance behaviors.

To achieve the goal of evaluative decision support in our conversational XAI interface, we adopted guiding messages in the customized buttons with user decision criteria (*i.e.,* the top-3 features the user selected in initial decision making). For XAI methods that require user input (*i.e.,* PDP and WhatIf), we adapted the guiding message with user decision criteria. For example, instead of selecting one option for PDP, users have an extra option to directly explore how one of the selected features influences credit worthiness. We believe that doing so can help them to explore how the selected features will affect the model prediction. After obtaining the XAI response, the conversational assistant sends a message to check whether the user wants to continue exploring the current XAI method by either modifying or selecting a feature randomly sampled from user decision criteria. In the case of SHAP, MACE, and Decision Tree (*i.e.,* XAI methods which do not require user input), the conversational assistant sends a message about how their initial decision criteria work in current XAI methods, serving as evaluative feedback. Similarly, this message helps them to check how their decision criteria differ from the AI system (as reflected by explanations provided via the XAI methods). After users obtain the SHAP, MACE, or Decision Tree XAI response, the conversational assistant provides an extra option message to guide them to explore the PDP (*i.e.,* global explanation on feature variation) response with one randomly selected feature from their initial set of top-3 features.

Conversational XAI Interface with LLM Agents.[6] While rule-based agents can inform the flow in conversational interactions, they lack the flexibility to deal with user needs in a bilateral human-like conversation. To address such concerns and further our understanding of the impact of flexible interaction and enhanced conversation quality in the conversational XAI interface, we built another conversational XAI interface powered by LLM agents. The benefits of introducing LLM agents are two-fold: (1) LLMs have shown promising user query understanding capability, which enables understanding user information needs and generating coherent and high-quality personalized conversation

[6]To notice that, the conversational XAI interface supported with LLM agents was adopted as a follow-up comparison with other conditions. In the pre-registration, we only include samples and hypotheses associated with other XAI interfaces.

responses [13]. (2) When equipped with XAI methods as potential tools, LLM agents can provide suitable XAI responses on demand, which may provide a better user experience (*e.g.,* more flexible expression of information needs and high-quality text responses based on XAI outcomes).

Apart from the difference in agents (LLM agents in this case), the entire procedure is identical to the basic conversational XAI interface. Our implementation of the LLM agent is based on autogen [308] and GPT-4. Given user queries, the LLM agent-based conversational XAI transforms user intents into pre-defined explainers and elaborates on the generated explanations to generate coherent text responses. We also provide the five hint questions (as shown in Table 6.1) to trigger potential XAI responses during the conversation in a randomized order on every task. Users can ask the LLM agent any questions using textual input. For more implementation details of our LLM agent-based conversational XAI interface, readers can refer to our supplementary materials.

6.3.3 Hypotheses

Our experiment was designed to answer questions surrounding the impact of conversational XAI interfaces on user understanding, trust, and reliance on AI systems. XAI dashboards, which can switch between different XAI methods with a navigation bar, have been recognized as a promising interactive interface to present explanations towards model decisions [263, 309, 310]. Considering its wide application for model explainability, we consider it a strong baseline in our study. As shown in prior work, conversational user interfaces have the advantages of more human-like interaction [264] and simplified understanding of complex tasks with filtered information [265] over graphical user interfaces. Compared with the XAI dashboard (where users interact with the dashboard in a uni-lateral fashion), the conversational XAI interface has the potential to increase user engagement, and provides a more natural bi-directional way for users to explore their information needs and develop an understanding of the AI system. As a result, users with a conversational XAI interface may develop a better understanding of the AI system. Thus, we hypothesize that:

> **(H1)**: Compared to the XAI dashboard, the conversational XAI interface creates a better understanding of the AI system among users.

Prior work has highlighted that humans show higher trust when interacting with intelligent systems using a conversational interface compared to conventional web interfaces [299]. Further, conversational user interfaces have been shown to increase worker engagement in microtask crowdsourcing [293] compared to a traditional GUI. Such increased engagement can potentially help users deliberate, reflect, and thereby make better decisions, relying on the AI system more critically. Conversational XAI interfaces can help users explore and address different information needs, which may bring a higher trust in the AI system. Thus, we hypothesize:

(H2): Compared to the XAI dashboard, the conversational XAI interface will help users exhibit a relatively higher trust in the underlying AI system.
(H3): Compared to the XAI dashboard, the conversational XAI interface will help users exhibit a relatively more appropriate reliance on the underlying AI system.

Evaluative decision support in the XAI interface may further help users reassess their initial thoughts about the AI system and AI advice. By revealing the difference among their decision criteria and providing explanations for the AI system's advice, users can obtain a better understanding of the AI system and make more critical decisions [307]. This can in turn facilitate critical thinking about the AI system, leading to a potential calibration of user trust and increased appropriate reliance on the AI system. Thus, we hypothesize that:

(H4): Adaptive steering of conversations for evaluative decision support in the conversational XAI interface will increase user trust and appropriate reliance on an AI system.

6.4 Study Design

This section describes our experimental conditions, variables, and procedures related to our study. This study was approved by the human research ethics committee of our institution.

6.4.1 Experimental Conditions

The main aspects of our research questions and hypotheses concern the effect of different XAI interfaces. In our study, all participants worked on the loan approval tasks with a two-stage setup (described in Section 6.3.1), where AI advice is provided in the second stage of decision making. The only difference is the nature of the interface through which AI advice is explained. Considering this factor as the sole independent variable in our study, we designed a between-subjects study with five experimental conditions:

- `Control`: no XAI interface.
- `Dashboard`: with XAI dashboard interface (as described in Section 6.3.2).
- `CXAI`: with a conversational XAI interface (as described in Section 6.3.2).
- `ECXAI`: with a evaluative conversational XAI interface (as described in Section 6.3.2).
- `LLM Agent`: with a conversational XAI interface powered by LLM agents (as described in Section 6.3.2).

6.4.2 Measures and Variables

Our hypotheses mainly considered five types of dependent variables: user understanding, user trust, performance, reliance, and appropriate reliance on the AI system.

User Understanding of the AI System. This chapter focuses on analyzing the impact of the XAI interfaces instead of evaluating the quality of explanations [311]. In our study, user understanding of the AI system is a function of interactive exploration with the XAI interfaces, which can evolve while working on tasks. Note that we consider and describe perceived explanation utility as a separate construct below. Based on existing literature [312–315], we synthesized and adopted four dimensions to assess user understanding of the AI system. As a result of practice through our study, users can potentially learn across tasks and understand the system. We aim to capture this through the dimensions of *Perceived Feature Understanding*, *Learning Effect* across tasks, and *Understanding of the System*. All questionnaires used to assess user understanding can be found in supplementary materials. To objectively quantify user understanding of the features, we calculated nDCG [316] of users' top-3 features and the SHAP feature importance ranking as *Objective Feature Understanding*. For the relevance scores, we adopted a decreasing relevance for the SHAP feature order (based on the abstract value of SHAP values) with an interval of 1. Thus the relevance scores range from [1, 11] for the 11 features we used. Besides, *Perceived Feature Understanding* is also used as an indicator of perceived user understanding.

Explanation Utility. Alongside user understanding, the perceived explanation utility is an important aspect identified in the existing literature on human-centered XAI [27, 124, 127, 290]. We synthesized and adopted four dimensions based on existing literature to evaluate the explanation utility provided in conditions with XAI interface. According to Jacovi *et al.* [259], effective explanations can provide users with a coherent and complete mental model to explain the current AI prediction. Thus, we adopted the dimensions of *Explanation Completeness* and *Explanation Coherence* in our post-task questionnaires. According to Hsiao *et al.* [317], perceived *Explanation Clarity* and *Explanation Usefulness* are also important dimensions for assessing perceived explanation goodness.

User Trust. Mohseni *et al.* [318] showed that understandability and predictability are desired properties for trustworthy intelligent systems. Moreover, the perceived competence of the AI system (*i.e.,* users' confidence about the system's capabilities) and reliability of the AI system (*i.e.,* the extent to which the system is perceived not suffer from unexpected errors) are also identified as essential constructs to establish trust [319, 320]. In addition to capturing these attributes, we also captured subjective trust of users by adopting three validated subscales from the trust in automation questionnaire [321]. These are TiA-Reliability/Competence (TiA-R/C), TiA-Understanding/Predictability (TiA-U/P), and TiA-Trust in Automation (TiA-Trust). Each subscale is calculated as the average score (5-point Likert) across related questions. These measures have been shown to be meaningful to use in empirical studies of human-AI decision making [22, 234].

Performance and Reliance. As has been argued by prior work, assessing user reliance on the AI system when users agree with AI advice can be inaccurate [29]. Thus, we measure both performance and user reliance from two distinct standpoints. Besides the global user performance (*i.e.,* overall *Accuracy*), we also considered user performance when their initial choice disagreed with AI advice (*i.e., Accuracy-wid*). Similarly, we consider *Agreement Fraction* (*i.e.,* how often users agree with AI advice in their final decisions) as a global measure of reliance. We consider *Switch Fraction* (*i.e.,* how often users adopt AI advice in cases of initial disagreement) as another precise indicator of user reliance. To assess ap-

propriate reliance, we followed Schemer *et al.* [29] to adopt Relative positive AI reliance (*RAIR*) and Relative positive self-reliance (*RSR*) metrics. These measures enumerate all cases when the user initially disagrees with AI advice, but the correct decision is present in one of them. By calculating the positive reliance patterns among all potential actions, *RAIR* and *RSR* assess whether users know when they should rely on the AI system and themselves, respectively. To our knowledge, they are the most representative objective measures of appropriate reliance.

Other Variables. To dive deep into the impact of different XAI interfaces, we also considered other variables in our study. User confidence has been identified as an important factor in human-AI decision making [65, 114, 170]. In our study, we recorded user confidence in each stage of decision making tasks with the question–"*What is your confidence level while making this decision?*." As described in Section 6.3.3, the conversational XAI interface may benefit human-AI decision making with higher user engagement. To quantitatively analyze such impact, we adopted the UES-SF [322] questionnaire in our study and considered the average score across all dimensions as an indicator of user engagement.

6.4.3 Participants

Sample Size Estimation. To ensure that our empirical study has a sufficient sample size for statistical analysis, we computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [92]. To correct for testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$, a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and that we will investigate four different experimental conditions/groups. This resulted in a required sample size of 244 participants. We thereby recruited participants from the crowdsourcing platform Prolific.[7] As illustrated in Figure 6.3, participants were recruited continuously and randomly assigned to an experimental condition, simultaneously accommodating for potential exclusion until the required sample size was reached (as described below). As a result, 352 participants were recruited for conditions `Control`, `Dashboard`, `CXAI`, and `ECXAI`, of which 107 were excluded. In the experiment process, the `LLM Agent` condition was considered as a follow-up study, which is not included in the initial sample size estimation. For ease of comparison with other conditions, we recruited 61 valid participants for `LLM Agent` condition.

Compensation. All participants were rewarded with £4, amounting to an hourly wage of £8 deemed to be a "*good*" payment by the platform (estimated completion time was 30 minutes). On top of this basic payment, we rewarded participants with extra bonuses of £0.05 for every correct decision in the ten loan approval tasks. This bonus setting encourages participants to reach a correct decision to the best of their ability, which is also a contextual requirement to encourage appropriate system reliance [40].

Filter Criteria. All participants were proficient English speakers above the age of 18, and had finished over 40 tasks while maintaing an approval rate of over 90% on the Prolific platform. To ensure reliable participation, we employed attention check questions (one

[7] `https://www.prolific.co`

for decision making, three for questionnaires) in our study. All attention check questions explicitly direct participants to select a specific option. They were designed to look similar to the questions or decision making tasks they were embedded in [93]. If users read our instructions and engaged genuinely with the task, passing these attention check questions is straightforward. We excluded participants from our analysis if they failed at least one attention check or if we found any missing data. The resulting sample of 306 participants had an average age of 32 (SD = 7.8) and a gender distribution (53.6% female, 46.4% male).

6.4.4 Procedure

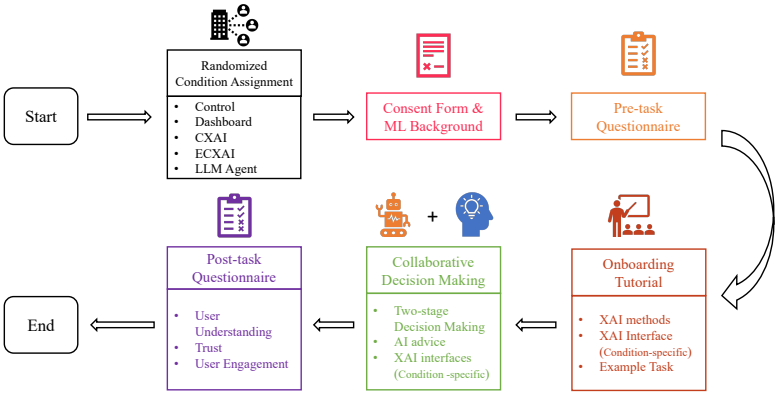


Figure 6.3: Illustration of the procedure that participants followed in our study. This flow chart describes the experimental condition `CXAI`.

The complete procedure participants followed in our study is illustrated in Figure 6.3. All participants will be first randomly assigned to one experimental condition. To proceed with participation, all participants were first asked to sign an informed consent form by clicking a button and also indicate their prior experience with machine learning. Next, participants were asked to complete a pre-task questionnaire to measure their affinity for technology interaction (*i.e.,* ATI). Then, an onboarding tutorial and a practice example were provided to help participants get familiar with the two-stage decision making setup and the corresponding XAI interface depending on the experimental condition.[8] At this stage, participants in the `Control` condition only see one practice example to get familiar with the loan approval task. Participants then worked on the ten selected tasks within a two-stage decision making setup. Finally, they were asked to fill in post-task questionnaires (including the TiA questionnaire and questions pertaining to user understanding of the AI system via the XAI methods).

6.5 Experimental Results

In this section, we present the results of our empirical study. In addition to the main results, we carried out exploratory analyses to draw nuanced interpretations of our key

[8]More details pertaining to the onboarding tutorial can be found in the supplementary material.

insights. Readers can refer to the appendix. Our code and data can be found at Github.[9]

6.5.1 Descriptive Statistics

To ensure the reliability of our results and interpretations, we only consider participants who passed all attention checks. Finally, the participants considered for analysis were distributed in a balanced manner across the four experimental conditions: 61 (`Control`), 61 (`Dashboard`), 62 (`CXAI`), 61 (`ECXAI`), 61 (`LLM Agent`). On average, each task consumes 13 API calls to obtain responses in `LLM Agent` condition, including generating reply messages and XAI usage. The average time (mins) spent across conditions are: 22 (`Control`), 34 (`Dashboard`), 52 (`CXAI`), 45 (`ECXAI`), 62 (`LLM Agent`). With Kruskal-Wallis H-tests and post-hoc Mann–Whitney test, we confirmed significance: `Control` < `Dashboard` < `CXAI`, `ECXAI` < `LLM Agent`.

Distribution of Covariates. The covariates' distribution is as follows: *ML Background* (22.5% with machine learning background knowledge, 77.5% without machine learning background knowledge), *ATI* (M = 3.99, SD = 0.90; 6-point Likert scale, *1: low, 6: high*), *TiA-Propensity to Trust* (M = 2.88, SD = 0.71; 5-point Likert scale, *1: tend to distrust, 5: tend to trust*), and *TiA-Familiarity* (M = 2.67, SD = 1.10; 5-point Likert scale, *1: unfamiliar, 5: very familiar*).

Performance Overview. On average across all conditions, participants achieved an accuracy of 64.5% (SD = 0.11), which is still lower than the AI accuracy (70%). The agreement fraction is 0.847 (SD = 0.16), and the switching fraction is 0.522 (SD = 0.41). With these measures, we confirm that when users disagree with AI advice, they do not always blindly rely on AI advice. As all dependent variables are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

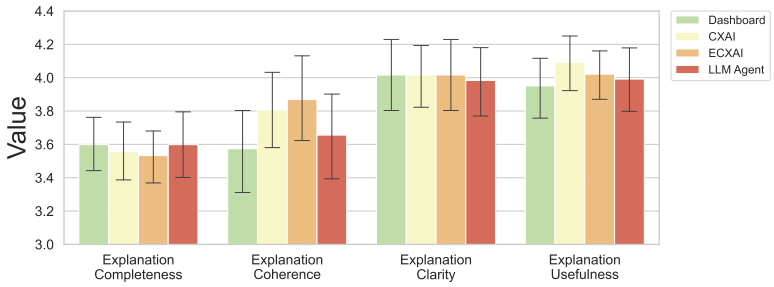


Figure 6.4: Bar plot illustrating the explanation utility across conditions. Error bars represent the 95% confidence interval.

Explanation Utility. To illustrate how the XAI interface will affect the perceived explanation utility, we adopted a bar plot of explanation utility across conditions. As shown in Figure 6.4, participants achieved similar level of *Explanation Completeness* and *Explanation Clarity*. Meanwhile, participants with conversational XAI interfaces (*i.e.,* condition `CXAI`, `ECXAI`, and `LLM Agent`) achieved slightly higher *Explanation Coherence* and *Explana-*

[9] `https://github.com/delftcrowd/IUI2025_ConvXAI`

tion Usefulness. Based on one-way ANOVA, we analyzed the impact of XAI interfaces in perceived explanation utility. There is no significant difference across conditions.

6.5.2 Hypothesis Tests

For the convenience of the readers, we have provided concise insights in the main body of this section and placed additional tables and figures (*e.g.,* estimation plots) that provide further details in the supplementary materials.[10]

Table 6.3: Kruskal-Wallis H-test results for XAI interfaces (**H3** and **H4**) on reliance-based dependent variables. The post-hoc results are based on Mann–Whitney tests. "††" indicates the effect of the variable is significant at the level of 0.0125.

Dependent Variables	H	p	$M \pm SD$					Post-hoc results
			Control	Dashboard	CXAI	ECXAI	LLM Agent	
Accuracy	9.09	.059	0.62	0.65	0.67	0.64	0.63	-
Agreement Fraction	33.66	**.000**††	0.74	0.86	0.89	0.85	0.89	Control < Dashboard, CXAI, ECXAI, LLM Agent
Switch Fraction	19.14	**.001**††	0.31	0.57	0.58	0.57	0.57	Control < Dashboard, CXAI, ECXAI, LLM Agent
Accuracy-wid	5.06	.281	0.46	0.50	0.52	0.55	0.42	-
RAIR	11.01	**.026**†	0.35	0.50	0.60	0.52	0.48	Control < CXAI
RSR	38.26	**.000**††	0.57	0.29	0.23	0.26	0.11	Control > Dashboard, CXAI, ECXAI, LLM Agent Dashboard > LLM Agent

H1: effect of XAI interfaces on user understanding

To analyze the main effect of the XAI interfaces on user understanding of the AI system, we conducted an *Analysis of Covariance* (ANCOVA) with the *experimental condition* as between-subjects factor and *TiA-Propensity to Trust*, *TiA-Familiarity*, *ATI*, and *ML Background* as covariates. While our data may not be normally distributed, we still adopted AN(C)OVAs for analysis because these analyses have been shown to be robust to Likert-type ordinal data [94]. For this analysis, we considered all participants across three experimental conditions with XAI (*i.e.,* `Dashboard`, `CXAI`, and `ECXAI`). We found no significant differences resulting from the different XAI interfaces (*i.e.,* experimental condition). However, the *TiA-Propensity to Trust* showed a significant impact on all dimensions of user understanding. For the objective feature understanding (continuous value, non-normal distribution), we conducted Kruskal-Wallis H-tests by considering different XAI interfaces. A significant difference ($H = 16.19, p = .001$) was found between participants with different XAI interfaces. Through post-hoc Mann–Whitney U test, we found that `LLM Agent` condition achieved significantly worse *objective feature understanding* than the `Dashboard`, `CXAI`, and `ECXAI` conditions. Thus, we did not find any support for **H1**.

H2: effect of XAI interfaces on user trust

To verify **H2** (*i.e.,* the impact of XAI interface on user trust), we conducted an *Analysis of Covariance* (ANCOVA) with the *experimental condition* as between-subjects factor and *TiA-Propensity to Trust*, *TiA-Familiarity*, *ATI*, and *ML Background* as covariates. This allows us to explore the main effects of the XAI interface on subjective trust as measured by the three subscales of the Trust in Automation questionnaire [321].

As we found, the experimental condition (*i.e.,* XAI interface) only showed a significant impact in **TiA-U/P**. With post-hoc Tukey's HSD test, we found that participants

[10] https://github.com/delftcrowd/IUI2025_ConvXAI/blob/main/supplementary_materials.pdf

who received XAI showed significantly higher trust in **Understandability/Predictability** (*i.e.,* `Control` < `Dashboard`, `CXAI`, `ECXAI`). Besides the significant results, participants in the `LLM Agent` condition showed a consistent but non-significant trend across all measures: `Control` < `LLM Agent` < `Dashboard`, `CXAI`, `ECXAI`. However, no significant difference is found between the `Dashboard` condition and conditions with conversational XAI. At the same time, there is no significant impact of the experimental conditions observed on the dependent variables of **TiA-R/C** and **TiA-Trust**. Meanwhile, we found that **TiA-Propensity to Trust** had a significant impact on all trust-related dependent variables, and that users' affinity to technology interaction (**ATI**) also had a significant impact on **TiA-U/P**.

To better understand effect sizes in terms of the TiA-U/P and go beyond *p*-values, we adopted an estimation plot [323] (shown in supplementary materials, Figure 3). As reflected by the swarm plot, participants with conversational XAI interface (*i.e.,* condition `CXAI` and `ECXAI`) exhibited a marginally higher TiA-U/P in comparison with condition `Dashboard`. Thus, we found partial support for **H2**.

H3: effect of XAI interfaces on appropriate reliance

To verify **H3**, we conducted a Kruskal-Wallis H-test to compare the performance, reliance, and appropriate reliance measures of participants across four experimental conditions. As shown in Table 6.3, participants showed significantly higher reliance (*i.e., Agreement Fraction* and *Switch Fraction*) with access to the XAI dashboard or conversational XAI interface. However, the increased reliance is not necessarily appropriate reliance. Only participants with access to conversational XAI interface (*i.e.,* condition `CXAI`) showed significantly better *RAIR* in comparison with the condition `Control`. We also found that participants showed significantly worse *RSR* with access to the XAI dashboard or conversational XAI interface. We also notice that participants in the `LLM Agent` condition showed significantly worse RSR compared to the `Control` and `Dashboard` conditions, which indicates that the `LLM Agent` condition led to severe over-reliance on the AI advice. Thus, **H3** is not supported by our experimental results.

There is no significant difference in team performance (*i.e., Accuracy* and *Accuracy-wid*). To interpret our data beyond *p*-values and better understand effect sizes in terms of the overall team performance, we adopted estimation plots [323] (shown in supplementary materials, Figure 4). Based on the normal distribution sampled for these measures, we can infer the reliance difference based on the mean difference of the estimated distribution. We found that: (1) Compared to the `Control` condition, participants in the `CXAI` condition showed a clearly higher mean accuracy. (2) Participants in the `ECXAI` condition showed slightly better *Accuracy-wid* than the `Dashboard` condition and the `CXAI` condition. Similarly, we adopted estimation plots [323] (cf. supplementary materials, Figure 4) to draw meaningful interpretations related to our appropriate reliance measures. We found that: (1) Compared to the `Control` condition, participants in the `CXAI` condition showed a significantly higher *RAIR*. At the same time, participants in the `CXAI` condition showed a slightly higher *RAIR* compared with participants in `Dashboard` and `ECXAI` conditions. (2) Participants in the `Dashboard` and `ECXAI` conditions showed slightly better *RSR* than the `CXAI` condition.

H4: effect of evaluative conversation on user trust and appropriate reliance
According to results reported for **H2** and **H3**, no significant difference in user trust and appropriate reliance was found between experimental condition `CXAI` and `ECXAI`. Thus, **H4** is not supported.

6.5.3 Additional Exploratory Analyses

Impact of Covariates

As shown in the analysis for **H2** (cf. Table 6.4), covariates like TiA-Propensity to Trust and ATI have shown some impact on user trust. To further analyze the impact of covariates on human-AI decision making, we conducted Spearman rank-order tests between covariates and all categories of dependent variables. The results are shown in Table 6.4. We have the following main findings: (1) Overall, *TiA-Propensity to Trust* significantly positively impacted most dependent variables in user understanding, trust, and reliance categories. (2) While the propensity to trust positively correlated with user reliance (*i.e., Agreement Fraction* and *Switch Fraction*), it negatively affects *RSR*. In other words, some participants with a higher propensity to trust tend to over-rely on the AI system. (3) *TiA-Familiarity* and *ATI* only showed some positive impact on user understanding and user trust. No significant correlation was found for user reliance. (4) *ML background* showed positive correlation with user trust. Meanwhile, some dimensions of explanation understanding also show a borderline positive correlation

Table 6.4: Correlation of covariates and dependent variables. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.0125, respectively.

Covariates	**Propensity to Trust**		**TiA-Familiarity**		**ATI**		**ML background**	
Dependent Variables	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Perceived Feature Understanding	0.344	**.000**††	0.131	.041†	0.148	.021†	0.049	.444
Explanation Completeness	0.366	**.000**††	0.106	.097	0.073	.254	0.152	.017†
Explanation Coherence	0.387	**.000**††	0.131	.040†	0.087	.175	0.135	.035†
Explanation Clarity	0.427	**.000**††	0.069	.285	0.129	.044†	0.142	.026†
Learning Effect Across Tasks	0.232	**.000**††	0.173	**.007**††	0.115	.072	0.147	.021†
Understanding of System	0.343	**.000**††	0.082	.202	0.146	.022†	0.080	.210
Explanation Usefulness	0.423	**.000**††	0.166	**.009**††	0.172	**.007**††	0.083	.196
Objective Feature Understanding	0.108	.092	-0.152	.017†	0.013	.844	-0.024	.714
TiA-R/C	0.677	**.000**††	0.126	.028†	0.171	**.003**††	0.153	**.008**††
TiA-U/P	0.472	**.000**††	0.083	.150	0.243	**.000**††	0.158	**.006**††
TiA-Trust	0.774	**.000**††	0.235	**.000**††	0.154	**.007**††	0.164	**.004**††
Accuracy	0.091	.111	0.073	.202	-0.039	.502	-0.019	.740
Agreement Fraction	0.223	**.000**††	0.055	.335	0.030	.598	-0.039	.499
Switch Fraction	0.137	.016†	-0.030	.595	-0.001	.982	0.037	.518
Accuracy-wid	0.056	.326	0.032	.582	-0.045	.434	0.057	.322
RAIR	0.118	.040†	-0.001	.980	-0.026	.648	0.026	.654
RSR	-0.186	**.001**††	-0.024	.674	-0.080	.162	-0.038	.505

The Impact of User Perceptions on Their Behavior

Prior work has shown that user trust can substantially affect user reliance behaviors [40, 87]. To further analyze how perception-based variables (*i.e.,* user trust, user understanding, and explanation utility) affect team performance and user reliance behaviors,

we conducted Spearman rank-order tests between corresponding categories of variables. The results are presented in Table 6.5.

Table 6.5: Correlation between perception-based variables (*i.e.,* user understanding, explanation utility, and user trust) and behavior-based variables. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.0125, respectively.

Behavior-based Variables	**Accuracy**		**Accuracy-wid**		**Agreement Fraction**		**Switch Fraction**		**RAIR**		**RSR**	
Perception-based Variables	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Perceived Feature Understanding	0.045	.484	-0.024	.709	0.254	**.000**††	0.117	.067	0.096	.135	-0.293	**.000**††
Objective Feature Understanding	0.332	**.000**††	0.195	**.002**††	0.469	**.000**††	0.322	**.000**††	0.269	**.000**††	-0.297	**.000**††
Learning Effect Across Tasks	0.084	.192	-0.085	.184	0.170	**.008**††	-0.006	.931	0.007	.913	-0.135	.035†
Understanding of System	0.114	.076	-0.083	.197	0.157	.014†	0.010	.877	-0.017	.795	-0.153	.016†
Explanation Completeness	0.050	.435	0.056	.387	0.146	.022†	0.142	.026†	0.157	.014†	-0.170	**.007**††
Explanation Coherence	0.107	.095	-0.030	.643	0.270	**.000**††	0.068	.286	0.005	.935	-0.218	**.001**††
Explanation Clarity	0.002	.973	-0.111	.083	0.190	**.003**††	0.081	.204	0.042	.514	-0.235	**.000**††
Explanation Usefulness	0.125	.051	0.081	.206	0.361	**.000**††	0.266	**.000**††	0.229	**.000**††	-0.300	**.000**††
TiA-R/C	0.127	.047†	0.090	.162	0.224	**.000**††	0.195	**.002**††	0.175	**.006**††	-0.200	**.002**††
TiA-U/P	0.099	.123	0.051	.430	0.210	**.001**††	0.132	.038†	0.125	.051	-0.182	**.004**††
TiA-Trust	0.145	.024†	0.032	.617	0.254	**.000**††	0.164	**.010**††	0.152	.017†	-0.203	**.001**††

We found that: (1) *Agreement Fraction* and *RSR* are significantly correlated with most dimensions of user understanding, explanation utility, and user trust. However, these dimensions are positively correlated with *Agreement Fraction* but negatively correlated with *RSR*. This suggests that the improved user understanding, explanation utility, and user trust with XAI interfaces can partially explain the increased over-reliance on the AI system. (2) While user trust dimension *TiA-R/C* and *TiA-Trust* positively correlated with reliance measures (*Agreement Fraction* and *Switch Fraction*), and *RAIR*, they negatively correlated with *RSR*. As a result, they do not show a significant correlation with *Accuracy-wid*. This corroborates that higher user trust in the AI system does not necessarily translate into appropriate reliance behaviors. (3) Overall, *Objective Feature Understanding* seems useful to facilitate appropriate reliance. With a higher *objective Feature Understanding*, participants demonstrate better team performance and higher reliance. Although it still contributes to over-reliance (reflected by negative correlation with *RSR*), it shows a more positive impact on appropriate reliance (*i.e., Accuracy-wid* and *RAIR*). In comparison, the positive impact of *Explanation Usefulness*, *TiA-R/C*, and *TiA-Trust* on mitigating under-reliance (*i.e.,* positive correlation with *RAIR*) get canceled by the side effect of over-reliance (*i.e.,* negative correlation with *RSR*). As a result, these variables do not significantly contribute to team performance.

Confidence Dynamics

As shown in Figure 6.5, we illustrate the confidence dynamics of participants in each condition along with the task order. In general, we found that participants reported a higher confidence after being exposed to AI advice and explanations. While participants in the `Control` condition, the `Dashboard` condition, and the `ECXAI` condition reported a fluctuating trend of confidence along the task order, participants in the `CXAI` condition reported a relatively clear ascending trend of confidence both before and after the AI advice (and explanations). Participants in the `LLM Agent` condition showed a clear upward and then downward trend in their confidence related to their final decisions. This suggests that participants in this condition first developed over-confidence in the AI system and then calibrated their confidence. Interestingly, we observed that the confidence dynamics of

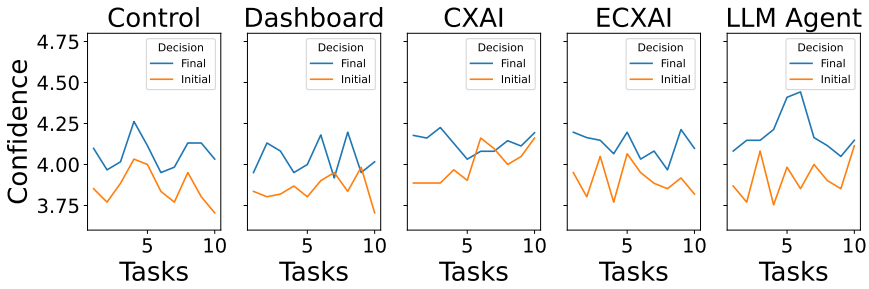


Figure 6.5: Line plot illustrating the confidence dynamics among users after receiving the AI advice (and explanations). The orange line and blue line illustrate the confidence dynamics before and after receiving AI advice (and explanations), respectively.

participants in the CXAI condition converge after a few tasks. The narrow confidence gap before and after receiving AI advice may indicate that participants in the CXAI condition calibrate their confidence in the AI advice, which reflects a better understanding of the AI system. To compare the confidence across conditions, we conducted ANOVA tests for both initial confidence (average across tasks) and final confidence (average across tasks). Although the CXAI, ECXAI, and LLM Agent conditions showed slightly better user confidence on average, we found no significant differences across conditions.

Further Analysis of User Engagement

We measured subjective user engagement reported by each participant in our study using the UES-SF questionnaire [322]. The distribution of user engagement across the different experimental conditions was as follows: Control ($M = 3.15, SD = 0.72$), Dashboard ($M = 3.33, SD = 0.66$), CXAI ($M = 3.20, SD = 0.63$), ECXAI ($M = 3.28, SD = 0.67$), LLM Agent ($M = 3.44, SD = 0.71$). While participants in the LLM Agent condition reported slightly higher engagement with the XAI interface, we found this to be non-significant (based on ANOVA analysis).

Further Analysis of Enhanced Conversation and XAI Usage

To compare how enhanced conversation (*i.e.,* adaptive steering for evaluative decision support and more flexible conversational interactions with LLM agents) affects user interaction with the conversational interface, we analyzed the usage of the XAI methods. To compare the usage of each XAI method, we conducted a Kruskal-Wallis H-test for total usage per participant. Across all five XAI methods, no significant differences in usage frequency were found between the CXAI and ECXAI conditions. The most obvious difference is that participants in the CXAI and ECXAI conditions used PDP method significantly more frequently: CXAI($M = 13.5$), ECXAI($M = 14.1$), LLM Agent($M = 3.6$). Meanwhile, participants in the LLM Agent condition showed significantly more usage of WhatIF, MACE, and SHAP methods than the CXAI and ECXAI conditions. The reason for such difference in the usage of XAI methods can be caused by the design of the rule-based conversational agent in the CXAI and ECXAI conditions. In the rule-based conversation agents, all messages are pre-defined, and users see them in a fixed order. Such fixed order may have biased user selection of the XAI responses. In comparison, the hint questions are randomized in con-

dition LLM Agent, and users can also use the free text input to ask anything they prefer. As a result, participants in the LLM Agent condition may have more flexible access to explore personalized information needs.

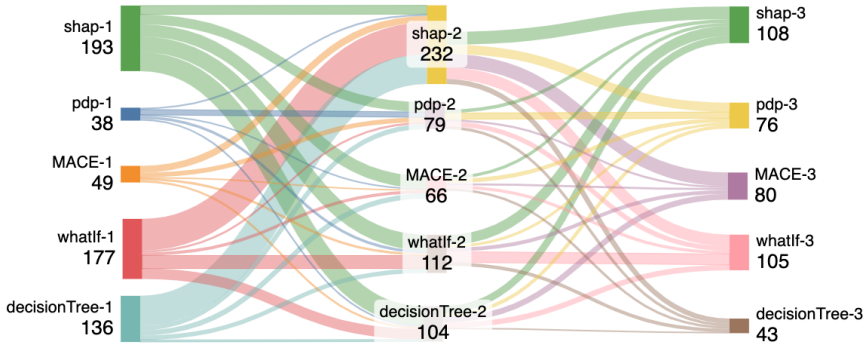


Figure 6.6: Illustration of the XAI usage used in our study. This Sankey diagram describes the sequence of interactions with XAI methods by users in the LLM Agent experimental condition.

To obtain further insights, we explored the user conversation history in the LLM Agent condition. Among all 61 users in LLM Agent , 1,946 user queries are asked in total. Among them, around 40% are based on the hint questions (5 questions we provide to trigger XAI responses, see Table 6.1). The valid user queries mainly consist of three types of intent: user queries to obtain XAI responses (*e.g.,* hint questions and some similar questions), greetings (*e.g.,* "Hi", "Thank you"), and opinion-seeking queries to the conversational agent (*e.g.,* "Do you think the loan application is creditworthy?"). When meaningless user queries are fired (such as gibberish, random strings or something irrelevant to our task context), the LLM agent-based conversational interface can handle them properly (*e.g.,* "I do not understand this. Please check information related to the current task."). To visualize the dynamics of user information needs along with exploring conversation, we adopted the Sankey diagram (Figure 6.6) to show the dynamic flow of XAI usage. Only a few participants in the LLM Agent condition asked for more than three XAI responses in each task, so we only considered the first three usages of XAI methods. As we can see, after using one XAI method, participants tend to use a different XAI method in the next step, which indicates that most participants explored diverse information needs in the LLM condition LLM Agent.

6.6 Discussion

6.6.1 Key Findings

Our experimental results show that participants with an interactive XAI interface (*i.e.,* either an XAI dashboard or a conversational XAI interface) can obtain a relatively high degree of perceived understanding, trust, and reliance on the AI system. However, the increase in trust and reliance may potentially stem from an illusion of their understanding of explanatory depth [62, 324]. As a result, they do not necessarily know when the AI advice is trustworthy and worth relying on. This is reflected by the over-reliance we observed (see Table 6.3) in all conditions with interactive XAI interfaces. with an LLM

agent-based conversational XAI interface (Section 6.5.2), we observed that over-reliance was further reinforced (*i.e.,* worse *RSR*) and users obtained significantly worse *objective feature understanding* compared to other conditions with XAI interfaces. This indicates that instead of calibrating user trust and reliance on the AI system, enhancing the conversation quality may further induce the illusion of explanatory depth. While no significant results are observed to support the superiority of conversational XAI interface over XAI dashboard, our exploratory analyses revealed the potential of conversational XAI interfaces (powered by LLMs) in increasing user exploration of the explanation methods. Participants with the conversational XAI interface reported a slightly better perceived user understanding and perceived explanation utility. As for trust and appropriate reliance, we see that participants showed a slightly higher trust (cf. Section 6.5.2), team performance (cf. Section 6.5.2), and relatively higher *RAIR* (cf. Table 6.3). We also found that participants with a conversational XAI interface (`CXAI`, `ECXAI`, and `LLM Agent` conditions) did not report a higher user engagement than participants with an XAI dashboard, suggesting that both the interactive interfaces are equally effective in engaging the participants.

Positioning in Existing Literature. In our study, we found that interactive XAI interfaces can have a negative impact of increasing over-reliance on the AI system. This is consistent with the findings of previous empirical studies of human-AI collaboration [22, 66, 107]. Our results indicate that participants perceive the conversational XAI interface to lead to a relatively better user understanding and team performance than the XAI dashboard. This is in line with findings of Slack *et al.* [263], where they found TalktoModel (a conversational XAI interface) was preferred by most participants and achieved better team performance when collaborating with users. We extend existing empirical work by going one step further to explore the impact of conversational XAI interfaces on trust and appropriate reliance. We found that users tend to show relatively higher trust and appropriate reliance on the conversational XAI interface. Further enhancement of the conversation (*i.e.,* adaptive steering for evaluative decision support) does not necessarily help further improve user understanding, user trust, and appropriate reliance on the AI system (*i.e.,* the `ECXAI` and `LLM Agent` conditions). Instead, we found that it can even be harmful (cf. Section 6.5.2), which is reflected by a decreased user understanding of the AI system, user trust, and appropriate reliance in the `LLM Agent` condition. Our exploratory findings suggest promising avenues for future research — further exploring how conversational XAI interfaces can affect user trust and reliance on the AI system through additional confirmatory studies in different contexts. Our work is an important first exploration to this end, and more empirical studies are required to corroborate and further contextualize these observations. As we strive towards optimal human-AI decision making, we highlight an important trade-off that needs to be managed between creating user-friendly, seamless, and plausible conversational XAI interfaces and simultaneously fostering critical consideration of AI advice.

6.6.2 Implications of Our Work

Interactive XAI Interfaces Can Amplify Illusions of Explanatory Paths. Our work has important theoretical implications for promoting appropriate reliance on AI systems with XAI methods. In our study, participants with the XAI dashboard as well as the con-

versational XAI interfaces showed obvious over-reliance on the AI system. The reason behind this can be that participants with XAI interfaces developed illusions of the intelligence level of the AI system. Prior work has shown that conversational interfaces can build user trust [299], and XAI can bring about an illusion of explanatory depth [62]. Both can contribute to uncalibrated trust in the AI system and cause over-reliance. Their combination could potentially amplify users' over-reliance depending on other task, human, and system factors. As our results suggested, participants with conversational XAI interface (*i.e.,* `CXAI`) showed slightly better perceived user understanding across multiple dimensions (non-significant results) and trust (*i.e.,* Understanding/Predictability) than participants with XAI dashboard. At the same time, participants in condition `CXAI` also showed the best *RAIR* and relatively worse *RSR* (see Table 6.3), while participants in the `LLM Agent` condition showed the worst *RSR* (see Section 6.5.2). Combined with exploratory findings in Table 6.5 — user understanding, explanation utility, and user trust is positively correlated with over-reliance. This indicates that the conversational XAI interface appears to be more persuasive to users and leads to relatively more over-reliance on the AI system. Thus, optimizing the XAI interfaces as a persuasive technology [325] may not be the ideal approach to promoting appropriate reliance on AI systems. In extreme cases, persuasive technology can even help untrustworthy AI systems deceive end users to gain their trust [326]. Instead, we should focus on developing methods and interfaces that can ensure that the XAI responses provided will not mislead users by creating an illusion of system intelligence or explanatory depth.

Why boosted conversations did not work as expected. In contrast to our expectation, boosted conversations (*i.e.,* in the `ECXAI` and `LLM Agent` conditions) did not provide further benefits in user understanding, trust, and appropriate reliance. According to the confidence dynamics (see Figure 6.5), enhanced conversation quality in condition `LLM Agent` seems to enlarge the confidence gap between the two stages of decision making (*i.e.,* before and after checking AI advice and XAI responses), especially when comparing the `LLM Agent` condition with the `CXAI` condition. Although the LLM-powered condition of `LLM Agent` was expected to lead to the most natural and personalized XAI responses among all conditions with XAI interfaces, participants in the `LLM Agent` condition demonstrated the least objective feature understanding, subjective trust, and appropriate reliance. Combined with the findings of confidence dynamics, we infer that introducing LLM agents to a conversational XAI interface may amplify the illusion of explanatory depth. As a result, participants in the `LLM Agent` condition exhibit high over-reliance on the AI system. Based on these findings, we argue it would be more important to align the plausibility of XAI responses with the trustworthiness of the AI system rather than solely improving the interactional quality and experiences with the XAI responses. This is in line with existing work on plausibility in XAI [327]: "a plausible but unfaithful interpretation may be the worst-case scenario." In comparison, the evaluative conversation enhances user self-reflection of their decision criteria. As a result, participants in condition `ECXAI` indicate a relatively lower *Agreement Fraction* and *RAIR* than condition `CXAI` (cf. Table 6.3). Thus, we can infer that the evaluative conversation brings about some side effects — under-reliance on the AI system. At the same time, the evaluative conversations fail to facilitate user understanding, calibrate user trust in the AI system, or mitigate over-reliance. Further research is required to understand how to provide suitable evaluative decision support in

conversational human-AI interactions.

Towards more effective conversational XAI interfaces. Our work has important implications for designing effective conversational XAI interfaces. Rather than being persuasive, we expect effective XAI interfaces to be accessible and low-barrier interfaces that can enhance user engagement and guide users to explore their information and explanation needs. As a result, users can have a better user experience, and a more comprehensive understanding of the AI system (*e.g.,* including both strengths and weaknesses), resulting in more appropriate reliance on the AI system. In our study, the conversational XAI interface failed to facilitate a significantly better user understanding, trust, and appropriate reliance. Based on our findings, there are multiple potential approaches to improve the effectiveness of the conversational XAI interface.

Firstly, the trustworthiness of AI advice should be calibrated within the conversation. As we found, the improved user experience and conversation quality do not necessarily translate into appropriate reliance. To that end, users need to be supported with faithful conversations, which may help them realize whether AI advice is trustworthy. To tackle the vulnerability of improved plausibility (*e.g.,* introducing LLMs or other persuasive technology), future work can explore how to align the trustworthiness of AI advice with the plausibility of conversational XAI responses. Secondly, conversational XAI interfaces could be used to address potential issues associated with AI literacy. Conversational interactions have been proven to be effective in supporting novice and low-literacy users in using mobile interfaces [328]. Prior work has shown that AI literacy plays an important role in calibrating user trust and reliance behavior [108]. Thus, leveraging conversational XAI interfaces to narrow down the literacy gap when working with AI systems can also be a promising future direction to explore. Thridly, although adaptive evaluative steering for evaluative decision support fails to facilitate optimal human-AI decision making, it leads to substantial impacts on user perception and user reliance behavior. For example, participants in condition `ECXAI` achieved slightly higher *Explanation Coherence*, slightly higher *Accuracy-wid* and decreased *Agreement Fraction* compared to condition `CXAI`. Such an evaluative AI [307] conceptual framework could still be a promising approach to facilitating human-AI interaction within a conversational manner. Future work can further combine such evaluative conversational XAI with cognitive forcing functions [55] through the dialogue to help calibrate user trust and reliance. Similarly, Ehsan *et al.* [253] proposed the framework of Seamful XAI to augment explainability and user agency in human-AI collaboration by revealing the "seams" (*i.e.,* imperfections of the AI system). Combined with these ideas, we can guide users to explore both the strengths and weaknesses of the AI system. Such a conversation may be more engaging and may potentially achieve similar functions as cognitive forcing functions [55] to help participants make decisions more critically. This is an important direction for future work.

6.6.3 Caveats and Limitations

In our study, we selected the most representative five XAI methods as the basis to form our interactive XAI interfaces. We cannot overrule that this design choice may have been a bottleneck for some participants in our study, as they may have had information needs that are not covered by the XAI methods. Once users find that their queries cannot be answered properly based on pre-defined XAI methods, their trust and reliance on the AI

system may decrease. Having said that, our setup is representative of current state-of-the-art AI-assisted decision making methods. In our study, the conversational XAI interfaces in the CXAI and ECXAI conditions are built upon rule-based dialogue systems. All conversations are guided in a pre-defined manner, which lacks flexibility in communication. We developed an LLM agent-based conversational XAI interface (*i.e.,* the LLM Agent condition) to select XAI methods on demand, improve the scope and quality of user interactions, and flexibly communicate the corresponding explanations. We found that more flexible and plausible conversations did not necessarily help further improve user trust and appropriate reliance on the AI system. Instead, it amplified over-reliance and negatively impacted user understanding of the AI system. Based on these results, we can infer that, improving the conversational quality by using more human-like utterances may be more persuasive and strengthen the illusion of explanatory depth.

According to prior studies about crowdsourcing [93], some participants can rush through the study and provide low-effort results. To alleviate participants with low-effort results, we adopted attention checks in the questionnaire and tasks in our study. Meanwhile, it would be challenging to keep participants engaged in the XAI interface and highly motivated to learn from the explanations of XAI responses. To ensure that participants spent enough effort to interact with the conversational XAI interface, participants were required to view at least two different types of XAI responses in each conversation. This was, however, not explicitly mentioned and participants were alerted to this only when they tried to proceed without engaging with the XAI methods.

Potential Bias. Our study is based on a crowdsourcing setup, which may be affected by cognitive biases introduced in the task design and workflow. With the help of the Cognitive Biases Checklist introduced by Draws *et al.* [157], we analyzed potential bias in our study. As crowd workers are motivated by monetary compensation, the *self-interest bias* is possible. As participants showed a relatively high degree of trust and *Agreement Fraction* with AI advice, *Confirmation Bias* may have also affected our results. The rule-based conversational agents in the CXAI and ECXAI conditions may bias the usage of XAI methods (see Section 6.5.3). As a result, the participants in the two conditions showed similar usage patterns of XAI methods, which may lead to similar user understanding and reliance patterns.

Broader Societal Implications. Our findings add to the urgency to be careful when employing AI-based decision support systems due to their tendency to act as persuasive technologies. Although evaluative conversations led to an increase in user trust and reliance in our study, contrary to expectations, this did not amount to an increased appropriate reliance. Future work can explore similar 'evaluative AI' [307] operationalizations in conversational human-AI interaction and decision support. We found that users' propensity to trust is strongly correlated with their subjective trust in the AI system and their appropriate reliance (cf. Section 6.5.2 and covariate analysis in supplementary materials). Participants with a higher propensity to trust showed significantly higher trust and reliance (*i.e., Agreement Fraction* and *Switch Fraction*) on the AI system. As a result, they were more likely to develop an illusion of explanatory depth and over-rely on misleading AI advice. Such a tendency to trust may have originated from a lack of AI literacy [108] and a critical mindset [31]. These results, along with recent findings in the IUI commu-

nity [108] suggest that the development and deployment of AI systems and XAI interfaces can systematically favor individuals with higher AI literacy or critical mindsets, and therefore cause disparities to others. Further work is required to ensure that different types of users (with varying AI literacy or differing individual traits) can equally benefit from AI systems and related interfaces.

6.7 Conclusion

In this chapter, we presented a first-of-its-kind empirical study to understand the impact of an XAI dashboard and a conversation XAI interface on user understanding of the AI system, and their further impact on user trust and appropriate reliance. Compared to participants with the XAI dashboard, participants with the conversational XAI interface showed a slightly better understanding (**RQ1**), and demonstrated a slightly higher trust in the AI system (**RQ2**). However, our findings suggest that the XAI interfaces were persuasive and have the potential to bring about an illusion of the AI systems' capability, which in turn increased over-reliance on the AI system. Moreover, we found that evaluative conversational interactions do not work as expected in facilitating user trust and understanding. With experimental results associated with conversational XAI interfaces powered with LLM agents, we found that boosting the conversation quality and flexibility (*i.e.,* with LLM-based conversational agent) may further reinforce over-reliance and hurt user understanding and user trust. Our insights and observations can inform the future design of conversational XAI interfaces to promote complementary human-AI collaboration. Conversational XAI interfaces should balance user engagement with seamful design requirements that can promote decision making that is married with critical reflection.

Our results indicate that we should be careful in presenting XAI methods with an interactive XAI interface, which may cause over-reliance on the AI system. While our experimental results do not provide support to our original hypotheses, more work is required to further contextualize the effectiveness of conversational XAI interfaces in shaping user understanding, trust, and appropriate reliance. As opposed to further improving user experiences with conversational XAI interfaces in the context of human-AI decision making, future work should first focus on mitigating the illusion of explanatory depth brought by the XAI methods.

6.8 Appendix

Questionnaire. To assess the user understanding of the AI system and explanation utility, we collected questionnaires shown below from participants:

- **Perceived Feature Understanding**:
  1. The explanations helped you improve and/or reinforce your understanding of the influential features.
  ☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
- **Understanding of the System**
  1. I can understand why the system provided specific explanations.
  ☐ Strongly disagree ☐ Disagree ☐ Neutral ☐ Agree ☐ Strongly Agree
- **Learning Effect across Tasks**
  1. My understanding of AI system and decision criteria improve over the tasks.

□ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree

To assess the explanation utility, we collected questionnaires shown below from participants:

- **Explanation Completeness**
  1. The explanations provide a sufficient rationale that supports the AI advice.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
  2. The explanations sufficiently express the uncertainty of the AI advice.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Explanation Coherence**
  1. The explanations you received are consistent with your initial expectations.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Explanation Usefulness**
  1. The provided explanations are useful in making final decision.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree
- **Explanation Clarity**
  1. Explanations are clear enough to inform my final decision.
  □ Strongly disagree □ Disagree □ Neutral □ Agree □ Strongly Agree

III

Enhancing User Control with Collaborative Workflows

7

Fine-grained Transparency and Appropriate Reliance

In recent years, the rapid development of AI systems has brought about the benefits of intelligent services but also concerns about security and reliability. By fostering appropriate user reliance on an AI system, both complementary team performance and reduced human workload can be achieved. Previous empirical studies have extensively analyzed the impact of factors ranging from task, system, and human behavior on user trust and appropriate reliance in the context of one-step decision making. However, user reliance on AI systems in tasks with complex semantics that require multi-step workflows remains under-explored. Inspired by recent work on task decomposition with large language models, we propose to investigate the impact of a novel Multi-Step Transparent (MST) decision workflow on user reliance behaviors. We conducted an empirical study (N = 233) of AI-assisted decision making in composite fact-checking tasks (i.e., fact-checking tasks that entail multiple sub-fact verification steps). Our findings demonstrate that human-AI collaboration with an MST decision workflow can outperform one-step collaboration in specific contexts (e.g., when advice from an AI system is misleading). Further analysis of the appropriate reliance at fine-grained levels indicates that an MST decision workflow can be effective when users demonstrate a relatively high consideration of the intermediate steps. Our work highlights that there is no one-size-fits-all decision workflow that can help obtain optimal human-AI collaboration. Our insights help deepen the understanding of the role of decision workflows in facilitating appropriate reliance. We synthesize important implications for designing effective means to facilitate appropriate reliance on AI systems in composite tasks, positioning opportunities for the human-centered AI and broader HCI communities.

7.1 Introduction

With the rapid development of artificial intelligence (AI) in recent years, there is a growing recognition of the promising value of AI assistance [329, 330]. AI systems have

This chapter is based on a peer-reviewed paper: **Gaole He**, Patrick Hemmer, Michael Vössing, Max Schemmer, Ujwal Gadiraju. *Fine-Grained Appropriate Reliance: Human-AI Collaboration with a Multi-Step Transparent Decision Workflow for Complex Task Decomposition.* Under review at CSCW 2026. https://arxiv.org/abs/2501.10909.

been used to answer knowledge-intensive questions [331], provide recommendations in e-commerce platforms [332], and even help make critical decisions [333]. While AI systems promise high effectiveness and use across domains, there is no guarantee of their correctness [22, 28]. Thus, accountability and verifiability become a major concern before adopting such systems into existing workflows. To address these concerns, researchers and practitioners are actively exploring the potential of human-AI collaboration [22, 123, 334, 335]. However, human-AI collaboration is not always effective, and there is a growing body of evidence suggesting that in many contexts human-AI team performance is inferior to AI performance alone [70, 123]. To address such issues and ensure complementary team performance (*i.e.,* where the team performance can exceed the individual performance of both team members), users should accept advice from the AI system when it is correct and be able to override it when AI advice is incorrect. Such reliance patterns are denoted as appropriate reliance [29], which has become a focal research topic at the intersection of AI and human-computer interaction.

In this context, existing work has explored how user trust and reliance are shaped by different aspects surrounding task characteristics [301], AI systems [111], and user factors [22]. However, most of the research focuses on decision making or data annotation tasks which can be solved in a so-called *one-step* manner [301]. In such a setting, a decision making task can be solved without requiring any intermediate steps. Herein human-AI collaboration allows humans to contrast their individual decisions against that of the AI, often enriched by further information, *e.g.,* its confidence or explanations on how a decision was derived [22]. The ultimate goal here, is to enable humans to (hopefully) derive correct final decisions, leading to optimal team performance. In contrast to the one-step decision making setting, human-AI collaboration in complex multi-step decision making situations that require a composite semantic understanding and a multi-step workflow (*e.g.,* composite fact checking [336]) is still under-explored.

In this chapter, we address this research gap by investigating the potential benefits and pitfalls of asking decision makers to follow the same multi-step workflow as that of an advisory AI system (*i.e.,* completing a sequence of decomposed sub-tasks) along with fine-grained transparency of AI systems. We consider the context of composite fact-checking due to its growing relevance in the age of LLMs, allowing us to simultaneously draw insights in a timely real-world task. The benefits of such a setup are two-fold. Firstly, such a workflow-based decision process enables us to analyze multi-step user decision making, where user decisions at the intermediate steps affect their final decision and reliance on the AI system. The key idea here is that following the same workflow as the AI system can provide global transparency — an overview of the process of the AI system (*i.e.,* task decomposition) — which allows users to check and verify intermediate steps of the AI system and better inform their reliance on AI advice. Secondly, each intermediate step can be viewed as a sub-task. Compared to global transparency, the sub-task information (*i.e.,* local decision criteria and evidence) entails local transparency (*i.e.,* at the level of a specific sub-task) of the intermediate decisions of the AI system. User reliance on the sub-task information (which is also input to the AI system) provides a fine-grained view to analyze appropriate reliance on the AI system. With the fine-grained transparency by design [337, 338], we denote such a multi-step workflow in our study as *multi-step transparent (MST)* decision workflow. In this spirit, a multi-step transparent (MST) decision workflow can potentially

facilitate appropriate reliance on the AI system and help advance our understanding of fine-grained user reliance.

Although appropriate reliance has been extensively studied in relatively simple tasks [22], it is still unclear how user reliance is shaped by a multi-step decision workflow to solve complex tasks. When intermediate steps are adopted to solve complex tasks, users of the AI system may have more decisions to make sequentially. For example, to verify the claim "*General Agreement on Trade in Services is a treaty created to extend the multilateral trading system to service sector and all members of the WTO are parties to the GATS.*", workers need to verify three sub-facts: (1) General Agreement on Trade in Services is a treaty; (2) General Agreement on Trade in Services is created to extend the multilateral trading system to service sector; (3) All members of the WTO are parties to the GATS. In such a multi-step decision workflow, accurate decisions at the intermediate steps can be important.

The intermediate steps and intermediate answers generated by the AI system can provide **global transparency** — overall logic of the AI system (*i.e.,* complex fact-checking with task decomposition and answer aggregation). At the same time, the retrieved evidence at each intermediate step enables users to verify the intermediate answers generated by the AI system. In this way, it can increase the transparency of the AI system's intermediate decisions through verifiability [221], which is denoted as **local transparency**. In this context, we propose to explore appropriate reliance on AI systems at the fine-grained level of intermediate steps and the level of task input of each step. In this chapter, we address the following research questions:

- **RQ1**: How does a multi-step decision workflow shape user reliance on an AI system?
- **RQ2**: How do global transparency and local transparency shape user reliance in a multi-step decision workflow?

To this end, we conducted an empirical study (N = 233) in a composite fact-checking task (*i.e.,* identifying the factual accuracy of claims based on supporting documents). On the one hand, our findings provide empirical evidence that fine-grained appropriate reliance positively contributes to appropriate reliance at the level of overall task. With an MST decision workflow, users developed a fine-grained appropriate reliance on the intermediate steps, which enabled them to detect misleading AI advice. On the other hand, we found that an MST workflow does not improve human-AI team performance and appropriate reliance on AI advice in comparison to a one-step decision workflow. In contrast to facilitating appropriate reliance globally, the MST decision workflow was effective only in a relatively challenging context, where AI advice is misleading. To encourage more precise intermediate decisions, we asked participants to reflect on the usefulness of supporting documents, which nudge users to carefully work on sub-tasks based on local transparency. We found such an intervention to increase user consideration in the intermediate steps brought about worse team performance and reliance patterns. Combined with the cognitive load feedback across experimental conditions, we infer that such an intervention imposes a high cognitive load on users, limiting its expected impact. However, we found that the MST workflow can help users develop a critical mindset when making final decisions. This can partially explain why participants using an MST workflow showed

decreased reliance on the AI system and their confidence decreased after access to the AI advice.

Our results highlight that the multi-step transparent decision workflow in complex tasks did have some positive impact in facilitating appropriate reliance. Appropriate reliance at the intermediate steps may be a prerequisite to making the MST decision workflow effective. While an MST workflow can help mitigate over-reliance in the presence of misleading AI advice, it may also cause under-reliance without enough explicit considerations in the intermediate steps. We infer that there is no one-size-fits-all decision workflow to achieve optimal team performance in complex tasks. To this end, future work in the human-centered AI and relevant research communities should explore how to dynamically adapt and combine multiple decision workflows according to the contextual requirements of human-AI collaboration. Our findings suggest that apart from the benefits of improving user consideration of the fine-grained transparency with specific interventions, it is important to consider potential trade-offs with concomitant side effects (*e.g.,* a high cognitive load caused by such interventions). Finally, we identify promising future directions that explore how to improve human-AI team performance and promote global appropriate reliance by characterizing fine-grained appropriate reliance. Our work has important theoretical implications for promoting appropriate reliance on AI systems in complex tasks and practical implications for the effective use of interventions to support human-AI collaboration.

7.2 Related Work

Our work proposes to analyze fine-grained appropriate reliance on AI systems in handling complex tasks with a multi-step transparent decision workflow. Thus, we position our work in four realms of related literature: trust calibration and appropriate reliance in AI-assisted decision making (7.2.1), multi-step hybrid workflows for complex tasks (7.2.2), transparency and verifiability of AI systems in human-AI collaboration (7.2.3), misinformation and fact-checking (7.2.4).

7.2.1 Trust Calibration and Appropriate Reliance in AI-assisted Decision Making

Existing empirical studies [22, 49, 70] and theoretical frameworks related to user trust [40] and reliance behavior [29] highlight that users of an AI system need to identify when an AI system is accurate to rely on and when it is inaccurate and should be overridden. Such ideal reliance patterns are recognized as appropriate reliance on the AI system, but have proven to be extremely hard to obtain even by leveraging explainable AI methods [107]. Prior literature has adopted different definitions of trust; interpreting trust as either a subjective attitude or as objective user behavior in different contexts. Following the growing interpretation in AI-assisted decision making [22, 40], we operationalize user trust as a subjective attitude and user reliance as objective behavior in this chapter. In most empirical studies [22, 57, 234] where AI systems outperform human decision makers by a margin, the team performance has been reported to be typically worse than that of the AI alone. Addressing such challenges, empirical studies in one-step decision making contexts have been proposed to mitigate under-reliance [107] (*i.e.,* disuse of accurate AI advice) and

over-reliance [55, 108] (*i.e.,* misuse of misleading AI advice).

Trust calibration has been extensively analyzed in interactions with AI systems [339, 340] and automation systems [37, 341–343]. The primary goal is to align or adjust the level of trust that a human places in an AI system or automated technology based on the actual capabilities of that system. Prior work [343–345] has shown that transparency of the system (*e.g.,* pertaining to uncertainty or the reasoning process behind AI advice) can provide users with more situation awareness, and contribute to trust calibration. In particular, existing research has explored how information about AI performance [111], uncertainty of AI advice [121, 220, 346], and reasoning process [347] affects user trust. As pointed out by Lee *et al.* [40], trust can substantially impact user reliance behaviors. Trust calibration has been shown to play an important role in facilitating appropriate reliance [348], aligning these lines of research.

Across multiple domains and diverse setups, researchers have found that many aspects surrounding user factors (like AI literacy [108] and cognitive bias [30]), task characteristics (*e.g.,* task complexity [301] and proxy task [68]), and AI transparency (*e.g.,* explainable AI [107]) have a substantial impact on user reliance. To mitigate the negative impact of these factors, researchers have proposed effective user interventions. User tutorials have been proposed as an intervention that aims at educating users to fill in the knowledge gap [108, 117] and recognize the weaknesses of an AI system [50]. Others have suggested performance feedback [30, 111] through training sessions to calibrate user perceptions of the accuracy of an AI system. Buccina *et al.* [55] proposed cognitive forcing functions to mitigate the illusion of explanatory depth [62] brought about by explainable AI methods.

While existing work has explored how to evaluate and promote appropriate reliance on a global level, little is understood about user reliance behaviors on fine-grained levels in decision making contexts that go beyond one-step decisions and require sequential decisions. In this chapter, we consider a composite fact-checking task as a test bed to explore how users leverage the intermediate steps and supporting documents in a multi-step transparent workflow. Although multi-step workflows have been widely adopted in crowdsourcing [349–351] and crowd-AI hybrid systems [352], they have been under-explored in the context of AI-assisted decision making.

7.2.2 Multi-step Hybrid Workflows for Effective Task Completion

With the goal to obtain high-quality human annotations in complex tasks, prior crowdsourcing literature [349–351] has explored how to decompose complex tasks into multiple microtasks. To ensure text generation quality, Bernstein *et al.* [353] proposed the "Find-Fix-Verify" workflow, which splits complex text writing and editing tasks into a series of generation and review stages. Through empirical studies on writing, brainstorming, and transcription, Little *et al.* [354] found that both iteration and multiple votes can increase the average quality of responses, which is referred to as the "Iterate-and-Vote" workflow. With the rise of conversational agents in recent years, Qiu *et al.* [293] leveraged conversational microtask workflows to improve worker engagement. However, Retelny *et al.* [355] argued that workflows can be a bottleneck to the effectiveness of crowdsourcing in complex tasks.

Inspired by such crowdsourcing literature, researchers have also proposed to build Crowd-AI Hybrid workflows [352] to obtain high-quality data services. For example, in-

stead of obtaining fully manual annotations, asking crowd workers to follow a "Find-Fix-Verify" workflow may boost work efficiency and ensure high-quality outcomes [356, 357]. Similar to the "Iterate-and-Vote" workflow, in a hybrid crowd-AI system, votes from crowd workers and AI systems can also improve outcome quality [358]. With the rise of large language models (LLMs), there is an increasing exploration of how conversational interaction can boost crowd-AI hybrid intelligence [263, 269]. For instance, users can obtain writing suggestions for a scientific paper using an LLM-powered conversational interface [269]. With a conversational human-AI interaction, users are involved in an implicit multi-step workflow to complete a task. Existing research has explored LLMs to automate exploratory conversations [359] and plan daily tasks [274]. Chaining multiple LLMs can achieve even complex functions entailed in music chatbots and writing assistants [360]. For example, Wu *et al.* [361] defined primitive operations based on LLMs and chained them to synthesize controllable workflows dynamically. Such AI chains can also be adapted from crowdsourcing workflows [362].

We draw inspiration from existing literature on workflows for accomplishing tasks, and propose a multi-step transparent workflow for decision making in a complex fact-checking task. In our study, participants were required to go through intermediate steps of the AI (indicating the step-wise process of the AI), and verify the correctness of the final AI advice. Such a process allows users to develop an understanding of AI advice in a step-wise manner, and make a final decision based on both AI advice and their initial decision. The multi-step transparent workflow is generated and executed by the AI system [363] (*i.e.,* LLMs coupled with retrieval-augmented generation[364]) to provide advice and support participants in the task. Such human-AI collaboration increases the transparency of the AI system. We aim to explore whether the increased transparency in such a process can facilitate appropriate reliance.

7.2.3 Transparency and Verifiability in Human-AI Collaboration

Transparency has been recognized as an important goal towards building trustworthy AI systems [28, 365–367]. Existing work has explored the transparency of AI systems from different angles — transparency in the reasoning process [368], transparency of data collection/curation [369], transparency of limitations (*e.g.,* uncertainty) [370], transparency of social context [110, 123] etc. Explainable AI (XAI) methods, which may be independent of the actual AI system, are also widely adopted to increase the transparency of AI systems in human-AI collaboration [27, 107, 123, 126]. Besides incorporating XAI to increase system transparency, AI transparency is more explored theoretically [366]. Relatively few works have attempted to empirically verify the impact of transparency on human-AI collaboration. With an empirical study, Vossing *et al.* [347] found that providing the transparency of the reasoning process can increase user trust, while providing transparency of system uncertainty can decrease user trust [347].

Different from the transparency of AI systems, verifiability is typically associated with specific AI advice. Within the context of human-AI collaboration, explainable AI methods [371, 372] are widely used to assist human decision makers by providing evidence (*e.g.,* highlighting a part of task input [159]) to support/oppose AI advice [107]. Among the explainable AI methods, causal explanations [373, 374] propose to reason about the causal relationships between the task input and AI advice, which provides a strong verifi-

ability of AI advice. Recently, retrieval-augmented generation [364, 375] has emerged as one popular paradigm to enhance the verifiability of LLMs. With the retrieved evidence (*e.g.,* documents or relevant structure knowledge) as a reference, humans can verify the factual correctness of LLM generation.

In this chapter, we followed the idea of transparency by design [337, 338] to modularize the complex fact-checking task into a series of sub-fact verification steps. With the decomposed sub-facts and sub-fact verification results, we provided users with global transparency of the AI system's overall process of task decomposition. At the same time, we also provide the retrieved documents in each sub-fact verification, which are input to the LLM-based fact verification system. These documents provide local transparency of the intermediate steps (*i.e.,* sub-tasks). Thus, we provided fine-grained transparency of the AI system and explored how user reliance is shaped through the multi-step transparent (MST) decision workflow. To the best of our knowledge, this is the first empirical effort to understand user reliance on an AI system with fine-grained transparency.

7.2.4 Misinformation and Fact-checking

From a data mining perspective [376], misinformation is mainly detected based on two criteria: *veracity* and *intentionality*. Veracity mainly focuses on whether referred media or an online post is factually false or inaccurate, regardless of intent. 'Fact-checking' is a task mainly based on veracity, which assesses whether claims made in written or spoken language are factually correct [377–379]. Intentionality is another dimension based on the intent of the information creator/provider. For example, hate speech [380] and 'fake news' in the political election [381]. Such misinformation, which often uses inflammatory and sensational language to alter people's emotions [382], can be harmful and widespread online [383, 384]. Based on these criteria, different communities have developed deep learning-based methods [378, 385] to automate checking the massive amount of information online. In this chapter, we focus on the veracity of factual claims and conducted fact-checking tasks in a human-AI collaborative setting.

While deep learning has been widely adopted to manage misinformation online, human partnership is still a crucial factor in this task [386]. In addition to domain experts who are capable of detecting inaccurate or false information, researchers have explored and showcased crowdsourcing as an effective means to conduct fact-checking [387–392]. Typically, crowdsourced fact-checking involves three steps in a complex workflow [391]: (1) claim selection, which targets selecting check-worthy claims; (2) evidence retrieval, which obtains necessary information sources (*e.g.,* with a search engine); and (3) claim verification, which includes discussion and aggregation of judgment across different crowd workers and further produces explainable, convincing verdicts (*i.e.,* justification production). Prior to the rapid adoption of LLMs, the AI assistant in each stage of the complex workflow was typically trained independently and served different purposes. Such disparity between AI systems in different stages prevents humans from building a coherent and unified mental model when working with these sub-tasks. Recent advances have led researchers to explore leveraging LLMs to enhance all sub-tasks and provide an end-to-end workflow by chaining LLMs [363].

It is evident that LLMs bring new opportunities and challenges to the fact-checking task [393–395]. On the one hand, LLMs have shown powerful natural language under-

standing and generation capabilities that can help tackle sub-tasks of fact-checking systems [396]. For example, LLMs can retrieve highly relevant information sources [375] and generate explanations to justify the verification process or the results [363]. Furthermore, LLMs can provide an easy way for humans to communicate with the AI system, offering further potential for human-AI interaction in fact-checking tasks [393]. On the other hand, LLMs are known to hallucinate [397], *i.e.,* generating seemingly plausible but incoherent or factually incorrect content. LLMs have been shown to suffer from out-of-distribution data issues [398] and evolving knowledge without external contextual input (*e.g.,* retrieved documents) [375]. Due to the uncertainty brought about by these prevalent flaws and the lack of accountability, human-AI collaborative fact-checking (comprising at least human oversight) is of fundamental importance in the era of LLMs.

In a user study of AI-assisted fact-checking, Nguyen *et al.* [386] found that crowd workers can be easily misled by wrong model predictions, but such errors can be reduced given interactions with the AI system. With dynamic user input and updated AI system predictions, crowd workers make much fewer errors misled by wrong AI predictions. Thus, Nguyen *et al.* [386] argued that 'transparent models are key to facilitating effective human interaction with fallible AI models.' Contributing to existing literature in the area of human-AI collaboration for fact-checking, our work provides a multi-step transparent decision workflow in assisting humans conduct fact-checking with fine-grained retrieved evidence and decomposed sub-steps. Through this, we aim to provide fine-grained transparency and facilitate appropriate reliance of humans on the AI system. Our insights add further empirical evidence and advance our understanding of how transparency of the AI system and decision workflow affects human-AI interaction.

7.3 Task and Hypothesis

In this section, we describe the composite fact-checking task (*i.e.,* identifying the factual accuracy of claims based on supporting documents), the multi-step transparent workflow (MST), and present our hypotheses, which have all been preregistered before any data collection.

7.3.1 Composite Fact-checking Task

To analyze how the MST decision workflow impacts human-AI collaboration in complex tasks, we consider a composite fact-checking task. An example of solving a composite fact-checking task based on the mutli-step transparent workflow is shown in Figure 7.1. This task asks participants to decide whether a factual claim is **True** or **False** using the supporting documents retrieved from Wikipedia. The reasons for selecting the composite fact-checking task as our test bed are three-fold. Firstly, it contains tasks that require composite semantic understanding and can be solved with a workflow. Secondly, the fact-checking task requires evidence-based verification, which provides verifiability in the intermediate steps. Thirdly, due to the practical need for content moderation online (*e.g.,* hate speech, rumors, and hallucinated content from generative AI systems), it is a timely and relevant scenario for human-AI collaboration.

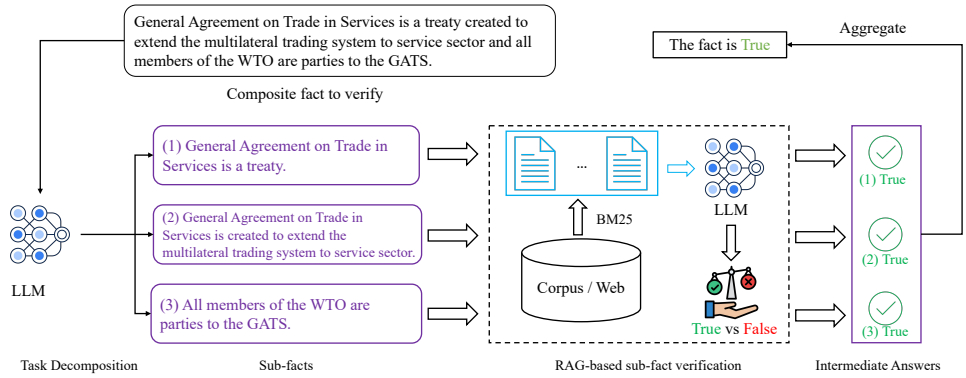


Figure 7.1: Illustration of the multi-step workflow on the composite fact-checking tasks using the ProgramFC method [363]. The sub-facts and intermediate answers (in the purple box) provide global transparency in our MST workflow. The retrieved documents (in the blue box) serve as local transparency in our MST workflow.

7.3.2 Multi-step Transparent Workflow

AI System Setup. In our study, we adopted an LLM-based method called *ProgramFC* [363] to serve as our AI system. Figure 7.1 illustrates how *ProgramFC* provides global transparency and local transparency. The *ProgramFC* method conducts fact-checking with two stages: (1) Using GPT-3.5 to generate decomposed steps to conduct composite fact-checking. (2) After generation of the decomposed steps, these steps are executed using another LLM, *flan-t5-xl* [399]. Using the generated decomposed steps (*i.e.,* sub-facts to verify), the execution step generates intermediate answers based on retrieved supporting documents for each sub-fact. The documents are retrieved based on the popular BM25 algorithm [400], which leverages the query terms frequency appearing in documents to achieve a ranking function. All source documents are from Wikipedia, which is provided with the implementation of *ProgramFC*.[1] Finally, *ProgramFC* aggregates the intermediate answers to obtain a final prediction of the factual accuracy for the composite fact. The generated decomposed steps, intermediate answers, and retrieved supporting documents form the basis for the multi-step transparent workflow in our study. In our implementation, we selected the aforementioned LLMs due to two reasons: (1) *flan-t5-xl* are representative open-sourced LLMs that are widely adopted in question answering and fact-checking practice [399], (2) GPT-3.5 is representative of the performance of most open-sourced and commercial LLMs at the time of data collection (*i.e.,* Jan 2024), offering transferable findings and implications within the scope of our empirical study.

Decision Workflow. In our study, all workflows follow a two-stage decision making setup, a widely adopted design in AI-assisted decision making [30, 49, 107, 108]. In the first stage, participants work on the fact-checking tasks based on the provided supporting documents and the decision workflow. Next, they were given a chance to alter their initial choice following AI advice. In multi-step decision workflows, decomposed steps and intermediate AI predictions are also shown to support user decisions. In the first stage

[1] https://github.com/teacherpeterpan/ProgramFC

of decision making, if participants do not find useful supporting documents to support or refute the sub-fact / fact, they can choose 'Uncertain' beside the label 'True' and 'False'. In the second stage of decision making, participants are asked to make a binary decision between 'True' and 'False.'

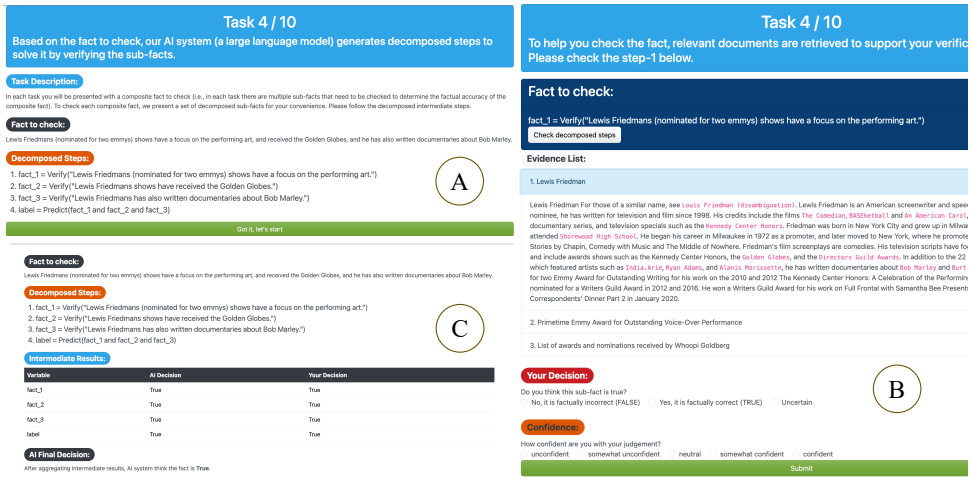


Figure 7.2: Screenshots of the composite fact-checking task interface with the MST workflow. (A) The starting point of the MST workflow, where the fact to check and decomposed steps are shown to users. (B) An intermediate step in the MST workflow. (C) Final decision making page, where decomposed steps and intermediate answers are provided as an explanation to the AI advice.

7

User Interface. The user interface of our study is shown in Figure 7.2. At the beginning of the MST workflow, we show participants the composite fact to check and decomposed steps (Figure 7.2 (A)). Next, participants are asked to follow the decomposed steps to verify the sub-facts (Figure 7.2 (B)) based on supporting documents. In our study, each step is provided with three relevant documents retrieved using the BM25 algorithm [400]. Finally, participants receive the final AI advice on the factual accuracy of the composite fact. The decomposed steps and the corresponding AI advice at intermediate steps are also provided as explanations for the final AI advice. As a baseline comparison to the MST workflow, we also adopted a basic one-step fact-checking workflow, where participants were asked to identify the factual accuracy of composite facts directly without explicitly decomposed steps. To ensure a fair comparison in terms of the evidence presented to participants across the two conditions, we gathered all documents retrieved in the three steps of the MST workflow and showed them on one page for the basic fact-checking workflow.

7.3.3 Hypotheses

Our study mainly aims to contribute to understanding user reliance behaviors on AI systems in the context of solving complex tasks in a sequence of decomposed sub-tasks generated by LLMs. To this end, we devised a multi-step decision workflow and catered to transparency along the decomposed steps and AI advice at the intermediate steps. The intermediate steps and answers work in facilitating a global transparency of the AI system,

rendering the decision making process visible [347]. Within our study, the decomposed steps generated by LLMs show the overall step-wise process of the AI system, which can potentially cause user trust to increase [347]. Meanwhile, user trust has been found to substantially impact user reliance behaviors [40]. Combining these findings from existing work, we infer that providing decomposed steps generated by LLMs (*i.e.,* global transparency) can increase user reliance on the AI system. Thus, we hypothesize that:

> **(H1)** Compared to only providing final AI advice, providing the decomposed steps generated by LLMs (*i.e.,* global transparency) will increase user reliance on the AI system.

With the multi-step transparent workflow, users follow the same process (*i.e.,* decomposed steps in the same order) to verify how the AI system works on the composite fact-checking task. Throughout this process, the retrieved documents are provided to increase the transparency of each step (*i.e.,* sub-task). Based on local transparency, users make independent judgments about the intermediate steps before being exposed to final AI advice and intermediate answers predicted by the AI system. In comparison, users with a one-step decision workflow do not have the chance to work on the decomposed sub-tasks following the same step-wise process of the AI system. Thus, users with a multi-step transparent workflow may develop a more critical mindset when adopting the final AI advice supported with intermediate steps and answers. In this way, they may be better equipped to recognize when the AI system provides correct advice and when they should rely on their own decisions. Thus, we hypothesize that:

> **(H2)** Providing users with a multi-step transparent decision workflow of the AI system will result in relatively more appropriate reliance on the AI system, in comparison to a one-step decision workflow with AI advice.

Although the intermediate steps and task input at each step (*e.g.,* retrieved documents) are designed to provide benefits in the decision making process, adequate consideration and appropriate use can be a prerequisite for their effectiveness. If users can properly leverage fine-grained transparency (*i.e.,* developing precise decisions and reflections at the level of intermediate steps and the level of task input in each step), they can benefit from the multi-step decision workflow, thereby calibrating their reliance behaviors on the AI system. Thus, we hypothesize that:

> **(H3)** Within a multi-step decision workflow, more accurate intermediate user decisions will result in relatively more appropriate reliance on the final AI advice.

7.4 Study Design

This study was approved by the human research ethics committee of our institution. Our hypotheses and experimental setup had all been preregistered before any data collection.

7.4.1 Experimental Conditions

Addressing the aforementioned RQs, we aim to explore the impact of a transparent decision workflow on user reliance on an AI system in a composite decision making task. Considering transparency of the decision workflow as the sole independent variable in our study, we designed a between-subjects study with four experimental conditions (see Table 7.1). In all conditions, participants follow a two-stage decision making setup (described in Section 7.3.1). The different experimental conditions are presented below, with each successive condition being a variant of the previous condition by a single factor.

1. `Control` — In this condition, participants follow a one-step fact-checking workflow in the first stage and only have access to the final AI advice in the second stage.

2. `MST-GT` — In this condition, participants can additionally check the intermediate steps from the AI system as global transparency in the second stage.

3. `MSTworkflow` — In this condition, participants follow a multi-step transparent workflow in the first stage, where they follow the same working logic (*i.e.,* decomposed steps in the same order) of the AI system, and check the retrieved documents (*i.e.,* part of AI input) at each sub-task. In the second stage of decision making, they will be shown the intermediate steps and intermediate answers from both AI systems and themselves (cf. Figure 7.2).

4. `MSTworkflow+` — On top of condition `MSTworkflow`, participants in this condition are asked to annotate the usefulness of the supporting documents in each intermediate step. Such annotation encourages users to carefully check each retrieved document at the intermediate steps and indicate their usefulness in informing their intermediate decisions. This is designed to function similarly to cognitive forcing functions [55], which nudge users towards critical use of AI advice.

Table 7.1: Differences between experimental conditions. The intermediate steps and answers are regarded as global transparency. Users have access to local transparency through the multi-step transparent workflow.

Exp Condition	Decision Workflow	AI Assistance
`Control`	one-step workflow	AI Advice
`MST-GT`	one-step workflow	AI advice + global transparency
`MSTworkflow`	multi-step transparent workflow	AI advice + global transparency
`MSTworkflow+`	multi-step transparent workflow + document usefulness annotation	AI advice + global transparency

7.4.2 Task Selection

As described earlier, composite fact-checking is an important avenue for human-AI collaboration (*e.g.,* credibility assessment systems [109]). All data used in our study is from a public fact-checking dataset – FEVEROUS-S [336]. This dataset is widely used in composite fact-checking, which leverages documents as evidence. The ten selected tasks in our study are shown in Table 7.2.

Table 7.2: Selected tasks in our study. 'Extra Notes' provides special cases of the correctness of intermediate AI advice and verifiability of intermediate steps.

ID	Decomposed Steps generated by LLMs	Ground Truth	System Advice	Extra Notes
1	(1) In 2014, both Orient Express and its holding company were renamed Belmond and Belmond Ltd, respectively. (2) Orient Express is a hospitality and leisure company that operates luxury hotels, train services and river cruises worldwide. (3) Belmond Ltd partnered with Irish Rail in 2015 to launch the luxury train Belmond Grand Hibernarian in Ireland.	True	False	two misleading intermediate advice
2	(1) Adrian Haynes is a Wampanoag chief. (2) Adrian Haynes served in the United States Navy during WWII from 1943 to 1947. (3) Adrian Haynes had a stint with the Naval Supply Ninth Amphibian Force that took part in the 1944 Anzio invasion in Italy.	True	True	one intermediate step is not verifiable
3	(1) Lewis Friedmans (nominated for two emmys) shows have a focus on the performing art. (2) Lewis Friedmans shows have received the Golden Globes. (3) Lewis Friedmans has also written documentaries about Bob Marley.	True	True	-
4	(1) General Agreement on Trade in Services is a treaty. (2) General Agreement on Trade in Services is created to extend the multilateral trading system to service sector. (3) All members of the WTO are parties to the GATS.	True	True	-
5	(1) Edgar McInnis wrote poetry in his spare time. (2) Edgar McInnis won the Newdigate Prize in 1925 for his poem "Byron". (3) Edgar McInnis received Master of Arts degree in 1930 from Oxford University.	True	True	-
6	(1) Teldenia strigosa was described by Warren in 1903. (2) Teldenia strigosa was found in New Guinea and Goodenough Island (in the Solomon Sea). (3) The length of the forewings of Teldenia strigosa is 12.5–15 mm.	False	True	one misleading intermediate advice
7	(1) The Travelling Church emigrants did not take any slaves with them when they traveled. (2) The Travelling Church emigrants traveled over the frozen and danger-filled Cumberland Gap. (3) The Cumberland Gap is a pass through the long ridge of the Cumberland Mountains, and within the Appalachian Mountains.	False	True	one misleading intermediate advice
8	(1) Tosi Fasinro finished fourth at the 1990 World Junior Championships. (2) Tosi Fasinro won the 1993 UK Championships. (3) Tosi Fasinro took one gold and one bronze at the AAA Championships.	False	False	-
9	(1) Stephanie Flanders was BBC's economics editor for five years. (2) Stephanie Flanders presented the docu-series Masters of Money. (3) Iain Duncan Smith praised Stephanie Flanders because of her pro-Labour stand in the coverage of unemployment figures.	False	False	one intermediate step is not verifiable
10	(1) The wild water buffalo or Asian buffalo is an endangered species. (2) The wild water buffalo or Asian buffalo is likely to become extinct shortly. (3) The wild water buffalo or Asian buffalo has a population of less than 1,000, of which the majority is found in India.	False	False	-

Selection Process. First, we generate decomposed steps for all tasks in the evaluation set of the FEVEROUS-S dataset. The task decomposition is achieved with prompting LLMs (GPT-3.5 in our study, but this can be easily replaced with other LLMs). The prompt is based on the implementation of ProgramFC [363].[2] Next, we considered and retained all tasks that can be solved by verifying 3 sub-facts. This resulted in 1,127 candidate composite tasks. The *ProgramFC* algorithm achieved 67.7% accuracy on these tasks. Estimating that each fact-checking task could take around 2-3 minutes for participants to complete, we selected ten tasks from these candidates. Considering all possible cases of (Ground Truth, AI Prediction) pairs, we randomly sample 10 tasks for each case (resulting in 40 tasks as candidates). An author of this chapter then annotated the correctness of the decomposed steps and manually followed the decomposed steps to annotate both the usefulness of each supporting document and the factual accuracy of each sub-fact in the decomposed steps. After that, 15 tasks, where the decomposed steps were correct to verify the composite fact, were reserved.

To balance the label distribution (True/False for the answer of each task), we selected five tasks with ground truth "True" and five tasks with ground truth "False" (ten tasks intotal). 70% accuracy is adopted when selecting the tasks, the rationale behind is: (1) it is very close to the actual AI accuracy 67.7% (2) With such an accuracy level, the AI system is compatible with crowd workers to provide decision support without risking optimal performance with over-reliance, which makes it suitable to analyze user (appropriate) reliance patterns. To control the difficulty of tasks where AI prediction is wrong, tasks 1, 6, and 7 contain one or two incorrect intermediate steps. Besides, tasks 2 and 9 contain one intermediate step where the supporting documents are not enough to conclude the factual accuracy. In the two tasks, the AI final advice and the intermediate steps are all correct.

7.4.3 Measures and Variables

Reliance-based dependent variables

In condition `MSTworkflow` and `MSTworkflow+`, participants have to assess the intermediate correctness of each sub-fact. Each task is decomposed into three sub-tasks, which can facilitate valid comparison across conditions. Based on the user assessments of the factual correctness of intermediate steps and the ground truth (obtained through expert annotation), we can measure appropriate reliance at intermediate steps (**AR-Intermediate**) as average accuracy of user intermediate decisions. Participants in condition `MSTworkflow+` were asked to annotate the usefulness of supporting documents when verifying each sub-fact with a question: "*Does this excerpt contain necessary information to verify the sub-fact?*". There are four potential responses: "*Useless: it does not contain any useful information to verify the fact*", "*Partial support: it contains some information partially support the sub-fact, but not fully support*", "*Full support: it contains all necessary information to support the sub-fact*", "*Contradiction: it contains necessary information to contradict with the sub-fact*". To analyze how users appropriately leverage the intermediate supporting documents (**AR-Evidence**), we adopted expert annotation of the usefulness of each supporting document as ground truth and calculated users' agreement ratio. Both **AR-Intermediate** and **AR-Evidence** are averaged across intermediate steps of the workflow.

[2] `https://github.com/teacherpeterpan/ProgramFC/blob/main/models/prompts.py`

Since we aim to analyze the impact of different decision workflows on team performance and (appropriate) reliance, we leveraged measures introduced by prior work and typically used in this context due to their suitability [29, 30, 57, 66]; we adopted *Team Performance* (*i.e.,* average user accuracy based on their final decision) and *Team Performance-wid* (*i.e.,* average user accuracy where their initial decision disagrees with AI advice) to measure user performance in our study. Following previous work by Yin *et al.* [57], Zhang *et al.* [66], we measured user reliance by using the **Agreement Fraction** and the **Switch Fraction**. These measures consider the degree to which user final decisions agree with AI advice, and how often they switch to AI advice when their initial decision disagrees with the AI advice. Following prior work [29] on the evaluation of appropriate reliance, we adopted *Relative positive AI reliance* (**RAIR**) and *Relative positive self-reliance* (**RSR**) as appropriate reliance measures. A low **RAIR** indicates under-reliance on the AI system, while a low **RSR** indicates over-reliance on the AI system. To provide a precise definition of reliance-based measures used in our study, we provide further details of calculation formula in Appendix.

Other Variables
We also adopted measures for user trust, user confidence, cognitive load, and relevant covariates to further our understanding of the impact of different decision workflows. The confidence is collected along with decision making based on 5-point Likert scale.

User Trust. Motivated by existing work on user trust in automation [40], we assessed user subjective trust with a post-task questionnaire. We adopted four sub-scales from the trust in automation questionnaire [321]: Reliability/Competence (TiA-R/C), Understanding/Predictability (TiA-U/P), Intention of Developers (IoD), and Trust in Automation (TiA-Trust).

Cognitive Load. As the decision workflows in our study provide different levels of transparency of the AI system, there is a potential for the workflows to pose varying cognitive load among users. We assessed user cognitive load using the NASA-TLX questionnaire [242].

Covariates. For a deeper analysis of our results, and to account for potential confounds based on exisiting literature, we considered the following covariates:

- Familiarity with the AI system (*Familiarity*) and general propensity to trust (*Propensity to Trust*) from the trust in automation questionnaire.
- User expertise in large language models (*LLM Expertise*) is assessed with a question "To what extent are you familiar with large language models?". Responses were gathered on a 5-point Likert scale from 1 (No prior experience/knowledge) to 5 (Extensive prior experience/knowledge).
- User expertise in fact-checking tasks (*Fact Checking Expertise*) is assessed with a question "Do you have any experience or knowledge with fact-checking?". Responses were gathered on a 5-point Likert scale from 1 (No prior experience/knowledge) to 5 (Extensive prior experience/knowledge).

Table 7.3 presents an overview of all the variables considered in our study.

Table 7.3: The different variables considered in our experimental study. "DV" refers to the dependent variable. **RAIR**, **RSR**, and **Accuracy-wid** are indicators of appropriate reliance.

Variable Type	Variable Name	Value Type	Value Scale
Performance (DV)	Team Performance	Continuous, Interval	[0.0, 1.0]
	Team Performance-wid	Continuous	[0.0, 1.0]
Reliance (DV)	Agreement Fraction	Continuous, Interval	[0.0, 1.0]
	Switch Fraction	Continuous	[0.0, 1.0]
Appropriate Reliance (DV)	RAIR	Continuous	[0.0, 1.0]
	RSR	Continuous	[0.0, 1.0]
	AR-Intermediate	Continuous	[0.0, 1.0]
	AR-Evidence	Continuous	[0.0, 1.0]
Trust (DV)	Reliability/Competence	Likert	5-point, 1: poor, 5: very good
	Understanding/Predictability	Likert	5-point, 1: poor, 5: very good
	Intention of Developers	Likert	5-point, 1: poor, 5: very good
	Trust in Automation	Likert	5-point, 1:strong distrust, 5: strong trust
Cognitive Load (DV)	Mental Demand	Likert	-7: very low, 7: very high
	Physical Demand	Likert	-7: very low, 7: very high
	Temporal Demand	Likert	-7: very low, 7: very high
	Performance	Likert	-7: Perfect, 7: Failure
	Effort	Likert	-7: very low, 7: very high
	Frustration	Likert	-7: very low, 7: very high
Covariates	Propensity to Trust	Likert	5-point, 1: tend to distrust, 5: tend to trust
	Familiarity	Likert	5-point, 1: unfamiliar, 5: very familiar
	LLM Expertise	Likert	5-point, 1: No expertise, 5: Extensive expertise
	Fact Checking Expertise	Likert	5-point, 1: No expertise, 5: Extensive expertise
Other	Usefulness of Evidence	Category	{useless, partial support, support, contradiction}
	Confidence	Likert	5-point, 1: inconfident, 5: confident

7.4.4 Participants

Sample Size Estimation. We computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [92]. In our experimental analysis, we applied a Bonferroni correction to correct for testing multiple hypotheses. The significance threshold decreased to $\frac{0.05}{3} = 0.017$, and is applied to all statistical analyses. We specified the default effect size $f = 0.25$, a significance threshold $\alpha = 0.017$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and that we will investigate four different experimental conditions/groups. This resulted in a required sample size of 230 participants. We thereby recruited 284 participants from the crowdsourcing platform Prolific,[3] accommodating potential exclusion.

Compensation. To ensure a fair comparison across conditions, we set the basic payment for all participants to £4. This payment is based on the time estimation of condition `MSTworkflow+` (30 minutes) and a "*Fair*" payment criteria (£8 per hour) by the platform. To motivate participants to reach correct decisions with their best ability, we rewarded each correct decision with a bonus of £0.05. Such a setup is also regarded as a contextual requirement to achieve appropriate trust in automation [40].

Filter Criteria. All participants were proficient English speakers aged between 18 and 50, and they had finished more than 40 tasks and maintained an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one attention check or if we found any missing data. In our study, frequently switching from initial agreement to opposite AI advice is treated as an indicator of potentially unreliable behavior. We excluded the participants with three or more such indications and participants who finished the study in a very short time (less than 15 minutes). These fil-

[3] `https://www.prolific.co`

ter criteria ensure the quality of collected data by removing low-effort submissions. After these filter criteria, we have 233 participants reserved for analysis. These participants had an average age of 34 (SD = 7.5) and a reasonably balanced gender distribution (54.1% male, 45.9% female).

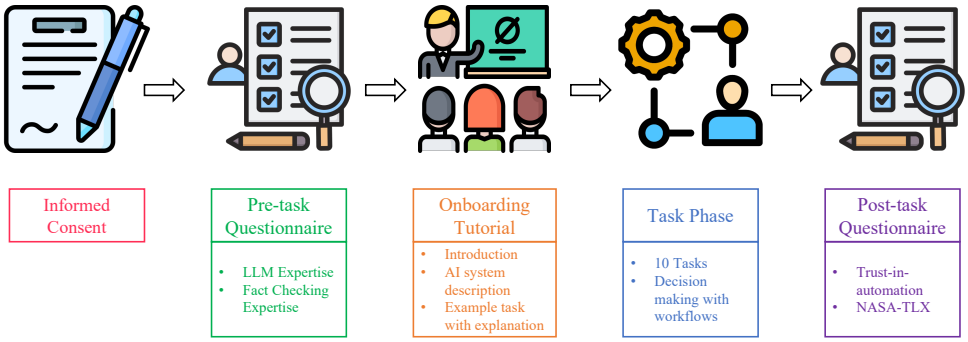


Figure 7.3: An illustration of the procedure that participants followed in our study.

7.4.5 Procedure

At the start of our study, all participants were asked to provide us with informed consent if they wished to proceed. Next, we gathered user self-reported expertise of large language models and fact-checking tasks using two questions. Before formally working on the tasks, we provide an onboarding tutorial to help participants get familiar with all elements shown in our study and understand how to work on the composite fact-checking tasks with an example. At this stage, participants in the condition `MST-GT`, `MSTworkflow`, and `MSTworkflow+` also have access to the decomposed steps and intermediate answers for the example task. Next, participants worked on the ten selected tasks based on the decision workflow of the corresponding condition. Finally, they were asked to fill in post-task questionnaires (including the trust in automation questionnaire and NASA-TLX questionnaire). We employed two attention check questions (one in the task phase and one in the post-task questionnaire) to ensure the quality of collected data. Figure 7.3 illustrates the procedure participants followed in our study.

7.5 Results

In this section, we will present the main experimental results and exploratory analysis for our study. In the spirit of open science, our code and data can be found in OSF repository.[4]

7.5.1 Descriptive Statistics

To ensure the reliability of our results and interpretations, we only consider 233 participants who passed all attention checks. These participants were distributed across four experimental conditions in a reasonably balanced manner — 54 (`Control`), 62 (`MST-GT`), 60 (`MSTworkflow`), 57 (`MSTworkflow+`).

[4] `https://osf.io/s4he5/?view_only=2810286e1e5c4573b723c4785d5fe45c`

Distribution of Covariates. The covariates' distribution is as follows: *LLM Expertise* (16.7% with good/extensive background knowledge about LLMs, 83.3% without any knowledge or with limited knowledge about LLMs), *Fact Checking Expertise* (28.8% with good / extensive background knowledge with fact-checking tasks, 71.2% without any knowledge or with limited background knowledge in fact-checking tasks), *Propensity to Trust* (M = 2.82, SD = 0.71; 5-point Likert scale, *1: tend to distrust*, *5: tend to trust*), and *Familiarity* (M = 2.54, SD = 1.11; 5-point Likert scale, *1: unfamiliar*, *5: very familiar*).

Table 7.4: Participant performance on fact-checking tasks. 'Accuracy' is reported in percent (%). We use **bold** and underlined fonts to denote the best and second-best performance in each task, respectively.

Task-ID	System Advice	Ground Truth	Accuracy				
			Control	MST-GT	MSTworkflow	MSTworkflow+	Avg
1	False	True	**33.3**	15.0	14.5	28.1	22.3
2	True	True	**79.6**	63.3	50.0	54.4	61.4
3	True	True	**70.4**	68.3	59.7	66.7	66.1
4	True	True	94.4	**100**	98.4	87.7	95.3
5	True	True	98.1	**100**	91.9	94.7	96.1
6	True	False	9.3	8.3	**25.8**	24.6	17.2
7	True	False	59.3	50.0	**62.9**	43.9	54.1
8	False	False	85.2	**86.7**	83.9	70.2	81.5
9	False	False	87.0	**90.0**	83.9	73.7	83.7
10	False	False	90.7	**91.7**	90.3	73.7	86.7

Performance Overview. On average across all conditions, participants achieved a *Team Performance* of 66% (SD = 0.12), which is still lower than the AI accuracy (70%). The average *Agreement Fraction* is 0.78 (SD = 0.14), and the average *Switch Fraction* is 0.50 (SD = 0.30). As for the appropriate reliance at fine-grained levels, participants in `MSTworkflow` and `MSTworkflow+` conditions achieved an average *AR-Intermediate* of 0.73 (SD = 0.12); participants in `MSTworkflow+` condition achieved an average *AR-Evidence* of 0.64 (SD = 0.24). With these measures, we confirm that (1) participants in our study do not always blindly rely on AI advice, and (2) participants put some effort into the intermediate steps and supporting documents. As all behavior-based dependent variables (*i.e.,* performance, reliance, and appropriate reliance) are not normally distributed, we used non-parametric statistical tests to verify our hypotheses.

Table 7.4 shows the accuracy across conditions in the ten selected fact-checking tasks. Among the seven tasks where system advice is aligned with the ground truth, participants in `Control` and `MST-GT` conditions showed higher accuracy. However, participants in `MSTworkflow` and `MSTworkflow+` conditions also showed competitive or even better performance in the three tasks where system advice was misleading.

Cognitive Load. Based on the NASA-TLX questionnaire, we assessed participants' cognitive load based on six dimensions. As the workflows used in our study are of different levels of complexity and annotation effort. We visualized their distribution across conditions in Figure 7.4. To compare the cognitive load across conditions, we conducted a one-way ANOVA test and post-hoc pairwise Tukey's HSD test. The results indicate that participants in `MSTworkflow+` showed significantly higher *Mental Load* than other conditions, and also showed much higher *Frustration* when compared to the `Control` condition.

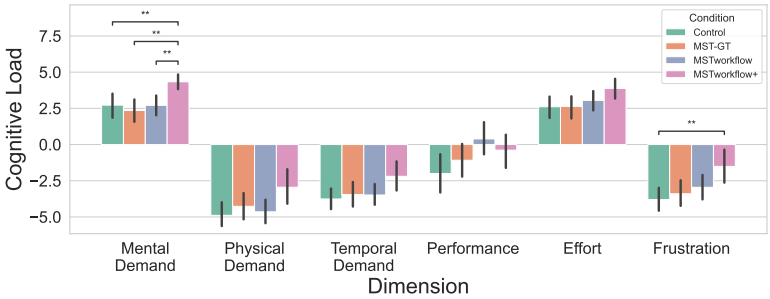


Figure 7.4: Bar plot illustrating the distribution of the cognitive load across different experimental conditions in our study. **: $p < 0.017$

While for the reserved four dimensions, the difference is non-significant, we still observe that participants in `MSTworkflow+` condition showed higher cognitive load.

7.5.2 Hypothesis Tests

H1 and H2: effect of different workflows

To verify **H1** and **H2**, we conducted Kruskal-Wallis H-tests to compare the performance, reliance, and appropriate reliance measures of participants across the four experimental conditions. The results are shown in Table 7.5. For the measures where significant differences exist $p < 0.017$, we conducted a post-hoc Mann-Whitney test to obtain pairwise comparisons.

Table 7.5: Kruskal-Wallis H-test results for workflow on reliance-based dependent variables. "††" indicates the effect of the variable is significant at the level of 0.017.

Dependent Variables	H	p	$M \pm SD$				Post-hoc
			Control	MST-GT	MSTworkflow	MSTworkflow+	
Team Performance	15.13	**.002**††	0.71 ± 0.11	0.67 ± 0.10	0.66 ± 0.13	0.62 ± 0.14	Control > MST-GT, MSTworkflow > MSTworkflow+
Agreement Fraction	16.37	**.001**††	0.80 ± 0.11	0.83 ± 0.13	0.75 ± 0.13	0.72 ± 0.16	Control, MST-GT > MSTworkflow, MSTworkflow+
Switch Fraction	3.07	.381	0.47 ± 0.27	0.54 ± 0.34	0.46 ± 0.26	0.52 ± 0.30	-
Team Performance-wid	12.20	**.007**††	0.64 ± 0.27	0.52 ± 0.26	0.52 ± 0.24	0.49 ± 0.21	Control > MST-GT, MSTworkflow, MSTworkflow+
RAIR	3.64	.302	0.60 ± 0.39	0.54 ± 0.42	0.48 ± 0.34	0.52 ± 0.33	-
RSR	13.77	**.003**††	0.73 ± 0.43	0.47 ± 0.47	0.65 ± 0.46	0.46 ± 0.48	Control, MSTworkflow > MST-GT, MSTworkflow+

Impact on user reliance. As shown in Table 7.5, compared to `Control` condition, `MST-GT` condition showed significantly higher *Agreement Fraction* and a non-significantly higher *Switch Fraction*. This supports that providing transparency in the AI system's intermediate steps (*i.e.,* global transparency) increases user reliance on the AI system. Thus, we find support for **H1**. However, with global transparency in final decision making, participants in conditions with a multi-step decision workflow (*i.e.,* `MSTworkflow` and `MSTworkflow+`) showed significantly lower *Agreement Fraction*.

Impact on appropriate reliance. Although there is no significant difference on *RAIR* across conditions. It is clear that participants in `Control` condition showed the highest *RAIR* and *RSR* corresponding to the highest level of appropriate reliance. Participants in `MST-GT` and `MSTworkflow+` conditions showed significantly worse *RSR* than `Control` and `MSTworkflow` conditions, which is a reflection of over-reliance. Meanwhile, participants in the `MSTworkflow` condition showed the worst *RAIR*, which reflects a sub-optimal human-

AI collaboration due to under-reliance. The reliance pattern differences between MST-GT and MSTworkflow conditions are a result of the decision workflow. Thus, we can infer that the MST decision workflow can help mitigate the over-reliance caused by providing global transparency. At the same time, it also introduces new issues of under-reliance. Thus, we do not find support for **H2**.

H3: The impact of consideration in the intermediate steps
As Table 7.5 shows, participants in MSTworkflow+ condition showed less reliance on the AI system and worst self-reliance (*RSR* measure) and team performance compared to MSTworkflow condition. In contrast to our expectation, the document usefulness annotation intervention fails to bring higher appropriate reliance at the intermediate steps (*i.e., AR-Intermediate*) – MSTworkflow condition achieved better *AR-Intermediate* than MSTworkflow+ condition. Thus, the results from Table 7.5 are not sufficient to verify **H3**.

To analyze how explicit consideration in the intermediate steps shapes user reliance, we evenly re-split participants in MSTworkflow and MSTworkflow+ conditions based on *AR-Intermediate* – high consideration (with higher *AR-intermediate*, top 50%) and low consideration (with lower *AR-intermediate*, bottom 50%). Similar to the statistical analysis for **H1** and **H2**, we conducted Kruskal-Wallis H-test to compare the performance, reliance, and appropriate reliance measures of participants across the groups of participants. The results are shown in Table 7.6.

Table 7.6: Kruskal-Wallis H-test results for **H3**. "††" indicates the effect of the variable is significant at the level of 0.017.

Dependent Variables	H	p	$M \pm SD$		Post-hoc Results
			high consideration	low consideration	
Team Performance	21.90	**.000**††	0.70 ± 0.11	0.58 ± 0.13	high > low
Agreement Fraction	7.43	**.006**††	0.78 ± 0.14	0.70 ± 0.14	high > low
Switch Fraction	2.42	.120	0.54 ± 0.27	0.44 ± 0.28	-
Team Performance-wid	7.97	**.005**††	0.56 ± 0.22	0.45 ± 0.21	high > low
RAIR	7.71	**.005**††	0.59 ± 0.33	0.41 ± 0.31	high > low
RSR	0.01	0.917	0.56 ± 0.48	0.55 ± 0.48	-

For the measures we found a significant difference between the two groups, we conducted a post-hoc Mann-Whitney test to reach a conclusion. As we can see, participants with higher explicit considerations of the intermediate steps in MST workflow will achieve higher *RAIR* and a similar level of *RSR*. It infers that low explicit considerations of the intermediate steps may cause under-reliance. Thus **H3** is supported by our experimental results.

We also found that participants with a high *AR-Intermediate* achieved comparable performance across all conditions, which provides support for the effectiveness of the MST workflow. Meanwhile, participants in group low consideration showed much worse team performance, *Agreement Fraction* and *RAIR* (non-significant). This indicates that precise decisions at the intermediate steps play a critical role in making the MST workflow effective in human-AI collaboration.

7.5.3 Exploratory Analysis

Impact of Trust

Inspired by prior work [40, 234], we analyzed how subjective user trust differs across conditions to provide further insights about the impact of different decision workflows. We conducted an *Analysis of Covariance* (ANCOVA) with the *decision workflow* as independent variable and *TiA-Propensity to Trust*, *TiA-Familiarity*, *LLM Expertise*, and *Fact Checking Expertise* as covariates. This allows us to explore the main effects of the decision workflow on subjective trust measured with the Trust in Automation questionnaire [321]. Table 7.7 shows the ANCOVA results of the trust-related dependent variables. While there exists a borderline impact of decision workflow in *TiA-R/C* and *TiA-U/P*, the results are non-significant. Thus, we found that user trust in the AI system was not influenced by the decision workflow. However, we found that participants' general *Propensity to Trust* had a significant impact on their trust. *Propensity to Trust* shows a strong positive correlation with user trust, which will be detailed in Section 7.5.3.

Table 7.7: ANCOVA test results on trust-related dependent variables. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.017, respectively.

Dependent Variables	**TiA-R/C**			**TiA-U/P**			**TiA-IoD**			**TiA-Trust**		
Variables	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
Exp Condition	2.78	.042†	0.02	3.31	.021†	0.04	1.63	.183	0.01	0.85	.470	0.01
LLM Expertise	2.33	.128	0.01	1.50	.222	0.01	1.00	.319	0.00	2.13	.146	0.00
Fact Checking Expertise	1.83	.178	0.00	0.01	.934	0.00	0.17	.676	0.00	0.63	.427	0.00
TiA-Propensity to Trust	211.76	**.000**††	0.47	36.31	**.000**††	0.13	96.24	**.000**††	0.29	211.82	**.000**††	0.48
TiA-Familiarity	3.59	.059	0.01	0.01	.922	0.00	5.06	.025†	0.02	2.08	.151	0.00

Impact of covariates

In our study, we considered user expertise in large language models (*LLM Expertise*), user expertise in fact-checking tasks (*Fact Checking Expertise*), *Familiarity*, and *Propensity to Trust* obtained from questionnaires as covariates. These covariates may have a substantial impact on user trust and user reliance. To analyze how the covariates impact the dependent variables used in our study, we conducted Spearman rank-order tests between covariates and all dependent variables (*i.e.,* performance, reliance, appropriate reliance, trust, and cognitive load). The corresponding results are presented in Table 7.8. Our findings suggest that — (1) Participants with a relatively higher *LLM Expertise*, *Fact Checking Expertise*, or *Familiarity* reported a higher cognitive load. (2) Participants with a relatively higher *LLM Expertise*, *Familiarity*, or *Propensity to Trust* also reported a higher level of trust in the AI system. (3) Participants with a relatively higher *Propensity to Trust* exhibited a a higher *Agreement Fraction* and *Switch Fraction*, indicating more reliance on the AI system. However, the increased reliance may translate into over-reliance, as this corresponds with a significant negative correlation with *RSR*. (4) Interestingly, higher *LLM Expertise* does not necessarily help improve team performance in this task. Instead, the weak negative correlation between *LLM Expertise* and *Team Performance* suggests that participants with higher *LLM Expertise* performed worse.

Appropriate Reliance at Intermediate Steps

To further explore the relationship between the appropriate reliance in fine-grained levels (*i.e., AR-Intermediate* and *AR-Evidence*) and global (appropriate) reliance, we conducted

Table 7.8: Spearman rank-order correlation coefficient for covariates level on dependent variables. "††" indicates the effect of the variable is significant at the level of 0.017.

Covariates	LLM expertise		Fact checking expertise		Familiarity		Propensity to Trust	
Dependent Variables	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Mental demand	0.073	.270	0.043	.517	-0.094	.152	0.072	.275
Physical demand	0.326	**.000**††	0.233	**.000**††	0.250	**.000**††	0.131	.046
Temporal demand	0.158	**.016**††	0.093	.157	0.178	**.007**††	0.053	.424
Performance	-0.050	.452	-0.104	.112	-0.056	.397	-0.020	.763
Effort	0.152	.020	0.061	.357	-0.072	.272	0.109	.098
Frustration	-0.020	.757	-0.029	.660	0.017	.795	-0.145	.027
Reliability/Competence	0.215	**.001**††	0.035	.597	0.268	**.000**††	0.699	**.000**††
Understanding/Predictability	0.146	.026	0.054	.412	0.106	.106	0.371	**.000**††
Intention of Developers	0.272	**.000**††	0.138	.035	0.311	**.000**††	0.580	**.000**††
Trust in Automation (TiA)	0.247	**.000**††	0.079	.229	0.273	**.000**††	0.725	**.000**††
Team performance	-0.158	**.016**††	-0.083	.204	-0.143	.029	-0.077	.240
Agreement Fraction	-0.017	.791	-0.002	.977	0.006	.924	0.174	**.008**††
Switch Faction	0.049	.453	0.036	.582	0.054	.414	0.217	**.001**††
Team Performance-wid	-0.127	.052	-0.070	.285	-0.149	.023	-0.086	.191
RAIR	-0.046	.480	-0.022	.736	-0.015	.823	0.136	.039
RSR	-0.120	.068	-0.087	.187	-0.095	.149	-0.235	**.000**††

the Spearman rank-order test separately for participants in `MSTworkflow` and `MSTworkflow+` conditions. The results are shown in Table 7.9. We found a strong positive monotonic relationship between the fine-grained appropriate reliance (*i.e., AR-Intermediate* and *AR-Evidence*) and performance-based measures (*Team Performance* and *Team Performance-wid*). *AR-Intermediate* also shows some positive impact on *Agreement Fraction* and *RAIR*, which indicates that participants with higher *AR-Intermediate* have less chance to be impacted by the under-reliance issues. In contrast to what one can intuitively expect, we found that *AR-Evidence* does not necessarily show a significant positive correlation with appropriate reliance. We will further discuss these findings in section 7.6.

Table 7.9: Spearman rank-order correlation coefficient for AR-Intermediate and AR-Evidence on dependent variables. "††" indicates the effect of the variable is significant at the level of 0.017.

Dependent Variables	Team performance		Agreement fraction		Switch faction		Team performance-wid		RAIR		RSR	
Fine-grained AR	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
AR-Intermediate	0.477	**.000**††	0.300	**.000**††	0.114	.217	0.308	**.001**††	0.270	**.003**††	0.089	.337
AR-Evidence	0.474	**.000**††	0.227	.090	-0.066	.624	0.300	.023	0.132	.329	0.123	.362

Confidence Dynamics

User confidence in their decision also plays an important role in shaping their reliance on AI systems. We illustrated user confidence dynamics using a line plot (Figure 7.5) based on their average confidence along with task order. For participants in `MSTworkflow` and `MSTworkflow+` conditions, we calculated the average confidence (*i.e.,* Initial-avg) and minimum confidence (*i.e.,* Initial-min) of the three intermediate steps for their initial decision confidence. Overall, participants were confident with both their initial and final decisions, which is around 4.0 (corresponding to "*somewhat confident*"). It is evident that the gap in confidence for participants in `Control` and `MST-GT` conditions is relatively smaller than the participants in `MSTworkflow` and `MSTworkflow+` conditions. We also found that the gap between the average initial confidence and minimum initial confidence is relatively stable

(around 0.5 on a 5 point scale). Compared with `MSTworkflow` condition, the participants in `MSTworkflow+` condition showed relatively lower initial confidence and final confidence. This reflects that participants indicated more uncertainty about their decisions in the presence of the document usefulness annotation in `MSTworkflow+` condition.

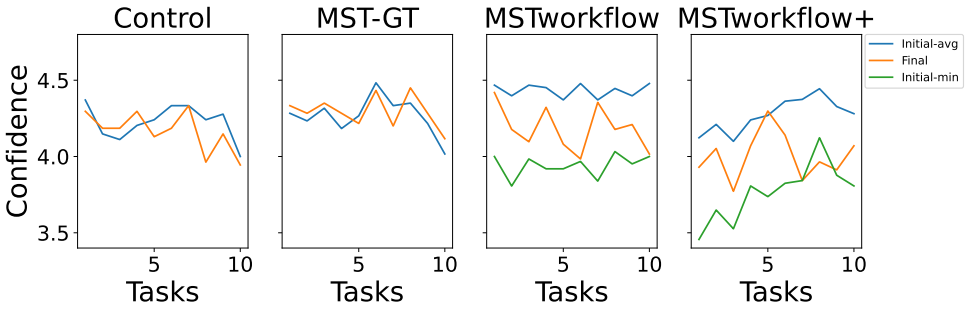


Figure 7.5: Line plot illustrating the confidence dynamics among users after receiving the AI advice (and explanations). The orange line and blue line illustrate the confidence dynamics before and after receiving AI advice (and explanations), respectively.

7.6 Discussion

7.6.1 Key Findings

Our experimental results show that the multi-step transparent workflow can be effective in specific contexts (*e.g.,* in challenging tasks where AI advice is misleading) and that appropriate reliance at intermediate steps is important to ensure effective human-AI collaboration and team performance. Our analysis of appropriate reliance at the intermediate steps highlights that the MST workflow can facilitate human-AI collaboration when users make explicit considerations of the intermediate steps (cf. Table 7.6). We also found that participants who do not demonstrate fine-grained appropriate reliance at the intermediate steps may exhibit under-reliance behavior on the final AI advice. Based on the user confidence dynamics, we found that participants with the MST workflow reported higher confidence in their initial decision, but also showed a decreased confidence after checking the AI advice and in the presence of transparency cues such as intermediate steps and intermediate answers. These findings can be explained as follows—participants with a MST workflow obtained more verifiability of the AI advice, which contributes to developing higher self-confidence (reflected by initial confidence) and a critical mindset on AI advice. These benefits may help mitigate over-reliance caused by the potential illusion of explanatory depth shown in condition `MST-GT`, but they may also decrease user reliance on the AI system and cause under-reliance when they make errors in the intermediate steps.

Impact of the Multi-step Transparent (MST) Workflow on User Reliance. In our study, we found that global transparency (*i.e.,* overall logic of the AI system — task decomposition and intermediate answers) can have the negative impact of increasing over-reliance on the AI system (cf. Table 7.4 and Table 6.3). This is consistent with findings in explainable AI literature — explainable AI can cause over-reliance on the AI system [55, 401].

Although the MST workflow does not always work as expected to facilitate appropriate reliance on the AI system, our results suggest that it works well in some specific contexts (*e.g.,* challenging tasks where AI advice can be misleading). Similar findings have been pointed out in crowdsourcing literature – crowdsourcing workflows can be useful in specific contexts but may also constrain complex work [355]. While the multi-step workflow can help address over-reliance on the AI system, it can also introduce under-reliance when users make mistakes at the intermediate steps. Such impact is highly similar to the impact of second opinion in AI-assisted decision making [277]. This is because, when there is no performance feedback, users may doubt the accuracy level of the AI assistance after disagreements with AI advice (caused by misleading second opinion / intermediate decision). As a result, users may decrease their trust and reduce reliance on the AI system.

Document Usefulness Annotation Did Not Work as Expected. The cognitive forcing function of document usefulness annotation (condition `MSTworkflow+`) did not show the effectiveness to increase *AR-Intermediate.* In contrast to our expectations, the intervention to increase user consideration of the evidence on top of the MST workflow resulted in decreased team performance and relatively lower appropriate reliance. With such an intervention, users demonstrated less confidence in themselves at intermediate steps and final decisions. Consistent with our expectations, we found that participants in the condition `MSTworkflow+` reported much higher cognitive load than other conditions (cf. Figure 7.4). A potential explanation is that the byproduct of a high cognitive load overrides the benefits of the multi-step transparent workflow. The decreased self-reported confidence in decisions can be a signal of uncertainty. As a result, users may turn to rely more on the AI system, which helps explain the over-reliance on misleading AI advice (cf. Table 6.3).

Our analysis of the impact of covariates (Section 7.5.3) also provides interesting insights: (1) participants who reported a relatively higher LLM expertise tended to show higher subjective trust but performed worse (cf. Table 6.4); (2) participants with a higher propensity to trust showed higher subjective trust, and higher reliance, in fact resulting in over-reliance. (3) Different from other covariates, the fact checking expertise does not show a strong correlation with trust and reliance on the AI system. Based on these observations, we can infer some user factors (*i.e.,* LLM expertise, familiarity, and propensity to trust in our study) can potentially increase user trust. However, the increased user trust is not calibrated according to the actual AI performance, which hinders effective human-AI collaboration. In summary, these user factors may lead to uncalibrated trust in the AI system, which causes over-reliance. These findings are consistent with previous empirical studies of AI-assisted decision making [30, 234], and our work sheds light on how they extend to the context of human-AI decision making using a multi-step transparent workflow.

This is the first work that has explored how a multi-step transparent workflow shapes user reliance and when such a workflow can be effective. While previous work has extensively studied user reliance and appropriate reliance at a global level [22], our work is the first to explore the fine-grained levels of appropriate reliance — appropriate reliance in the intermediate steps aligned with evidence. Our results indicate that participants who made better use of the intermediate steps and evidence achieved better team performance (cf. Table 7.6 and Table 7.9). Their reliance patterns were also positively impacted by their consideration of the intermediate steps (*i.e.,* positive correlation between *AR-Intermediate*

and *RAIR* in Table 7.9). These findings suggest promising future directions to explore in the context of decision making with *RAG*-based AI systems and human-AI collaboration with decision workflows.

7.6.2 Implications

Suitable workflows and user interventions can ensure effective human-AI collaboration. Our work has important implications for designing effective human-AI collaboration workflows. In our study, we found that participants who followed a basic one-step decision workflow performed better on relatively easy tasks. In comparison, participants who adopted a multi-step transparent workflow performed better on challenging tasks where AI advice was misleading. Our findings suggest that there is no one-size-fits-all solution for human-AI collaboration workflows. This echoes findings in previous analyses of workflows in crowdsourcing [355]. Multiple aspects in task characteristics (*e.g.,* task complexity [301]), user factors (*e.g.,* cognitive bias [30, 84]), and system transparency (*e.g.,* reasoning process [337]) may impact the final decision outcome. As opposed to seeking and designing for optimal human-AI workflows that can always lead to high effectiveness, future work can explore how to combine multiple human-AI workflows depending on different contextual requirements. For example, in tasks where an AI system demonstrates low confidence, we may expect more useful and independent input from human decision makers. We would then need to adopt a suitable workflow (such as an MST workflow) to improve team performance, a critical mindset for critical consideration of AI advice and the verifiability of AI advice.

We also found that some interventions (*i.e.,* condition `MSTworkflow+`) can pose a high cognitive load on users of the AI system, which can result in the side effects of user frustration and decreased effectiveness. This is consistent with prior findings of unforeseen negative impacts of user interventions [22, 218]. Such a phenomenon has also been observed in prior empirical studies [30, 84, 107, 277, 401] of AI-assisted decision making. For example, some prior work [84, 107, 401] found that explainable AI can help address under-reliance, but also simultaneously led to a higher over-reliance. Providing a second opinion shows a similar impact [277] on mitigating over-reliance while increasing under-reliance. These observations reveal that user interventions can be effective only in specific contexts. Therefore, a trade-off between the benefits and harms of user interventions can be a prerequisite for their effectiveness. Identifying the target audience and contextual requirements for user interventions based on behavioral and psychological patterns can be a promising direction to explore.

Fine-grained Analysis to Promote Appropriate Reliance. Our experimental results suggest that appropriate reliance at a global level and complementary team performance may be dependent on more fine-grained appropriate reliance on the intermediate steps (*i.e.,* global transparency) and supporting documents (*i.e.,* local transparency). While most existing work has explored a *one-step* decision workflow, the user decision making process and user decision criteria are not accessible for analysis. Existing empirical studies typically set up experiments with several conditions by controlling factors about user, task, and AI system. In such a setup, the user reliance on more fine-grained task input and the surrounding context (*e.g.,* relevant documents) are typically not considered for

analysis. We argue that this limits us, as a community, from developing an insightful understanding of appropriate reliance on AI advice. In this spirit, our work has important methodological implications for both studying and promoting appropriate reliance with a fine-grained analysis. Our findings and implications can help develop human-centered AI systems for complex tasks highlighting accountability, like medical diagnosis, loan prediction, supply chain optimization, etc. The users would benefit from the critical mindset and intermediate results of a multi-step transparent workflow. In the future, human-centered AI studies exploring appropriate reliance could consider more structured workflows for decision making and operationalize fine-grained user reliance.

7.6.3 Caveats and Limitations

Transferability Concerns. In our study, we used the lens of a single complex task — composite fact-checking supported with retrieved documents, and a specific AI system — an LLM-based system that first decomposes a complex fact and then verifies sub-facts by leveraging retrieval-augmented generation. It is unclear how our findings and implications can transfer to a different context where task characteristics (*e.g.,* difficulty, uncertainty and risks) and system characteristics (*e.g.,* transparency and system accuracy) are different [220, 301]. It is noteworthy that in multi-step decision workflows, there can be dependencies between sub-tasks. In our study, we considered sub-tasks that are largely independent (*i.e.,* sub-facts could be treated independently). This presents a limited view of multi-step decision workflows and future research is needed to extend our work to workflows with dependencies. Our setup of fine-grained transparency and multi-step decision workflow offers a relatively general framing to analyze multi-step human-AI collaboration and fine-grained appropriate reliance. Future work can follow this framing to explore how different aspects surrounding task characteristics, AI systems and user factors affect user trust and reliance in a multi-step human-AI collaboration.

Impact of cognitive load. We found significant differences in the perceived cognitive load of participants across different experimental conditions, suggesting that the `MSTworkflow+` condition required participants to consistently exert relatively more effort. A qualitative follow-up or an in-person user study with a selection of participants may provide deeper insights. More work is needed to understand the effectiveness of such workflows that are demanding of users during the decision making process itself. Such workflows may be less suited in low-stakes compared to relatively high-stakes contexts, where cognitive effort can be considered as a viable trade-off for better team performance.

Potential Bias. As our study is carried out with crowd workers, participants with a MST workflow may spend more effort and feel more temporal demand. Their performance may be impacted by *self-interest bias* [157]. To ensure the collected data is of high quality, we made sure that participants in each condition received a fair payment according to platform standards and provide bonuses to motivate correct decisions.

7.7 Conclusion

In this chapter, we conducted an empirical study to analyze how a multi-step transparent (MST) decision workflow shapes user reliance in a composite fact-checking task. Com-

pared to a basic workflow and one-step human-AI collaboration, participants with an MST workflow showed a more critical mindset in making decisions while relying on AI advice. As a result, the MST workflow tackled over-reliance on AI advice to some extent but also led to a decrease in user reliance on the final AI advice and user confidence, which may cause under-reliance (**RQ1**). When participants demonstrate explicit considerations of AI advice in the intermediate steps (*i.e.,* when they can achieve high appropriate reliance at fine-grained levels), such under-reliance can be mitigated. Then the MST workflow can facilitate effective human-AI collaboration. Increasing the transparency of the AI system by providing intermediate steps and answers caused over-reliance on the AI system. At the same time, we found that the MST workflow with an additional task that attempts to cognitively engage participants by annotating the supporting documents and increase critical reflection may pose a demanding cognitive load on participants. This resulted in harming the overall human-AI collaboration, as reflected by the user experience, team performance, and appropriate reliance of participants in that experimental condition. Having said that, through further analysis (cf. Table 7.6), we found appropriate reliance at the level of intermediate steps may be required to ensure the effectiveness of the MST workflow. We also found that appropriate reliance on the retrieved evidence is positively correlated with team performance. Based on this finding, we infer that the transparency of the AI system at the level of task input can also play a positive role in facilitating appropriate reliance and complementary human-AI collaboration (**RQ2**). More work is required to further advance our understanding of this problem.

Our results indicate that the MST workflow can be effective in specific contexts, and there is no one-size-fits-all decision workflow to achieve optimal human-AI collaboration. A trade-off between the benefits (*e.g.,* fine-grained transparency and critical consideration of AI advice, more verifiability) and side effects (*e.g.,* higher cognitive load) of decision workflows should be considered in human-AI collaboration. Our findings have important implications for designing effective decision workflows to facilitate appropriate reliance and better human-AI collaboration.

8

Plan-then-execute LLM Agent Workflow

Since the explosion in popularity of ChatGPT, large language models (LLMs) have continued to impact our everyday lives. Equipped with external tools that are designed for a specific purpose (e.g., for flight booking or an alarm clock), LLM agents exercise an increasing capability to assist humans in their daily work. Although LLM agents have shown a promising blueprint as daily assistants, there is a limited understanding of how they can provide daily assistance based on planning and sequential decision making capabilities. We draw inspiration from recent work that has highlighted the value of 'LLM-modulo' setups in conjunction with humans-in-the-loop for planning tasks. We conducted an empirical study (N = 248) of LLM agents as daily assistants in six commonly occurring tasks with different levels of risk typically associated with them (e.g., flight ticket booking and credit card payments). To ensure user agency and control over the LLM agent, we adopted LLM agents in a plan-then-execute manner, wherein the agents conducted step-wise planning and step-by-step execution in a simulation environment. We analyzed how user involvement at each stage affects their trust and collaborative team performance. Our findings demonstrate that LLM agents can be a double-edged sword — (1) they can work well when a high-quality plan and necessary user involvement in execution are available, and (2) users can easily mistrust the LLM agents with plans that seem plausible. We synthesized key insights for using LLM agents as daily assistants to calibrate user trust and achieve better overall task outcomes. Our work has important implications for the future design of daily assistants and human-AI collaboration with LLM agents.

8.1 Introduction

Autonomous agents have been regarded as a research focus for artificial intelligence (AI) over the last century [402]. With the wish that autonomous agents can make our life better,

This chapter is based on a peer-reviewed paper: **Gaole He**, , Gianluca Demartini, Ujwal Gadiraju. *Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant.* CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan. https://doi.org/10.1145/3706598.3713218.

many autonomous agents have been designed as virtual personal assistants [403]. These AI assistants (*e.g.,* Siri) perform well (albeit imperfectly) in following user instructions to execute low-risk tasks like playing a song, reporting weather forecasts, or searching for an image to support everyday tasks. However, on tasks entailing potential risks (*e.g.,* monetary payments or hiring an employee), humans hesitate to trust such AI systems due to loss aversion [404] and algorithmic aversion [33, 113, 138, 405]. Only when users can obtain a sense of control by being able to modify the outcomes of imperfect AI can they overcome such algorithm aversion and be willing to collaborate with imperfect AI systems [35].

With the recent rise of large language models (LLMs) in natural language understanding and generation [13], researchers have started to analyze LLM-based agents and their applicability in a plethora of tasks [406, 407]. The term '*LLM agent*' refers to an artificial entity based on LLMs that perceives its context, makes decisions, and then takes actions in response [406]. Compared to existing deep learning and LLM-based methods (*e.g.,* chaining multiple LLMs [361]), LLM agents provide more flexibility in task solving and user interaction, which makes them suitable for daily assistance. This is primarily due to three reasons. First, with a planning module, LLM agents can generate a dynamic plan based on the tools provided [406, 407]. Such plans are typically defined in a logical structure — step-wise plans, which can be easily understood by humans. Second, with LLMs as a core control module, users can access and interact with external toolkits via a more natural interaction (*i.e.,* conversation) with LLM agents [13, 408], reducing manual control efforts over function-specific tools. For example, LLM agents can complete time-consuming jobs like information seeking and information filtering (*e.g.,* searching for a flight in itinerary planning) based on specific user needs. Third, the Markov decision process of LLM agents can generate a sequence of actions (*i.e.,* using external toolkits) as output.[1] Paired with an understanding of actions and necessary parameters for the interaction with the LLM agents, users can get involved in the real-time execution of tasks with LLM agents and fix potential problems while benefiting from task delegation.Based on an intuitive framework for task delegation, Lubars *et al.* [409] found that user trust can play an important role in human delegation behaviors to AI systems. However, there is a relatively limited understanding of user trust development and calibration in collaboration with LLM agents.

There is also a growing debate in the machine learning and AI research communities about whether LLMs can be truly considered as planning and reasoning agents [410]. With this in the backdrop, existing work on automated task completion has revealed that LLM agents can exhibit promising performance in handling complex tasks like playing games [411], answering complex questions [412], and in simulating social behavior [413]. However, such agents are still far from perfect [410, 414]. Due to the probabilistic nature of LLMs, there is much uncertainty in automating LLM agents for tasks with high risks attached. To avoid unintended or unexpected consequences, there is a need for user control over the real-time execution process. Through an empirical study of LLM planning capabilities, planning experts found that "LLMs' ability to generate executable plans autonomously is rather limited" [414]. However, when combined with a sound planner in an 'LLM-Modulo' mode, "the LLM-generated plans can improve the search process for underlying sound planners" [414]. Humans can potentially be the 'sound planners' who

[1] In our study, the usage of one tool is the same as executing one action. Therefore, we refer to a tool and an action interchangeably.

can work in conjunction and optimize plans drafted by LLMs, which can then be executed by LLM agents. Such human-AI collaboration can reduce human efforts in generating a reliable plan from scratch.

Drawn by the promise of LLM agents, there have been some early explorations [415] of adopting them in human-AI collaboration. However, existing works have primarily analyzed how LLM agents can serve specific use cases (*e.g.,* design creation [415]), while others have conducted structured interviews to obtain expert insights [416, 417]. Yet, little is known about how well LLM agents can work as general purpose daily assistants—to assist users in everyday tasks with varying stakes—and how user trust and team performance evolve by interacting with LLM agents.

In our work, we address this research gap and adopt LLM agents to assist humans in everyday tasks by following a plan-then-execute workflow [418]. First, the LLM agent generates a step-wise plan formulated with a hierarchical structure. Then, the LLM agent executes the generated plan by transforming it into a sequence of actions (leveraging external toolkits). The benefits of such a plan-then-execute framing are three-fold: (1) Compared to a dynamic process where planning and execution are bound closely, separating planning and execution into two stages provides more task clarity to the users, which reduces user cognitive load and contributes to the quality of task outcomes [419]. (2) With planning at the beginning of the task, users can develop a global understanding of how the LLM agents will execute the task. Based on a follow-up step-by-step execution, it would be straightforward for users to be involved in such a process and control the outcomes of task execution. (3) Planning and execution are representative abstractions of how LLM agents work. The findings of such an empirical study can be generalized to human-AI collaboration with other kinds of LLM agents (*e.g.,* dynamic planning-execution). To this end, we propose the following research questions:

- **RQ1**: How does human involvement in the high-level planning and real-time execution shape their trust in an AI system powered by LLM agents?
- **RQ2**: How does human involvement in the high-level planning and real-time execution of tasks with an AI system powered by LLM agents affect the overall task performance?

Addressing these research questions, we carried out an empirical study (N = 248) of human-AI collaboration in six different everyday scenarios with varying stakes and risks attached (*e.g.,* credit card payment and itinerary planning). We found that user involvement in the planning and execution can be beneficial in addressing imperfect plans and fixing execution errors. As a result, LLM agents can achieve better task performance. However, we also found that user involvement in the planning and execution stages of the LLM agent fails to calibrate user trust in corresponding task outcomes. A potential reason here is that the plausible plans generated by the LLMs can mislead users into trusting the LLM agents when they are in fact wrong. Our findings highlight that user involvement can also bring about additional trade-offs to consider: (1) user involvement in the planning and execution poses a high cognitive load on users and decreases user confidence in their decisions; (2) user involvement can be harmful in some task contexts (*e.g.,* user involvement reduces plan quality). Further research is required to understand when to provide necessary user involvement. Our key insight is that as opposed to following a fixed mode of user involvement, it is prudent to explore how user involvement in planning and exe-

cution can be tailored to fit the task and the user. Based on our quantitative and qualitative findings, we share insights for designing effective LLM agents as daily assistants and synthesize promising directions for further research around LLM agents in the context of human-AI collaboration. Our work has important theoretical implications for human-AI collaboration with LLM assistance and design implications for plan-then-execute LLM agents to support human-AI collaboration.

8.2 Background and Related Work

Our work proposes to analyze how user involvement in the planning and execution stages of LLM agents shapes user trust in the LLM agents and the overall task performance of LLM agents. Thus, we position our work in three realms of related literature: human-AI collaboration (§ 8.2.1), trust and reliance on AI systems (§ 8.2.2), task support with LLMs and LLM agents (§ 8.2.3).

8.2.1 Human-AI Collaboration

In recent decades, deep learning-based AI systems have shown promising performance across various domains [420, 421] and applications [10, 422]. However, such AI systems are not good at dealing with out-of-distribution data [423, 424], and their intrinsic probabilistic nature brings much uncertainty in practice [425]. Such observations raise wide concerns about the accountability and reliability of AI systems [365]. Under such circumstances, human-AI collaboration has been recognized as a well-suited approach to taking advantage of their promising predictive power and ensuring trustworthy outcomes [22, 426]. While humans can provide more reliable and accountable task outcomes, too much user involvement to check and control AI outcomes is undesirable [427]. It goes against the premise that AI systems are introduced to reduce human workload. In that context, researchers have theorized and empirically analyzed when and where users could and should delegate to AI systems [409, 427].

Task Delegation. While humans prefer to play the leading role in human-AI collaboration [409], delegating to AI systems can bring benefits like cost-saving and higher efficiency. Apart from manual delegation decisions, it is common to apply automatic rules for human delegation (*e.g.,* heuristics obtained from domain expertise or manually crafted rules [427]). Many user factors like trust [409], human expertise domain [219], and AI knowledge [428]) have a substantial impact on human delegation behaviors. Another relevant stream of recent research has explored AI delegation to humans [428–430]. Researchers have investigated the conditions under which AI systems should defer to a human decision maker, which may bring benefits of improved fairness [429], accuracy [431], and complementary teaming [432]. Compared to human delegation, AI delegation has been observed to achieve more consistent benefits in team performance [430, 433]. In collaboration with LLM agents, users need to determine when they should be involved in high-level planning and real-time execution. Such involvement decisions are similar to the delegation choices made by users. While task delegation is not the focus of our study, future work can explore this further.

AI-assisted Decision Making has attracted a lot of research focus in human-AI collaboration literature. Most existing work has conducted empirical studies [22] and structured

interviews [426] to understand how factors surrounding the user, task, and AI systems affect human-AI collaboration. User factors like AI literacy [108], cognitive bias [434], peer input [278], and risk perception [110, 120] have been observed to substantially impact user trust and reliance on the AI system. Similarly, task characteristics like task complexity and uncertainty [220, 301] and factors of the AI system (*e.g.,* performance feedback [49, 119], AI transparency [347], stated accuracy [234], and confidence of AI advice [66, 121]) also affect user trust and reliance on the AI system. For a more comprehensive survey of existing work on AI-assisted decision making, readers can refer to [22].

While machine learning and deep learning methods have been extensively analyzed in existing human-AI collaboration literature, to our knowledge, human-AI collaboration with LLM agents is still under-explored. Unlike previous studies where AI systems only follow a fixed mode to generate advice, LLM agents can be equipped with more logical clarity and can provide a step-wise plan and can follow a step-by-step execution. With such a plan-then-execute setup, LLM agents can bring high flexibility as well as uncertainty in high-level planning and real-time execution. Little is known about how well LLM agents can work as daily assistants while handling tasks entailing varying stakes and potential risks. In our study, we analyzed the impact of user involvement in such AI systems by adjusting their intermediate outcomes (plan and step-by-step execution) to calibrate their trust and improve task outcomes. Our findings and implications can help advance the understanding of the effectiveness of LLM agents in human-AI collaboration.

8.2.2 Trust and Reliance on AI systems

Trust and reliance have been important research topics since human adoption of automation systems [40, 435]. Due to the widespread integration of AI systems and LLMs in all walks of society, there has been a growing interest in understanding user trust [275, 276, 436] and reliance [32] on AI systems. User trust in the context of human-AI collaboration is typically operationalized as a subjective attitude toward AI systems/AI advice [40]. In comparison, user reliance on AI systems is based on user behaviors (*e.g.,* adoption of AI advice and modification of AI outcomes). The two constructs have been shown to be highly related [40, 437]: for example, user trust can substantially affect user reliance [40]. However, they are intrinsically different and cannot be viewed as a direct reflection of each other [348]. Most existing work has, therefore, studied the two constructs separately in terms of subjective trust and objective reliance.

Earlier work exploring human-AI trust primarily focused on the impact of different contextual factors surrounding user (*e.g.,* risk perception [110]), task (*e.g.,* task complexity [301]), and system (*e.g.,* stated accuracy [57, 66]). Empirical studies have shown that most users tend to trust AI systems that are perceived to be highly accurate [57]. Such trust is vulnerable, as the AI system may provide an illusion of competence with persuasive technology (*e.g.,* explanations [62, 359]) or overclaimed performance [57]. Even if the AI systems are accurate on specific datasets, they still suffer from out-of-distribution data [50, 438]. The misplaced trust in the AI systems can lead to misuse of the systems. Several empirical studies [87] have shown that once users realize the AI system errs or performs worse than expected, their trust in the AI system can be violated, even resulting in the disuse of the AI system. Both the misuse and disuse of the AI system hinder optimal human-AI collaboration.

To address such concerns, researchers have explored how to help users calibrate their trust in the AI system. Different techniques to help users realize the trustworthiness of the AI system have been proposed [111, 365, 439]. For example, increasing the transparency of AI systems by providing confidence scores [66], explanations [107], trustworthiness cues [440], and uncertainty communication [441]. However, the actual trustworthiness of the AI system does not always align with user perception. As found by Banovic *et al.* [326], untrustworthy AI systems can deceive end users to gain their trust. Another example is that users can develop an illusion of explanatory depth brought by explainable AI techniques [62], which leads to uncalibrated trust in the AI system. Even if users have indicated trust in the AI system, they may turn to rely more on themselves in final decision-making. The reasons are complex, and many factors, such as accountability concerns [442, 443] and cognitive bias [30], may affect user reliance behaviors.

While trust calibration is an important goal in human-AI collaboration, it may be not enough to ensure complementary team performance. Through empirical user studies with different confidence levels of AI predictions, Zhang *et al.* [66] found that "trust calibration alone is not sufficient to improve AI-assisted decision making". To achieve optimal human-AI collaboration, humans and AI systems need to play complementary roles [334, 444], and humans need to know when they should adopt AI assistance. In other words, humans should rely on AI advice when AI systems are correct and outperform them, and override AI advice when AI systems are incorrect or less capable than humans. Such user reliance patterns are denoted as *appropriate reliance* [29, 255], which is the key to achieving complementary team performance.

The main issues that lead to sub-optimal human-AI collaboration are: under-reliance (*i.e.,* disuse AI assistance when AI systems outperform humans) and over-reliance (*i.e.,* misuse AI assistance when AI systems are wrong or perform worse than humans) [29]. Users with an uncalibrated trust in the AI system can be easily misled to disuse or misuse AI systems [445]. Researchers have proposed various interventions to promote appropriate reliance [30, 49, 50, 108, 277, 446] and calibrate user trust in AI systems [55, 66]. For example, explainable AI methods have been shown to help reduce over-reliance [256] and under-reliance [107] in different scenarios albeit with little consistency across contexts. Another example is tutorial interventions, which have shown effectiveness in user onboarding [117], mitigating cognitive biases [30] and developing AI literacy [108]. For a more comprehensive overview of interventions to facilitate trust calibration and appropriate reliance, readers can refer to [22, 32, 348, 436].

LLM agents [407] have gained much popularity in recent years, distinguishing them from most prior AI systems. They can communicate through conversation, plan logically, and can be built to leverage powerful external tools to achieve complex functions. While trust and reliance have been extensively analyzed in existing human-AI collaboration literature, it is still unclear how users trust and rely on AI systems powered by LLM agents. In our work, calibrated trust is adopted as an important goal for human-AI collaboration in the planning and execution stage. Meanwhile, users are expected to fix potential errors in the planning and execution stages, reflecting their reliance on the AI system. Our work can substantially advance the understanding of trust and reliance on plan-then-execute LLM agents.

8.2.3 Task Support with LLMs and LLM Agents

LLMs and LLM agents bring new opportunities and challenges to human-AI collaboration [408]. It is evident that their generation capabilities can help reduce the cognitive effort from humans. But LLMs are also riddled with challenges such as hallucination [397] (*i.e.,* generated text seems plausible but is factually incorrect). Failure to handle such issues may bring fatal errors with unaffordable costs depending on the context (*e.g.,* medical diagnosis).

Due to the capability of generating coherent, knowledgeable, and high-quality responses to diverse human input [447], a wide community of human-computer interaction researchers has paid attention to large language models [366]. Researchers have actively explored how LLMs can assist users in various tasks like data annotation [448, 449], programming [450], writing [269, 273], and fact verification [451]. All the above functions can be achieved with elaborate prompt engineering using a single LLM. By chaining multiple LLMs with different functions, humans can customize task-specific workflows to solve complex tasks [361]. Apart from obtaining answers with a one-shot text generation, LLMs also provide convenient conversational interactions. Through empirical studies, such conversational interactions have been shown to be effective in human-AI collaboration with multiple applications, such as decision making [263, 452, 453], scientific swriting [269], and mental health support [454]. With the growing popularity of LLMs, more and more humans have begun to adopt LLMs (*e.g.,* ChatGPT) to boost their work efficiency and productivity [13].

LLM agents have been shown to have good planning, memory, and toolkit usage capabilities [406, 407]. When suitable toolkits are provided, LLM agents can readily generate a task-specific plan and solve the tasks using toolkits. Attracted by the promise of LLM agents, there have been some early explorations [415–417] of adopting them in human-AI collaboration contexts. These works were mostly analyzed in specific use cases (*e.g.,* design creation [415]). It is unclear how user trust and team performance are affected by user interactions with LLM agents in a sequential decision making setup (*i.e.,* solving a task by executing a sequence of actions) where users can be in control of the execution. To fill this research gap and advance our understanding of user control over LLM agents, we carried out a quantitative empirical study.

8.3 Method

8.3.1 Overview of User Involvement in Plan-then-execute LLM Agents

In our study, we adopted plan-then-execute LLM Agents [418] as assistants to help users handle daily tasks, *e.g.,* itinerary planning and currency transactions. Figure 8.1 illustrates how users collaborate with plan-then-execute LLM agents. First, the LLM agents will generate a step-wise plan based on a prompt specifying the plan format adopted from [455]. Then, users will make necessary edits to the plan based on the provided edit tools (will be further detailed in Section 8.3.2). After the user edit, we obtained the step-wise plan as outcomes of the planning stage. Next, the LLM agents will transform the step-wise plan into a sequence of action predictions, which will be served in a step-by-step manner. Users will join the real-time execution process and check whether they approve the current

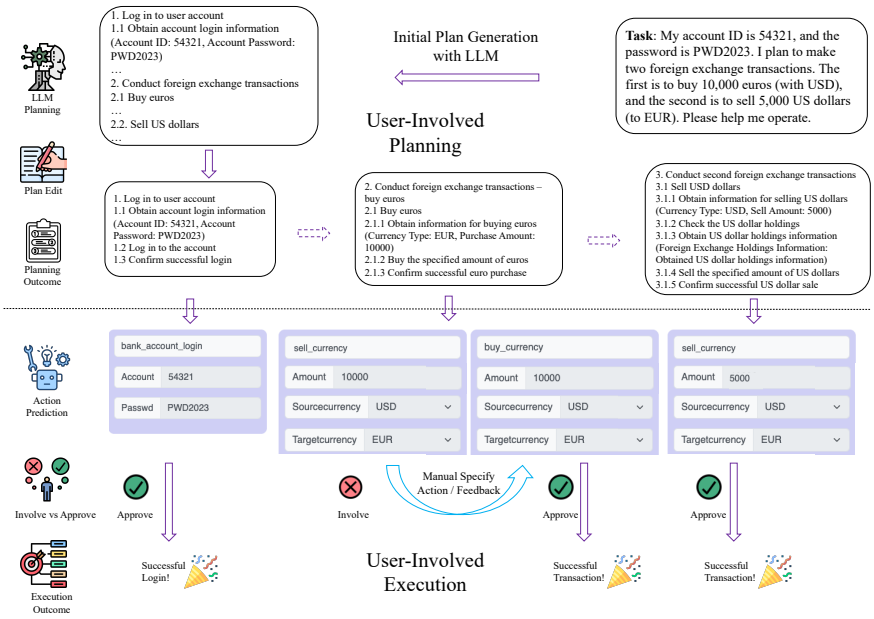


Figure 8.1: Illustration of the human-AI collaboration with plan-then-execute LLM agents.

predicted action (*i.e.,* blue card shown in Figure 8.1) or they would like to modify the current action prediction. The user involvement in execution stages will be introduced in Section 8.3.3. After the iterative execution of all steps, the task is solved. The evaluation of task performance is mainly based on the plan quality and execution accuracy of the action sequences.

Implementation details. In our study, we adopted GPT-3.5-turbo as the backbone LLM to serve the plan-then-execute LLM agent. The backend LLM agent implementation is mainly based on the Langchain plan and execute agent.[2] The execution of tasks are based on a simulation environment, where all tools/actions of the LLM agents are pre-defined as backend APIs hosted with Flask[3]. In the spirit of open science, all code and data analysis results can be found at Github.[4]

8.3.2 Planning

While LLMs can generate high-quality plans, there is no guarantee of their correctness and their further impact on the execution of the plan. Thus, involving users in the planning stage and controlling the plan quality would be essential to ensure successful subsequent execution.

Plan Format. The step-wise plan in our study followed a hierarchical structure, adapted

[2] https://api.python.langchain.com/en/latest/plan_and_execute/langchain_experimental.plan_and_execute.agent_executor.PlanAndExecute.html
[3] https://github.com/pallets/flask
[4] https://github.com/RichardHGL/CHI2025_Plan-then-Execute_LLMAgent

from a benchmark for LLM agents toolkit usage [455]. The whole plan consists of multiple sub-steps, which are at most three levels (*e.g.,* 1., 1.x, 1.x.y where x,y are integers). All sub-steps started with the same prefix index are denoted as one primary step (*e.g.,* the three blocks of planning outcome in Figure 8.1). A high-level step (*e.g.,* 1.) will provide high-level instruction of the current primary step, while low-level steps (*e.g.,* 1.x, 1.x.y) will provide subsequent details. In the execution stage, each primary step will be used as the execution unit. The LLM agent will transform one primary step into a predicted action filled with parameters. Thus, we ask participants to provide all necessary details in sub-steps of each primary step. Each primary step will be transformed into **single action** in the follow-up execution stage. If one primary step requires two actions to accomplish, it may cause a potential loss of one action. Thus, when a plan contains one primary step that contains information about two potential actions (*e.g.,* the initial plan in Figure 8.1), we consider it as a low-quality plan with 'grammar errors.'[5] All these plan format designs are informed in our onboarding tutorial.

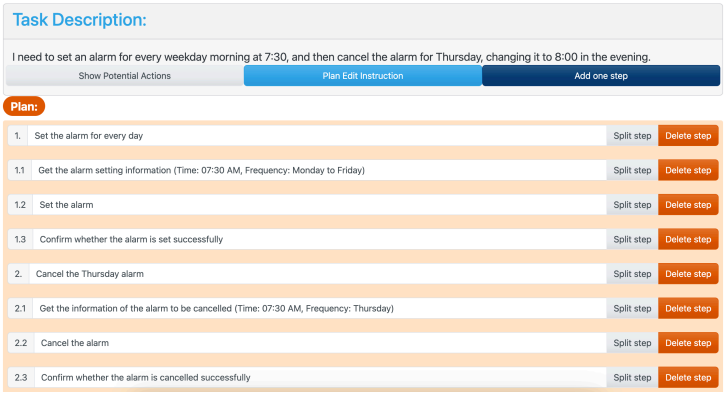


Figure 8.2: Screenshot of user-involved planning interface.

User-involved Planning. Figure 8.2 shows one screenshot of user-involved planning in our study. At the top of the interface, we provide a task description along with three buttons: 'Show Potential Actions', 'Plan Edit Instruction', and 'Add one step'. By clicking 'Show Potential Actions', we provide a prompt window to show concrete documentary descriptions of all potential actions (including action purpose and parameters) to be used in the execution stage. All instructions used in our tutorial are accessible with clicking the button 'Plan Edit Instruction'. After users join the planning stage, an initial plan generated by LLM will be presented in the orange area. We allow users to edit the plan with following interactions:

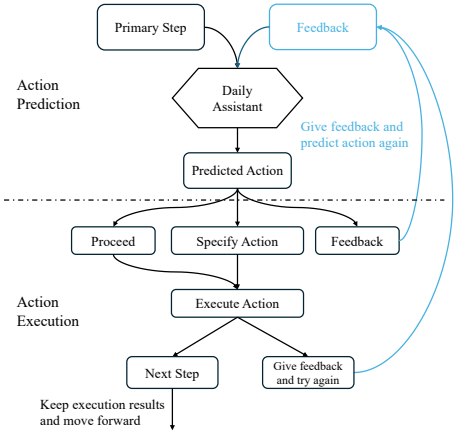
- Add step. By clicking 'Add one step' button, users can insert a valid sub-step index into the whole plan, and then they can edit the plan text.
- Delete step. By clicking the 'Delete step' button at the end of one step, all sub-steps associated with that step will be deleted from the plan.

[5]Note that this is not to be confused with the notion of grammar in language.

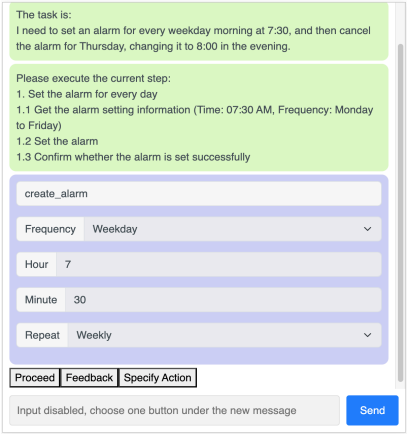
- Edit step. By clicking the text input area in each step, users are allowed to edit the text with keyboard input.

- Split step. By clicking the 'Split step' button associated with one step, we will split the original primary step into two primary steps. A new primary step will start the current step and contain all follow-up sub-steps. For example, if we click 'Split step' for the plan show in Figure 8.2 at index '2.2'. We will generate a new blank step '3.' (where user input is expected) and re-index all sub-steps with '2.2.x' to '3.1.x'. At the same time, the original plan steps behind it will be automatically updated. Through this action, users can easily split one step that contains too much information into two primary steps. Figure 8.1 shows an example of plan edit with 'split step'.

8.3.3 Execution



(a) Illustration of user-involved execution of one primary step.



(b) Screenshot of conversation interface for user-involved execution.

Figure 8.3: User-involved execution flow chart and interface. Panel (a): a flow chart illustrating how each primary step is executed with two stages: action prediction and action execution. Panel (b): a screenshot of the conversation interface for user-involved execution.

After the planning stage, we obtain a plan with a step-wise structure. In the execution stage, the LLM agent executes the outcome of the planning stage (*i.e.,* a step-wise plan in text) in a step-by-step manner. In each step, the LLM agent translates a single step of the plan into one action, which is implemented with an API call in the backend. This setup is a simulation of real-world applications, which provide services with API calls (commonly implemented as langchain tools[6]). Such a simulation setup is effective in developing and validating theory [456] and has been widely adopted in existing research on agent-based modeling and HCI studies [457]. To provide a smooth user experience, we adopted a conversational interface to present the execution process. Figure 8.3b shows one screenshot

[6] https://python.langchain.com/v0.1/docs/modules/tools/

of user-involved execution in our study. As we can see, after a message of the first primary step of the plan, the LLM agent predicts one action 'create_alarm'. In our study, to provide a tidy view of the action prediction, we wrap the predicted action as one card (the blue area in Figure 8.3b).

User-involved Execution. Figure 8.3a presents a flow-chart to illustrate a primary step executed by the daily assistant (*i.e.,* LLM agent). First, given one primary step, the daily assistant predicts an action based on a given list of prepared actions (*i.e.,* pre-defined APIs in the backend). After users check the predicted action, they can choose from one of the following three buttons to respond. **(1)'Proceed':** It indicates users agree that the predicted action is correct. After clicking this button, the LLM agent moves forward to execute it and shows the execution result of this action. **(2) 'Feedback':** Users can give text feedback based on the message input area at the bottom of the conversational interface. This triggers another action prediction based on the current primary step and user feedback. Then, users are provided with the three options to proceed again. **(3) 'Specify Action':** Users can override the current action prediction with the manual specification of one action. If users choose this response, they are first asked to choose one action from the prepared action list and then fill in the parameters manually. The LLM agent directly executes the user-specified action. After one action is executed, if users are not satisfied with the results, they can choose to re-execute this step by providing text feedback (*i.e.,* by clicking button 'Give feedback and try again'), which works similarly to the 'Feedback' option. If users are satisfied with the execution results, they can click the 'Next Step' button and move to execute the next primary step. By iterating over this process through the step-wise plan, users can choose to either approve or get involved in modifying the execution outcomes in each step. All actions are predicted and executed in the backend (*i.e.,* the respective API calls are triggered).

8.3.4 Hypotheses

Our experiment is designed to answer questions of how human involvement in the planning and execution stages will shape their trust and overall task performance. To analyze such impact, we regulate the levels of automation in the LLM agent through the planning and execution stage as baselines for comparison. The automatic planning and execution denotes that the LLM agent directly generates the task outcomes without user involvement.

With user involvement in the planning stage, users have the opportunity to fix potential mistakes or issues in the plan generated by LLMs. Working on such plan editing tasks is similar to debugging, which has been argued to bring about a critical mindset [31] to the generated plan. With a critical mindset, users may better calibrate their trust in the planning outcome. We also consider user involvement in planning to be beneficial to the plan quality, which can then contribute to the overall task performance. Thus, we hypothesize that:

(H1): Compared to automatic planning, user-involved planning will result in a higher calibrated trust in the plan.

(H2): Compared to automatic planning, user-involved planning will result in better overall task performance.

In the user-involved execution process, users manually check the action prediction and execution results of each primary step. Such user involvement increases the chances of discovering potential mistakes of LLM agents. Once users realize that the LLM agent made mistakes, they can get involved in modifying the execution outcome of the current step. By fixing these mistakes, the overall task performance gets improved. With such involvement in fixing potential errors, users will be more critical of trusting the task outcome. Therefore, we hypothesize that:

(H3): Compared to automatic execution, user-involved execution will result in a higher calibrated trust in execution outcome.
(H4): Compared to automatic execution, user-involved execution will result in better overall task performance.

8.4 Study Design

This section describes our experimental conditions, tasks, variables, procedure, and participants in our study. Our study was approved by the human research ethics committee of our institution.

8.4.1 Experimental Conditions

In our study, users collaborate with LLM agent-based daily assistants in two stages: planning and execution. To comprehensively understand the effect of user involvement at each stage, we considered a 2 × 2 factorial design with four experimental conditions: (1) automatic planning, automatic execution (represented as AP-AE), (2) automatic planning, user-involved execution (represented as AP-UE), (3) user-involved planning, automatic execution (represented as UP-AE), (4) user-involved planning, user-involved execution (represented as UP-UE). In conditions with user-involved planning, users are allowed to edit the plan generated by LLM with the actions of edit/add/delete/split step. By comparison, in conditions with automatic planning, users will directly adopt the plan generated by the daily assistant. In conditions with user-involved execution, users can interact with the step-by-step execution LLM agent (cf. Section 8.3.3) and refine execution results with text feedback or manual specification. By comparison, in conditions with automatic execution, users will directly accept the automatic execution results.

8.4.2 Tasks

To analyze how LLM agents can serve as daily assistants, we adopted tasks from a planning dataset designed for LLM agents — UltraTool [455]. We selected daily scenarios: currency transactions, credit card payments, repair service appointments, alarm setting, flight ticket booking, and trip itinerary planning. The selected tasks are shown in Table 8.1. For more details about how the plan-then-execute LLM agent works on the selected tasks

Table 8.1: Selected tasks in our study. The 'Risk' is based on the risk feedback obtained with pilot study. #A and #C refer to the number of actions and the number of named concepts in each task, respectively.

ID	Risk	Domain	Task Description	#A	#C	Notes
1	High	Finance	My account ID is 54321, and the password is PWD2023. I plan to make two foreign exchange transactions. The first is to buy 10,000 euros (with USD), and the second is to sell 5,000 US dollars (to EUR). Please help me operate.	4	4	simple task, imperfect plan
2	High	Finance	Please inquire about the current debt amount of my credit card with the last five digits 12345, and deduct the corresponding 12000 USD from my savings card number 6212345678900011 to repay this debt, then help me check the amount of the outstanding bill for the same credit card within 30 days after today.	4	6	complex task, imperfect plan
3	High	Repair	I need to schedule a repair for my TV at 6 PM tomorrow evening. The brand is Sony, model X800H, and there is an issue with the screen. Please book the repair service and tell me the reservation number.	4	7	complex task, imperfect plan
4	Low	Alarm	I need to set an alarm for every weekday morning at 7:30, and then cancel the alarm for Thursday, changing it to 8:00 in the evening.	2	3	simple task, correct plan
5	Low	Flight	I have an important meeting to attend next Wednesday, and I need to book a flight ticket from London to Amsterdam for tomorrow, it must be a morning flight, and then return from Amsterdam to London tomorrow night, please handle it for me.	2	6	simple task, correct plan
6	Low	Travel	Please plan a trip for me departing on October 1st at 8:00 AM to Japan, returning on October 7th at 11:00 PM, including Tokyo Disneyland, Sensoji Temple, Ginza, Mount Fuji, Kyoto cultural experience, Universal Studios Osaka, and visiting the Nara Deer Park on October 4th, and help me find hotels where the nightly cost does not exceed 10,000 Japanese yen.	3	11	complex task, correct plan

(*e.g.,* automatic plan, pre-defined actions, automatic evaluation, and explanation for errors in automation), please refer to the appendix. All tasks in UltraTool dataset are annotated with the step-wise plan format described in Section 8.3.2. The execution of these tasks is based on a simulation environment (described in Section 8.3.3) where all required actions are implemented as backend APIs. In our study, all tasks are executed in a simulation setup, which has been a popular method for orchestrating meaningful human-centered AI studies [183, 220].

Task Selection. First, based on the domain distribution of the UltraTool dataset, we selected seven domains: Finance, Alarm, Travel, Tracking, Restaurant, Flight, and Repair. For each domain, we only consider tasks that contain more than ten steps (including all sub-steps) and require at least three uses of actions. Then, we manually selected ten tasks: four from the finance domain and one for each of the others. With a pilot study, we tested how users work on the ten tasks. We recruited 10 participants from the Prolific platform and only considered the feedback of 9 participants who passed all attention checks. Using the question "How much risk do you perceive in this task when relying on this daily AI assistant?", we collected the perceived risk of working with the LLM agents on each task using a 5-point Likert scale, ranging from *1: not risky at all*—to—*5:very risky*. We categorize the ten tasks into a high-risk group (top 5) and a low-risk group (bottom 5). We selected three tasks from each group while balancing the complexity of the task description (three simple tasks and three complex tasks) and the correctness of the provided plan (three correct plans and three imperfect plans). Based on existing literature on task complexity [301, 458], we considered component complexity to inform our selection. This is assessed as the 'total number of distinct information cues that need to be processed to

perform the task'. Here, we considered the number of unique actions and the number of named concepts provided in each task. According to prior work [459], most people can only handle 5 to 9 concepts at the same time. The component complexity of all complex tasks in our study is more than nine. The six tasks selected are shown in Table 8.1. Besides the six tasks, we used one simple task (*i.e.,* checking bank account balance) as the example in the onboarding tutorial.

8.4.3 Measures and Variables

The variables and measures used in our study refer to existing empirical studies of human-AI collaboration [22]. All measures adopted in our study can be summarized in Table 8.2.

Table 8.2: The different variables considered in our experimental study. "DV" refers to the dependent variable.

Variable Type	Variable Name	Value Type	Value Scale
Calibrated Trust (DV)	Calibrated Trust in planning (CT_p)	Binary	0: miscalibrated trust, 1: calibrated trust
	Calibrated Trust in execution (CT_e)	Binary	0: miscalibrated trust, 1: calibrated trust
Task Performance (DV)	Plan Quality	Likert	5-point, 1: low, 5: high
	Action Sequence Accuracy (ACC_s)	Binary	0: mismatch, 1: exact match with ground truth
	Execution Accuracy (ACC_e)	Binary	0: wrong execution result, 1: correct execution result
Trust	Reliability/Competence	Likert	5-point, 1: poor, 5: good
	Understanding/Predictability	Likert	5-point, 1: poor, 5: good
	Intention of Developers	Likert	5-point, 1: poor, 5: good
	Trust in Automation	Likert	5-point, 1:strong distrust, 5: strong trust
Covariates	LLM Expertise	Likert	5-point, 1: No experience, 5: Extensive experience
	Automatic Assistant Expertise	Likert	5-point, 1: No experience, 5: Extensive experience
	Propensity to Trust	Likert	5-point, 1: tend to distrust, 5: tend to trust
	Familiarity	Likert	1: unfamiliar, 5: very familiar
Exploratory	Confidence	Likert	5-point, 1: unconfident, 5: confident
	Risk Perception	Likert	5-point, 1: not risky at all, 5: very risky
	Open Feedback on Planning	Text	Open Text
	Open Feedback on Execution	Text	Open Text
	Other Open Feedback	Text	Open Text
Cognitive Load	Mental Demand	Likert	-7: very low, 7: very high
	Physical Demand	Likert	-7: very low, 7: very high
	Temporal Demand	Likert	-7: very low, 7: very high
	Performance	Likert	-7: Perfect, 7: Failure
	Effort	Likert	-7: very low, 7: very high
	Frustration	Likert	-7: very low, 7: very high

Calibrated Trust. To assess calibrated trust in the planning stage and execution stage, we assessed user trust at each stage with a question "Do you trust that [the execution of this plan / the execution process] can provide a correct outcome based on the task instructions?". Based on the plan quality evaluation (5-point Likert), the calibrated trust in the planning (CT_p) is calculated based on the frequency at which users trusted the high-quality plan (expert annotation with 5) and users distrusted the plan with other evaluation results. Similarly, for the calibrated trust in execution (CT_e), we calculated the frequency at which users trusted the correct execution results and distrusted the wrong execution results. The two measures can be calculated as:

$$\begin{aligned} CT_p = &\mathbb{I}\big(\text{trust} = \text{'Yes'}, \text{plan quality} = 5\big) \\ &+ \mathbb{I}\big(\text{trust} = \text{'No'}, \text{plan quality} < 5\big) \end{aligned} \tag{8.1}$$

$$CT_e = \mathbb{I}(\text{trust} = \text{'Yes'}, ACC_e = 1) + \mathbb{I}(\text{trust} = \text{'No'}, ACC_e = 0) \tag{8.2}$$

To assess the task performance, we mainly considered the task outcome from the planning and execution stages.

Plan Quality. As for the planning outcome, we evaluate the plan quality based on a 5-point Likert scale: 1. low-quality plan, task requirements not covered; 2. low-quality plan, task requirements covered but with grammar errors; 3. medium-quality plan, task requirement covered but with at least one action intent mismatch with ground truth action sequence; 4. medium-quality plan, task requirements covered but miss or have wrong details for action parameters; 5. high-quality plan, covering all task requirements and providing all necessary details.

Execution Performance. The execution of the step-wise plan will result in an action sequence. We provide a ground truth action sequence as a reference to evaluate the generated action sequence. We measure the action sequence accuracy (ACC_s) as the strict match of the action sequence and ground truth. Meanwhile, if one action sequence contains some redundant actions that are not harmful (*e.g.,* searching for flights), the execution results should still be correct. Thus, we also consider execution accuracy (ACC_e) as a task performance measure.

Subjective Trust and Covariates. To enrich our analysis of user trust, we followed existing work to adopt the six subscales from the Trust-in-automation questionnaire [321]. The four subscales — *Reliability/Competence*, *Understanding/Predictability*, *Intention of Developers*, *Trust in Automation* are used as subjective measures of user trust in the LLM agent. Meanwhile, the *Familiarity* and *Propensity to Trust* are also used as covariates. Besides them, we considered user expertise in LLMs and user expertise in automatic assistants as covariates.

Exploratory Variables. To enrich our understanding of LLM agent as daily assistant, we assessed user confidence (both planning and execution) and risk perception along with each task. After users finish the study, we also ask for their open-text feedback on the planning and execution stages as well as other comments. To check the cognitive load of user involvement in our study, we adopted the NASA-TLX questionnaire [242], which contains six subscales.

8.4.4 Participants

Sample Size Estimation. To ensure sufficient statistical power, we estimated the required sample size for a 2 × 2 factorial design based on G*Power [92]. To correct for testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.8$, and that we will investigate 4 different experimental conditions/groups. This resulted in a required sample size of 244 participants. We thereby recruited 347 participants from the crowdsourcing platform Prolific[7], to accommodate potential exclusion.

Compensation. All participants were rewarded with an hourly wage of £8.1 deemed

[7] https://www.prolific.co

to be "*Fair*" payment by the platform (estimated completion time was 30 minutes). As participants in condition UP-UE spent longer in the study, we paid each participant a commensurate bonus accounting for an extra 10 minutes. We rewarded participants with extra bonuses of £0.05 for every high-quality plan and correct execution result. According to existing literature [40], such a bonus setup can help incentivize participants to reach a correct decision. In comparison with existing literature exploring human-AI decision making [22], our reward setup is above the average payment and can be considered as being sufficient to elicit ecologically valid behavior among participants (*i.e.,* aiming to arrive at accurate execution results). Moreover, similar bonus structures akin to our setup have been effective in incentivizing reliable participant behavior and improving data quality across different studies with crowdsourced participants [220, 460–462].

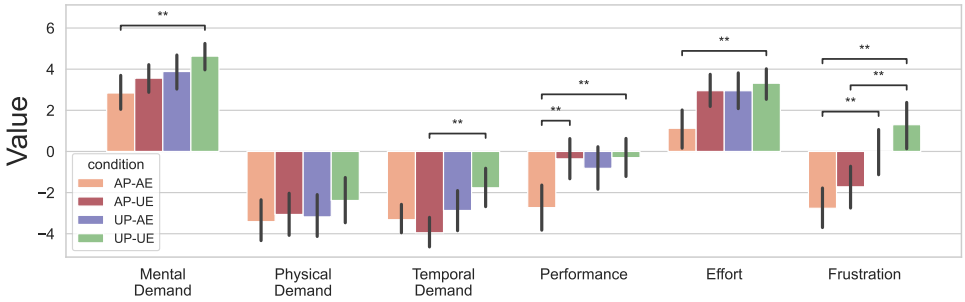


Figure 8.4: Bar plot for cognitive load across all conditions. ** indicates significance ($p < 0.0125$) through post-hoc Tukey HSD test. The error bars represent the 95% confidence interval.

Filter Criteria. All participants were proficient English speakers between the ages of 18 - 50. We also constrained their prior experience (at least 40 successful submissions) and had an approval rate of above 90% on the Prolific platform. We excluded participants from our analysis if they failed any attention check, or represented an outlier regarding the plan quality. Outliers were 4 participants who generated more than three low-quality plans among six tasks. The reserved 248 participants had an average age of 32.5 (SD= 8.1) and a balanced gender distribution (50%, 49.6% female, 0.4% other).

8.4.5 Procedure

Participants were first presented with a study description and an informed consent for data collection. Only those who signed the informed consent were allowed to continue onto our study. Next, participants were asked to complete a pre-task questionnaire to measure their expertise on LLMs and automatic assistants.

Participants were then assigned to one of the experimental conditions, which differed in the level of user involvement in the planning stage and execution stage. With an onboarding tutorial, we showcased the necessary interactions that participants were expected to perform in the planning and execution stages. We used an example task to help participants understand how to work with the plan-then-execute LLM agent. After the onboarding tutorial, participants worked on the selected tasks, which were shuffled at random for every participant to prevent task ordering effects. After the participants

finished the task batch, we measured their perceived cognitive load using the NASA-TLX questionnaire [242], their overall trust in the daily assistant using the trust in automation questionnaire [321], and we gathered their feedback on our system (related to planning, execution, and other aspects) using open-ended text.

8.5 Results

In this section, we will present the main experimental results and exploratory analysis for our study.

8.5.1 Descriptive Statistics

In total, our analysis is based on 248 participants, who are balanced across conditions: AP-AE (63), AP-UE (64), UP-AE (61), and UP-UE (60). All edited plans in user-involved planning conditions are evaluated by the authors following the plan quality criteria described in Section 8.4.3.

Distribution of Covariates. In our study, most participants claimed to have some experience with using large language models ($M = 3.6, SD = 1.0$) and automatic assistants ($M = 3.4, SD = 1.1$). In the trust in automation questionnaire, participants indicated a medium level of *Familiarity* ($M = 2.9, SD = 1.2$) and *Propensity to Trust* ($M = 3.0, SD = 0.7$).

Performance Overview. Overall, users show calibrated trust in the planning ($M = 0.50, SD = 0.13$) and calibrated trust in the execution ($M = 0.64, SD = 0.19$). For the execution outcome, we find that although it is tricky to obtain a ground truth action sequence ($M = 0.48, SD = 0.17$), the action sequence has a relatively high recall of ground truth actions ($M = 0.77, SD = 0.11$). The successful rate for correct execution ($M = 0.52, SD = 0.18$) is higher than the strict evaluation of the action sequence. We also collected user subjective trust with four subscales of the trust in automation questionnaire: *Reliability/Competence* ($M = 3.49, SD = 0.77$), *Understanding/Predictability* ($M = 3.30, SD = 0.56$), *Intention of Developers* ($M = 3.61, SD = 0.81$), *Trust in Automation* ($M = 3.52, SD = 1.01$). With a two-way ANOVA analysis considering user involvement in planning and execution, we do not find any significant impact of user involvement on subjective user trust in AI systems across conditions.

Cognitive Load. The cognitive load of participants across the four experimental conditions is shown in Figure 8.4. Based on two-way ANOVA, we analyzed the impact of user involvement in planning and execution affect user cognitive load. User involvement in planning shows a significant impact on *Mental Demand*, *Temporal Demand*, and *Frustration*. User involvement in execution shows a significant impact on *Performance* and *Effort*. With post-hoc Tukey HSD test, we confirmed such impact — involvement in both planning and execution posed a higher cognitive load on participants.

User Involvement. Among 121 participants in conditions with user-involved planning, 104 participants edited at least one task plan. Meanwhile, 90 participants used the provided buttons (*i.e.,* add/delete/split step) in our study. In total, *delete step* is used 394 times, *add step* is used 183 times, *split step* is used 126 times. Among 124 participants in conditions with user-involved execution, 114 participants interacted with the conversation interface to change action prediction (*i.e.,* have at least one task where they choose to give feedback

Table 8.3: Task-specific evaluation results for user-involvement in planning on calibrated trust in planning (CT_p) and plan quality. We also report the mean value for each measure on each condition.

Tasks	CT_p					Plan Quality				
	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results
Avg	0.51	0.50	0.50	0.50	-	3.8	3.8	3.6	3.7	AP > UP
task-1	0.11	0.20	0.13	0.27	-	2.0	2.0	2.3	2.4	AP < UP
task-2	0.21	0.11	0.20	0.17	-	3.0	3.0	2.9	2.9	-
task-3	0.10	0.03	0.10	0.07	-	3.0	3.0	2.7	2.9	AP > UP
task-4	0.94	0.97	0.80	0.90	AP > UP	5.0	5.0	4.3	4.8	AP > UP
task-5	0.87	0.84	0.90	0.82	-	5.0	5.0	4.6	4.8	AP > UP
task-6	0.81	0.81	0.85	0.75	-	5.0	5.0	4.7	4.6	AP > UP

Table 8.4: Task-specific evaluation results for user-involvement in planning on task performance. ACC_s denotes the strict accuracy of an action sequence, and ACC_e denotes the correctness of execution results. Bold fonts are used to highlight the best performance across conditions.

Tasks	ACC_s					ACC_e				
	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results
Avg	**0.53**	0.46	0.46	0.48	-	0.54	0.53	0.47	**0.56**	-
task-1	0.00	0.00	0.10	**0.12**	AP < UP	0.00	0.00	0.10	**0.13**	AP < UP
task-2	**0.78**	0.64	0.61	0.57	-	**0.78**	0.72	0.66	0.75	-
task-3	**0.44**	0.12	0.36	0.28	-	0.44	0.42	0.36	**0.52**	-
task-4	**0.95**	0.89	0.75	0.82	AP > UP	**0.95**	0.89	0.75	0.82	AP > UP
task-5	**0.98**	0.91	0.90	0.90	-	**0.98**	0.91	0.92	0.90	-
task-6	0.05	**0.22**	0.02	0.18	-	0.06	**0.23**	0.03	0.22	-

or override predicted action). Meanwhile, 105 participants specified at least one action in the task batch. In total, *Specify Action* is used 445 times, feedback to the LLM agent is used 91 times before action execution, and feedback to the LLM agent is used 163 times after execution.

8.5.2 Hypothesis Verification

As the tasks selected in our study are of different initial plan quality and risk levels, we conducted a task-specific analysis in each hypothesis verification.

The Impact of User Involvement in Planning on Calibrated Trust

To verify **H1**, we adopted the one-way ANOVA test and post-hoc Tukey HSD test on the calibrated user trust in planning (*i.e.,* CT_p). The results are shown in Table 8.3. Only in task-4, we found user involvement in planning will have a negative impact on calibrated trust in planning. To avoid a potential impact of user involvement in the execution stage, we conducted a two-way ANOVA test to confirm the findings. We only find a significant difference in task-4. Post-hoc Tukey HSD results show that participants in conditions with automatic planning (AP) showed significantly higher calibrated trust in planning outcomes than those in conditions with user-involved planning (UP). Thus, our experimental results do not support **H1**.

We noticed that the calibrated trust in planning is quite low in the high-risk tasks where all initial plans are imperfect. This indicates that many users across all conditions consider the generated plan trustworthy. On tasks with low risk, where the initial plan is of high quality, users achieved much higher calibrated trust in the planning outcome. We also find that conditions with user-involved execution (UE) show slightly higher CT_p in

task-1 and task-4 than conditions with automatic execution (AE). With the same statistical test as **H1** analysis, such differences are not significant.

The Impact of User Involvement in Planning on Task Performance

To verify **H2**, we considered plan quality, the accuracy of action sequences (ACC_s), and the execution accuracy of the plan (ACC_e) for analysis. For plan quality (cf. Table 8.3), we conducted one-way ANOVA on plan quality considering the user involvement in the planning stage. We found that overall user involvement in the planning stage caused a decrease plan quality, especially on tasks with a perfect plan (*i.e.,* task 4, 5, 6, where plan quality = 5) and task-3. However, in task-1, where the original plan contains a grammar error, we find that user involvement in planning can improve the plan quality. As the action sequence accuracy (ACC_s) and execution accuracy (ACC_e) are not normally distributed, we conducted the Kruskal-Wallis H-test by considering the user involvement in the planning as the independent variable. The results are shown in Table 8.4. With further post-hoc Mann-Whitney tests, we found that while participants achieved a relatively higher accuracy of action sequences in condition AP-AE, the condition UP-UE achieved the best execution accuracy. In most tasks, condition UP-UE achieved better or compatible performance as other conditions. The only exception is task-4, where user involvement in the planning caused a significantly worse performance (both ACC_s and ACC_e). As user involvement does not consistently lead to improved performance, these results are not enough to support **H2**.

We found that in task-1 and task-6 most participants in the AP-AE condition achieved a very low success rate. This is mainly due to the imperfect plans and imperfect execution generated by LLMs. In task-1, the plan generated by LLMs includes one step which contains two actions to execute. Due to the inability to edit the plan, the LLM agent execution missed one transaction in conditions with automatic planning. In task-6, the plan generated by LLMs is correct. However, in the automatic execution of step 2 of the plan (*i.e.,* selecting an itinerary suggested), the LLM agent has a high probability of choosing an itinerary that does not match the task description. If the participants do not carefully check the task description, and correct this agent behavior, the execution results would be wrong. This also helps explain why user involvement substantially improves the task outcome accuracy in task-6. More details about tasks can be found in the appendix.

Table 8.5: Task-specific evaluation results for user-involvement in execution on task performance. Bold fonts are used to highlight the best performance across conditions.

Tasks	ACC_s					ACC_e				
	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results
Avg	**0.53**	0.46	0.50	0.51	-	0.54	0.53	0.50	**0.58**	-
task-1	0.00	0.00	0.10	**0.12**	-	0.00	0.00	0.10	**0.14**	-
task-2	**0.78**	0.64	0.67	0.62	-	**0.78**	0.72	0.69	**0.78**	-
task-3	**0.44**	0.12	0.42	0.29	AE > UE	0.44	0.42	0.42	**0.53**	-
task-4	**0.95**	0.89	0.94	0.88	-	**0.95**	0.89	0.94	0.88	-
task-5	0.98	0.91	**1.00**	0.98	-	0.98	0.91	**1.00**	0.98	-
task-6	0.05	**0.22**	0.02	0.19	AE < UE	0.06	**0.23**	0.04	**0.23**	AE < UE

The Impact of User Involvement in Execution on Calibrated Trust in Execution Outcome

As we observe in Table 8.3, user involvement in planning can have some negative impact on the plan quality, which further impacts the execution stage. To control such impact, we filtered out the tasks where plan quality decreased after user-involved planning in the analysis of user involvement in the execution stage. To verify **H3**, we conducted one-way ANOVA on calibrated trust in execution outcome (CT_e). The results are shown in Table 8.6. We found that user involvement in execution causes no significant difference across conditions. Thus, **H3** is not supported by our experimental results.

Table 8.6: Task-specific evaluation results for user-involvement in execution on calibrated trust in execution (CT_e). We also report the mean value for each measure on each condition.

Tasks	CT_e				
	AP-AE	AP-UE	UP-AE	UP-UE	Post-hoc results
Avg	**0.66**	0.65	0.64	0.65	-
task-1	0.48	0.44	**0.51**	0.49	-
task-2	0.78	**0.83**	0.71	0.80	-
task-3	0.51	0.41	**0.60**	0.47	-
task-4	**0.94**	0.92	0.88	0.86	-
task-5	0.89	0.92	**0.96**	0.94	-
task-6	0.37	0.38	0.28	**0.42**	-

The Impact of User Involvement on Overall Task Performance

Similar to the verification of **H3**, we excluded the tasks where plan quality decreased after user-involved planning in this analysis. As the plan is generated before user involvement in the execution, we only considered ACC_s and ACC_e in the analysis of user involvement in the execution stage. To verify **H4**, we conducted Kruskal-Wallis H-test by considering the user involvement in the execution as the independent variable. The results are shown in Table 8.5. With post-hoc Mann-Whitney tests, we found that user involvement in the execution stage showed significantly higher ACC_s and ACC_e in task-6 (where the LLM assistant mainly failed to choose the most suitable itinerary plan). We found that participants in the AP-AE condition achieved the best accuracy of action sequences (*i.e.,* ACC_s), and participants in condition UP-UE achieved the best execution accuracy (*i.e.,* ACC_e). In other words, the executed action sequence in condition AP-AE is more aligned with the ground truth action sequence annotated by the authors. However, with user involvement in the execution stage, participants in condition UP-UE have a better opportunity to obtain correct task outcomes by correcting potentially flawed actions. Such a difference is due to our measure of ACC_e, which tolerates the non-risky actions (*e.g.,* search flight) and failure of action predictions. In contrast, our measure of ACC_s considers this as a wrong action sequence. Thus, in task-3, even if we find automatic execution achieved significantly better ACC_s than user-involved execution, participants in condition AP-UE and UP-UE obtained comparable or higher execution accuracy (*i.e.,* ACC_e) than conditions with automatic execution. While user involvement shows some positive impact on the execution accuracy, such impact is not significant and consistent across all tasks. Only in task-6, where users can correct the errors made by the LLM agent (*i.e.,* the wrong itinerary

selection mentioned in Section 8.5.2), user involvement in the execution shows a significant contribution to the task performance. Thus, these results are not enough to strictly support **H4**.

8.5.3 Exploratory Analysis

The Impact of Covariates

For further insights into all user factors on user trust and team performance, we calculated Spearman rank-order correlation coefficients for user trust, calibrated trust, risk perception, and task performance. As can be seen in Table 8.7, we found these covariates mainly show correlations with subjective user trust, calibrated trust in execution, and risk perception. First, all covariates (*i.e.,* user factors) positively correlated with user trust (four subscales in the trust in automation questionnaire [321]) and negatively correlated with perceived risk (average over six tasks). It indicates that users with more expertise or familiarity with such systems tend to trust the daily assistant and show less perceived risk when using it. Meanwhile, users with a general propensity to trust also tend to trust the AI system. Besides user trust, *Assistant Expertise* and *Propensity to Trust* show a significant negative correlation with calibrated trust in the execution outcome. Apart from the above correlation, these user factors do not significantly correlate with task performance measures or calibrated trust in the planning outcome.

Table 8.7: Spearman rank-order correlation coefficient for covariates level on dependent variables. All measures are calculated based on average over task batch. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.0125, respectively.

Covariates		llm expertise		assistant expertise		Familiarity		Propensity to Trust	
Category	**Variables**	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
User Trust	Reliability/Competence	0.334	**.000**††	0.245	**.000**††	0.321	**.000**††	0.679	**.000**††
	Understanding/Predictability	0.307	**.000**††	0.164	**.010**††	0.208	**.001**††	0.380	**.000**††
	Intention of Developers	0.406	**.000**††	0.324	**.000**††	0.362	**.000**††	0.517	**.000**††
	Trust in Automation	0.380	**.000**††	0.278	**.000**††	0.356	**.000**††	0.698	**.000**††
Calibrated Trust	CT_p	0.053	.404	0.053	.402	0.056	.378	0.037	.566
	CT_e	-0.120	.059	-0.195	**.002**††	-0.032	.621	-0.174	**.006**††
Risk Perception	Perceived Risk	-0.187	**.003**††	-0.180	**.004**††	-0.237	**.000**††	-0.363	**.000**††
Task Performance	ACC_s	0.037	.560	-0.014	.823	0.110	.085	0.018	.772
	ACC_e	-0.000	.995	-0.037	.567	0.085	.184	0.007	.911
	Plan Quality	-0.035	.587	-0.037	.560	0.080	.211	-0.032	.611

Impact of Plan Quality and Risk Percetion.

Besides the measures calculated over task batch, a task-level analysis of plan quality and risk perception can deepen our understanding of their impacts. Besides measures adopted in Table 8.7, we include task-level confidence in this analysis and exclude the subscales from the trust in automation questionnaire. Thus, we calculated Spearman rank-order correlation coefficients for task-level measures across all groups of participants (shown in Table 8.8). As we can see, both plan quality and risk perception significantly correlate with user trust, calibrated trust, task performance, and user confidence. The *plan quality* shows a significant positive correlation with most measures, which indicates users perform better and calibrate their trust in the LLM agents in tasks with a high-quality plan. By contrast, the *risk perceptions* shows a negative correlation with most measures and also a negative correlation with the plan quality.

Table 8.8: Task-specific spearman rank-order correlation coefficient for plan quality and risk perception. "†" and "††" indicate the effect of the variable is significant at the level of 0.05 and 0.0125, respectively.

Category	Variables	Plan Quality		Risk Perception	
		r	*p*	*r*	*p*
User Trust	Trust-p	0.056	.032†	-0.293	**.000**††
	Trust-e	0.258	**.000**††	-0.160	**.000**††
Calibrated Trust	CT_p	0.723	**.000**††	-0.102	**.000**††
	CT_e	0.221	**.000**††	0.000	.995
Task Performance	Plan Quality	-	-	-0.141	**.000**††
	ACC_e	0.400	**.000**††	-0.110	**.000**††
	ACC_s	0.446	**.000**††	-0.096	**.000**††
Confidence	Confidence-p	0.137	**.000**††	-0.532	**.000**††
	Confidence-e	0.225	**.000**††	-0.271	**.000**††

Failure Analysis

As we find that plan quality substantially affects task execution accuracy, we look into task performance across different plan qualities. For the tasks with low-quality plans (plans fail to cover task information or plan with grammar errors, *i.e.,* plan quality=1, 2), the execution accuracy is 1.8%. While for tasks with a plan that may mislead action prediction (plan quality = 3, 4), our LLM agent-based daily assistant achieved 59% execution accuracy. The average execution accuracy for tasks with a high-quality plan (plan quality =5) is 66.7%.

We further check 717 tasks where a high-quality plan (plan quality = 5) is provided. Among them, 235 tasks provide wrong execution results. The main causes are: (1) Wrong action parameter prediction (48.9%). While action names match, one or more parameters mismatch the expected value at some step of the action sequence. (2) Invalid actions (48.5%). Given a perfect plan, the LLM agent failed to predict one valid action (failed to predict one action name or failed to predict some action parameter value) to execute in some steps. (3) Wrong action name prediction (2.6%). The generated action sequence has at least one action name prediction that mismatches the ground truth.

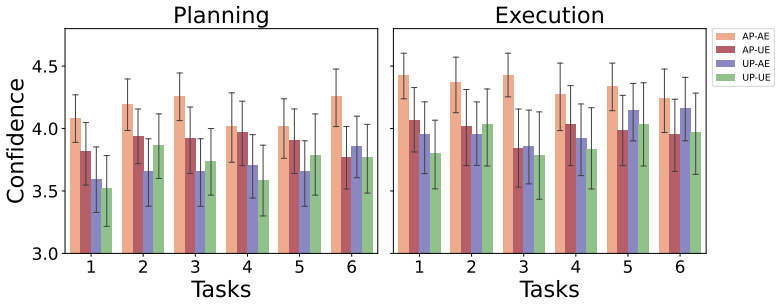


Figure 8.5: Bar plot for confidence dynamics, the x-axis denotes the task ordering index (shuffled for every participant). The error bars represent the 95% confidence interval.

Table 8.9: Excerpts from participants' responses to open questions soliciting their opinions.

Opinions towards Planning	**Sentiment**	**Reason**
I really like how organized it is. The step by step and numerical planning allows it to make sense in a clear and structured way, meaning there is less room for errors or misinformation	Positive	Helpful with reducing error
It was remarkable how quickly. It was able to achieve the goals which was set out in the tasks. I quite liked it I would definitely want something like this in my life as It would my my life much easier	Positive	Effective and make life easy
As I said previously, it's far, far too detailed in an unnecessary way. I'm not sure people need the entire plan of what the AI will do, as long as the job gets done.	Negative	Too detailed
I found it really helpful, but made me slightly nervous thinking all my plans being successful are in the hands of ai tech	Mixed	Helpful assistant, agency concerns
Opinions towards Execution	**Sentiment**	**Reason**
The execution stage was amazing. I feel like this could be the future and we wont need to call or talk to people to get this kind of thing done ever again.	Positive	Promising future
The execution stage went smoothly, except for a few rare instances of an error response before also saying the AI's automatic reply (which was correct).	Mixed	Smooth user experience, error response
I found it clunky and nit that user friendly	Negative	Clunky, not user-friendly
This bit is user friendly, but very robotic, which makes it difficult to trust	Mixed	User-friendly, distrust due to robotic nature

Confidence Dynamics

To visualize the user confidence in the planning and execution stage, we draw point plots (see Figure 8.5) for user confidence in the task order. Overall, condition AP-AE shows the highest confidence in both the planning and execution stages. To verify the impact of user involvement in confidence, we adopted two-way ANOVA and post-hoc Tukey HSD test. We find that: (1) with user involvement in the planning, participants showed significantly lower confidence in planning (AP-AE > UP-AE, UP-UE); (2) with user involvement in the execution, participants showed a significantly lower confidence in execution (AP-AE > AP-UE, UP-UE). Meanwhile, users typically showed a higher confidence in the execution stage. Compared with conditions with automation execution (*i.e.,* condition AP-AE and UP-AE), the confidence gap narrows down in the conditions with user-involved execution (*i.e.,* condition AP-UE and UP-UE).

8.5.4 Analysis of Open Feedback

At the end of our study, we collected open feedback regarding the planning stage, execution stage, and any other feedback using the following question: '*Please share any comments, remarks or suggestions regarding the planning/execution stage of LLM Assistant*' and '*Do you have any other comments, remarks or suggestions regarding the study?*'. Overall, we analyzed all the feedback based on user opinions (positive, negative, mixed, neutral) and their suggestions. In our analysis, we ignored all responses that did not directly lead to useful input such as 'None', 'N/A', and 'No comment'.

Feedback and Suggestions. While most comments tended to demonstrate positive opinions towards LLM agents as daily assistants (more than 80%), there were also negative

opinions regarding the difficulty, expertise, and trust. We provide example excerpts from participants in Table 8.9. Besides opinions towards the system, some participants also appreciated our user-centric setup:

"*The study does a good job of emphasizing user experience by asking about perceptions of risk, trust, and confidence. This approach ensures that the evaluation is user-centric, which is important for assessing the real-world applicability of the LLM Assistant.*"

Some participants also provided suggestions on how to further improve the general design of LLM agent-based daily assistants. Regarding the plan edit, participants hoped for the provision of more convenient edit operations like 'drop/drag' to adjust the plan-related text ordering and an 'undo' operation to tolerate unexpected mistakes. Some participants also found the plans too detailed, which could increase their perceived cognitive load (cf. Table 8.9 except 3). As for the execution, many participants found it to be smooth. At the same time, some believed that additional verification in each step could further enhance the reliability of daily assistants:

"*For the execution stage, I commend it for creating an input formatting box to execute the user's request validating each requirement.*"

Other comments from participants reflected on the entire plan-then-execute workflow:

"*The planning was really challenging, and I mostly left the default plans (they looked fine). This worked in the main, but a couple clearly needed revisiting. I would approach this iteratively: plan, test, observe, back to planning, then another test, before reaching the desired outcome.*"

Our findings suggest open research opportunities to explore more effective ways to provide an overview of plans that trade-off user cognitive load resulting from granular descriptions, with the need to provide details to help users identify flaws. For example, we can consider developing methods to interactively allow users to flesh out further details in a plan.

8.6 Discussion

8.6.1 Key Findings

Our experimental results show that user involvement in the plan-then-execute workflow with LLM agents can help fix imperfect plans in planning and wrong action predictions in the step-by-step execution. However, user involvement does not ensure a consistently positive impact on calibrated trust and overall task performance across different tasks.

User Involvement Fails to Calibrate User Trust. Overall, user involvement in the planning and execution does not significantly impact user trust and calibrated trust in planning and execution outcomes. As Table 8.3 shows, user involvement in planning can harm plan quality in tasks with a high-quality initial plan, which may potentially cause worse task performance in the subsequent execution stage. Our experimental results do not support **H1** or **H3**, which indicates user involvement does not necessarily help calibrate user trust in our study. Instead, with a task-specific correlation analysis (cf. Table 8.8), we found that the plan quality has a significant positive correlation with calibrated trust in both planning and execution outcomes. Combined with task-specific user trust and task-specific confidence, we can infer that users tend to trust the LLM agent overall. Such trust can be expected and calibrated in tasks with a high-quality plan. In contrast, users fail to calibrate

their trust in the tasks where a low-quality plan is provided. A potential cause of such miscalibrated trust is the plausibility of plans generated by LLMs (i.e., plans that appear to be likely correct). In our study, all initial plans are formulated with a clear, logical structure, which covers most of the task requirements. At first glance, such high-quality text pieces seem quite plausible and trustworthy. We also received some open text feedback such as, — "The plans look nice, I do not find any space for improvement" and "the planning stage of the LLM assistant was helpful and trustworthy." Findings from recent work on LLM-assisted fact checking corroborate this, wherein authors found that convincing explanations provided by LLMs can cause over-reliance when LLMs are wrong [451].

User Involvement can Benefit Task Performance. User involvement in planning and execution can positively impact overall task performance, especially execution accuracy. As the results in Table 8.3 and Table 8.4 show, user involvement in planning can help address imperfect plans (*e.g.,* task-1 with grammar error). Doing so further contributes to improvements in the execution accuracy. After controlling the plan quality, we found that user involvement in the execution can provide the best execution accuracy among most tasks considered in our study (cf. Table 8.5). Based on the failure analysis (Section 8.5.3), LLM agents can make mistakes in executing high-quality plans, which can be attributed to prediction errors (*i.e.,* wrong action name or action parameters) and prediction failures (*i.e.,* failure to provide valid action prediction). In practice with deployed LLM services, there is no reliability guarantee for the generated plan in planning or predicted actions in execution. User involvement can play an important role in the plan quality control and risky action control, ensuring that only correct and safe actions are executed to obtain desirable task outcomes.

Other Findings. We also found some user factors and perceptions that affect user trust and task performance. As seen in Table 8.7, nearly all covariates show a significant positive correlation with user trust in the AI system. Some of these covariates also impact user trust in the planning and execution outcomes. Overall, these findings indicate that users who are familiar with such systems tend to show higher user trust. However, some factors also correlate negatively with the calibrated trust in the execution outcomes and risk perception of using the LLM agents as daily assistants. This reflects that these factors can cause miscalibrated trust and reduced risk perception when working with the LLM agent. While we found that risk perception negatively correlated with user trust, calibrated trust, task performance, and confidence (cf. Table 8.8), it does not mean risk perception is harmful in the human-LLM agent collaboration. The main cause is that users may only notice the risks of using LLM agents when the task is provided with a relatively low-quality plan. Risk perception is important to calibrate user trust in the planning and execution outcomes. Collaborative workflows should support users with the provision to take over control of planning and/or execution stages based on their perceived risk.

8.6.2 Implications

The Impact of Convincingly Wrong LLM Outcomes. As our study follows a plan-then-execute workflow for users to collaborate with LLM agents, users were not offered a chance to revise the plan after starting with execution. Users following a wrong plan can lead to negative outcomes. Combined with existing work on algorithm aversion [33] and

the impact of negative first impressions on user trust [87], we can infer that such convincingly wrong content [451] can bias user trust and reliance towards the extremes. Before users take notice, they may develop an uncalibrated trust in the AI system, as observed through our findings in high-risk tasks (*i.e.,* tasks 1,2,3) and corroborating work by Si *et al.* [451]. As a result, users over-rely on AI assistance, which is misuse akin to behavior that resonates with algorithm appreciation [34]. Once users notice such phenomena, their trust in the LLM-based systems may sharply decrease, resulting in disuse due to algorithm aversion. This can be a result of the misalignment between perceived AI performance and actual AI performance. Existing human-AI collaboration literature has provided potential solutions for such problems, ranging from performance feedback interventions [30] to agreement-in-confidence heuristic [49, 463]. Future work can combine these insights to explore effective interventions for user trust calibration with convincingly wrong LLM outcomes.

Insights for Effective Collaboration with Plan-then-execute LLM Agents. Our work has important theoretical implications for effective human-AI collaboration with plan-then-execute LLM agents. On the one hand, user involvement can be necessary to achieve complementary team performance. Although LLM agents have shown promising planning and execution capabilities, they are never perfect due to probabilistic uncertainty. With user involvement in the planning, users can fix imperfect plans with grammar errors (cf. Table 8.3 task-1). With user involvement in the execution, users can fix uncertainty issues (*e.g.,* LLM agent predicts invalid actions) and prevent risky actions (*e.g.,* LLM agents choose an itinerary conflicting with task requirements, cf. Table 8.5, task-6). On the other hand, user involvement may also bring uncertainty and even harm LLM agent performance. In tasks where the LLM agent provides a high-quality plan (cf. task 4, 5, 6 in Table 8.3 and Table 8.4), user involvement can harm the plan quality, which further negatively impacts the execution accuracy. Moreover, user involvement in planning and execution poses a significantly higher cognitive load on users (cf. Figure 8.4) and negatively impacts user confidence (cf. Figure 8.5). Thus, too much human involvement in collaboration with plan-then-execute LLM agents can be undesirable. User involvement in the execution process brings more consistent benefits than user involvement in the planning stage. As suggested by the participants, iterative LLM agent simulation may be one potential way to decide when users should be involved. The LLM agent may first conduct a plan-then-execute round to obtain a clear plan and execution results. With humans checking the whole process and simulated outcomes, humans can decide whether to be involved in revising the plan or the execution process. In this way, we can minimize user involvement while keeping highly effective task outcomes through LLM agents.

Human Oversight and Designing Flexible Collaborative Workflows. In our study, we found that human oversight does not consistently lead to improved outcomes. One potential cause can be the disparity between the planning and execution of LLM agents. Specifically, it is unclear how one plan step will be transformed into one action. When users realize one plan step can be wrong during the execution stage, they may need to articulate it or manually override the agent action, posing a high cognitive load. Even worse, when users realize the LLM agent missed one action due to limited steps designed in the plan (in task-1), they do not have a chance to change the plan or add one extra step.

To address such concerns, we may need a more flexible collaborative workflow where humans can fix planning and execution simultaneously. In this way, users can exercise more flexible control over the workflow and the task outcomes. For instance, the action prediction from the LLM agent can be provided along with each step in the planning stage. Users can thereby be informed of the potential impact of their edited plan, which provides more straightforward feedback and helps users adjust the plan according to the expected actions.

8.6.3 Caveats and Limitations

Limitations and Potentail Biases. To ensure reliable task outcomes, humans are expected to fix imperfect plans (*e.g.,* grammar errors, misleading action intents) in the planning stage. However, not everyone in conditions UP-AE and UP-UE noticed such grammar errors and split the plan in task-1. Similarly, not everyone in conditions AP-UE and UP-UE noticed that the LLM agent chose an itinerary that conflicted with task requirements. As discussed earlier, LLM agents can generate plausible plans, which may mislead user trust in the planning and execution outcomes. In that case, participants in our study may have easily ignored some convincingly wrong plan steps or execution actions. In our study, one primary step in the plan is only transformed into a single action. In practice, LLM agents can generate multiple actions for one specific goal. However, such action generation and execution modes are challenging for humans to get involved in and control, as the execution of the action sequences is automated by the LLM agent within one goal. Furthermore, using multiple actions to achieve one primary step (*i.e.,* goal) also results in higher task complexity and reduced task clarity, which may impact the task outcomes [419].

Transferability Concerns. Although we selected representative tasks for daily scenarios, our study may not sufficiently cover all potential cases of daily assistance with LLM agents. Some task characteristics (*e.g.,* task complexity, time consumption) may also impact how users are willing to rely on AI assistance. Moreover, complete control over the plan-then-execute LLM agents may not be desirable for some simple tasks (*e.g.,* setting alarms). Once the efforts to control/interact with LLM agents are greater than the efforts to execute the tasks themselves, users will be unwilling to adopt such "assistance." Future work can look into what daily user needs are suitable for LLM agents to support. In our study, the execution of plans is conducted in a simulation environment. While it has been proven to be effective in prior work of agent-based modeling and HCI studies [457], more work is needed to understand how execution of tasks in real-world environments with additional dependencies and complexities can influence our findings.

Participants in our study only followed a relatively fixed mode in collaboration with LLM agents, and they could determine when to get involved in the planning and execution stages. The experimental conditions considered in our study range from full automation (*i.e.,* AP-AE) to complete user control (*i.e.,* UP-UE). Such a setup provides good flexibility, and simulates the spectrum of real-world practice. Our findings and implications provide valuable insights to guide future research on human-AI collaboration with LLM agents.

8.7 Conclusion

In this chapter, we empirically studied human-AI collaboration using plan-then-execute LLM agents. Adopting such LLM agents in various everyday scenarios, we analyzed the impact of user involvement in the planning and execution stages on user trust and overall task performance. We provide various interactions in each stage to help users fix imperfect plans and modify execution outcomes. Our results suggest that the LLM agents can provide plausible plans (in text) to cover task requirements, which can be convincingly wrong. As a result, users develop uncalibrated trust in the planning and execution outcomes, and user involvement in the planning and execution stages fails to calibrate user trust (**RQ1**). We also found that the plan quality substantially affects the subsequent execution accuracy. Thus, when user involvement in planning can fix imperfect plans, the overall task performance (*i.e.,* plan quality, accuracy of action sequence, and execution accuracy) is improved. However, user involvement in planning can also harm task plan quality where the original plan is good to begin with. As a result, the LLM agents demonstrate worse task performance in these tasks. In contrast, user involvement in execution brings about a more stable positive impact on task performance (**RQ2**). Our results suggest that plausible but wrong LLM outcomes can be detrimental to user trust calibration and overall task performance. We discussed the impact of convincingly wrong LLM outcomes and provided potential solutions and insights for future work. Furthermore, we synthesized key insights for better control and effective collaboration with plan-then-execute LLM agents. We also shed light on opportunities to design flexible collaborative workflows with human oversight for effective collaboration with LLM agents.

Our results indicate that user involvement in the LLM agent workflow can be important in ensuring reliable task outcomes. Future work can further investigate how to detect and handle plausible but imperfect LLM outcomes and design effective interventions to fix such problems. We hope that our key findings and implications reported in this chapter will inspire further research on human-AI collaboration with LLM agents.

9

Conclusions

Our research focuses on not only performance-related outcomes (*e.g.,* team performance, appropriate reliance, and calibrated trust), but also experience-related outcomes (*e.g.,* user satisfaction, work load, and user agency). To approach the goal of appropriate reliance, we explored calibrating user perception of competence (Part I), facilitating user understanding with human-centered XAI methods (Part II), and enhancing user control with collaborative workflows (Part III). The findings and implications obtained from empirical studies help advance the understanding of fundamental aspects of human-AI collaboration. This dissertation can also inspire future research by offering insights into intervention methodologies, experimental design, and theoretical foundations.

In this concluding chapter, we revisit the aforementioned research questions, summarize key findings, and discuss the implications and limitations of the research work described in this dissertation. Furthermore, based on recent advances in relevant technologies and methodologies, we point out promising future directions.

9.1 Summary of Findings

Part I: Calibrating User Perception of Competence

In human-AI collaboration, users may easily develop uncalibrated trust in AI systems, potentially originating from a miscalibrated estimation of AI competence and self-competence. Thus, we mainly focus on the following research questions in Part I:

> **RQ1-a**: How does the perceived performance of humans and AI systems shape user reliance on AI systems?
> **RQ1-b**: How to mitigate the impact of cognitive bias associated with misperception on user reliance on AI systems?

To answer these research questions, we conducted a series of quantitative empirical studies. We started by analyzing the impact of human understanding of system accuracy on their reliance behaviors (Chapter 2). Based on two empirical studies, we found that

explaining the AI system's stated accuracy with analogies was insufficient to facilitate appropriate reliance on the AI system. However, based on these results, we reasoned that under-reliance on the AI system may be a result of users' overestimation of their ability to solve the given task. With a further empirical analysis of Dunning-Kruger effect (DKE) [60] (Chapter 3), we found that participants who overestimate their performance tend to exhibit under-reliance on AI systems, which hinders optimal team performance. To mitigate such an effect, we proposed a tutorial intervention that gives performance feedback to users about the performance of both the AI system and themselves. While we found the effectiveness of our tutorial intervention in mitigating DKE, we also found that the tutorial intervention can mislead some other participants (*i.e.,* participants who underestimated themselves) to overestimate themselves.

To further understand the impact of user perception of AI competence and the effectiveness of proposed debugging intervention, we conducted an empirical user study (Chapter 4) in deceptive review detection tasks. Based on the experimental results, we found that our proposed debugging intervention does not work as expected to facilitate appropriate reliance. Instead, we observe a decrease in reliance on the AI system after the intervention — potentially resulting from an early exposure to the AI system's weakness. Through an exploratory analysis based on different performance quartiles, we found that participants who performed worse in our study tended to underestimate AI performance. Thus, they achieved suboptimal team performance, which is largely impacted by the under-reliance on the AI system.

Part II: Facilitating User Understanding with Human-centered XAI

For effective human-AI collaboration, users need to develop a good understanding of the AI system. With a recent trend of human-centered explainable AI [27, 124, 126], researchers advocate putting user information needs as the focus of explainable AI. We recognized one empirical gap in tasks that require domain expertise (*e.g.,* medical diagnosis) – the explanations for AI advice may be difficult for laypeople to understand. Meanwhile, digesting such explanations also poses too much cognitive load on users. To fill such research gaps, we proposed to adopt analogy-based concept-level explanations (Chapter 5). Specifically, we focused on the following research questions:

RQ2-a: How do analogies for concept-level explanations shape the understanding of an AI system among non-expert users?
RQ2-b: How do analogy-based explanations affect user reliance on AI systems?

To answer these research questions, we conducted an empirical study (N = 280) on a skin cancer detection task with non-expert humans and an imperfect AI system. The results of our study confirmed that a knowledge gap can prevent participants from understanding concept-level explanations. Although we did not find quantitative support for our hypotheses around the benefits of using analogies, we found considerable qualitative evidence suggesting the potential of high-quality analogies in aiding non-expert users in their decision making with AI-assistance.

Apart from the explainable AI methods or explanations with AI advice, the interface to present explainable AI methods may also substantially affect user exploration of informa-

tion needs. Inspired by prior work on conversational user interfaces [263–265], we argued that augmenting existing XAI methods with conversational user interfaces can increase user engagement and boost user understanding of the AI system. To this end, we explored the following research questions:

> **RQ3-a**: How does a conversational XAI interface shape user understanding of an AI system, in comparison with an XAI Dashboard?
> **RQ3-b**: How does a conversational XAI interface influence user reliance on an AI system, in comparison with an XAI Dashboard?

To answer these research questions, we conducted an empirical study (N = 306) by comparing several variants of conversational XAI interfaces with the XAI dashboard in a loan approval task (Chapter 6). Compared to an XAI dashboard, we observed limited improvements in user understanding and trust brought forth by the conversational XAI interface. Overall, we found that users with conversational XAI interfaces tended to rely more on the AI system. However, such increased reliance did not always translate into appropriate reliance. Instead, it was characterized by clear patterns of over-reliance. Furthermore, with an LLM agent-based conversational XAI interface, we observed that over-reliance was further reinforced, and users obtained a worse understanding of AI decision criteria. Our findings suggest that the XAI interfaces were persuasive and have the potential to bring about an illusion of the AI systems' capability, which in turn increased over-reliance on the AI system.

Part III: enhancing user control with collaborative workflows

While most prior work has extensively analyzed human-AI collaboration in a one-step decision making setup, the decision making for complex tasks typically follows a multi-step manner. It is unclear how fine-grained transparency of the AI systems and multi-step decision workflow will impact user reliance. Thus, we analyzed the following research questions:

> **RQ4-a**: How does a multi-step decision workflow shape user reliance on an AI system?
> **RQ4-b**: How do global transparency and local transparency shape user reliance in a multi-step decision workflow?

To answer these research questions, we conducted an empirical study (N = 233) in composite fact-checking tasks (Chapter 7). Our findings demonstrate that human-AI collaboration with a multi-step transparent decision workflow can outperform one-step collaboration in specific contexts (*e.g.,* when advice from an AI system is misleading). Further analysis of the appropriate reliance at fine-grained levels indicates that a multi-step transparent decision workflow can be effective when users demonstrate a relatively high consideration of the intermediate steps. We also found that participants who do not demonstrate fine-grained appropriate reliance at the intermediate steps may exhibit under-reliance be-

havior on the final AI advice. Based on these results, we infer that the multi-step transparent workflow provides better verifiability of the AI advice, which contributes to developing higher self-confidence and a critical mindset on AI advice. While it may help mitigate over-reliance caused by the potential illusion of explanatory depth, it may also decrease user reliance on the AI system and cause under-reliance issues.

Although LLM agents have shown a promising blueprint as daily assistants, there is a limited understanding of how they can provide daily assistance based on planning and sequential decision making capabilities. To ensure the reliability of task outcomes and user agency in collaboration with LLM agents, we explore the impact of user involvement in the planning and execution stages. Thus, we proposed to answer the following questions:

> **RQ5-a**: How does human involvement in the high-level planning and real-time execution shape their trust in an AI system powered by LLM agents?
> **RQ5-b**: How does human involvement in the high-level planning and real-time execution of tasks with an AI system powered by LLM agents affect the overall task performance?

To answer these research questions, we conducted an empirical study (N=248) of LLM agents as daily assistants in six commonly occurring tasks (*e.g.,* flight ticket booking and credit card payments) in Chapter 8. Our experimental results show that user involvement in the plan-then-execute workflow with LLM agents can help fix imperfect plans in planning and wrong action predictions in the step-by-step execution. However, user involvement does not ensure a consistently positive impact on calibrated trust and overall task performance across different tasks. Our findings demonstrate that LLM agents can be a double-edged sword – (1) they can work well when a high-quality plan and necessary user involvement in execution are available, and (2) users can easily mistrust the LLM agents with plans that seem plausible.

9.2 Implications

Our work has important implications for designing effective user interventions and AI assistants to facilitate appropriate reliance and improve human-AI collaboration.

Guidelines for Effective Tutorial Interventions. To calibrate user perception of competence, we adopted several tutorial interventions (*i.e.,* performance feedback tutorial in Chapter 3 and debugging tutorial in Chapter 4). While they have been observed to be effective in calibrating user perception of competence, such calibrated self-assessment/assessment of AI competence does not necessarily translate to optimal appropriate reliance. One possible cause is that while the tutorial makes such users aware that they underestimated themselves and they can make correct decisions when the AI system is wrong in the task, users may have an illusion of superior capability than the AI system. Meanwhile, we also noticed that such tutorial interventions may also bring negative impacts in specific contexts. For example, participants who underestimated their own competence showed worse reliance patterns on AI systems after tutorial intervention. To avoid such side effects, tutorials designed to mitigate a specific kind of bias should be carefully checked before subjecting them to broad participant pools. This also implies that tutorials

designed for promoting appropriate reliance should not only reveal the shortcomings of users or AI systems (*i.e.,* when they are less capable of making the right decision), but also their strengths (*i.e.,* when they are capable or more capable). With such a comprehensive understanding, human decision makers can potentially have a better chance to understand when they should rely on AI systems, and when they should rely on themselves, ultimately leading to (more) appropriate reliance. Our findings in Chapter 4 suggest that the debugging intervention and similar interventions with training purposes (*e.g.,* user tutorials) may suffer from the cognitive bias brought by the ordering effect within such interventions. While using such interventions to demonstrate the strength and weakness of AI systems, we should be careful not to leave users with a bad first impression, highlighting the weakness of the AI system.

Align Plausibility of Explanations with AI Trustworthiness. According to prior work [182, 464], users of explainable AI systems implicitly hold the belief that "plausible explanations typically imply correct decisions, and vice versa". In Chapter 5, the participants who may have found the analogies to be implausible may have perceived certain AI advice as untrustworthy and thereby relied less on the AI system. Such under-reliance could result in sub-optimal team performance. Similarly, when users perceive the explanations to be persuasive and plausible, they may tend to exhibit over-reliance on AI systems. In Chapter 6, participants with both conversational XAI interface and XAI dashboard showed over-reliance on the AI systems. Meanwhile, boosting conversation quality and flexibility (*i.e.,* with LLM-based conversational agent) may further reinforce over-reliance and hurt user understanding and user trust. Based on these findings, we argue it would be more important to align the plausibility of XAI responses with the trustworthiness of the AI system rather than solely improving the interactional quality and experiences with the XAI responses. This is in line with existing work on plausibility in XAI [327]: "a plausible but unfaithful interpretation may be the worst-case scenario."

Fine-grained Analysis to Promote Appropriate Reliance. Our experimental results in Chapter 7 suggest that appropriate reliance at a global level and complementary team performance may be dependent on more fine-grained appropriate reliance on the intermediate steps (*i.e.,* global transparency) and supporting documents (*i.e.,* local transparency). While most existing work has explored a *one-step* decision workflow, the user decision making process and user decision criteria are not accessible for analysis. Existing empirical studies typically set up experiments with several conditions by controlling factors about user, task, and AI system. In such a setup, the user reliance on more fine-grained task input and the surrounding context (*e.g.,* relevant documents) are typically not considered for analysis. We argue that this limits us, as a community, from developing an insightful understanding of appropriate reliance on AI advice. In this spirit, Chapter 7 has important methodological implications for both studying and promoting appropriate reliance with a fine-grained analysis. Our findings and implications can help develop human-centered AI systems for complex tasks highlighting accountability, like medical diagnosis, loan prediction, supply chain optimization, etc. The users would benefit from the critical mindset and intermediate results of a multi-step transparent workflow.

Insights for Effective Collaboration with Plan-then-execute LLM Agents. Chapter 8 has important theoretical implications for effective human-AI collaboration with

plan-then-execute LLM agents. On the one hand, user involvement can be necessary to achieve complementary team performance. Although LLM agents have shown promising planning and execution capabilities, they are never perfect due to probabilistic uncertainty. With user involvement in the planning, users can fix imperfect plans with grammar errors. With user involvement in the execution, users can fix uncertainty issues and prevent risky actions. On the other hand, user involvement may also bring uncertainty and even harm LLM agent performance. Moreover, user involvement in planning and execution poses a significantly higher cognitive load on users and negatively impacts user confidence. Thus, too much human involvement in collaboration with plan-then-execute LLM agents can be undesirable. User involvement in the execution process brings more consistent benefits than user involvement in the planning stage. As suggested by the participants, iterative LLM agent simulation may be one potential way to decide when users should be involved. The LLM agent may first conduct a plan-then-execute round to obtain a clear plan and execution results. With humans checking the whole process and simulated outcomes, humans can decide whether to be involved in revising the plan or the execution process. In this way, we can minimize user involvement while keeping highly effective task outcomes through LLM agents.

9.3 Limitations and Future Work

Recent years have witnessed rapid growth in research on facilitating effective human–AI collaboration. While the studies presented in this dissertation followed rigorous designs, some limitations may stem from the technological constraints and conceptual understanding available at the time of their implementation. This section acknowledges these limitations and outlines promising directions for future research in this evolving field.

In Part I, the implementation of AI systems mainly involves simple classifier models based on raw input features and hidden vectors obtained from transformer models. Meanwhile, we only provide explanations for AI advice with feature attribution (*i.e.,* highlighting the contribution of model input). Such explanations may not be interpretable enough for users to understand, potentially affecting the expected tutorial interventions. With the recent progress of large language models (LLMs), we can obtain more accurate AI predictive power and more coherent and understandable explanations. Future work can consider examining our findings and implications with improved technical implementation. Meanwhile, we also realized that the analysis in Part I only adopted relatively simple tasks that do not require domain expertise. Future work may confirm their impacts in different contexts by adjusting task-related factors (*e.g.,* task complexity). Our implications also provide useful guidelines for creating effective tutorial interventions. Within such a context, future work can consider providing a more user-friendly interaction design (*e.g.,* with a dashboard to visualize the human and AI differences across tasks) to help calibrate user perception of competence and promote appropriate reliance.

In part II, we found analogy-based explanations (Chapter 5) and evaluative decision support (Chapter 6) did not perform as expected. These results may reflect limitations in the implementation of the XAI methods. For example, neither approach provided sufficiently informative signals to help users distinguish between correct and incorrect AI outputs [250]. In addition, the cognitive effort required to engage in evaluative conversation may have been too demanding in a crowdsourcing setting, diverting participants'

attention from the core task of judging AI correctness.

In Chapter 2 and Chapter 5, the analogies are manually selected or generated with a template-based crowd computing method. However, according to the recent advances in LLMs, we can now rely on powerful LLMs (*e.g.,* GPT-4o) to generate high-quality analogies. Such analogy generation may be more cost-effective and can be adjusted according to user needs. Meanwhile, our findings also highlight the importance of personalized and flexible presentation of analogies. Instead of providing prepared analogies in human-AI collaboration, we may consider setting up high-quality prompts tailored to user information needs to provide analogies on demand. To our knowledge, this area is still underexplored and deserves more research efforts. Furthermore, the idea of using analogical inference to help interpret AI predictions deserves further exploration. Such analogy-based concept-level explanations have the potential to facilitate user understanding in complex tasks that require domain expertise. Future work can consider applying it to knowledge-intensive tasks where external knowledge sources (like a knowledge base and supporting documents) can help generate high-quality analogy-based explanations.

In Chapter 6, we selected the most representative five XAI methods as the basis to form our interactive XAI interfaces. We cannot overrule that this design choice may have been a bottleneck for some participants in our study, as they may have had information needs not covered by the XAI methods. Once users find that their queries cannot be answered properly based on pre-defined XAI methods, their trust and reliance on the AI system may decrease. While LLMs can 'understand' user information needs to some extent, their effectiveness is not guaranteed. Thus, there is a substantial need to create a more comprehensive and precise approach to identifying user information needs and explainability intent. Conversational XAI interface for decision support may substantially benefit from such kind of research. Meanwhile, we would argue that evaluative decision support can be a promising avenue to develop a critical mindset in human-AI collaboration. The findings in this dissertation also confirm that the XAI decision support can provide users with an illusion of explanatory depth, reinforcing over-reliance. Developing a critical mindset on users of AI systems has been proven to be effective in facilitating appropriate reliance. The evaluative XAI decision support may be a natural and low-barrier approach to advancing user understanding while keeping a critical mindset in adopting AI advice.

A clear limitation of Part III is that users were still required to complete the entire workflow alongside AI assistance, which may impose a substantial cognitive load on them. Once the efforts to control/interact with complex workflow are greater than the efforts to execute the tasks themselves, users will be unwilling to adopt such "assistance". Future work can explore how to provide necessary user involvement on demand, which can be implemented with heuristics (*e.g.,* based on AI uncertainty or the potential loss of wrong outcomes at each step). Going one step further, future work can train a predictive system to decide when to incorporate human oversight. An alternative to reduce user cognitive load in such human-AI collaboration can be providing better visualization and user-friendly interactive interfaces. Future work can design more effective user involvement and low-barrier user interfaces for user control of agentic workflow.

Another limitation of part III is the potential risks brought by LLMs. In both Chapter 7 and Chapter 8, LLMs generate convincingly wrong content (*i.e.,* workflow and plans, respectively) that may cause over-reliance. The perceived plausibility of AI-generated con-

tent can foster an illusion of intelligence, prompting users to defer critical decisions to AI systems inappropriately. Future work can combine the insights and findings from this dissertation to explore effective interventions for user trust calibration with convincingly wrong LLM outcomes. Meanwhile, LLMs still suffer from the inherent opaqueness and uncertainty issues of deep neural networks. Even the explanations and reasoning process of the LLMs have been observed to be insufficient to reflect their rationale for response generation. Therefore, future work can further explore the transparency and rationale of LLM functioning. We have noticed that there have been some proposals for the mechanistic interpretability of AI systems. Future work may follow this thought to provide more transparent AI assistance with powerful LLMs.

Based on the open feedback collected from human-AI collaboration with agentic workflow in Chapter 8, we also find that some users expressed concerns about losing agency to AI systems. Moving forward, future research should not only focus on advancing performance-related outcomes but also prioritize experience-based outcomes (*e.g.,* user satisfaction, workload, and user agency). Ethical frameworks must be embedded into the development lifecycle of AI systems, ensuring that these tools augment rather than undermine human well-being.

A general limitation of this dissertation is that all empirical studies are based on crowdsourcing. The findings may be biased due to the contextual factors associated with crowd workers and the online working environment. Meanwhile, all human-AI collaboration is operationalized in a simulation environment (most existing work in human-AI collaboration also suffers from it). Although we provide monetary bonuses to incentivize active engagement, there can be some differences in practical human-AI collaboration. Future work can aim to minimize such impacts and increase the transferability of findings/implications obtained with such a setup.

Bibliography

References

[1] Stanley H Ambrose. Paleolithic technology and human evolution. *Science*, 291(5509):1748–1753, 2001.

[2] Christophe Boesch and Hedwige Boesch. Tool use and tool making in wild chimpanzees. *Folia primatologica*, 54(1-2):86–99, 1990.

[3] Michael Tomasello and Josep Call. *Primate cognition*. Oxford University Press, 1997.

[4] Vaclav Smil. *Energy and civilization: a history*. MIT press, 2017.

[5] George Basalla. *The evolution of technology*. Cambridge University Press, 1988.

[6] Andrew Feenberg. *Transforming technology: A critical theory revisited*. Oxford University Press, 2002.

[7] A Norman Donald. *The design of everyday things*. MIT Press, 2013.

[8] Langdon Winner. *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press, 2020.

[9] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. pearson, 2016.

[10] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM computing surveys (CSUR)*, 51(5):1–36, 2018.

[11] Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38, 2020.

[12] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.

[13] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[14] Martin Adam, Michael Wessel, and Alexander Benlian. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 31(2):427–445, 2021.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[16] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–20, 2023.

[17] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504, 2023.

[18] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30, 2018.

[19] Dan W Meyrowitsch, Andreas K Jensen, Jane B Sørensen, and Tibor V Varga. Ai chatbots and (mis) information in public health: impact on vulnerable communities. *Frontiers in Public Health*, 11:1226776, 2023.

[20] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[21] Ben Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[22] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. page 1369–1385, 2023.

[23] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. Ai for social good: unlocking the opportunity for positive impact. *Nature Communications*, 11(1):2468, 2020.

[24] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2024.

[25] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021.

[26] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

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

[28] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[29] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making. In *ACM Conference on Human Factors in Computing Systems (CHI'22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*, 2022.

[30] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2023.

[31] Gaole He, Abri Bharos, and Ujwal Gadiraju. To err is ai! debugging as an intervention to facilitate appropriate reliance on ai systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, pages 98–105, 2024.

[32] Sven Eckhardt, Niklas Kühl, Mateusz Dolata, and Gerhard Schwabe. A survey of ai reliance. *arXiv preprint arXiv:2408.03948*, 2024.

[33] 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(1):114, 2015.

[34] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[35] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170, 2018.

[36] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253, 1997.

[37] John D Lee and Neville Moray. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies*, 40(1):153–184, 1994.

[38] Gerhard Fischer. User modeling in human–computer interaction. *User modeling and user-adapted interaction*, 11:65–86, 2001.

[39] Albert Bandura et al. Social foundations of thought and action. *Englewood Cliffs, NJ*, 1986(23-28):2, 1986.

[40] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[41] Mihai Mutascu. Artificial intelligence and unemployment: New insights. *Economic Analysis and Policy*, 69:653–667, 2021.

[42] Zaixuan Zhang, Zhansheng Chen, and Liying Xu. Artificial intelligence and moral dilemmas: Perception of ethical decision-making in ai. *Journal of Experimental Social Psychology*, 101:104327, 2022.

[43] Jorge Galindo and Pablo Tamayo. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, 15(1):107–143, 2000.

[44] Gang Kou, Yi Peng, and Guoxun Wang. Evaluation of clustering algorithms for financial risk analysis using mcdm methods. *Information Sciences*, 275:1–12, 2014.

[45] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

[46] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. Risk stratification of lung nodules using 3d cnn-based multi-task learning. In *International conference on information processing in medical imaging*, pages 249–260. Springer, 2017.

[47] Sam Desiere, Kristine Langenbucher, and Ludo Struyven. Statistical profiling in public employment services: An international comparison. 2019.

[48] Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *IJCAI*, pages 4070–4073, 2016.

[49] Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[50] Chun-Wei Chiang and Ming Yin. You'd better stop! understanding human reliance on machine learning models under covariate shift. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 120–129, 2021.

[51] Gaole He and Ujwal Gadiraju. Walking on eggshells: Using analogies to promote appropriate reliance in human-ai decision making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*, 2022.

[52] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 8, pages 43–52, 2020.

[53] 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(5):1–42, 2018.

[54] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[55] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[56] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. How model accuracy and explanation fidelity influence user trust in ai. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019*, 2019.

[57] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[58] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170, 1983.

[59] Douglas R Hofstadter and Emmanuel Sander. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books, 2013.

[60] Justin Kruger and David Dunning. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121, 1999.

[61] Matan Mazor and Stephen M Fleming. The dunning-kruger effect revisited. *Nature Human Behaviour*, 5(6):677–678, 2021.

[62] 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. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 307–317, 2021.

[63] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94, 2002.

[64] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

[65] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[66] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In

Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 295–305. ACM, 2020.

[67] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 401–413. ACM, 2021.

[68] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 454–464, 2020.

[69] Anuschka Schmitt, Thiemo Wambsganss, Matthias Söllner, and Andreas Janson. Towards a trust reliance paradox? exploring the gap between perceived trust in and reliance on algorithmic advice. In *International Conference on Information Systems (ICIS)*, 2021.

[70] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021.

[71] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Does stated accuracy affect trust in machine learning algorithms. In *Proceedings of ICML2018 Workshop on Human Interpretability in Machine Learning (WHI 2018)*, volume 7, 2018.

[72] 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. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 97–105, 2019.

[73] Michael A Martin. "it's like... you know": The use of analogies and heuristics in teaching introductory statistical methods. *Journal of Statistics Education*, 11(2), 2003.

[74] Eric Mintz and Truls Østbye. Teaching statistics to health professionals: the legal analogy. *Medical Teacher*, 14(4):371–374, 1992.

[75] Mirta Galesic and Rocio Garcia-Retamero. Communicating consequences of risky behaviors: Life expectancy versus risk of disease. *Patient education and counseling*, 82(1):30–35, 2011.

[76] Mirta Galesic and Rocio Garcia-Retamero. Using analogies to communicate information about health risks. *Applied Cognitive Psychology*, 27(1):33–42, 2013.

[77] Pradeep Sopory and James Price Dillard. The persuasive effects of metaphor: A meta-analysis. *Human communication research*, 28(3):382–419, 2002.

[78] Stephanie K Van Stee. Meta-analysis of the persuasive effects of metaphorical vs. literal messages. *Communication Studies*, 69(5):545–566, 2018.

[79] Elisa Barilli, Lucia Savadori, Stefania Pighin, Sara Bonalumi, Augusto Ferrari, Maurizio Ferrari, and Laura Cremonesi. From chance to choice: The use of a verbal analogy in the communication of risk. *Health, Risk & Society*, 12(6):546–559, 2010.

[80] Stefania Pighin, Lucia Savadori, Elisa Barilli, Rino Rumiati, Sara Bonalumi, Maurizio Ferrari, and Laura Cremonesi. Using comparison scenarios to improve prenatal risk communication. *Medical Decision Making*, 33(1):48–58, 2013.

[81] Carmen Keller, Michael Siegrist, and Vivianne Visschers. Effect of risk ladder format on risk perception in high-and low-numerate individuals. *Risk Analysis: An International Journal*, 29(9):1255–1264, 2009.

[82] Rocio Garcia-Retamero, Agata Sobkow, Dafina Petrova, Dunia Garrido, and Jakub Traczyk. Numeracy and risk literacy: What have we learned so far? *The Spanish journal of psychology*, 22, 2019.

[83] Marianne Bertrand, Sendhil Mullainathan, and Eldar Shafir. Behavioral economics and marketing in aid of decision making among the poor. *Journal of Public Policy & Marketing*, 25(1):8–23, 2006.

[84] 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. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021.

[85] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 90–99, 2019.

[86] Katharine Sanderson. Covid vaccines protect against delta, but their effectiveness wanes. *Nature*, 2021.

[87] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization*, pages 77–87, 2021.

[88] Brian J Zikmund-Fisher, Dylan M Smith, Peter A Ubel, and Angela Fagerlin. Validation of the subjective numeracy scale: Effects of low numeracy on comprehension of risk communications and utility elicitations. *Medical Decision Making*, 27:663–671, 2007.

[89] Angela Fagerlin, Brian J Zikmund-Fisher, Peter A Ubel, Aleksandra Jankovic, Holly A Derry, and Dylan M Smith. Measuring numeracy without a math test: development of the subjective numeracy scale. *Medical Decision Making*, 27(5):672–680, 2007.

[90] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer, 2018.

[91] Thomas Franke, Christiane Attig, and Daniel Wessel. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2019.

[92] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009.

[93] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1631–1640, 2015.

[94] Geoff Norman. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education*, 15(5):625–632, 2010.

[95] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[96] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(4):1–26, 2017.

[97] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. I can do better than your ai: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 240–251, New York, NY, USA, 2019. Association for Computing Machinery.

[98] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 89–101, 2022.

[99] Mahsan Nourani, Joanie King, and Eric Ragan. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121, 2020.

[100] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14, 2018.

[101] Niels Van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[102] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, 2018.

[103] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

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

[105] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[106] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

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

[108] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In Giulio Jacucci, Samuel Kaski, Cristina Conati, Simone Stumpf, Tuukka Ruotsalo, and Krzysztof Gajos, editors, *IUI 2022: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, March 22 - 25, 2022*, pages 148–161. ACM, 2022.

[109] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. Understanding the role of explanation modality in ai-assisted decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 223–233, 2022.

[110] Ben Green and Yiling Chen. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–33, 2021.

[111] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*, pages 1–14, 2022.

[112] Ilan Yaniv and Eli Kleinberger. Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational behavior and human decision processes*, 83(2):260–281, 2000.

[113] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with ai systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[114] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018, 2022.

[115] Joyce Ehrlinger, Kerri Johnson, Matthew Banner, David Dunning, and Justin Kruger. Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes*, 105(1):98–121, 2008.

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

[117] Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[118] David Dunning, Chip Heath, and Jerry M Suls. Flawed self-assessment: Implications for health, education, and the workplace. *Psychological science in the public interest*, 5(3):69–106, 2004.

[119] 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. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11, 2019.

[120] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, 2021.

[121] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4):100049, 2020.

[122] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

[123] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.

[124] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O Riedl. Operationalizing human-centered perspectives in explainable ai. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2021.

[125] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In Sorelle A. Friedler and Christo Wilson, editors, *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, page 48. PMLR, 2018.

[126] Upol Ehsan and Mark O Riedl. Human-centered explainable ai: towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*, pages 449–466. Springer, 2020.

[127] Upol Ehsan, Philipp Wintersberger, Q Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. Human-centered explainable ai (hcxai): beyond opening the black-box of ai. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

[128] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239, 2020.

[129] Jan Gogoll and Matthias Uhl. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 74:97–103, 2018.

[130] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650, 2019.

[131] Nicole Dillen, Marko Ilievski, Edith Law, Lennart E Nacke, Krzysztof Czarnecki, and Oliver Schneider. Keep calm and ride along: Passenger comfort and anxiety as physiological responses to autonomous driving styles. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[132] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. In Frantz Rowe, Redouane El Amrani, Moez Limayem, Sue Newell, Nancy Pouloudi, Eric van Heck, and Ali El Quammah, editors, *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020*, 2020.

[133] Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

[134] Donghee Shin, Bouziane Zaid, and Mohammed Ibahrine. Algorithm appreciation: Algorithmic performance, developmental processes, and user interactions. In *2020 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*, pages 1–5. IEEE, 2020.

[135] Max F Kramer, Jana Schaich Borg, Vincent Conitzer, and Walter Sinnott-Armstrong. When do people want ai to make decisions? In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 204–209, 2018.

[136] Donghee Shin, Kerk F Kee, and Emily Y Shin. Algorithm awareness: Why user awareness is critical for personal privacy in the adoption of algorithmic platforms? *International Journal of Information Management*, 65:102494, 2022.

[137] Sangseok You, Cathy Liu Yang, and Xitong Li. Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, 39(2):336–365, 2022.

[138] Yoyo Tsung-Yu Hou and Malte F Jung. Who is the expert? reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[139] Rachel A Jansen, Anna N Rafferty, and Thomas L Griffiths. A rational model of the dunning–kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, 5(6):756–763, 2021.

[140] Cynthia Sherraden Bradley, Kristina Thomas Dreifuerst, Brandon Kyle Johnson, and Ann Loomis. More than a meme: The dunning-kruger effect as an opportunity for positive change in nursing education. *Clinical Simulation in Nursing*, 66:58–65, 2022.

[141] James Sawler. Economics 101-ism and the dunning-kruger effect: Reducing overconfidence among introductory macroeconomics students. *International Review of Economics Education*, 36:100208, 2021.

[142] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April 2020.

[143] Catherine C Marshall and Frank M Shipman. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 234–243, 2013.

[144] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai. *IEEE Internet Computing*, 25(5):19–24, 2021.

[145] Muhammad Bilal Zafar, Philipp Schmidt, Michele Donini, Cédric Archambeau, Felix Biessmann, Sanjiv Ranjan Das, and Krishnaram Kenthapadi. More than words: Towards better quality interpretations of text classifiers. *arXiv preprint arXiv:2112.12444*, 2021.

[146] Hanqi Yan, Lin Gui, and Yulan He. Hierarchical interpretation of neural text classification. *Computational Linguistics*, 48(4):987–1020, 2022.

[147] Fangzhi Xu, Jun Liu, Qika Lin, Yudai Pan, and Lingling Zhang. Logiformer: A two-branch graph transformer network for interpretable logical reasoning. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1055–1065. ACM, 2022.

[148] Ricardo Baeza-Yates. Bias on the web. *Communications of the ACM*, 61(6):54–61, 2018.

[149] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

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

[151] Jennifer Fereday and Eimear Muir-Cochrane. The role of performance feedback in the self-assessment of competence: a research study with nursing clinicians. *Collegian*, 13(1):10–15, 2006.

[152] Conor Thomas McKevitt. Engaging students with self-assessment and tutor feedback to improve performance and support assessment capacity. *Journal of University Teaching & Learning Practice*, 13(1):2, 2016.

[153] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[154] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. No risk no trust: Investigating perceived risk in highly automated driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 177–185, 2019.

[155] Peter Lipton. Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27:247–266, 1990.

[156] Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36, 2021.

[157] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. A checklist to combat cognitive biases in crowdsourcing. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 48–59, 2021.

[158] Ioannis Petros Samiotis, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, and Alessandro Bozzon. Exploring the music perception skills of crowd workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 9, pages 108–119, 2021.

[159] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, 2016.

[160] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 272–283, 2020.

[161] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042, 2018.

[162] Piyawat Lertvittayakumjorn and Francesca Toni. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528, 2021.

[163] Agathe Balayn, Natasa Rikalo, Christoph Lofi, Jie Yang, and Alessandro Bozzon. How can explainability methods be used to support bug identification in computer vision models? In *CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2022.

[164] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[165] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020.

[166] Andrew Selbst and Julia Powles. "meaningful information" and the right to explanation. In *conference on fairness, accountability and transparency*, pages 48–48. PMLR, 2018.

[167] 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. 2020.

[168] Murat Dikmen and Catherine Burns. The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792, 2022.

[169] Hussein Mozannar, Arvind Satyanarayan, and David Sontag. Teaching humans when to defer to a classifier via exemplars. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5323–5331, 2022.

[170] Samir Passi and Mihaela Vorvoreanu. Overreliance on ai literature review. *Microsoft Research*, 2022.

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

[172] 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(CSCW):1–24, 2019.

[173] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, pages 126–137, 2015.

[174] Anthony Liu, Santiago Guerra, Isaac Fung, Gabriel Matute, Ece Kamar, and Walter Lasecki. Towards hybrid human-ai workflows for unknown unknown detection. In *Proceedings of The Web Conference 2020*, pages 2432–2442, 2020.

[175] Shahin Sharifi Noorian, Sihang Qiu, Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. In *Proceedings of the ACM Web Conference 2022*, pages 882–892, 2022.

[176] Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K Thiruvathukal, and Ming Yin. Crowdsourcing detection of sampling biases in image datasets. In *Proceedings of The Web Conference 2020*, pages 2955–2961, 2020.

[177] Gamze Türkmen and Sonay Caner. The investigation of novice programmers'debugging behaviors to inform intelligent e-learning environments: A case study. *Turkish Online Journal of Distance Education*, 21(3):142–155, 2020.

[178] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[179] Kacper Sokol and Peter Flach. One explanation does not fit all. *KI-Künstliche Intelligenz*, 34(2):235–250, 2020.

[180] Anne-Marie Nussberger, Lan Luo, L Elisa Celis, and Molly J Crockett. Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nature Communications*, 13(1):1–13, 2022.

[181] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[182] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Rethinking ai explainability and plausibility. *arXiv preprint arXiv:2303.17707*, 2023.

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

[184] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9, 2019.

[185] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

[186] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[187] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR, 2018.

[188] Esther I Verhoef, Wiggert A van Cappellen, Johan A Slotman, Gert-Jan Kremers, Patricia C Ewing-Graham, Adriaan B Houtsmuller, Martin E van Royen, and Geert JLH van Leenders. Three-dimensional analysis reveals two major architectural subgroups of prostate cancer growth patterns. *Modern Pathology*, 32(7):1032–1041, 2019.

[189] Francesco Sovrano, Salvatore Sapienza, Monica Palmirani, and Fabio Vitali. Metrics, explainability and the european ai act proposal. *J*, 5(1):126–138, 2022.

[190] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y Lim. Cogam: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14, 2020.

[191] Daniel E Ehrmann, Sara N Gallant, Sujay Nagaraj, Sebastian D Goodfellow, Danny Eytan, Anna Goldenberg, and Mjaye L Mazwi. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. *Nature Medicine*, pages 1–2, 2022.

[192] Filip Ilievski, Alessandro Oltramari, Kaixin Ma, Bin Zhang, Deborah L. McGuinness, and Pedro A. Szekely. Dimensions of commonsense knowledge. *Knowl. Based Syst.*, 229:107347, 2021.

[193] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2829–2839. Association for Computational Linguistics, 2019.

[194] Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. Augmenting end-to-end dialogue systems with commonsense knowledge. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4970–4977. AAAI Press, 2018.

[195] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE, 2019.

[196] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the general public. In Robert Meersman and Zahir Tari, editors, *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002 Irvine, California, USA, October 30 - November 1, 2002, Proceedings*, volume 2519 of *Lecture Notes in Computer Science*, pages 1223–1237. Springer, 2002.

[197] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press, 2017.

[198] Luis von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: a game for collecting common-sense facts. In Rebecca E. Grinter, Tom Rodden, Paul M. Aoki, Edward Cutrell, Robin Jeffries, and Gary M. Olson, editors, *Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006, Montréal, Québec, Canada, April 22-27, 2006*, pages 75–78. ACM, 2006.

[199] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. Ready player one! eliciting diverse knowledge using A configurable game. In Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini, editors, *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 1709–1719. ACM, 2022.

[200] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4932–4942. Association for Computational Linguistics, 2019.

[201] Haozhe Ji, Pei Ke, Shaohan Huang, Furu Wei, and Minlie Huang. Generating commonsense explanation by extracting bridge concepts from reasoning paths. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 248–257. Association for Computational Linguistics, 2020.

[202] Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian J. McAuley. Rationale-inspired natural language explanations with commonsense. *CoRR*, abs/2106.13876, 2021.

[203] John K Gilbert and Rosária Justi. Analogies in modelling-based teaching and learning. In *Modelling-based teaching in science education*, pages 149–169. Springer, 2016.

[204] Samson Madera Nashon. The nature of analogical explanations: High school physics teachers use in kenya. *Research in Science Education*, 34(4):475–502, 2004.

[205] David Geelan. Teacher explanations. *Second international handbook of science education*, pages 987–999, 2012.

[206] Nilmara Braga Mozzer and Rosária Justi. Students' pre-and post-teaching analogical reasoning when they draw their analogies. *International Journal of Science Education*, 34(3):429–458, 2012.

[207] Hanxiao Liu, Yuexin Wu, and Yiming Yang. Analogical inference for multi-relational embeddings. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2168–2178. PMLR, 2017.

[208] Myriam Bounhas, Marc Pirlot, Henri Prade, and Olivier Sobrie. Comparison of analogy-based methods for predicting preferences. In Nahla Ben Amor, Benjamin Quost, and Martin Theobald, editors, *Scalable Uncertainty Management - 13th International Conference, SUM 2019, Compiègne, France, December 16-18, 2019, Proceedings*, volume 11940 of *Lecture Notes in Computer Science*, pages 339–354. Springer, 2019.

[209] Marc T. Law, Nicolas Thome, and Matthieu Cord. Learning a distance metric from relative comparisons between quadruplets of images. *Int. J. Comput. Vis.*, 121(1):65–94, 2017.

[210] Henri Prade and Gilles Richard. Analogical proportions: Why they are useful in AI. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4568–4576. ijcai.org, 2021.

[211] Eyke Hüllermeier. Towards analogy-based explanations in machine learning. In Vicenç Torra, Yasuo Narukawa, Jordi Nin, and Núria Agell, editors, *Modeling Decisions for Artificial Intelligence - 17th International Conference, MDAI 2020, Sant Cugat, Spain, September 2-4, 2020, Proceedings*, volume 12256 of *Lecture Notes in Computer Science*, pages 205–217. Springer, 2020.

[212] Reinders Duit, Wolff-Michael Roth, Michael Komorek, and Jens Wilbers. Fostering conceptual change by analogies—between scylla and charybdis. *Learning and Instruction*, 11(4-5):283–303, 2001.

[213] Mark Cosgrove. A study of science-in-the-making as students generate an analogy for electricity. *International journal of science education*, 17(3):295–310, 1995.

[214] Tony Veale. Analogy generation with hownet. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pages 1148–1153. Professional Book Center, 2005.

[215] Zhendong Dong and Qiang Dong. Hownet-a hybrid language and knowledge resource. In *International conference on natural language processing and knowledge engineering, 2003. Proceedings. 2003*, pages 820–824. IEEE, 2003.

[216] Andy Chiu, Pascal Poupart, and Chrysanne DiMarco. Generating lexical analogies using dependency relations. In Jason Eisner, editor, *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 561–570. ACL, 2007.

[217] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[218] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. How cognitive biases affect xai-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 78–91, 2022.

[219] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. Understanding choice independence and error types in human-ai collaboration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

[220] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in

human-ai decision making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

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

[222] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, page 103952, 2023.

[223] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*, 2018.

[224] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021.

[225] Diane F Halpern, Carol Hansen, and David Riefer. Analogies as an aid to understanding and memory. *Journal of educational psychology*, 82(2):298, 1990.

[226] Paul Bartha. Analogy and Analogical Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

[227] Keith J. Holyoak and Paul Thagard. Analogical mapping by constraint satisfaction. *Cogn. Sci.*, 13(3):295–355, 1989.

[228] Bernhard Thalheim. The theory of conceptual models, the theory of conceptual modelling and foundations of conceptual modelling. In David W. Embley and Bernhard Thalheim, editors, *Handbook of Conceptual Modeling - Theory, Practice, and Research Challenges*, pages 543–577. Springer, 2011.

[229] Dedre Gentner and Arthur B Markman. Structure mapping in analogy and similarity. *American psychologist*, 52(1):45, 1997.

[230] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464, 2018.

[231] Samuel J Stratton. Population research: convenience sampling strategies. *Prehospital and disaster Medicine*, 36(4):373–374, 2021.

[232] Anselm L Strauss. *Qualitative analysis for social scientists*. Cambridge university press, 1987.

[233] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.

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

[235] Lindsey E Richland and Janice Hansen. Reducing cognitive load in learning by analogy. *International Journal of Psychological Studies*, 5(4):69, 2013.

[236] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[237] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

[238] Troy L Adams, Yuanxia Li, and Hao Liu. A replication of beyond the turk: Alternative platforms for crowdsourcing behavioral research–sometimes preferable to student groups. *AIS Transactions on Replication Research*, 6(1):15, 2020.

[239] Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720, 2023.

[240] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Computer Methods and Programs in Biomedicine*, 215:106620, 2022.

[241] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

[242] Lacey Colligan, Henry WW Potts, Chelsea T Finn, and Robert A Sinkin. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics*, 84(7):469–476, 2015.

[243] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

[244] 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. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2020.

[245] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[246] Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In *International semantic web conference*, pages 486–504. Springer, 2014.

[247] Mike Schaekermann, Graeme Beaton, Minahz Habib, Andrew Lim, Kate Larson, and Edith Law. Understanding expert disagreement in medical data analysis through structured adjudication. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[248] Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In Simone D. J. Barbosa, Cliff Lampe, Caroline Appert, David A. Shamma, Steven Mark Drucker, Julie R. Williamson, and Koji Yatani, editors, *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 115:1–115:19. ACM, 2022.

[249] Chacha Chen, Shi Feng, Amit Sharma, and Chenhao Tan. Machine explanations and human understanding. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1–1, 2023.

[250] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction*, 7(CSCW2):1–32, 2023.

[251] Roger Parloff. Why deep learning is suddenly changing your life. *Fortune. New York: Time Inc*, 2016.

[252] Terrence J Sejnowski. *The deep learning revolution.* MIT press, 2018.

[253] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–29, 2024.

[254] Karl de Fine Licht and Jenny de Fine Licht. Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & society*, 35:917–926, 2020.

[255] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on ai advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422, 2023.

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

[257] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–15, 2020.

[258] Wenzhuo Yang, Jia Li, Caiming Xiong, and Steven CH Hoi. Mace: An efficient model-agnostic framework for counterfactual explanation. *arXiv preprint arXiv:2205.15540*, 2022.

[259] Alon Jacovi, Jasmijn Bastings, Sebastian Gehrmann, Yoav Goldberg, and Katja Filippova. Diagnosing ai explanation methods with folk concepts of behavior. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 247–247, 2023.

[260] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[261] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

[262] Wenzhuo Yang, Hung Le, Silvio Savarese, and Steven Hoi. Omnixai: A library for explainable ai. 2022.

[263] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, 5(8):873–883, 2023.

[264] Robert J Moore, Raphael Arar, Guang-Jie Ren, and Margaret H Szymanski. Conversational ux design. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*, pages 492–497, 2017.

[265] Susan E Brennan. Conversation as direct manipulation: An iconoclastic view. *The art of human-computer interface design*, pages 393–404, 1990.

[266] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Talktomodel: Explaining machine learning models with interactive natural language conversations. 2022.

[267] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner's perspective. In *NeurIPS Workshop on Human Centered AI*, 2022.

[268] Anjana Wijekoon, David Corsar, and Nirmalie Wiratunga. Behaviour trees for conversational explanation experiences. *arXiv preprint arXiv:2211.06402*, 2022.

[269] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao Kenneth Huang. Convxai: Delivering heterogeneous ai explanations via conversations to support human-ai scientific writing. pages 384–387, 2023.

[270] Dimitry Mindlin, Amelie Robrecht, Michael Morasch, and Philipp Cimiano. Measuring user understanding in dialogue-based xai systems. In *ECAI 2024. 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, 2024.

[271] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[272] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.

[273] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. Mind the gap! choice independence in using multilingual llms for persuasive co-writing tasks in different languages. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

[274] Gaole He, Gianluca Demartini, and Ujwal Gadiraju. Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2025.

[275] Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. Unpacking trust dynamics in the llm supply chain: An empirical exploration to foster trustworthy llm production & use. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025.

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

[277] Zhuoran Lu, Dakuo Wang, and Ming Yin. Does more advice help? the effects of second opinions in ai-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–31, 2024.

[278] Sara Salimzadeh and Ujwal Gadiraju. When in doubt! understanding the role of task characteristics on peer decision-making with ai assistance. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 89–101, 2024.

[279] Greta Warren, Ruth MJ Byrne, and Mark T Keane. Categorical and continuous features in counterfactual explanations of ai systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 171–187, 2023.

[280] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. Directive explanations for monitoring the risk of diabetes onset: Introducing directive data-centric explanations and combinations to support what-if explorations. In *Proceed-*

ings of the 28th International Conference on Intelligent User Interfaces, pages 204–219, 2023.

[281] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Supporting high-uncertainty decisions through ai and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 251–263, 2023.

[282] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

[283] Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. Contrastive explanations for model interpretability. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, 2021.

[284] Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, 2022.

[285] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[286] Scott Cheng-Hsin Yang, Nils Erik Tomas Folke, and Patrick Shafto. A psychological theory of explainability. In *International Conference on Machine Learning*, pages 25007–25021. PMLR, 2022.

[287] Philip N Johnson-Laird. Mental models in cognitive science. *Cognitive science*, 4(1):71–115, 1980.

[288] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 1–10, 2012.

[289] 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. In *2013 IEEE Symposium on visual languages and human centric computing*, pages 3–10. IEEE, 2013.

[290] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[291] Wikipedia. Conversational user interface — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Conversational%20user%20interface, 2023. [Online; accessed 05-September-2023].

[292] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[293] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[294] Michael F McTear. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169, 2002.

[295] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 117–126, 2017.

[296] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

[297] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. Crslab: An open-source toolkit for building conversational recommender system. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 185–193, 2021.

[298] Shrikanth Narayanan and Alexandros Potamianos. Creating conversational interfaces for children. *IEEE Transactions on Speech and Audio Processing*, 10(2):65–78, 2002.

[299] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. To trust or not to trust: How a conversational interface affects trust in a decision support system. In *Proceedings of the ACM Web Conference 2022*, pages 3531–3540, 2022.

[300] Sumit Srivastava, Mariët Theune, and Alejandro Catala. The role of lexical alignment in human understanding of explanations by conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 423–435, 2023.

[301] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. A missing piece in the puzzle: Considering the role of task complexity in human-ai decision making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 215–227, 2023.

[302] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[303] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[304] Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–11, 2019.

[305] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Ticktalkturk: Conversational crowdsourcing made easy. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, pages 53–57, 2020.

[306] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*, pages 555–565, 2017.

[307] Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 333–342, 2023.

[308] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

[309] Oege Dijk, oegesam, Ray Bell, Lily, Simon-Free, Brandon Serna, rajgupt, yanhongzhao ef, Achim Gädke, Hugo, and Tunay Okumus. oegedijk/explainerdashboard: v0.3.8.2: reverses set_shap_values bug introduced in 0.3.8.1, April 2022.

[310] 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*, 26(1):1064–1074, 2019.

[311] Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the quality of explanations: the system causability scale (scs) comparing human and machine explanations. *KI-Künstliche Intelligenz*, 34(2):193–198, 2020.

[312] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International journal of human-computer studies*, 146:102551, 2021.

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

[314] 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. In *Proceedings of the 27th international conference on intelligent user interfaces*, pages 807–819, 2022.

[315] Timothée Schmude, Laura Koesten, Torsten Möller, and Sebastian Tschiatschek. On the impact of explanations on understanding of algorithmic decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–970, 2023.

[316] Kalervo Järvelin and Jaana Kekäläinen. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA, 2017.

[317] Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. Roadmap of designing cognitive metrics for explainable artificial intelligence (xai). *arXiv preprint arXiv:2108.01737*, 2021.

[318] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4):1–45, 2021.

[319] Mark Ryan. In ai we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5):2749–2767, 2020.

[320] Matt Twyman, Nigel Harvey, and Clare Harries. Trust in motives, trust in competence: Separate factors determining the effectiveness of risk communication. *Judgment and Decision Making*, 3(1):111–120, 2008.

[321] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*, pages 13–30. Springer, 2019.

[322] Heather L O'Brien, Paul Cairns, and Mark Hall. A practical approach to measuring user engagement with the refined user engagement scale (ues) and new ues short form. *International Journal of Human-Computer Studies*, 112:28–39, 2018.

[323] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. Moving beyond p values: data analysis with estimation graphics. *Nature methods*, 16(7):565–566, 2019.

[324] Leonid Rozenblit and Frank Keil. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science*, 26(5):521–562, 2002.

[325] Brian J Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):2, 2002.

[326] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–17, 2023.

[327] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020.

[328] Indrani Medhi, Somani Patnaik, Emma Brunskill, SN Nagasena Gautama, William Thies, and Kentaro Toyama. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(1):1–28, 2011.

[329] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[330] Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Aslı Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, et al. Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 34(3):660–671, 2021.

[331] Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. Complex knowledge base question answering: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[332] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

[333] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. " the human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 99–109, 2020.

[334] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, page 78, 2021.

[335] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 617–626, 2022.

[336] Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[337] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4942–4950, 2018.

[338] Heike Felzmann, Eduard Fosch-Villaronga, Christoph Lutz, and Aurelia Tamò-Larrieux. Towards transparency by design for artificial intelligence. *Science and engineering ethics*, 26(6):3333–3361, 2020.

[339] Yugang Li, Baizhou Wu, Yuqi Huang, and Shenghua Luan. Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-ai trust. *Frontiers in Psychology*, 15:1382693, 2024.

[340] Alexandra D Kaplan, Theresa T Kessler, J Christopher Brill, and Peter A Hancock. Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359, 2023.

[341] Ewart J De Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478, 2020.

[342] Kumar Akash, Griffon McMahon, Tahira Reid, and Neera Jain. Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Systems Magazine*, 40(6):98–116, 2020.

[343] Christopher A Miller. Trust, transparency, explanation, and planning: Why we need a lifecycle perspective on human-automation interaction. In *Trust in human-robot interaction*, pages 233–257. Elsevier, 2021.

[344] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, 19(3):259–282, 2018.

[345] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, 58(3):401–415, 2016.

[346] Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835, 2024.

[347] Michael Vössing, Niklas Kühl, Matteo Lind, and Gerhard Satzger. Designing transparency for effective human-ai collaboration. *Information Systems Frontiers*, 24(3):877–895, 2022.

[348] Patricia K Kahr, Gerrit Rooks, Martijn C Willemsen, and Chris CP Snijders. Understanding trust and reliance development in ai advice: Assessing model accuracy, model explanations, and experiences from previous interactions. *ACM Transactions on Interactive Intelligent Systems*, 2024.

[349] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 43–52, 2011.

[350] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. Crowdweaver: visually managing complex crowd work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 1033–1036, 2012.

[351] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. A taxonomy of microtasks on the web. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 218–223, 2014.

[352] António Correia, Andrea Grover, Daniel Schneider, Ana Paula Pimentel, Ramon Chaves, Marcos Antonio De Almeida, and Benjamim Fonseca. Designing for hybrid intelligence: A taxonomy and survey of crowd-machine interaction. *Applied Sciences*, 13(4):2198, 2023.

[353] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322, 2010.

[354] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 68–76, 2010.

[355] Daniela Retelny, Michael S Bernstein, and Melissa A Valentine. No workflow can ever be enough: How crowdsourcing workflows constrain complex work. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–23, 2017.

[356] Ece Kamar and Lydia Manikonda. Complementing the execution of ai systems with human computation. In *AAAI Workshops*, 2017.

[357] Tongshuang Wu, Haiyi Zhu, Maya Albayrak, Alexis Axon, Amanda Bertsch, Wenxing Deng, Ziqi Ding, Bill Guo, Sireesh Gururaja, Tzu-Sheng Kuo, et al. Llms as workers in human-computational algorithms? replicating crowdsourcing pipelines with llms. *arXiv preprint arXiv:2307.10168*, 2023.

[358] Daniel Zhang, Yang Zhang, Qi Li, Thomas Plummer, and Dong Wang. Crowdlearn: A crowd-ai hybrid system for deep learning-based damage assessment applications. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 1221–1232. IEEE, 2019.

[359] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. Is conversational xai all you need? human-ai decision making with a conversational xai assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 2025.

[360] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. Promptchainer: Chaining large language model prompts through visual programming. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–10, 2022.

[361] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–22, 2022.

[362] Madeleine Grunde-McLaughlin, Michelle S Lam, Ranjay Krishna, Daniel Weld, and Jeffrey Heer. Designing llm chains by adapting techniques from crowdsourcing workflows. *ACM Transactions on Computer-Human Interaction*.

[363] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, 2023.

[364] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[365] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)*, 55(2):1–38, 2022.

[366] Q Vera Liao and Jennifer Wortman Vaughan. Ai transparency in the age of llms: A human-centered research roadmap. *Harvard Data Science Review*, (Special Issue 5), 2024.

[367] NIST AI. Artificial intelligence risk management framework (ai rmf 1.0), 2023.

[368] Joshua A Kroll. Outlining traceability: A principle for operationalizing accountability in computing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 758–771, 2021.

[369] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, 2021.

[370] Jessie J Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. Real ml: Recognizing, exploring, and articulating limitations of machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 587–597, 2022.

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

[372] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

[373] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[374] Yu-Liang Chou, Catarina Moreira, Peter Bruza, Chun Ouyang, and Joaquim Jorge. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion*, 81:59–83, 2022.

[375] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, 2024.

[376] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

[377] James Thorne and Andreas Vlachos. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, 2018.

[378] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[379] Jon Roozenbeek and Sander Van der Linden. *The psychology of misinformation*. Cambridge University Press, 2024.

[380] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 85–94, 2017.

[381] Andrew Guess, Brendan Nyhan, and Jason Reifler. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, 9(3):4, 2018.

[382] Shan Jiang and Christo Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

[383] Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 2: Short papers)*, pages 647–653, 2017.

[384] Svitlana Volkova and Jin Yea Jang. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583, 2018.

[385] Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232, 2023.

[386] An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. Believe it or not: designing a human-ai partnership for mixed-initiative fact-checking. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*, pages 189–199, 2018.

[387] An Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. An interpretable joint graphical model for fact-checking from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[388] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 324–332, 2018.

[389] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 155–168, 2017.

[390] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. Can the crowd identify misinformation objectively? the effects of judgment scale and assessor's background. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 439–448, 2020.

[391] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. Crowdsourced fact-checking at twitter: How does the crowd compare with experts? In *Proceedings of the 31st ACM international conference on information & knowledge management*, pages 1736–1746, 2022.

[392] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393, 2021.

[393] Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, pages 1–12, 2024.

[394] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda

Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics.

[395] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.

[396] Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. Self-checker: Plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 163–181, 2024.

[397] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[398] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023.

[399] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[400] Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer, 1994.

[401] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. On the influence of explainable ai on automation bias. In *Proceedings of the 30th European Conference on Information Systems (ECIS), Timişoara, RO, June 18-24, 2022*, 2022.

[402] Stefano V Albrecht and Peter Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.

[403] Veton Kepuska and Gamal Bohouta. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pages 99–103. IEEE, 2018.

[404] Amos Tversky and Daniel Kahneman. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics*, 106(4):1039–1061, 1991.

[405] Hasan Mahmud, AKM Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390, 2022.

[406] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

[407] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[408] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[409] Brian Lubars and Chenhao Tan. Ask not what ai can do, but what ai should do: Towards a framework of task delegability. *Advances in neural information processing systems*, 32, 2019.

[410] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Paul Saldyt, and Anil B Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Forty-first International Conference on Machine Learning*, 2024.

[411] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, Yitao Liang, and Team CraftJarvis. Describe, explain, plan and select: interactive planning with large language models enables open-world multi-task agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34153–34189, 2023.

[412] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36, 2024.

[413] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[414] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.

[415] Florian Geissler, Karsten Roscher, and Mario Trapp. Concept-guided llm agents for human-ai safety codesign. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 100–104, 2024.

[416] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents.

In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2024.

[417] Qingxiao Zheng, Zhongwei Xu, Abhinav Choudhary, Yuting Chen, Yongming Li, and Yun Huang. Synergizing human-ai agency: a guide of 23 heuristics for service co-creation with llm-based agents. *arXiv preprint arXiv:2310.15065*, 2023.

[418] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, 2023.

[419] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM conference on hypertext and social media*, pages 5–14, 2017.

[420] Yanming Yang, Xin Xia, David Lo, and John Grundy. A survey on deep learning for software engineering. *ACM Computing Surveys (CSUR)*, 54(10s):1–73, 2022.

[421] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

[422] Shi Dong, Ping Wang, and Khushnood Abbas. A survey on deep learning and its applications. *Computer Science Review*, 40:100379, 2021.

[423] Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, 2017.

[424] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics.

[425] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.

[426] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R Brubaker. Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

[427] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[428] Marc Pinski, Martin Adam, and Alexander Benlian. Ai knowledge: Improving ai delegation through human enablement. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–17, 2023.

[429] David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: improving fairness and accuracy by learning to defer. *Advances in neural information processing systems*, 31, 2018.

[430] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696, 2022.

[431] Harikrishna Narasimhan, Wittawat Jitkrittum, Aditya K Menon, Ankit Rawat, and Sanjiv Kumar. Post-hoc estimators for learning to defer to an expert. *Advances in Neural Information Processing Systems*, 35:29292–29304, 2022.

[432] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-ai teams: Building machine learning models that complement the capabilities of multiple experts. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2478–2484. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[433] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. Human-ai collaboration: the effect of ai delegation on human task performance and task satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 453–463, 2023.

[434] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22, 2022.

[435] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718, 2003.

[436] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4):1–45, 2024.

[437] John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.

[438] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021.

[439] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[440] Q Vera Liao and S Shyam Sundar. Designing for responsible trust in ai systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1257–1268, 2022.

[441] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 822–835, New York, NY, USA, 2024. Association for Computing Machinery.

[442] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–17, 2021.

[443] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.

[444] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. Complementarity in human-ai collaboration: Concept, sources, and evidence. *arXiv preprint arXiv:2404.00029*, 2024.

[445] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635, 2021.

[446] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. Opening the analogical portal to explainability: Can analogies help laypeople in ai-assisted decision making? *Journal of Artificial Intelligence Research*, 81:117–162, 2024.

[447] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

[448] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2024.

[449] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. If in a crowdsourced data annotation pipeline, a gpt-4.

In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25, 2024.

[450] Behrooz Omidvar Tehrani and Anmol Anubhai. Evaluating human-ai partnership for llm-based code migration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

[451] Chenglei Si, Navita Goyal, Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé Iii, and Jordan Boyd-Graber. Large language models help humans verify truthfulness–except when they are convincingly wrong. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1459–1474, 2024.

[452] Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. Decision-oriented dialogue for human-ai collaboration. *Transactions of the Association for Computational Linguistics*, 12:892–911, 2024.

[453] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. *arXiv preprint arXiv:2403.16812*, 2024.

[454] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.

[455] Shijue Huang, Wanjun Zhong, Jianqiao Lu, Qi Zhu, Jiahui Gao, Weiwen Liu, Yutai Hou, Xingshan Zeng, Yasheng Wang, Lifeng Shang, Xin Jiang, Ruifeng Xu, and Qun Liu. Planning, creation, usage: Benchmarking llms for comprehensive tool utilization in real-world complex scenarios, 2024.

[456] Jason P Davis, Kathleen M Eisenhardt, and Christopher B Bingham. Developing theory through simulation methods. *Academy of management review*, 32(2):480–499, 2007.

[457] Judith S Olson and Wendy A Kellogg. *Ways of Knowing in HCI*, volume 2. Springer, 2014.

[458] Robert E Wood. Task complexity: Definition of the construct. *Organizational behavior and human decision processes*, 37(1):60–82, 1986.

[459] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[460] Shaoyang Fan, Ujwal Gadiraju, Alessandro Checco, and Gianluca Demartini. Crowdco-op: Sharing risks and rewards in crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24, 2020.

[461] Zhuoyan Li and Ming Yin. Utilizing human behavior modeling to manipulate explanations in ai-assisted decision making: The good, the bad, and the scary. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[462] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024.

[463] Niccolò Pescetelli and Nicholas Yeung. The role of decision confidence in advice-taking and trust formation. *Journal of Experimental Psychology: General*, 150(3):507, 2021.

[464] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Why is plausibility surprisingly problematic as an xai criterion?, 2024.

Curriculum Vitæ

Gaole HE

26-06-1996 Born in Chongqing, China.

Education

2021–2025 Doctor of Philosophy (PhD), Computer Science
Delft University of Technology, the Netherlands

2018–2021 Master of Engineering (M.Eng.), Computer Application Technology
Renmin University of China, China

2014–2018 Bachelor of Engineering (B.Eng.), Computer Science
Bachelor of Science (B.Sc.), Applied Mathematics (minor)
Renmin University of China, China

Awards & Recognition

2023 Selected as a Heidelberg Laureate Forum Young Researcher
2023 Gary Marsden Travel Award at CHI'23
2022 WWW Best Paper Nomination
2022 Best Paper Award at the AAAI HCOMP conference
2021 Toloka Best Poster/Demo Award at the AAAI HCOMP conference

List of Publications

1. **Gaole He**, , Gianluca Demartini, Ujwal Gadiraju. *Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant.* CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan. https://doi.org/10.1145/3706598.3713218.

2. **Gaole He**, Nilay Aishwarya, Ujwal Gadiraju. *Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant.* 30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy. https://doi.org/10.1145/3708359.3712133

3. **Gaole He**, Patrick Hemmer, Michael Vössing, Max Schemmer, Ujwal Gadiraju. *Fine-Grained Appropriate Reliance: Human-AI Collaboration with a Multi-Step Transparent Decision Workflow for Complex Task Decomposition.* Revised after reviews from CSCW'25 and CHI'25, now under review.

4. **Gaole He**, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. *Opening the Analogical Portal to Explainability: Can Analogies Help Laypeople in AI-assisted Decision Making?* Journal of Artificial Intelligence Research 81 (2024): 117-162. https://doi.org/10.1613/jair.1.15118

5. **Gaole He**, Abri Bharos, and Ujwal Gadiraju. *To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems.* In Proceedings of the 35th ACM Conference on Hypertext and Social Media, pp. 98-105. 2024. https://doi.org/10.1145/3648188.3675130

6. **Gaole He***, Stefan Buijsman*, Ujwal Gadiraju. *How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System.* Proceedings of the ACM on Human-Computer Interaction 7, no. CSCW2 (2023): 1-29. https://doi.org/10.1145/ 3610067

7. **Gaole He**, Lucie Kuiper, and Ujwal Gadiraju. *Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems.* In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1-18. 2023. https://doi.org/10.1145/3544548.3581025

8. **Gaole He**, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. *It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge.* In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, vol. 10, pp. 89-101. 2022. https://doi.org/10.1609/hcomp.v10i1.21 990; **🏆 Best Paper Award**

9. **Gaole He**, and Ujwal Gadiraju. *Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making.* In Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22). 2022.

10. Sara Salimzadeh, **Gaole He**, Ujwal Gadiraju. *Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making.* In Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1-17. 2024. https://doi.org/ 10.1145/3613904.3641905

11. Sara Salimzadeh, **Gaole He**, Ujwal Gadiraju. *A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making.* In Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, pp. 215-227. 2023. https://doi.org/10.1145/3565472.3592959

12. Agathe Balayn*, **Gaole He***, Andrea Hu*, Jie Yang, and Ujwal Gadiraju. *Ready player one! eliciting diverse knowledge using a configurable game.* In Proceedings of the ACM Web Conference 2022, pp. 1709-1719. 2022. https://dl.acm.org/doi/abs/10.1145/3485447.3512241; **Best Paper Honorable Nomination**

13. Agathe Balayn*, **Gaole He***, Andrea Hu*, Jie Yang, and Ujwal Gadiraju. *Finditout: A multiplayer gwap for collecting plural knowledge.* In Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing, Online, pp. 14-18. 2021. 🏆 **Best Demo Award**

📄 Included in this thesis.
🏆 Won a best paper, tool demonstration, or proposal award.

SIKS Dissertation Series

2016
01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines
02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow
03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support
04 Laurens Rietveld (VUA), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UvA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VUA), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VUA), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UvA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VUA), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UvA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VUA), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VUA), Refining Statistical Data on the Web
19 Julia Efremova (TU/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UvA), Context & Semantics in News & Web Search
21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VUA), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UvA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakarathne (VUA), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (TiU), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UvA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TU/e), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UvA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design
39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect
40 Christian Detweiler (TUD), Accounting for Values in Design
41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance
42 Spyros Martzoukos (UvA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora
43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice
44 Thibault Sellam (UvA), Automatic Assistants for Database Exploration
45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control
46 Jorge Gallego Perez (UT), Robots to Make you Happy
47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks
48 Tanja Buttler (TUD), Collecting Lessons Learned
49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis
50 Yan Wang (TiU), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017
01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime
02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation
03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines
04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store
05 Mahdieh Shadi (UvA), Collaboration Behavior
06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search
07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly
08 Rob Konijn (VUA), Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery
09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text
10 Robby van Delden (UT), (Steering) Interactive Play Behavior
11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment
12 Sander Leemans (TU/e), Robust Process Mining with Guarantees
13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology
14 Shoshannah Tekofsky (TiU), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior
15 Peter Berck (RUN), Memory-Based Text Correction
16 Aleksandr Chuklin (UvA), Understanding and Modeling Users of Modern Search Engines
17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution
18 Ridho Reinanda (UvA), Entity Associations for Search
19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval
20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility
21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)
22 Sara Magliacane (VUA), Logics for causal inference under uncertainty
23 David Graus (UvA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VUA), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VUA), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (TiU), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (TiU), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics
35 Martine de Vos (VUA), Interpreting natural science spreadsheets
36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
37 Alejandro Montes Garcia (TU/e), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
38 Alex Kayal (TUD), Normative Social Applications
39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement
46 Jan Schneider (OU), Sensor-based Learning Support
47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
48 Angel Suarez (OU), Collaborative inquiry-based learning

2018 01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TU/e), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UvA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UvA), Enabling Framework for Service-oriented Collaborative Networks
12 Xixi Lu (TU/e), Using behavioral context in process mining
13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future
14 Bart Joosten (TiU), Detecting Social Signals with Spatiotemporal Gabor Filters
15 Naser Davarzani (UM), Biomarker discovery in heart failure
16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children
17 Jianpeng Zhang (TU/e), On Graph Sample Clustering
18 Henriette Nakad (UL), De Notaris en Private Rechtspraak
19 Minh Duc Pham (VUA), Emergent relational schemas for RDF
20 Manxia Liu (RUN), Time and Bayesian Networks
21 Aad Slootmaker (OU), EMERGO: a generic platform for authoring and playing scenario-based serious games
22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks
23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis
24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots
25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections
26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology
27 Maikel Leemans (TU/e), Hierarchical Process Mining for Scalable Software Analysis
28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel
29 Yu Gu (TiU), Emotion Recognition from Mandarin Speech
30 Wouter Beek (VUA), The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019 01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification
02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty
03 Eduardo Gonzalez Lopez de Murillas (TU/e), Process Mining on Databases: Extracting Event Data from Real Life Data Sources
04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data
05 Sebastiaan van Zelst (TU/e), Process Mining with Streaming Data
06 Chris Dijkshoorn (VUA), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets
07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms
08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes
09 Fahimeh Alizadeh Moghaddam (UvA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction
11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs
12 Jacqueline Heinerman (VUA), Better Together
13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation
14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses
15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments
16 Guangming Li (TU/e), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models
17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts
18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication
19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents
20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence
21 Cong Liu (TU/e), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection
22 Martin van den Berg (VUA),Improving IT Decisions with Enterprise Architecture
23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification
24 Anca Dumitrache (VUA), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing
25 Emiel van Miltenburg (VUA), Pragmatic factors in (automatic) image description
26 Prince Singh (UT), An Integration Platform for Synchromodal Transport
27 Alessandra Antonaci (OU), The Gamification Design Process applied to (Massive) Open Online Courses
28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations
29 Daniel Formolo (VUA), Using virtual agents for simulation and training of social skills in safety-critical circumstances
30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems
31 Milan Jelisavcic (VUA), Alive and Kicking: Baby Steps in Robotics
32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games
33 Anil Yaman (TU/e), Evolution of Biologically Inspired Learning in Artificial Neural Networks
34 Negar Ahmadi (TU/e), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES
35 Lisa Facey-Shaw (OU), Gamification with digital badges in learning programming

	36	Kevin Ackermans (OU), Designing Video-Enhanced Rubrics to Master Complex Skills
	37	Jian Fang (TUD), Database Acceleration on FPGAs
	38	Akos Kadar (OU), Learning visually grounded and multilingual representations
2020	01	Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour
	02	Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models
	03	Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding
	04	Maarten van Gompel (RUN), Context as Linguistic Bridges
	05	Yulong Pei (TU/e), On local and global structure mining
	06	Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support
	07	Wim van der Vegt (OU), Towards a software architecture for reusable game components
	08	Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search
	09	Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research
	10	Alifah Syamsiyah (TU/e), In-database Preprocessing for Process Mining
	11	Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models
	12	Ward van Breda (VUA), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment
	13	Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming
	14	Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases
	15	Konstantinos Georgiadis (OU), Smart CAT: Machine Learning for Configurable Assessments in Serious Games
	16	Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling
	17	Daniele Di Mitri (OU), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences
	18	Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems
	19	Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems
	20	Albert Hankel (VUA), Embedding Green ICT Maturity in Organisations
	21	Karine da Silva Miras de Araujo (VUA), Where is the robot?: Life as it could be
	22	Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar
	23	Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging
	24	Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots
	25	Xin Du (TU/e), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization
27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context
28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality
29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference
30 Bob Zadok Blok (UL), Creatief, Creatiever, Creatiefst
31 Gongjin Lan (VUA), Learning better – From Baby to Better
32 Jason Rhuggenaath (TU/e), Revenue management in online markets: pricing and online advertising
33 Rick Gilsing (TU/e), Supporting service-dominant business model evaluation in the context of business model innovation
34 Anna Bon (UM), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development
35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021 01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space
02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models
03 Seyyed Hadi Hashemi (UvA), Modeling Users Interacting with Smart Devices
04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning
05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems
06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot
07 Armel Lefebvre (UU), Research data management for open science
08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking
09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play
10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning
11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision
12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs
13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning
14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support
15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm
16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues
17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks
18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VUA), Architectural Technical Debt: Identification and Management
20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems
21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes
22 Sihang Qiu (TUD), Conversational Crowdsourcing
23 Hugo Manuel Proença (UL), Robust rules for prediction and description
24 Kaijie Zhu (TU/e), On Efficient Temporal Subgraph Query Processing
25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications
26 Benno Kruit (CWI/VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables
27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You
28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022 01 Judith van Stegeren (UT), Flavor text generation for role-playing video games
02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey
03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare
04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework
05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization
06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding
07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics
08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design
09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach
10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines
11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers' engagement with robots and tasks during second-language tutoring
12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases
13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge
14 Michiel Overeem (UU), Evolution of Low-Code Platforms
15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining
16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification
18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation
19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation
20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media
21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments
22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations
23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents
24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values
25 Anna L.D. Latour (UL), Optimal decision-making under constraints and uncertainty
26 Anne Dirkson (UL), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences
27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences
28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems
29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality
30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays
31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management
32 Cezara Pastrav (UU), Social simulation for socio-ecological systems
33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management
34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change
35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

---

2023 01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions
02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts
03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations
04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval
05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment
07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning
08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning
09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques
10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing
11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications
12 Tessel Bogaard (VUA), Using Metadata to Understand Search Behavior in Digital Libraries
13 Injy Sarhan (UU), Open Information Extraction for Knowledge Representation
14 Selma Čaušević (TUD), Energy resilience through self-organization
15 Alvaro Henrique Chaim Correia (TU/e), Insights on Learning Tractable Probabilistic Graphical Models
16 Peter Blomsma (TiU), Building Embodied Conversational Agents: Observations on human nonverbal behaviour as a resource for the development of artificial characters
17 Meike Nauta (UT), Explainable AI and Interpretable Computer Vision – From Oversight to Insight
18 Gustavo Penha (TUD), Designing and Diagnosing Models for Conversational Search and Recommendation
19 George Aalbers (TiU), Digital Traces of the Mind: Using Smartphones to Capture Signals of Well-Being in Individuals
20 Arkadiy Dushatskiy (TUD), Expensive Optimization with Model-Based Evolutionary Algorithms applied to Medical Image Segmentation using Deep Learning
21 Gerrit Jan de Bruin (UL), Network Analysis Methods for Smart Inspection in the Transport Domain
22 Alireza Shojaifar (UU), Volitional Cybersecurity
23 Theo Theunissen (UU), Documentation in Continuous Software Development
24 Agathe Balayn (TUD), Practices Towards Hazardous Failure Diagnosis in Machine Learning
25 Jurian Baas (UU), Entity Resolution on Historical Knowledge Graphs
26 Loek Tonnaer (TU/e), Linearly Symmetry-Based Disentangled Representations and their Out-of-Distribution Behaviour
27 Ghada Sokar (TU/e), Learning Continually Under Changing Data Distributions
28 Floris den Hengst (VUA), Learning to Behave: Reinforcement Learning in Human Contexts
29 Tim Draws (TUD), Understanding Viewpoint Biases in Web Search Results

2024 01 Daphne Miedema (TU/e), On Learning SQL: Disentangling concepts in data systems education
02 Emile van Krieken (VUA), Optimisation in Neurosymbolic Learning Systems

03 Feri Wijayanto (RUN), Automated Model Selection for Rasch and Mediation Analysis
04 Mike Huisman (UL), Understanding Deep Meta-Learning
05 Yiyong Gou (UM), Aerial Robotic Operations: Multi-environment Cooperative Inspection & Construction Crack Autonomous Repair
06 Azqa Nadeem (TUD), Understanding Adversary Behavior via XAI: Leveraging Sequence Clustering to Extract Threat Intelligence
07 Parisa Shayan (TiU), Modeling User Behavior in Learning Management Systems
08 Xin Zhou (UvA), From Empowering to Motivating: Enhancing Policy Enforcement through Process Design and Incentive Implementation
09 Giso Dal (UT), Probabilistic Inference Using Partitioned Bayesian Networks
10 Cristina-Iulia Bucur (VUA), Linkflows: Towards Genuine Semantic Publishing in Science
11 withdrawn
12 Peide Zhu (TUD), Towards Robust Automatic Question Generation For Learning
13 Enrico Liscio (TUD), Context-Specific Value Inference via Hybrid Intelligence
14 Larissa Capobianco Shimomura (TU/e), On Graph Generating Dependencies and their Applications in Data Profiling
15 Ting Liu (VUA), A Gut Feeling: Biomedical Knowledge Graphs for Interrelating the Gut Microbiome and Mental Health
16 Arthur Barbosa Câmara (TUD), Designing Search-as-Learning Systems
17 Razieh Alidoosti (VUA), Ethics-aware Software Architecture Design
18 Laurens Stoop (UU), Data Driven Understanding of Energy-Meteorological Variability and its Impact on Energy System Operations
19 Azadeh Mozafari Mehr (TU/e), Multi-perspective Conformance Checking: Identifying and Understanding Patterns of Anomalous Behavior
20 Ritsart Anne Plantenga (UL), Omgang met Regels
21 Federica Vinella (UU), Crowdsourcing User-Centered Teams
22 Zeynep Ozturk Yurt (TU/e), Beyond Routine: Extending BPM for Knowledge-Intensive Processes with Controllable Dynamic Contexts
23 Jie Luo (VUA), Lamarck's Revenge: Inheritance of Learned Traits Improves Robot Evolution
24 Nirmal Roy (TUD), Exploring the effects of interactive interfaces on user search behaviour
25 Alisa Rieger (TUD), Striving for Responsible Opinion Formation in Web Search on Debated Topics
26 Tim Gubner (CWI), Adaptively Generating Heterogeneous Execution Strategies using the VOILA Framework
27 Lincen Yang (UL), Information-theoretic Partition-based Models for Interpretable Machine Learning
28 Leon Helwerda (UL), Grip on Software: Understanding development progress of Scrum sprints and backlogs

29 David Wilson Romero Guzman (VUA), The Good, the Efficient and the Inductive Biases: Exploring Efficiency in Deep Learning Through the Use of Inductive Biases
30 Vijanti Ramautar (UU), Model-Driven Sustainability Accounting
31 Ziyu Li (TUD), On the Utility of Metadata to Optimize Machine Learning Workflows
32 Vinicius Stein Dani (UU), The Alpha and Omega of Process Mining
33 Siddharth Mehrotra (TUD), Designing for Appropriate Trust in Human-AI interaction
34 Robert Deckers (VUA), From Smallest Software Particle to System Specification - MuDForM: Multi-Domain Formalization Method
35 Sicui Zhang (TU/e), Methods of Detecting Clinical Deviations with Process Mining: a fuzzy set approach
36 Thomas Mulder (TU/e), Optimization of Recursive Queries on Graphs
37 James Graham Nevin (UvA), The Ramifications of Data Handling for Computational Models
38 Christos Koutras (TUD), Tabular Schema Matching for Modern Settings
39 Paola Lara Machado (TU/e), The Nexus between Business Models and Operating Models: From Conceptual Understanding to Actionable Guidance
40 Montijn van de Ven (TU/e), Guiding the Definition of Key Performance Indicators for Business Models
41 Georgios Siachamis (TUD), Adaptivity for Streaming Dataflow Engines
42 Emmeke Veltmeijer (VUA), Small Groups, Big Insights: Understanding the Crowd through Expressive Subgroup Analysis
43 Cedric Waterschoot (KNAW Meertens Instituut), The Constructive Conundrum: Computational Approaches to Facilitate Constructive Commenting on Online News Platforms
44 Marcel Schmitz (OU), Towards learning analytics-supported learning design
45 Sara Salimzadeh (TUDelft), Living in the Age of AI: Understanding Contextual Factors that Shape Human-AI Decision-Making
46 Georgios Stathis (Leiden University), Preventing Disputes: Preventive Logic, Law & Technology
47 Daniel Daza (VUA), Exploiting Subgraphs and Attributes for Representation Learning on Knowledge Graphs
48 Ioannis Petros Samiotis (TUD), Crowd-Assisted Annotation of Classical Music Compositions

2025 01 Max van Haastrecht (UL), Transdisciplinary Perspectives on Validity: Bridging the Gap Between Design and Implementation for Technology-Enhanced Learning Systems
02 Jurgen van den Hoogen (JADS), Time Series Analysis Using Convolutional Neural Networks
03 Andra-Denis Ionescu (TUD), Feature Discovery for Data-Centric AI
04 Rianne Schouten (TU/e), Exceptional Model Mining for Hierarchical Data
05 Nele Albers (TUD), Psychology-Informed Reinforcement Learning for Situated Virtual Coaching in Smoking Cessation

06 Daniël Vos (TUD), Decision Tree Learning: Algorithms for Robust Prediction and Policy Optimization
07 Ricky Maulana Fajri (TU/e), Towards Safer Active Learning: Dealing with Unwanted Biases, Graph-Structured Data, Adversary, and Data Imbalance
08 Stefan Bloemheuvel (TiU), Spatio-Temporal Analysis Through Graphs: Predictive Modeling and Graph Construction
09 Fadime Kaya (VUA), Decentralized Governance Design - A Model-Based Approach
10 Zhao Yang (UL), Enhancing Autonomy and Efficiency in Goal-Conditioned Reinforcement Learning
11 Shahin Sharifi Noorian (TUD), From Recognition to Understanding: Enriching Visual Models Through Multi-Modal Semantic Integration
12 Lijun Lyu (TUD), Interpretability in Neural Information Retrieval
13 Fuda van Diggelen (VUA), Robots Need Some Education: on the complexity of learning in evolutionary robotics
14 Gennaro Gala (TU/e), Probabilistic Generative Modeling with Latent Variable Hierarchies
15 Michiel van der Meer (UL), Opinion Diversity through Hybrid Intelligence
16 Monika Grewal (TU Delft), Deep Learning for Landmark Detection, Segmentation, and Multi-Objective Deformable Registration in Medical Imaging
17 Matteo De Carlo (VUA), Real Robot Reproduction: Towards Evolving Robotic Ecosystems
18 Anouk Neerincx (UU), Robots That Care: How Social Robots Can Boost Children's Mental Wellbeing
19 Fang Hou (UU), Trust in Software Ecosystems
20 Alexander Melchior (UU), Modelling for Policy is More Than Policy Modelling (The Useful Application of Agent-Based Modelling in Complex Policy Processes)
21 Mandani Ntekouli (UM), Bridging Individual and Group Perspectives in Psychopathology: Computational Modeling Approaches using Ecological Momentary Assessment Data
22 Hilde Weerts (TU/e), Decoding Algorithmic Fairness: Towards Interdisciplinary Understanding of Fairness and Discrimination in Algorithmic Decision-Making
23 Roderick van der Weerdt (VUA), IoT Measurement Knowledge Graphs: Constructing, Working and Learning with IoT Measurement Data as a Knowledge Graph
24 Zhong Li (UL), Trustworthy Anomaly Detection for Smart Manufacturing
25 Kyana van Eijndhoven (TiU), A Breakdown of Breakdowns: Multi-Level Team Coordination Dynamics under Stressful Conditions
26 Tom Pepels (UM), Monte-Carlo Tree Search is Work in Progress
27 Danil Provodin (JADS, TU/e), Sequential Decision Making Under Complex Feedback
28 Jinke He (TU Delft), Exploring Learned Abstract Models for Efficient Planning and Learning

29 Erik van Haeringen (VUA), Mixed Feelings: Simulating Emotion Contagion in Groups
30 Myrthe Reuver (VUA), A Puzzle of Perspectives: Interdisciplinary Language Technology for Responsible News Recommendation
31 Gebrekirstos Gebreselassie Gebremeskel (RUN), Spotlight on Recommender Systems: Contributions to Selected Components in the Recommendation Pipeline
32 Ryan Brate (UU), Words Matter: A Computational Toolkit for Charged Terms
33 Merle Reimann (VUA), Speaking the Same Language: Spoken Capability Communication in Human-Agent and Human-Robot Interaction
34 Eduard C. Groen (UU), Crowd-Based Requirements Engineering
35 Urja Khurana (VUA), From Concept To Impact: Toward More Robust Language Model Deployment
36 Anna Maria Wegmann (UU), Say the Same but Differently: Computational Approaches to Stylistic Variation and Paraphrasing
37 Chris Kamphuis (RUN), Exploring Relations and Graphs for Information Retrieval
38 Valentina Maccatrozzo (VUA), Break the Bubble: Semantic Patterns for Serendipity
39 Dimitrios Alivanistos (VUA), Knowledge Graphs & Transformers for Hypothesis Generation: Accelerating
40 Stefan Grafberger (UvA), Declarative Machine Learning Pipeline Management via Logical Query Plans
41 Mozhgan Vazifehdoostirani (TU/e), Leveraging Process Flexibility to Improve Process Outcome - From Descriptive Analytics to Actionable Insights
42 Margherita Martorana (VUA), Semantic Interpretation of Dataless Tables: a metadata-driven approach for findable, accessible, interoperable and reusable restricted access data
43 Krist Shingjergji (OU), Sense the Classroom - Using AI to Detect and Respond to Learning-Centered Affective States in Online Education
44 Robbert Reijnen (TU/e), Dynamic Algorithm Configuration for Machine Scheduling Using Deep Reinforcement Learning
45 Anjana Mohandas Sheeladevi (VUA), Occupant-Centric Energy Management: Balancing Privacy, Well-being and Sustainability in Smart Buildings
46 Ya Song (TU/e), Graph Neural Networks for Modeling Temporal and Spatial Dimensions in Industrial Decision-making
47 Tom Kouwenhoven (UL), Collaborative Meaning-Making. The Emergence of Novel Languages in Humans, Machines, and Human-Machine Interactions
48 Evy van Weelden (TiU), Integrating Virtual Reality and Neurophysiology in Flight Training
49 Selene Báez Santamaría (VUA), Knowledge-centered conversational agents with a drive to learn
50 Lea Krause (VUA), Contextualising Conversational AI
51 Jiaxu Zhao (TU/e), Understanding and Mitigating Unwanted Biases in Generative Language Models

52 Qiao Xiao (TU/e), Model, Data and Communication Sparsity for Efficient Training of Neural Networks
53 Gaole He (TUD), Towards Effective Human-AI Collaboration: Promoting Appropriate Reliance on AI Systems.