Taylor & Francis
Taylor & Francis Group

International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Don’t Be Fooled: The Misinformation Effect of Explanations in Human–AI Collaboration

Philipp Spitzer, Joshua Holstein, Katelyn Morrison, Kenneth Holstein, Gerhard Satzger & Niklas Kühl

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2025.2574511

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS   Check for updates

Don't Be Fooled: The Misinformation Effect of Explanations in Human–AI Collaboration

Philipp Spitzer[a], Joshua Holstein[a], Katelyn Morrison[b], Kenneth Holstein[b], Gerhard Satzger[a] and Niklas Kühl[c]

[a]Institute for Information Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany; [b]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA; [c]Information Systems and Human-centric Artificial Intelligence, University of Bayreuth, Germany

ABSTRACT

Across various applications, humans increasingly use black-box artificial intelligence (AI) systems without insight into these systems' reasoning. To counter this opacity, explainable AI (XAI) methods promise enhanced transparency and interpretability. While recent studies have explored how XAI affects human–AI collaboration, few have examined the potential pitfalls caused by incorrect explanations. The implications for humans can be far-reaching but have not been explored extensively. To investigate this, we conducted a study ($n = 160$) on AI-assisted decision-making in which humans were supported by XAI. Our findings reveal a *misinformation effect* when incorrect explanations accompany correct AI advice with implications post-collaboration. This effect causes humans to infer flawed reasoning strategies, hindering task execution and demonstrating impaired procedural knowledge. Additionally, incorrect explanations compromise human–AI team performance during collaboration. With our work, we contribute to HCI by providing empirical evidence for the negative consequences of incorrect explanations on humans post-collaboration and outlining guidelines for designers of AI.

CCS CONCEPTS

Human-centered computing → Empirical studies in HCI; Computing methodologies → Artificial intelligence; Computer vision tasks

KEYWORDS

Human–AI collaboration; explainable AI; XAI for computer vision

1. Introduction

Imagine you are studying for an art history exam and must know how to distinguish two architectural styles. You seek advice from an online artificial intelligence (AI) assistant for explanations to differentiate the two styles. Despite being plausible and providing you with correct architectural styles, the AI's explanation is incorrect. Yet, you learn from these incorrect explanations, and as a consequence, your understanding and capability to distinguish the architectural styles are impaired! You fail your exam.

Recent technological advancements have significantly bolstered both the capabilities and the adoption of AI (Bommasani et al., 2021; Dwivedi et al., 2023). For instance, large language models (LLMs) can assist in high-stakes classification tasks, such as categorizing legal cases based on case descriptions, by providing initial advice along with an explanation, even when humans have no prior expertise in legal analysis (Ma et al., 2024). In the medical field, AI has been used to accurately classify retinal disorders, leading to early diagnosis and providing critical support to ophthalmologists (Mukhtorov et al., 2023). However, as these AI technologies become more sophisticated, especially with the use of generative AI, the complexity and opacity of supported decision-making processes increase. This presents new challenges in ensuring that these systems remain transparent and interpretable for humans (Rudin et al., 2022). This increasing complexity necessitates a deeper understanding of the underlying factors that impact human–AI interaction, especially in scenarios where human oversight is critical (Amershi et al., 2019; Cabitza, Campagner, Natali, et al., 2023; Sterz et al., 2024). Regulatory frameworks, such as

CONTACT Philipp Spitzer philipp.spitzer@kit.edu Institute for Information Systems, Karlsruhe Institute of Technology, Karlsruhe, Germany

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2025.2574511.

the EU AI Act (European Commission, 2021), mandate human oversight to ensure AI systems operate ethically, legally, and safely. This underscores the importance of human–computer interaction (HCI) research to develop methods and tools that facilitate effective collaboration between humans and AI (Shneiderman, 2020). Ensuring that AI is not only accurate but also explainable is vital to foster appropriate reliance and complementary team performance (Hemmer et al., 2023; Schemmer, Kuehl, et al., 2023). The need for explainability in AI has been widely acknowledged, leading to substantial advancements in both research and practice (Barredo Arrieta et al., 2020; Silva et al., 2023).

Despite these advancements, an important potential pitfall in human–AI collaboration scenarios, incorrect explanations (Kayser et al., 2024; Lakkaraju & Bastani, 2020; Morrison et al., 2024) for accurate AI advice, remains underexplored in the XAI literature. This phenomenon has gained particular relevance with the widespread deployment of LLMs, which can generate self-explanations for their classifications that can diverge from their actual internal decision processes (Randl et al., 2024). Compared to intrinsic or post-hoc explanation methods, this systematic disconnect between generated self-explanations and underlying computational mechanisms creates conditions where humans may receive correct AI advice accompanied by factually incorrect or misleading explanations (Madsen et al., 2024). With the rapid integration of LLMs into decision-making tasks and educational contexts (Harvey et al., 2025; Hemmer et al., 2023), alongside emerging regulatory requirements (European Commission, 2021), empirical investigation of how this explanation-accuracy mismatch affects human cognition and task performance has become essential. Such incorrect explanations may compromise human performance in post-collaboration phases by degrading both *procedural knowledge*—the ability to perform specific tasks independently—and *reasoning capabilities*—the capacity to make informed decisions based on acquired understanding. Given the growing role of generative AI in various settings, for instance, in a learning context, humans often turn to AI to make sense of unfamiliar topics and try to learn something new. Before consulting experts or verified sources, humans rely on AI-generated explanations, making it essential to understand how these affect humans' understanding and learning outcomes. Exploring the effect of incorrect explanations on procedural knowledge and reasoning is crucial, as regulations (European Commission, 2021) or performance expectations (Hemmer et al., 2023) might require humans to complement AI capabilities with their domain knowledge and, therefore, maintain the ability to perform tasks on a superior performance level. The implications are especially critical for high-stakes domains such as healthcare, finance, and legal settings, where compromised decision-making can result in severe consequences (Rudin, 2019). Moreover, preserving human procedural knowledge becomes critical in scenarios where AI support may be temporarily available, such as in AI-assisted educational settings (Spitzer et al., 2023; Spitzer, Kühl, et al., 2024), but learners ultimately must demonstrate task competency on their own.

Understanding the impact of incorrect explanations in AI-assisted decision-making, even when the AI advice is correct, is essential for designing more effective collaborations between humans and AI. By identifying how and why incorrect explanations impact humans not only *during*, but in particular *post-collaboration* with AI, we can develop effective strategies to mitigate these effects (Chen et al., 2022; Ehsan et al., 2019; Lakkaraju & Bastani, 2020; Ribeiro et al., 2016). Thus, we ask the following research questions (RQs):

RQ1: How do incorrect explanations for correct AI advice impair humans' procedural knowledge in AI-assisted decision-making?

RQ2: How do incorrect explanations for correct AI advice impair humans' reasoning in AI-assisted decision-making?

RQ3: How do incorrect explanations for correct AI advice affect the human–AI team performance?

To address these questions, we first synthesize how to measure the impact of incorrect explanations in AI-assisted decision-making based on previous research. We pre-registered our hypotheses on AsPredicted.org. Through an online study with 160 participants, we examine how incorrect explanations influence humans' procedural knowledge (RQ1) and reasoning (RQ 2) in a task to classify buildings' architectural styles. We chose this architectural classification task because AI, and especially LLMs, are increasingly being used nowadays not only to generate content but also to support human reasoning and decision-making (Hadi et al., 2024). Thus, they provide advice to humans in scenarios where no prior knowledge is needed (e.g., patients seeking medical advice) to help them interpret and draw conclusions from complex data. Such settings remain crucial as they provide a controlled environment (clear ground truth and measurable outcomes) to study how humans are impacted by incorrect explanations. Furthermore, architectural classification is a domain that many people may encounter in everyday life, requiring a combination of visual perception, pattern recognition, and creative reasoning. This makes it a representative and engaging task for studying human–AI collaboration, particularly in scenarios where generative AI can provide explanatory support. In this task, we provide participants with different types of AI support: no support at all, AI advice without explanations, AI advice with correct explanations, and AI advice with incorrect explanations. We measure the effects on their task performance during and after collaboration to derive the impact on their procedural knowledge and qualitatively analyze their understanding of the task to derive their reasoning ability. Additionally, we analyze the effect on the human–AI team performance (RQ3).

The findings of our study reveal a *misinformation effect* in AI-assisted decision-making: incorrect explanations significantly impair humans, resulting in a notable decline in their procedural knowledge once they have to perform the task autonomously. We also find that humans' reasoning is impaired when they receive incorrect explanations. In fact, we also observe a negative effect during the collaboration in our study: human–AI teams perform worse when the AI provides incorrect explanations, curtailing the complementary benefits of this collaboration. These findings underscore the potential dangers of incorrect explanations and highlight the importance of developing robust and reliable explanatory support for humans.

In summary, our study makes several contributions to the field: first, it fills a critical gap in the literature by examining the impact of incorrect explanations on humans in AI-assisted decision-making. Second, it identifies and measures the misinformation effect within AI explanations and outlines its negative repercussions on humans—a decline in procedural knowledge and reasoning. Lastly, our research extends existing knowledge on the interplay between AI and human decision-makers, providing insights into the hazards of explanations on the human–AI team. Overall, this article sheds light on the potential pitfalls of incorrect explanations and their implications for humans post-collaboration. With this exploratory work, we hope to contribute to the development of concepts and hypotheses, helping to advance theoretical knowledge in XAI and AI-assisted decision-making, as well as to further advance the design of more effective AI-based decision support systems.

2. Background

With the rapid advancement of AI and its integration into diverse decision-making processes, XAI has emerged as a critical technique for enhancing transparency and assistance to help decision-makers understand AI's reasoning (Alufaisan et al., 2021; Cabitza, Natali, et al., 2024). Especially with applications based on generative AI (like Open AI's ChatGPT or Anthropic's Claude), explanations are being generated in natural language that provide human-understandable support for the AI's response (Singh et al., 2024; Zytek et al., 2024). Previous research on human–AI collaboration has predominantly focused on how correct AI explanations influence human decision-making (e.g., Lai & Tan, 2019; Schemmer et al., 2022; Subramanian et al., 2024; Yeung et al., 2020; Zhang et al., 2020). For example, Hemmer et al. (2021) examine the factors that impact human–AI team performance, suggesting that XAI can foster complementary collaboration between humans and AI.

Albeit the positive effects, studies also highlight how XAI can impair the collaboration between humans and AI (Binns, 2018; Ehsan et al., 2021; Miller, 2019; Schemmer et al., 2022). van der Waa et al. (2021) demonstrate that example-based explanations have the power to increase overreliance on AI advice. Similarly, Schoeffer et al. (2024) investigate how explanations impact fairness perceptions and humans' tendency to adhere to or overwrite AI recommendations, showing that feature-based explanations do not improve distributive fairness. However, the understanding of the negative consequences of XAI is still limited (Mohseni et al., 2021), especially empirical evidence for the negative impact of incorrect explanations is missing. In Table 1, we sort recent works in AI-assisted decision-making according to the correctness of AI advice and explanations. The table shows the under-

Table 1. HCI literature investigating impacts of the correctness of AI advice and explanations on human–AI collaboration.

	Correct AI explanations	Incorrect AI explanations
Correct AI advice	Adhikari et al. (2019); Hase and Bansal (2020); Hemmer et al. (2021); Lai and Tan (2019); Ribeiro et al. (2018); Schemmer et al. (2022); Schoeffer et al. (2024); van der Waa et al. (2021); Yeung et al. (2020); Zhang et al. (2020)	Cabitza, Fregosi, et al. (2024); Kayser et al. (2024); Morrison et al. (2024)
Incorrect AI advice	Alufaisan et al. (2021); Buçinca et al. (2021); Cabitza, Campagner, Ronzio, et al. (2023); Cau et al. (2023); Chen et al. (2023); Ehrlich et al. (2011); S. S. Y. Kim et al. (2023); Kocielnik et al. (2019); Sadeghi et al. (2024); Schmitt et al. (2024); Vicente and Matute (2023)	Kayser et al. (2024); Lakkaraju and Bastani (2020); Morrison et al. (2024); Papenmeier et al. (2019)

explored topic of incorrect explanations in HCI. In the remaining section, we review recent literature highlighting the risks and limitations of collaborative settings between humans and AI, motivating the need to explore further how incorrect explanations impair humans. While we introduced works studying correct explanations at the beginning of this section, we focus the remainder on prior work that shows implications of the incorrectness of advice and explanations.

2.1. Incorrect AI advice in human–AI collaboration

Research in HCI has explored the effects of incorrect AI advice on human–AI collaboration (e.g., Bansal et al., 2019; Kocielnik et al., 2019; Schmitt et al., 2024; Vicente & Matute, 2023; Yin et al., 2019). Kocielnik et al. (2019) investigate how such incorrect AI advice influences human satisfaction and acceptance. In their study, they demonstrate how interventions like accuracy indicators or performance control can maintain humans' perception and trust even when the AI-based scheduling assistant provides wrong advice.

Next to scheduling assistants, other recent studies have investigated how programmers collaborate with Copilot, an AI programming assistant that is not always accurate (e.g., Barke et al., 2023; Dakhel et al., 2023; Vasconcelos, Bansal, et al., 2023). For instance, Vasconcelos, Bansal, et al. (2023) focus on an AI supporting code completion. The work demonstrates that AI can assist programmers in arriving at solutions more efficiently. Additionally, Dakhel et al. (2023) conclude that such an AI assistance can support expert humans, while non-experts should exercise caution when using it due to the potential for errors. Other studies further examine how different levels of expertise among humans influence their interaction with AI-generated code (Barke et al., 2023). Similarly, Barke et al. (2023) find that expert humans can effectively navigate and correct AI-generated errors, while novices often struggle, leading to decreased efficiency and increased frustration. While previous research shows the impact of incorrect AI advice on humans with different levels of prior domain knowledge, Vicente and Matute (2023) conduct a study on medical diagnosis to investigate how humans are influenced during collaboration with an AI that provides erroneous advice. They show that humans inherit the error patterns of AI, thus impairing their ability to conduct the task themselves. Building on this, previous research also shows how non-expert humans are influenced by AI systems providing inaccurate advice (Schemmer, Bartos, et al., 2023). They find that humans with limited domain expertise were prone to overreliance on the AI's outputs, resulting in decision errors. Similarly, Schmitt et al. (2024) show that free-text explanations can increase non-expert performance, but XAI features (e.g., highlights) can lead to overreliance. Contrarily, Vasconcelos, Jörke, et al. (2023) show that explanations can reduce overreliance by introducing a cost-benefit framework. They outline that explanations to be effective need to diminish the costs of verifying AI advice.

This review of prior research illustrates the potential negative impact of incorrect AI advice on decision outcomes and highlights how people's existing knowledge shapes their ability to interpret and respond to AI output. Building on these findings, the next section explores the role of incorrect explanations and how they influence humans' decision-making when supported by AI.

2.2. Incorrect explanations in human–AI collaboration

Next to the AI advice, the understanding of how incorrect explanations can impact AI-assisted decision-making in the literature is limited. Only a few studies investigate the effect of explanations' incorrectness (Cabitza, Fregosi, et al., 2024; Kayser et al., 2024; Lakkaraju & Bastani, 2020; Morrison et al., 2024; Papenmeier et al., 2019). A recent study in HCI shows that not only incorrect advice but also incorrect explanations have the potential to deceive decision-makers (Morrison et al., 2024). Morrison et al. (2024) explore the negative impacts of incorrect explanations on humans' decision-making behavior. They extend the conceptualization of Schemmer, Kuehl, et al. (2023) by the explanation dimension and explore, in a bird classification study, how the correctness of explanations impacts humans' reliance on AI. They show that incorrect explanations can deceive decision-makers who possess no prior domain knowledge, leading to inappropriate reliance behavior. Cabitza, Fregosi, et al. (2024) explore the effects of explanations in a logic puzzle task. They also show that if advice and the accompanying explanation do not align, humans are misled, ultimately resulting in inappropriate reliance behavior. Papenmeier et al. (2019) conduct another study in AI-assisted decision-making with incorrect explanations. They study how explanations affect humans' trust in identifying offensive tweets. Similarly, Lakkaraju and Bastani (2020) find that incorrect explanations can affect humans' trust in AI by investigating their effects in law and criminal justice use cases. Lastly, Kayser et al. (2024) examine in a real-world use-case in the healthcare environment how the factual correctness of explanations impacts humans. The study finds that explanations' factual correctness influences explanations' usefulness.

These studies collectively underscore the complex dynamics of human–AI interaction, especially how incorrect advice and incorrect explanations affect humans' reliance behavior on AI. However, the HCI field lacks a deeper understanding of the effects of such impaired collaboration scenarios on humans themselves. Especially for scenarios in which not the advice but the explanation for the decision-maker—intended to foster interpretability—is incorrect. Studies like Morrison et al. (2024) and Cabitza, Fregosi, et al. (2024) build a promising starting point to inform HCI researchers and practitioners of the downsides of incorrect explanations. However, we still do not know anything about the impact of these AI shortcomings on humans' ability to perform the tasks autonomously (procedural knowledge) and to conclude about the underlying domain (reasoning) post-collaboration. In this study, we therefore focus on scenarios in which the AI provides correct advice to isolate and evaluate the specific effects of explanation correctness. Prior research has already examined scenarios where AI advice is incorrect but explanations are correct—often leading to overreliance on AI, as shown by Schemmer et al. (2022); however, less is known about cases where the correctness of advice and explanation diverge in other ways, which are especially relevant for understanding how explanation correctness shapes AI-assisted decision-making. This design choice allows us to disentangle the influence of explanations from confounding effects introduced by incorrect AI advice. By holding the advice constant and correct, we can rigorously examine how incorrect explanations impact humans' understanding and performance.

3. Theoretical development

In the evolving field of HCI, understanding the impact of AI explanations on decision-making has become critical. Explanations can serve as a bridge between AI and humans, influencing trust, reliance, and collaboration (Hemmer et al., 2023; Schemmer, Kuehl, et al., 2023; Schoeffer et al., 2024). This work investigates how incorrect explanations, when paired with accurate AI advice, can mislead humans, potentially impairing their procedural knowledge and reasoning capabilities. In Section 2, we present several works that investigate the impact of incorrect AI advice and incorrect explanations in decision-making. Building on prior research (Cabitza, Fregosi, et al., 2024; Morrison et al., 2024; Papenmeier et al., 2019), we derive several hypotheses grounded in related psychology and human behavior research to study the impact on humans' knowledge.

Research in behavioral science distinguishes between declarative knowledge (the "know-what") and procedural knowledge (the "know-how") in decision-making (Herz & Schultz, 1999). While both types of knowledge are interconnected, our study focuses on procedural knowledge as we want to determine the impact of incorrect explanations on humans' downstream task performance. Humans' ability to complete tasks effectively remains paramount, especially where AI support is not perfectly reliable or may not always be available (Laux, 2024). By measuring task performance, we can capture procedural knowledge (McCormick, 1997; Nahdi & Jatisunda, 2020) and the extent to which humans effectively apply their understanding in practice (Clark & Dumas, 2015). Exposure to incorrect explanations can significantly distort both declarative and procedural knowledge. As procedural knowledge is particularly vulnerable to cognitive disruptions (Sweller, 1988), incorrect explanations might increase cognitive load and impair the acquisition of procedural knowledge. Chi and Wylie (2014) provide crucial insights into this phenomenon, demonstrating how cognitive engagement can be systematically undermined by misleading information. Thus, we assume that incorrect explanations for correct AI advice impair humans' procedural knowledge. We hypothesize:

Hypothesis 1: Incorrect explanations for correct AI advice lead to lower procedural knowledge.

While our study primarily centers on procedural knowledge, we also assess participants' reasoning abilities to explore how individuals articulate the cognitive principles underlying their decision-making processes (Johnson & Seifert, 1994). This enables us to examine the extent to which incorrect explanations influence both task performance and conceptual understanding. Prior research has demonstrated the profound impact of misleading information on cognitive processes, revealing significant impairments in memory and comprehension (Loftus et al., 1978; Soon & Goh, 2018). Critically, these cognitive distortions persist even after subsequent correction attempts (Ecker et al., 2011; Kendeou et al., 2013), demonstrating the effect of misinformation on cognition and their potential to impair reasoning (Ecker et al., 2011). In the context of AI-assisted classification tasks, these prior findings suggest that incorrect explanations are likely to disrupt reasoning ability in classification tasks.

Hypothesis 2: Human reasoning is impaired by incorrect explanations for correct AI advice.

Seeber et al. (2020) emphasize the importance of communication between humans and AI for effective collaboration. They argue that misunderstandings or unclear communication can lead to disruptions in team performance, particularly when AI systems provide incorrect or unclear explanations. Furthermore, Dzindolet et al. (2003) investigate how trust in automated systems is impacted by incorrect feedback, leading to reduced team performance. Similarly, Eiband et al. (2018) provide insights into how human–AI collaboration is affected by the transparency of AI systems, suggesting that incorrect advice can hinder effective teamwork. Lastly, Bansal et al. (2021) explore the dynamics of human–AI interaction, showing that incorrect AI advice alongside explanations can create friction in collaboration, leading to poorer outcomes. Incorrect explanations in AI-assisted decision-making can impair the collaboration, leading to inappropriate reliance on AI (Morrison et al., 2024). With prior research showing the relationship between humans' reliance behavior on AI and the human–AI team performance (Schemmer, Kuehl, et al., 2023), we assume that the human–AI team performance drops when humans are provided with incorrect explanations. Thus, we extend prior research by directly assessing the impact of incorrect explanations for correct AI advice and hypothesize:

Hypothesis 3: Incorrect explanations for correct AI advice lead to a lower human–AI team performance.

4. Methodology

In this section, we describe our methodology to assess how incorrect explanations for correct AI advice influence humans *during* and *post* collaboration with an AI. We set up an online study and investigated how participants performed in a visual classification task on an architectural dataset. We outline the task domain, the study design, the recruitment of participants, the development of the AI, and finally, the metrics that we use to assess our RQs. Before we ran the study, we pre-registered the study on AsPredicted.org to report our hypotheses, our treatments, our planned analyses, and our exclusion strategy. An anonymized copy of the pre-registration is provided in the supplemental materials.

4.1. Task domain

To analyze the impact of explanations, we chose a task where most people have little experience (i.e., have no expert-level knowledge and typically cannot perform well themselves initially): the classification of the architectural style of buildings. This task represents various real-world scenarios in which AI is utilized: seeking advice and explanations for a task that is either unknown or humans might not possess enough knowledge to solve the task confidently. For example, humans increasingly seek AI assistance to understand unfamiliar topics such as historical concepts, financial terminology, or bureaucratic forms. Thus, the study of how humans integrate incorrect explanations into their decision-making likely extends to more complex generative AI interactions. We use the established dataset of Abdul et al. (2020) and Xu et al. (2014) containing images of buildings across 25 different architectural styles. In close discussion with architecture researchers of the local university, we chose three architectural styles that share similar features and are not easily distinguishable: Art Nouveau, Art Deco, and Georgian Architecture. For each architectural style, we selected 30 images that clearly represent the features of each architectural style and, thus, are appropriate instances for our study. We further made sure that the buildings were centered in each image and cropped all irrelevant information in the images, like other buildings.

4.2. Study design

Our research questions target the understanding of the impacts of incorrect explanations in AI-assisted decision-making on human procedural knowledge and reasoning and the resulting human–AI team performance. To address them, we employed a study combining between- and within-subjects design: between subjects, we analyze the impact of different AI support. Within subjects, we observe this impact in different stages of decision-making: before, during, and after collaboration with the AI. The study was approved by the university's institutional review board.

The online study was divided into five different parts (see Figure 1): in part (1), the participants had to give their consent to participate and were introduced to the study and its procedure. They also received context information about the three different architectural styles with a description of their main characteristics (see Supplementary Table 3 in Section A). Additionally, we included two attention checks, one of them in part (1): we asked participants what their task would be throughout the study, and they could select from three options. In part (2), a pretest assessed participants' task performance in classifying architectural styles of buildings: they were each shown six images randomly drawn from a

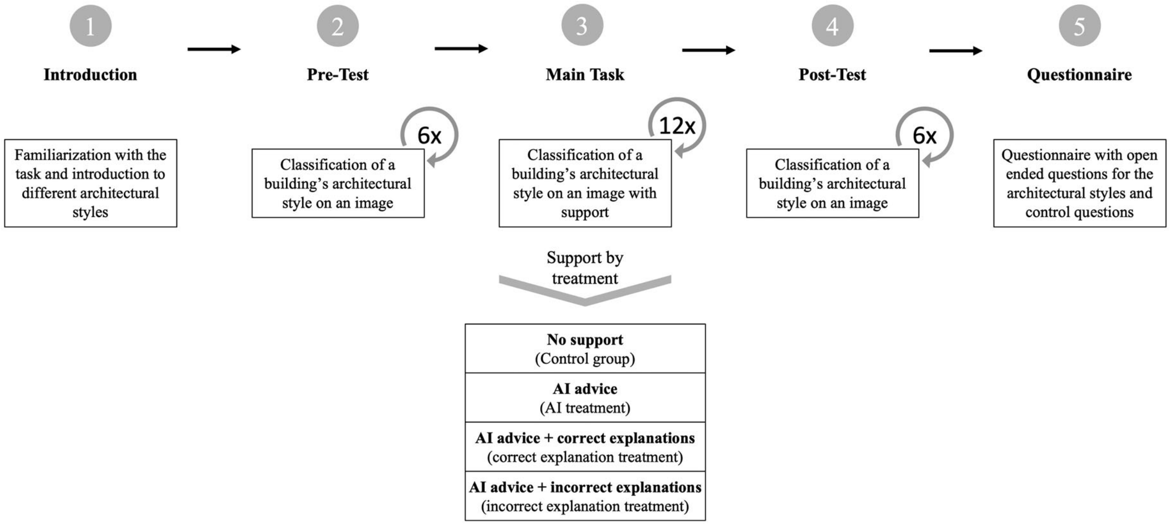
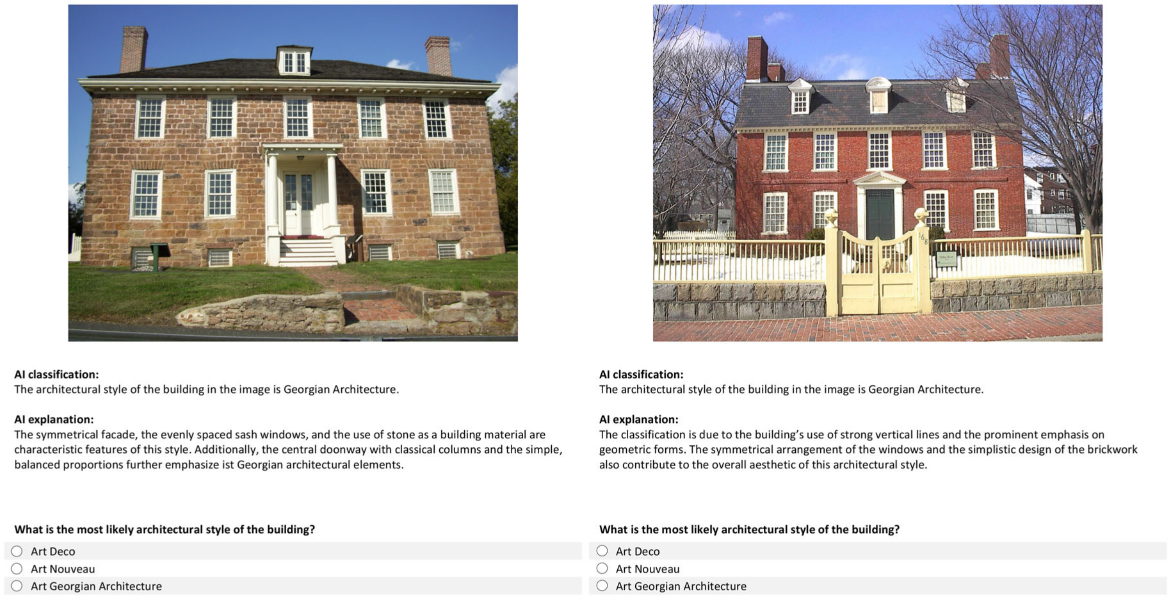

1
Introduction
Familiarization with the task and introduction to different architectural styles
2
Pre-Test
6x
Classification of a building's architectural style on an image
3
Main Task
12x
Classification of a building's architectural style on an image with support
4
Post-Test
6x
Classification of a building's architectural style on an image
5
Questionnaire
Questionnaire with open ended questions for the architectural styles and control questions
Support by treatment
No support
(Control group)
AI advice
(AI treatment)
AI advice + correct explanations
(correct explanation treatment)
AI advice + incorrect explanations
(incorrect explanation treatment)

Figure 1. The study design is outlined in five different parts: in part (1), participants were introduced to the study. In part (2), participants had to classify six images as a pretest. In part (3), participants were randomly assigned to a treatment and classified twelve images. In part (4), participants classified six images as post-test without support. Participants had to complete a questionnaire in the final part (5).



AI classification:
The architectural style of the building in the image is Georgian Architecture.

AI explanation:
The symmetrical facade, the evenly spaced sash windows, and the use of stone as a building material are characteristic features of this style. Additionally, the central doonway with classical columns and the simple, balanced proportions further emphasize ist Georgian architectural elements.

What is the most likely architectural style of the building?

○ Art Deco
○ Art Nouveau
○ Art Georgian Architecture

AI classification:
The architectural style of the building in the image is Georgian Architecture.

AI explanation:
The classification is due to the building's use of strong vertical lines and the prominent emphasis on geometric forms. The symmetrical arrangement of the windows and the simplistic design of the brickwork also contribute to the overall aesthetic of this architectural style.

What is the most likely architectural style of the building?

○ Art Deco
○ Art Nouveau
○ Art Georgian Architecture

(a) Correct explanation treatment.

(b) Incorrect explanation treatment.

Figure 2. Instances shown to participants in the main task with correct explanations (left) and incorrect explanations (right).

bucket of 18 pre-selected images of the dataset. We ensured that each class was balanced for each participant (two images per class) and that the order in which the classes were shown to participants was varied. By doing so, we minimized the risk of inducing biases in the order of images. It also ensured that our results were not dependent on the difficulty of pre-selected images. Moreover, the pretest allowed us to check that participants in each treatment did not differ in their prior expertise regarding the task. Following the pretest, participants were randomly assigned to a treatment in the main task in part (3), distinguishing the type of AI support received:

- *Control group:* Participants did not receive any AI support
- *AI treatment:* Participants received only the AI's classification
- *Correct explanation treatment:* Participants received the AI's classification and a correct explanation
- *Incorrect explanation treatment:* Participants received the AI's classification and an incorrect explanation

By distinguishing the groups in this way, we can infer the effect of incorrect explanations on participants' procedural knowledge and reasoning. After the assignment, participants had to classify buildings' architectural styles on twelve images (see Figure 2 for an example). All twelve images were randomly drawn from a bucket of 30 pre-selected images balanced in classes. Similar to the pretest, each participant was provided with four images of each class, and the order was randomized. Each image was displayed on a separate page in the study. During the main task, participants had the option to click on a button to show the context information for the architectures' characteristics and verify the support they received. In part (4), the post-test, participants of each treatment had to classify six different images without support. Similar to the pretest, the images were drawn from another bucket of 18 pre-selected images balanced in classes. Each participant was shown two images of each class on separate pages in randomized order (Supplementary Table 4 in Section A).

To prevent participants from randomly clicking through the study, participants could continue to the next page in the pretest, main task, and post-test only after a few seconds. This design choice followed the protocol of Spitzer, Holstein, et al. (2024) and was to ensure that participants focused on the actual task. In the final part (5), participants answered a questionnaire. The questionnaire asked them to describe and explain each architectural style. With this, we were able to analyze whether participants had learned the correct distinctions between the architectural styles and were able to conclude based on their understanding (e.g., their reasoning). On top of that, they had to rate several variables on seven-point Likert scales to measure for confounding factors. These included the AI usability, their experience with AI, their task load, and their trust in AI. All variables are presented in Section A in Supplementary Table 5. Throughout the questionnaire, we implemented a second attention check to ensure only valid results, as suggested by Abbey and Meloy (2017). In this attention check, we asked participants to select *"Likely"* for an item of AI usability (*"As an attention check, please choose "Likely" for this statement"*). The questionnaire ended with demographic questions on participants' age, gender, education, and employment.

4.3. AI development

We selected natural language explanations as our explanation modality due to the increasing prominence of generative AI systems that communicate directly with humans through language (Feuerriegel et al., 2024). In real-world applications such as the medical decision-making (Molinet et al., 2024) or wildlife classification (Hendricks et al., 2018), humans often receive explanations in natural language. Studying such rationales allows us to better understand how humans process explanations that align with the dominant interaction paradigm of large language models. Moreover, while prior work has examined the effects of incorrect explanations, much of this research has focused on visual explanation formats (e.g., example-based explanations or text highlighting) (Lakkaraju & Bastani, 2020; Morrison et al., 2024; Papenmeier et al., 2019).

To generate AI advice and explanations, we used OpenAI's GPT-4o model (model version 2024-05-13) through an Azure OpenAI Studio instance. The LLM provided participants with a classification (AI advice) and natural language-based self-explanations describing the reasoning behind its decisions, depending on the treatment participants were assigned to. The author team ensured that by comparing the advice against the ground truth, the architectural styles showed the correct advice. The explanations were examined to ensure that the correct ones only included the correct aspects of the corresponding architectural styles. Incorrect explanations, on the other hand, incorporated features of the other architectural classes, but did not include features that could be identified as incorrect through cross-checking visual cues. In the pre-phase of the study, we tested multiple prompt strategies in a workshop with three authors. The final prompts that were used are shown in Section A in Supplementary Table 4. The LLM was prompted such that the correct explanation treatment provided an explanation that corresponded to its prediction, while the explanation did not match the prediction in the incorrect explanation treatment. To determine the appropriate depth and length of the explanations, we gathered feedback on clarity and informativeness from participants in the tests of the pre-phase. Based on this input, we adjusted the prompts and explanations to balance between being sufficiently detailed while avoiding cognitive overload. We also assessed the comprehensibility of the explanations with the pre-phase participants.

We restricted our study to conditions in which AI advice was correct across both explanation treatments (correct and incorrect explanations). This choice enables a clean comparison of explanation effects without the added complexity of different correctness levels at AI advice, which can introduce confounding factors. While examining the interaction of incorrect advice and explanations is also important, we focused in this study on a setting where effects could be clearly attributed, with the expectation that future work can build on these findings.

The definition of an incorrect explanation follows Morrison et al. (2024) and Cabitza, Fregosi, et al. (2024) in that the AI does not provide correct justification for its predicted class (not coherent with the AI advice), independent of the ground truth of the image. To systematically characterize and manipulate the explanation correctness, we followed the definitions of recent works Morrison et al. (2024) and Cabitza, Fregosi, et al. (2024) and distinguished between three dimensions: coherence, factual correctness, and pertinence. Coherence refers to whether the explanation is consistent with the AI's predicted class (i.e., whether the features it explains align with the predicted class); factual correctness assesses whether the explanation is accurate in terms of architectural knowledge (i.e., whether the features it explains correctly describe the ground truth); and pertinence captures whether the explanation references visual features truly present in the image (i.e., whether the features it explains exist in the image). In our study, the incorrect explanation condition violated both coherence (misaligned with the predicted class) and factual correctness, while maintaining high pertinence by referencing visible features in the images. For instance, if the AI made a correct prediction for an Art Nouveau building, but the explanation did not conform to this class and did not properly describe the ground truth, it was defined as incorrect. To avoid information overload (Arnold et al., 2023), we designed the explanations to be no longer than three sentences. As outlined in Section 3, we analyze the impact of incorrect explanations for cases where the AI classification is correct.

4.4. Recruitment

We recruited 186 participants from the United States through the platform Prolific.co and ran the study on 12 August 2024. Previous research indicates that this platform is a reliable source of research data (Palan & Schitter, 2018; Peer et al., 2017). Several screening mechanisms were implemented through the Prolific platform. With the filters, we targeted individuals who were fluent in the English language and had shown high quality in previous studies (100% completion rate). Our recruitment strategy was designed not to focus on participants with specific backgrounds, but to admit participants without any further restrictions to be able to generalize our findings. Participants who met the stated criteria and completed the study's requirements received a base payment of 2.25£. Additionally, we implemented an incentive structure: participants are incentivized to conduct the task correctly by providing a bonus for each correctly classified image. This should ensure that participants paid attention during the task and did not provide random answers. The bonus was 4 pennies for each correct answer and led to a potential maximum payment of 3.21£. As stated in our pre-registration, we excluded participants who did not finish the main task on time (within 30 min) or did not finish the entire study. We also excluded participants with obvious misbehavior (e.g., clicking through the cases and always providing the same answer). Additionally, we computed the overall mean and standard deviation across all treatments and winsorized at 2.5 SD above/below the mean. Applying this exclusion strategy, we ended up with 160 participants equally assigned to the four treatments (40 participants for each treatment).

4.5. Metrics

Similar to previous work (e.g., Hemmer et al., 2023; Schoeffer et al., 2024), we assessed participants' task performance in classifying the architectural styles in pretest, main task, and post-test. Aligning with prior research (McCormick, 1997; Nahdi & Jatisunda, 2020), we use the task performance to infer participants' procedural knowledge. We measure human–AI team performance in the main task. As metrics for the task performance, we used accuracy and measured the ratio of correctly classified images over all images. Participants had to select one of the three different architectural styles for each image by selecting from a drop-down menu on each page of the task. This means that a random guess corresponded to 33.3% of performance. Aligning with Schoeffer et al. (2024), we also measured the correct adherence and detrimental overrides in the main task of the study. The correct adherence refers to situations in which participants correctly follow the AI advice. Detrimental overrides define situations in which participants wrongly override the AI advice and adhere to their own incorrect judgment.

We assessed participants' reasoning ability for the three architectural styles through open-ended questions by asking them to describe and explain each style, thereby following the procedure of Chi et al. (1989). We assessed participants' reasoning abilities in terms of correctness by comparing them to the correct characteristics and features for each architectural style. We followed the procedure of Huang et al. (2024) and used LLMs to evaluate the correctness of participants' answers and computed reasoning scores for participants: zero represents a completely incorrect or irrelevant answer, and five represents a completely correct and comprehensive answer. Based on those scores, we computed the average scores across all three architecture styles for each participant. These scores were then max-normalized. The prompt used for this analysis is in Supplementary Table 4 in Section A. By doing so, we were able to assess whether incorrect explanations impaired participants' reasoning. To ensure a high reliability in assessing the participants' answers, we provided the LLM with the description of the architectural styles in the prompt (see Supplementary Table 4 in Section A). Additionally, we assessed the inter-rater reliability (IRR) using the Intraclass Correlation Coefficient to reflect the absolute agreement between ChatGPT and two human coders. The resulting value of 0.755 indicates a good agreement between the LLM and the human raters.

Finally, we established several control variables to investigate the potential underlying factors that might influence AI-assisted decision-making. In particular, we controlled for participants' task load as previous research suggests that the information in explanations displayed to humans can affect their decision-making behavior (Abdul et al., 2020; Herm, 2023; Hudon et al., 2021; Spitzer, Holstein, et al., 2024). We measured participants' task load on a seven-point Likert scale by having them rate five validated items previously developed by Hart (1986) and that were already applied in the field (Senoner et al., 2024). In addition, we assessed participants' AI trust (Jian et al., 2000), AI usability (Davis, 1989), and experience with AI by using items proven in previous research (Senoner et al., 2024), all of them also rated on a seven-point Likert scale. All items are shown in Supplementary Table 5 in Section A.

5. Results

It took participants, on average, 14 min and 54 s to complete the study. Overall, 160 participants passed the attention check and finished the study according to the study protocol. Of these 160 participants, 79 were male, 77 were female, and four identified as diverse. To address the research questions, we first conduct several statistical analyses to answer RQ: 1 in subsection 5.1. Subsequently, we address RQ: 2 and qualitatively assess the open-ended questionnaires in subsection 5.2. In the final subsection 5.3, we evaluate the impact of incorrect explanations on the human–AI team performance to answer RQ: 3.

5.1. RQ1: Impact of incorrect explanations on procedural knowledge

We first compare the performance levels between treatments at the beginning of the study and then examine which further factors influence post-test performance when different correctness levels of explanations are provided. To establish a baseline and ensure that all treatments began on an equal performance level in classifying the architectural styles, we conduct a one-way ANOVA on the pretest performance scores across the four treatments ($F = .80, p = .498$). These results fail to reject the null hypothesis, indicating no significant differences in pretest performance among the four treatments and suggesting that participants started with comparable levels of procedural knowledge. This allows for a thorough interpretation of any differences observed in the post-test results, as they can be more readily attributed to the treatment effects rather than preexisting differences (Supplementary Tables 6 and 7 in Section A).

Post-test performance between treatments: Overall, the control group maintains the lowest performance at around 51.25%, while the correct explanation treatment has the highest performance with approximately 72.50% accuracy. The AI classification and incorrect explanation treatments show similar post-test performances of 65.42 and 63.33%, respectively. We show the difference in performance for the post-test in Figure 3.

To assess the significance of differences in post-test performance across treatments, we compare performance levels between treatments. A one-way ANOVA yields evidence of treatment effects ($F = 5.86, p = .001$), indicating that the type of AI support in the main task influences participant performance in the post-test. As we set out to explore the impact of incorrect explanations on participants' post-test performance, we further conduct pairwise comparisons using one-sided t-tests, assuming participants in the incorrect explanation treatment exhibit lower performance compared to participants in the other treatment. We report the corrected p-values according to the Holm-Bonferroni correction (Holm, 1979) (see Supplementary Table 8 in Section A). Interestingly, the data demonstrate that incorrect explanations do not lead to significantly lower post-test performance compared to the other treatments. While there is no significant difference in post-test performances between the incorrect explanation treatment and the control group ($t = 2.241, p = .986$) or the AI treatment ($t = -.371, p = .356$), there is a trend that incorrect explanations lead to a lower post-test performance than correct explanations ($t = -1.742, p = .085$). While we can see the tendency that participants provided with incorrect explanations perform worse, correct explanations lead to the highest post-test performance. Thus, the data does not provide evidence to support hypothesis 1.

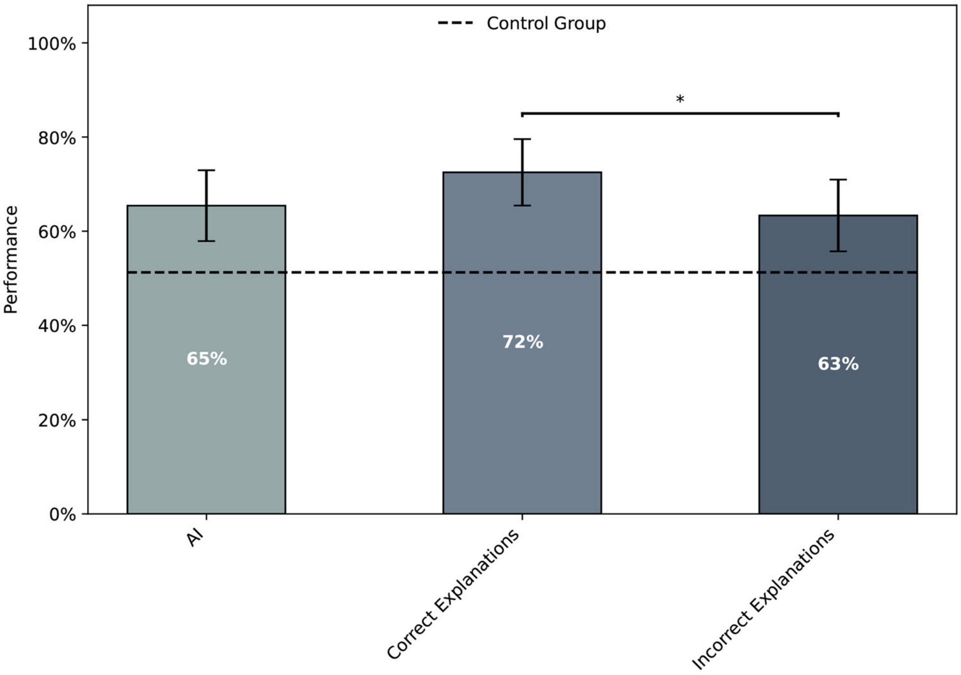


Figure 3. The post-test performances across treatments.

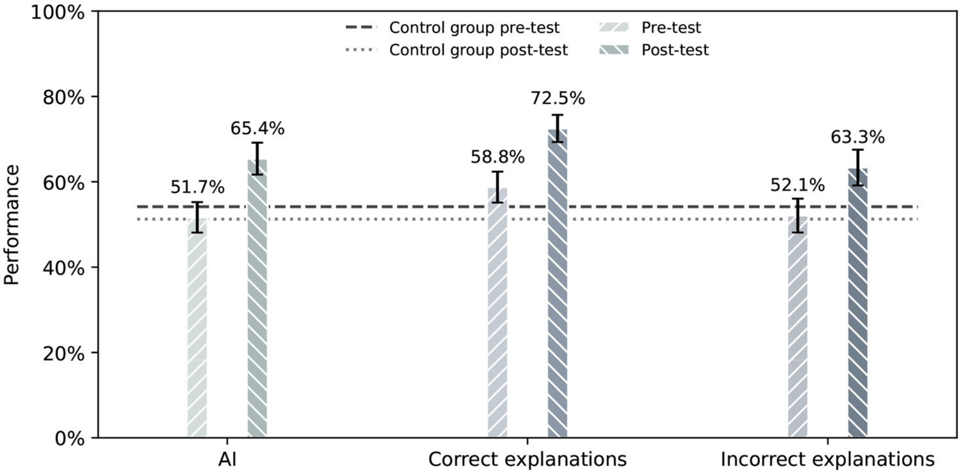


Figure 4. The procedural knowledge approximated by task performance in pretest and post-test across the different treatments. The control group is highlighted as line plots.

Performance development from pre- to post-test: We also illustrate the performance development from pretest to post-test in Figure 4. The figure illustrates whether procedural knowledge develops or degrades after withdrawing the AI support compared to the initial pretest performance. The AI treatment and correct explanation treatment show the highest procedural knowledge development, exceeding the development in the incorrect explanation treatment. For example, both the AI and correct explanation treatments show the strongest improvements (13.8 percentage points and 13.7 percentage points difference, respectively), clearly surpassing the control group's performance development (−2.9 percentage points). Even the incorrect explanation treatment shows a gain, albeit smaller (11.2 percentage points). Error bars (by standard deviation) are included in the figure (Supplementary Table 9 in Section A).

Furthermore, we conduct an ANOVA and Tukey HSD Post-Hoc Test, which reveals significant differences in procedural knowledge gains across treatments ($F = 4.04, p = .008$). We further conduct pair-wise comparisons between treatments (see Supplementary Table 10 in Section A). Notably, the incorrect explanation treatment does not significantly differ from the AI treatment or the correct explanation

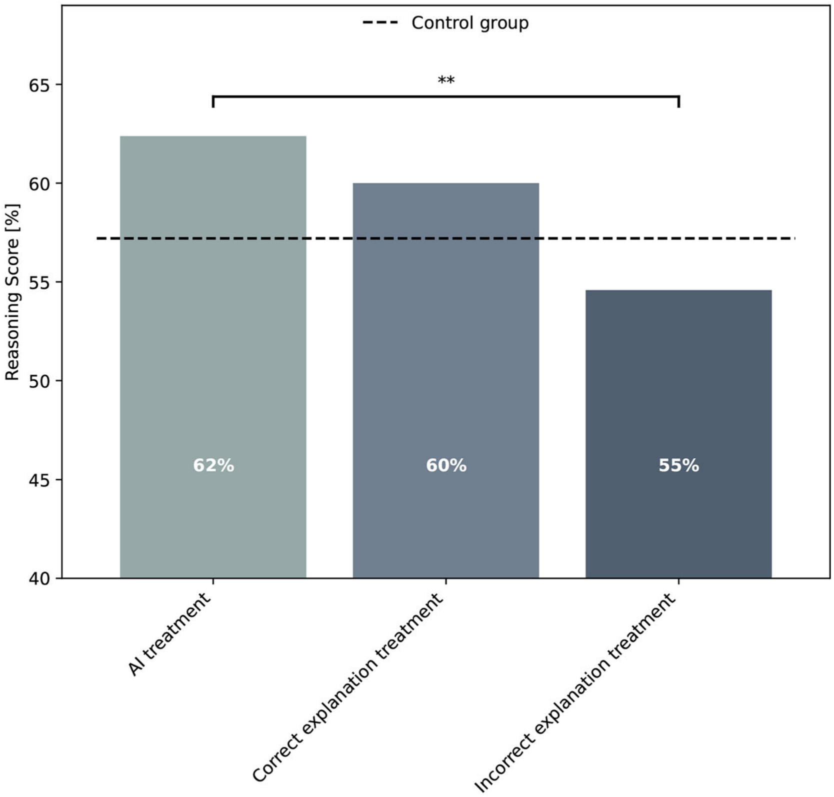


Figure 5. The reasoning scores for participants of the different treatments (*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$).

treatment ($p = .971$ for both comparisons), but it shows a trend toward higher procedural knowledge gains compared to the control group ($p = .061$). This analysis supports the previous findings showing that participants build procedural knowledge even when provided incorrect explanations (Figure 5).

Factors influencing the post-test performance: To identify the potential reasons behind these findings, we further analyze which factors influence the post-test performance. We perform a regression analysis with the post-test performance as the dependent variable and model the type of AI support as the independent variable. We define task load, AI trust, AI knowledge, and AI usability as control variables. The regression analysis reveals that the AI support significantly influences participants' performance (see Supplementary Table 9 in Section A). The model accounts for 12.7% of the variance in post-test performance ($R^2 = .127, p = .004$). Specifically, the largest coefficient can be observed for participants in the correct explanation treatment, with a significant positive effect ($coef = .198, p < .001$), followed by those in the AI treatment ($coef = .125, p = .020$) and the incorrect explanation treatment ($coef = .117, p = .025$). Interestingly, task load is negatively associated with post-test performance ($coef = -.197, p = .097$), suggesting that higher task load impedes procedural knowledge. We further analyze this finding by comparing the task load for each treatment. We can see that task load is highest for the control group (35.42%) and the incorrect explanation treatment (33.58%) compared to the other treatments (AI treatment = 28.75%, correct explanations treatment = 28.33%), illustrating that task load plays a role in participants' procedural knowledge in the post-test. Contrarily, trust, AI knowledge, and AI usability do not impact the post-test performance.

Overall, the analyses in this subsection show that the post-test performance in the incorrect explanation treatment is not significantly lower than in the other treatments and that participants provided with incorrect explanations still improve their procedural knowledge. Furthermore, we identify participants' task load as an impact factor on the post-test performance.

5.2. RQ2: Impact of incorrect explanations on reasoning

To analyze participants' reasoning capabilities as they transition from knowing "how" to understanding "why," we assess the answers they provided in the open-text questionnaire. The primary focus is on evaluating how the accuracy of AI-generated explanations influences the participants' reasoning abilities to understand how incorrect explanations impact their knowledge. We provide some randomly selected answers to these open-ended questions in the Section A in Table 11. All participants' results are provided in the supplementary materials.

The scores are then analyzed using one-sided independent t-tests to compare the performance between treatments, and the Holm-Bonferroni method is used to correct them.

In the incorrect explanation treatment ($score = 54.58\%$), participants' score is lower on average than those in the control group ($score = 57.20\%$). However, the difference is not significant ($t = -.706, p = .241, p_{corrected} = .241$). When comparing the incorrect explanation treatment to the AI treatment($score = 62.38\%$), participants exposed to incorrect explanations perform significantly worse ($t = -2.203, p = .015, p_{corrected} = .046$). The comparison between the incorrect explanation treatment and the correct explanation treatment ($score = 60.00\%$) shows a tendency for incorrect explanations to decrease the reasoning ability compared to correct explanations ($t = -1.645, p = .052, p_{corrected} = .104$). The data shows that participants exposed to incorrect explanations decrease their reasoning ability.

Additionally, we analyze further control variables that might influence participants' reasoning capabilities. To do so, we perform a regression analysis, similar to Section 5.1, using task load, trust, AI knowledge, and AI usability as control variables. We find that AI usability has a direct effect on the reasoning capabilities ($coef = .7443,\ p = .068$). We report the results in Section A in Supplementary Table 12. Thus, when participants perceive the AI as useful, they also build a better understanding of the underlying task domain. This is an interesting finding that we further discuss in Section 6.

Overall, the data shows that incorrect explanations lead to a lower reasoning ability compared to having no explanations, and there is a tendency that it is also lower compared to having correct explanations. Furthermore, we identify AI usability as an impact factor for participants' reasoning ability.

5.3. RQ3: Impact of incorrect explanations on human–AI team performance

In this subsection, we analyze the data of the study to derive insights into how incorrect explanations affect the human–AI team performance.

Figure 6 shows the main task performance across treatments and additionally illustrates participants' reliance behavior. The control group demonstrates the lowest performance at 60.83%. In contrast, all AI-assisted treatments show higher performance levels. The AI treatment achieves 87.92% accuracy, while the correct explanation treatment performs best at roughly 92.50%. Interestingly, the incorrect explanation treatment still outperforms the control group, reaching about 86.04% accuracy. The data shows that participants in the correct explanation treatment correctly adhere to the AI advice the most, followed by participants in the AI treatment and in the incorrect explanation treatment. These results suggest that AI support, regardless of the explanation accuracy, enhances performance during the main task, with incorrect explanations leading to the lowest gain.

To assess the significance of differences in main task performance across treatments, we conduct a one-way ANOVA. The results ($F = 27.80, p < .001$) indicate significant differences among the treatments. To assess the impact of incorrect explanations with the other treatments on main task performance, we conduct pairwise comparisons between the incorrect explanation treatment and other treatments using one-sided t-tests, assuming that participants in the incorrect explanation treatment exhibit lower performance compared to participants in the other treatments (see hypothesis 3). We also correct the tests using the Holm-Bonferroni method and report these p-values (see Supplementary Table 6 in Section A). The comparison between the incorrect explanation treatment and the control group shows no significant difference in performance ($t = 5.637, p = 1.000$), indicating that the performance for incorrect explanations is not below the control group's performance level. Similarly, the comparison with the AI treatment yields no significant difference ($t = -.546, p = .293$). However, when comparing the incorrect explanation treatment with the correct explanation treatment, there is a significant difference in main task performance ($t = -2.359, p = .021$), indicating that the incorrect explanations lead to a lower performance compared to the correct explanations. Thus, we find support for hypothesis 3.

Analog to the post-test performance, we perform a regression analysis to reveal the effect of further factors on the main task performance (see Supplementary Table 7 in Section A). The model

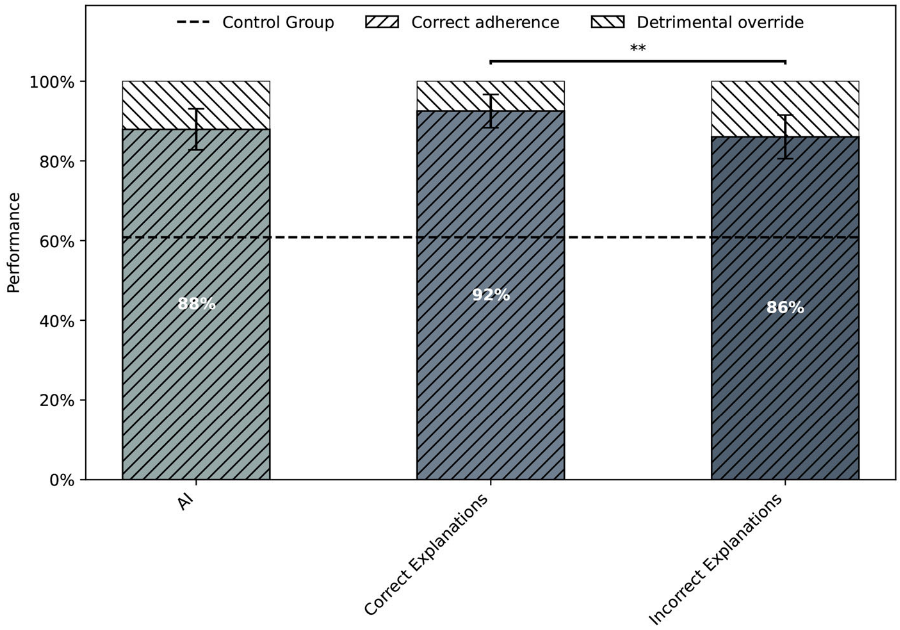


Figure 6. The main task performances across treatments.

investigating the main task performance explains 36.7% of the variance ($R^2 = .367, p < .001$), showing that participants in AI-supported treatments outperform those in the control group. Participants in the correct explanation treatment demonstrate the most substantial improvement over the control group, with a significant positive effect ($coef = .311, p < .001$). This is closely followed by the AI treatment ($coef = .259, p < .001$) and the incorrect explanation treatment ($coef = .250, p < .001$). Furthermore, AI trust shows a significant negative association with performance ($coef = -.237, p = .078$), indicating a possible nuanced effect of trust on how participants interact with the AI. We take further steps and analyze AI trust between the different treatments. AI trust is highest for the correct explanation treatment (48.75%), followed by the incorrect explanation treatment (47.08%), the control group (46.32%), and the AI treatment (44.72%). This finding shows that explanations increase trust independently of their correctness. Interestingly, as the score is higher in the control group than in the AI treatment, only providing AI advice does not seem to benefit participants' trust in the AI.

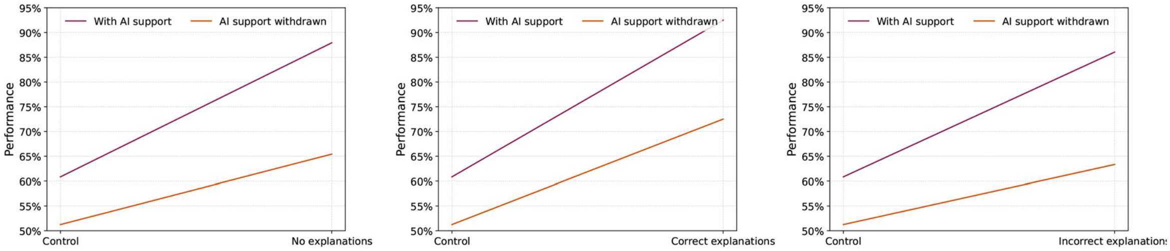
The findings suggest that while incorrect explanations can aid immediate task performance, they may hinder the retention of procedural knowledge (as the post-test performance is below the main-task performance). To analyze this aspect in more depth, we use a repeated measures approach and run a mixed-effects model to analyze how procedural knowledge is affected by first providing and then removing AI support (see Supplementary Table 2). We define performance as the dependent variable, the type of AI support as the independent variable, the different stages—with AI support in the main task and without AI support in the post-test—as the mediating factor, task load, AI trust, AI knowledge, and AI usability as control factors, and participant ID as a random factor. The binary categorical variable "withdraw AI support" represents the stages in which participants have to make decisions: the reference value 0 represents the decision-making stage in which AI is present to support participants (main-task), whereas the value 1 represents the stage in which AI is withdrawn and participants have to make decisions without the AI (post-task) (Table 2).

Participants in all AI treatments show significantly higher performance during the main task compared to the control group, with the correct explanation treatment leading to the highest performance ($coef = .306, p < .001$), followed by the AI treatment ($coef = .256, p < .001$), and the incorrect explanations ($coef = .249, p < .001$). However, when transitioning to the post-test, where AI support is removed, all AI-assisted treatments experience a significant decline in performance. The incorrect explanation treatment shows a substantial additional decline when AI support is withdrawn ($coef = -.131, p = .008$), indicating that while AI support with incorrect explanations initially boosts

Table 2. Mixed effects model analysis on performance.

	Performance	
Dependent variable	Coeff	SE
Intercept	0.695^{***}	0.084
AI support:		
- Control group (baseline)		
- No explanations	0.256^{***}	0.046
- Correct explanations	0.306^{***}	0.047
- Incorrect explanations	0.249^{***}	0.045
Withdraw AI support	-0.096^{***}	0.035
No explanations: withdraw AI support	$\mathbf{-0.129}^{***}$	0.050
Correct explanations: withdraw AI support	$\mathbf{-0.104}^{**}$	0.050
Incorrect explanations: withdraw AI support	$\mathbf{-0.131}^{***}$	0.050
task_load	−0.110	0.086
AI trust	−0.187	0.132
AI knowledge	−0.011	0.060
AI usability	0.063	0.095
participant_ID	0.016	0.029
Log-likelihood	49.9483	
Scale	0.0248	

Note: Bold values indicate significant interaction effects. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.



(a) The interaction effect on the AI treatment.

(b) The interaction effect on the correct explanation treatment.

(c) The interaction effect on the incorrect explanation treatment.

Figure 7. The sub-figures present the interaction effects of the transition from the main task to the post-test on the relationship of the type of AI support on performance.

the performance during the main task, it negatively impacts the retention and development of procedural knowledge when the AI support is removed compared to the control group. Similarly, no explanations show similar declines when AI is withdrawn ($coef = -.129, p = .009$), while correct explanations show significantly less additional decline ($coef = -.104, p = .036$). This suggests that correct explanations provide some protection against performance loss, while both incorrect explanations and no explanations are similarly detrimental to knowledge retention. The interaction effects in Figure 7 support this. The sub-figures illustrate that incorrect explanations can undermine procedural knowledge development, as evidenced by the greater performance drop in the post-test for participants who receive incorrect explanations (see Section 5.3) compared to those who receive correct explanations (see Section 5.3) or AI advice only (see Section 5.3) (lower slope in the lines for incorrect explanations).

Overall, the analyses in this subsection show that any type of AI support in this study increased the human–AI team performance over the control group. We further identify AI trust to have a negative impact on the main task performance and to be highest for treatments in which explanations are provided. Most interestingly, AI support improves performance in the short term during interaction. However, in all treatments, performance decreases post-collaboration, with the greatest decrease occurring for incorrect explanations.

6. Discussion

6.1. Summary of findings

With the rise of AI in decision-making domains, it is crucial to understand how the interaction with AI affects decision-makers. Prior research so far has either focused on the implications of incorrect AI advice (e.g., Bansal et al., 2021; Schemmer, Kuehl, et al., 2023) or explored how the correctness of explanations affects trust and reliance on AI (e.g., Cabitza, Fregosi, et al., 2024; Morrison et al., 2024). This work investigates how the correctness of AI explanations impacts humans through collaboration with an AI. In a classification task, participants were assigned to one of four treatments: no AI support, AI advice only, AI advice with correct explanations, and AI advice with incorrect explanations. We measured the task performance before, during, and after collaboration to derive humans' procedural knowledge, reasoning, and the human–AI team performance. The results show that all AI-assisted treatments led to significantly higher human–AI team performance during the main task compared to the control group, demonstrating the merits of AI support. However, the correctness of the explanations played a crucial role: Participants who had received correct explanations exhibited the highest main task performance, followed by those who received only AI advice without explanations, dominating those who had received incorrect explanations. Importantly, the findings reveal that the accuracy of the explanations has a lasting impact on procedural knowledge development. While incorrect explanations temporarily increase performance during the main task, the procedural knowledge was impaired post-collaboration, ultimately hindering knowledge retention compared to correct explanations or no explanations. Moreover, these results must be considered with caution. Even though there was an increase in procedural knowledge, participants' reasoning capabilities in the incorrect explanation treatment were below those of the control group. This indicates that incorrect explanations can undermine the durable benefits of human–AI collaboration despite initial performance improvements. These results contribute to the HCI literature by underscoring the nuanced effects of explanation accuracy on human–AI collaboration. The findings emphasize the importance of designing AI-based support that provides accurate explanations to not only enhance human decision-making but also maintain and improve humans' procedural knowledge and reasoning in the long run. Overall, the study's key takeaway is that, while AI support can generally improve performance, the correctness of the explanations provided by the AI is a crucial determinant of human–AI collaboration, influencing humans' ability to draw conclusions and perform the task autonomously post-collaboration with an AI. These insights can inform the design of more effective and human-centric AI-based decision support systems.

6.2. Theoretical implications

This work makes several contributions to the field of HCI by deepening the understanding of how the correctness of AI explanations influences humans' procedural knowledge and reasoning ability, as well as the human–AI team performance. Although previous research has often focused on the benefits of AI assistance in enhancing decision-making (Schemmer, Bartos, et al., 2023), this study offers a more nuanced perspective, highlighting the critical role of explanation correctness on humans' knowledge and understanding. It complements the large corpus of HCI literature on XAI (i.e. Morrison et al., 2024; Schoeffer et al., 2024; Speith, 2022) by explicitly investigating the influence of incorrect explanations for scenarios when the AI advice is correct and identifies the negative repercussions for humans.

6.2.1. The misinformation effect of explanations

The study highlights a crucial risk associated with AI explanations: the potential misinformation effect to impair procedural knowledge and reasoning through incorrect explanations. The misinformation effect, extensively studied by Loftus et al. (1978), Loftus and Palmer (1974), describes how exposure to incorrect information can distort an individual's memory of an event. This phenomenon extends to AI explanations, where incorrect explanations can similarly distort humans' understanding. The integration of incorrect information through explanations into existing knowledge structures, as discussed by Ayers and Reder (1998), can lead to significant changes in both procedural knowledge and reasoning. Incorrect explanations, while potentially unharmful in the short term, can create an illusion of understanding, resulting in poorer performance in subsequent tasks without AI support. This phenomenon is particularly concerning in high-stakes environments such as healthcare or legal decision-making, where the quality of decision-making has far-reaching consequences (Topol, 2019). For organizations, this implies that the long-term efficacy of AI hinges not only on their immediate performance but also on their ability to foster accurate knowledge. Incorrect explanations can lead to a misalignment between the AI's recommendations and humans' understanding, which may diminish their overall effectiveness (Morrison et al., 2024). Yet, it also has the potential to impair humans post-collaboration with an AI. Therefore, organizations must prioritize the development of AI that provides correct and transparent explanations to ensure sustained, high-quality decision-making and prevent detrimental impacts on organizational knowledge and performance.

6.2.2. Taking a human-centric perspective

In our study, we could see that the human–AI team performance improved over the participants' initial task performance. This pattern, also observed by prior research (Hemmer et al., 2024; Inkpen et al., 2023), showcases the potential of humans collaborating with AI. Even though the scenario in which the human–AI team performance exceeds the performance of human or AI alone—also referred to as complementary team performance (Bansal et al., 2019)—could not be reached (due to the study design choices this was not feasible as the AI advice was always correct), our findings showcase the two sides of explanations: while correct explanations improve the human–AI team performance compared to only receiving AI advice (bright side) and demonstrate the merits of XAI, incorrect explanations decrease the human–AI team performance compared to only receiving AI advice (dark side) and outline the pitfalls of XAI. Thus, it is important to implement mechanisms that allow humans to verify the correctness of explanations, for instance, through reflection mechanisms (Ehsan & Riedl, 2024). Furthermore, we could also identify AI trust as an influencing factor during the main task and task load during the post-test. While these findings align with prior research (Buçinca et al., 2021; Ueno et al., 2022; Westphal et al., 2023), they also emphasize the important role of individuals' characteristics and traits in human–AI collaboration. Therefore, research has to anticipate these factors in the design of robust and safe explanations.

6.2.3. The role of AI

While the focus of our research has mainly been on the humans in our AI-assisted decision-making study, the AI can take an important role in minimizing the risk of negative repercussions. In our study, these negative repercussions were expressed by humans' decreasing procedural knowledge and reasoning capabilities post-collaboration. Although this decline was seen in every AI-supported treatment, incorrect explanations led to the greatest decline, with reasoning ability falling below that of the control group. Recent research in HCI should, therefore, focus on developing methods in which the AI itself can warn human collaborators of a potentially incorrect explanation, for instance, by using uncertainty scores (Huang et al., 2023; Prabhudesai et al., 2023) or cognitive forcing functions (Buçinca et al., 2021).

6.2.4. Anchoring on explanations

We also take a critical view on the phenomenon prior research has identified in collaboration scenarios between humans and AI in which the AI provides explanations: the anchoring effect on explanations (Bansal et al., 2021; Wang et al., 2019). The effect occurs when human decision-makers fixate on the explanation and form an incorrect understanding. In our study, participants in the incorrect explanation treatment demonstrated this behavior by achieving lower procedural knowledge on the post-test compared to the other AI-supported treatments and also by showing lower reasoning capabilities than participants in the control group. As a result, they made the right decisions for the wrong reasons, a condition often referred to as the Clever Hans Effect (Kauffmann et al., 2020; Schramowski et al., 2020). Although research has adopted the term to describe AI behavior, in our study, humans demonstrate this effect. While this illustrates a major incision in the human knowledge structure, it is crucial for research and practice to develop mechanisms to counteract this effect.

6.2.5. Incorrect explanations can impair long-term knowledge

Our findings also have implications for AI-supported learning. As generative AI tools are increasingly used as learning aids, especially in situations where humans lack prior knowledge, the accuracy of explanations becomes essential not just for short-term performance but also for meaningful knowledge acquisition. Even when AI advice is correct, incorrect explanations can lead humans to develop flawed reasoning strategies that persist post-collaboration. This phenomenon, observed in our study, may have wide-ranging consequences for human knowledge development across domains and suggests that explanation design must be treated as a pedagogical concern—ensuring that humans not only complete tasks but also consider and reflect on the content they are engaging with Essien et al. (2024). Thus, if such AI systems are used as AI-based learning systems, they must be designed to minimize misinformation, scaffold critical thinking, and support reflective interaction (Lee et al., 2025).

6.2.6. The impact for organizations

Prior research has mainly explored effects that occur during the collaboration of humans and AI. For instance, Schemmer, Kuehl, et al. (2023) conceptualize the relationship between the appropriateness of reliance and how it relates to the human–AI team performance. Morrison et al. (2024) advance this view by the dimension of explanations and explore the effects of their correctness on humans' appropriate AI reliance. By outlining potential downsides of XAI, this work addresses the impact of explanations' correctness on humans post-collaboration. Our study demonstrates that their procedural knowledge and reasoning are impaired when they are provided with incorrect explanations, and the AI support is removed. Taking in a human-centric perspective (Horvatić & Lipic, 2021), this repercussion presents harm to not only humans' individual knowledge development (Bhatt, 2000) but also to their ability to provide meaningful assets to organizations (Davenport, 1998; Nonaka, 2009). Maintaining individuals as valuable assets to organizations is crucial because it directly influences organizational innovation, efficiency, and adaptability in a competitive market. Schemmer et al. (2021) emphasize that the design of decision-support systems can encourage automation bias and, consequently, deskill human humans. Sustaining humans' knowledge development can foster a more resilient and informed workforce capable of driving sustained success (Grant, 1996).

6.2.7. Reconsidering explanations with generative AI

Our work leverages LLM-generated explanations, which diverge from traditional XAI methods that aim to reveal a model's internal logic (e.g., intrinsic methods like decision trees (Mahbooba et al., 2021) or SHAP for deep learning models (Lundberg & Lee, 2017)). Instead, LLM explanations are generated through prompting and may not reflect the actual reasoning process behind a decision, bearing the potential to incorrectly reflect the underlying reasoning for the AI's decision. This distinction is important: our findings demonstrate that even when AI advice is correct, misleading explanations can negatively impact humans' procedural knowledge and reasoning. As generative AI becomes more common in collaborative contexts and everyday tasks—such as writing, ideation, and tutoring—the role of explanations is shifting. They no longer serve purely to justify decisions, but also influence how humans interpret, learn from, and rely on AI systems.

Recent HCI research has explored how humans engage with LLM-generated content, particularly in writing contexts. For instance, studies have examined how individuals perceive and utilize LLM-generated suggestions in writing tasks, focusing on writers' preferences when interacting with such generative AI tools (Qin et al., 2025). Similarly, research has investigated the integration of LLMs into everyday tasks and decision-making, revealing how humans incorporate AI-generated suggestions into their reasoning (E. Kim et al., 2025).

This body of work highlights a critical gap: while LLMs can produce outputs that appear contextually appropriate, they may not accurately represent the model's internal decision-making processes. This aligns with findings from Randl et al. (2024), who evaluated the reliability of self-explanations generated by LLMs. Their study found that these self-explanations often do not faithfully reflect the model's reasoning. Meske et al. (2025) describe explanations of generative AI as a shift from using explanations to make the AI advice transparent to the human, toward using the explanation as an explanatory support.

Our study highlights the need for caution in presenting LLM outputs as "explanations" and suggests that future work should develop design strategies that make the constructed and potentially misleading nature of these outputs more transparent to humans.

By addressing these complex dynamics, this study contributes to the advancement of HCI as a field, offering practical insights for the design of AI that are both effective for human–AI collaboration and beneficial for humans' procedural knowledge development.

6.3. Practical implications

Based on our findings, we propose design guidelines of AI systems for practice that aim to support human decision-making without impairing humans' knowledge development:

6.3.1. Prioritize the correctness of explanations to support long-term knowledge retention

Our results show that while AI support improves immediate task performance—even when explanations are incorrect—this boost can be short-term. Incorrect explanations negatively impact humans' procedural knowledge and reasoning after collaboration, leading to degraded performance when AI support is removed. This finding aligns with and extends prior work that highlights the short-term benefits of explanations for human–AI team performance (Bansal et al., 2019; Schemmer, Kuehl, et al., 2023), by illustrating their potential long-term downsides. In contexts where humans must decide autonomously—such as those governed by the EU AI Act (European Commission, 2021)—designers should ensure explanations are reliable and develop mechanisms that enable humans to critically evaluate explanation correctness, such as counterfactuals (Goyal et al., 2019; Schemmer, Bartos, et al., 2023; Wachter et al., 2017) or post-hoc explanation generation (Hartwig et al., 2024; Xu et al., 2023; Zhou et al., 2023). For example, designers could implement automated explanation verification pipelines that flag or revise explanations likely to be incorrect before presenting them to humans. One approach could be to use the uncertainty scores of the outputs.

6.3.2. Minimize cognitive overload through careful explanation design

Explanations should be informative but not cognitively overloading. Our findings indicate that high task load, particularly when explanations are incorrect, correlates with impaired procedural knowledge. This aligns with prior research (Ehsan et al., 2021) and suggests that overly complex explanations, even those intended to be helpful, can have the opposite effect. Developers of XAI systems should thus balance informativeness and simplicity, and provide just enough information to aid task performance without compromising interaction. A practical approach is to provide layered explanations, where humans can expand or collapse additional details based on their current needs and expertise level.

6.3.3. Account for individual differences in perceived AI usefulness

The participants with higher perceived AI usefulness demonstrated better reasoning—even in the presence of incorrect explanations. This indicates that human characteristics influence how explanations are processed. To enhance reasoning, explanation design could be tailored to individuals' perceived trust and usefulness of the AI, for instance, through adaptive explanation strategies. This insight adds to prior work on individualized explanations (Riefle et al., 2022; Spitzer, Kühl, et al., 2024) and transparency and reliance in XAI (Ehsan et al., 2018; Lai & Tan, 2019; Morrison et al., 2024), suggesting that explanation effectiveness is not one-size-fits-all. For instance, AI systems could include an initial calibration phase where humans' trust and perceived usefulness are assessed (similar to Mozannar et al., 2023), allowing the system to provide explanations depending on individual preferences.

Taken together, these guidelines highlight the need for a cautious and context-sensitive approach to explanation design. While explanations may enhance reliance and performance in the moment, they also carry the risk of misinforming humans and impairing long-term knowledge development. Especially in organizational or regulated settings, the trade-offs introduced by incorrect explanations can impact not only individual performance but also broader operational and safety outcomes.

6.4. Limitations and future work

Despite the valuable insights gained from this study, several limitations must be acknowledged, offering avenues for future research. First, the study explores how incorrect explanations affect humans in AI-assisted decision-making. In real-world applications, the interaction with AI that provides only correct or incorrect explanations is rather unlikely. It presents valuable means to take the first steps to investigate the impact of incorrect explanations, but it does not reflect the real world. In real-world settings, explanations are often partially correct, incomplete, or ambiguous rather than fully accurate or inaccurate. Future research should explore how such nuanced explanations influence human trust, learning, and decision-making, as this more closely resembles human–AI interaction in deployed systems. Additionally, future research could take on this aspect to extend our findings and evaluate how a mix of correct and incorrect explanations—a mix that is more realistic for deployed AI—affects humans' procedural knowledge and reasoning. In particular, different ratios of the correctness of explanations could provide further insights and advance the field.

On top of that, our results might not generalize to every human–AI collaboration scenario. We set up a study and gain empirical insights for a decision-making task in which we deploy an LLM to support humans in classifying architectural styles. Even though this task resembles various real-world use-cases in which generative AI is employed to provide advice and additional reasoning (e.g., patients seeking advice for making an initial medical diagnosis or receiving advice for financial investments), we cannot generalize our findings for every context. In different modalities with other task objectives (e.g., content creation) and conditions (e.g., duration of the study), incorrect explanations might show varying effects. In particular, our study involved a classification task in a relatively unfamiliar domain—architectural styles—in which participants were unlikely to possess prior knowledge. Through reference materials and pre-phase tests, we assessed the difficulty level to be appropriate. While this allowed for tight experimental control, it limits the generalizability of our findings to expert tasks or domains where humans bring substantial background knowledge or are provided with such during decision-making. Future work should evaluate whether the observed effects also replicate in more complex or knowledge-intensive tasks. In addition, our study used text-based explanations only. While textual formats are common in many AI applications, explanations can also be visual, interactive, or multi-modal. These different formats may shape how humans interpret and respond to incorrect information. Future research should explore how the explanation modality interacts with the explanation correctness to influence human outcomes. Finally, we restricted the task interface and measured variables to those most relevant to our research focus on explanation correctness. However, other interface elements could potentially influence human behavior and knowledge acquisition, such as AI confidence indicators, interactive feedback mechanisms, or task complexity variations. Future research should therefore investigate how these different system characteristics interact with explanation correctness to provide a more comprehensive understanding of human–AI collaboration.

Next, the focus of the study was on measuring task performance to derive insights into humans' procedural knowledge and inform about the impact on human–AI team performance. In addition, we used an LLM-based approach to derive the reasoning abilities of participants. First, such an approach to evaluate the findings poses certain challenges, as LLMs might incorporate biases (Hardy Chen et al., 2024; Shi et al., 2024). Second, in real-world scenarios, performance might not be the only metric relevant. Other measures, like appropriate reliance (Schemmer, Kuehl, et al., 2023) or fairness (Schoeffer et al., 2024) in the AI, might also be of high relevance in AI-assisted decision-making, as previous studies show (Bansal et al., 2021; Hemmer et al., 2023). It is of high importance to explore how these factors change over time and under the effect of incorrect explanations. Exploring the temporal implications can extend the views and offer new insights that support the robust and effective design of AI. Moreover, future research should also investigate strategies for mitigating the negative effects of incorrect explanations. While we show that misinformation can rapidly distort procedural knowledge, our study does not assess whether interventions—such as corrective feedback, explanation uncertainty indicators, or human training—can reduce or reverse these effects. Identifying and testing such mitigation approaches is crucial for safe and trustworthy AI integration.

Lastly, the study primarily relied on short-term measures based on task performance, which may not fully capture the long-term impact of AI support on human knowledge and how the correctness of explanations impacts the human–AI team performance. In our work, participants take, on average, 15 min to conduct the study. While such an interaction with AI reflects common real-world scenarios where humans briefly consult AI systems for immediate decision support (e.g., seeking advice for math equations in homework or requesting a classification for quality checks), it limits our ability to observe longer-term effects. Especially with AI systems being used on an everyday basis, understanding and limiting potential negative longitudinal effects is crucial. The insights we gained even in this brief interaction suggest that incorrect explanations can rapidly influence human understanding, though future work should examine whether these effects persist or evolve over extended periods of engagement. The short-term nature of our study may actually underestimate the full impact of incorrect explanations, as longer and constant exposure could lead to deeper entrenchment of misunderstandings. This could invoke further critical aspects (e.g., the ability for autonomous assessment and task execution), especially if the AI support is withdrawn. Future research could employ longitudinal designs to assess how incorrect explanations influence procedural knowledge development and reasoning over extended periods and across multiple tasks. This approach would offer a deeper understanding of how different types of explanations contribute to sustained knowledge development, aligning with the principles of human-centered AI.

7. Conclusion

This work sets out the first steps toward investigating the effect of incorrect explanations on the human and the human–AI team. By doing so, we take a human-centric perspective and analyze the repercussions of incorrect explanations on task performance to derive insights into humans' procedural knowledge and reasoning. In an online study, we assessed the impact of such explanations, specifically after the AI support is withdrawn, and humans must act autonomously.

With our work, we make several contributions to the HCI field: First, we identify a misinformation effect caused by incorrect explanations, which impairs humans' procedural knowledge and reasoning. Second, we offer insights into how such incorrect explanations limit human–AI team capabilities. Finally, we provide guidelines for the effective and safe design of explanations that can foster AI-assisted decision-making. *So we can eventually imagine: the AI provides a correct explanation for differentiating the architectural styles. You pass your exam.*

Acknowledgements

Generative AI tools were utilized throughout this work. Specifically, ChatGPT, Claude, and GitHub Copilot were employed to generate code for visualizations. Additionally, ChatGPT, DeepL Write, and Grammarly were used to enhance the writing quality of tutorials and explanations provided to participants during the experiments, as well as to improve the language across all sections of this article.

Author contributions

CRediT: **Philipp Spitzer**: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing; **Joshua Holstein**: Validation, Writing – original draft, Writing – review & editing; **Katelyn Morrison**: Methodology, Writing – original draft, Writing – review & editing; **Kenneth Holstein**: Supervision, Validation; **Gerhard Satzger**: Supervision, Writing – original draft; **Niklas Kühl**: Conceptualization, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Philipp Spitzer http://orcid.org/0000-0002-9378-0872
Joshua Holstein http://orcid.org/0009-0005-3885-8365
Katelyn Morrison http://orcid.org/0000-0002-2644-4422
Kenneth Holstein http://orcid.org/0000-0001-6730-922X
Gerhard Satzger http://orcid.org/0000-0001-8731-654X
Niklas Kühl http://orcid.org/0000-0001-6750-0876

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53*, 63–70. https://doi.org/10.1016/j.jom.2017.06.001

Abdul, A., von der Weth, C., Kankanhalli, M., & Lim, B. Y. (2020). COGAM: Measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). ACM.

Adhikari, A., Tax, D. M. J., Satta, R., & Faeth, M. (2019). LEAFAGE: Example-based and feature importance-based Explanations for Black-box ML models. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1–7). IEEE.

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., & Kantarcioglu, M. (2021). Does explainable artificial intelligence improve human decision-making? *Proceedings of the AAAI Conference on Artificial Intelligence, 35*, 6618–6626.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., & Inkpen, K. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM.

Arnold, M., Goldschmitt, M., & Rigotti, T. (2023). Dealing with information overload: A comprehensive review. *Frontiers in Psychology*, *14*(2023), 1122200. https://doi.org/10.3389/fpsyg.2023.1122200

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review*, *5*(1), 1–21. https://doi.org/10.3758/BF03209454

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*, 2–11.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM.

Barke, S., James, M. B., & Polikarpova, N. (2023). Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages, 7*(OOPSLA1), 85–111.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*(2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bhatt, G. D. (2000). Organizing knowledge in the knowledge development cycle. *Journal of Knowledge Management*, *4*(1), 15–26. https://doi.org/10.1108/13673270010315371

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency* (pp. 149–159). PMLR.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. V., Bernstein, M. S., Bohg, J., Bosselut, A., & Brunskill, E. (2021). *On the opportunities and risks of foundation models*. arXiv Preprint arXiv:2108.07258.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). 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.

Cabitza, F., Campagner, A., Natali, C., Parimbelli, E., Ronzio, L., & Cameli, M. (2023). Painting the black box white: Experimental findings from applying XAI to an ECG reading setting. *Machine Learning and Knowledge Extraction*, *5*(1), 269–286. https://doi.org/10.3390/make5010017

Cabitza, F., Campagner, A., Ronzio, L., Cameli, M., Mandoli, G. E., Pastore, M. C., Sconfienza, L. M., Folgado, D., Barandas, M., & Gamboa, H. (2023). Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine*, *138*(2023), 102506. https://doi.org/10.1016/j.artmed.2023.102506

Cabitza, F., Fregosi, C., Campagner, A., & Natali, C. (2024). Explanations considered harmful: The Impact of misleading explanations on accuracy in hybrid human-AI decision making. In *World Conference on Explainable Artificial Intelligence* (pp. 255–269). Springer.

Cabitza, F., Natali, C., Famiglini, L., Campagner, A., Caccavella, V., & Gallazzi, E. (2024). Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial Intelligence in Medicine*, *150*(2024), 102819. https://doi.org/10.1016/j.artmed.2024.102819

Cau, F. M., Hauptmann, H., Spano, L. D., & Tintarev, N. (2023). Effects of AI and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1–42. https://doi.org/10.1145/3588320

Chen, V., Li, J., Kim, J. S., Plumb, G., & Talwalkar, A. (2022). Interpretable machine learning: Moving from mythos to diagnostics. *Communications of the ACM*, *65*(8), 43–50. https://doi.org/10.1145/3546036

Chen, V., Liao, Q. V., Vaughan, J. W., & Bansal, G. (2023). 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.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182. https://doi.org/10.1016/0364-0213(89)90002-5

Clark, I., & Dumas, G. (2015). The regulation of task performance: A trans-disciplinary review. *Frontiers in Psychology*, *6*(2016), 1862. https://doi.org/10.3389/fpsyg.2015.01862

Dakhel, A. M., Majdinasab, V., Nikanjam, A., Khomh, F., Desmarais, M. C., & Jiang, Z. M. J. (2023). Github copilot AI pair programmer: Asset or liability? *Journal of Systems and Software*, *203*, 111734. https://doi.org/10.1016/j.jss.2023.111734

Davenport, T. H. (1998). *Working knowledge: How organizations manage what they know*. Harvard Business School.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., … Wright, R. (2023). Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*(2023), 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Ecker, U. K. H., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*(3), 570–578. https://doi.org/10.3758/s13423-011-0065-1

Ehrlich, K., Kirk, S. E., Patterson, J., Rasmussen, J. C., Ross, S. I., & Gruen, D. M. (2011). Taking advice from intelligent systems: The double-edged sword of explanations. In *Proceedings of the 16th International Conference on Intelligent User Interfaces* (pp. 125–134). ACM.

Ehsan, U., & Riedl, M. O. (2024). Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns*, *5*(6), 100971. https://doi.org/10.1016/j.patter.2024.100971

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018). Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 81–87). ACM.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–19). ACM.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: A technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 263–274). ACM.

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (pp. 211–223). ACM.

Essien, A., Bukoye, O. T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. *Studies in Higher Education*, *49*(5), 865–882. https://doi.org/10.1080/03075079.2024.2316881

European Commission. (2021). *Proposal for a regulation of the European Parliament and of the council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. In *International Conference on Machine Learning* (pp. 2376–2384). PMLR.

Grant, R. M. (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science*, *7*(4), 375–387. https://doi.org/10.1287/orsc.7.4.375

Hadi, M. U., Tashi, Q. A., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., & Wu, J. (2024). *Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects*. Authorea Preprints.

Hardy Chen, G., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). *Humans or LLMS as the judge? A study on judgement biases*. arXiv Preprint arXiv:2402.10669

Hart, S. G. (1986). *NASA task load index (TLX)*. https://ntrs.nasa.gov/api/citations/20000021488/downloads/20000021488.pdf

Hartwig, K., Doell, F., & Reuter, C. (2024). The landscape of user-centered misinformation interventions – A systematic literature review. *ACM Computing Surveys*, *56*(11), 1–36. https://doi.org/10.1145/3674724

Harvey, E., Koenecke, A., & Kizilcec, R. F. (2025). "Don't forget the teachers": Towards an educator-centered understanding of harms from large language models in education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)* (p. 19). Association for Computing Machinery.

Hase, P., & Bansal, M. (2020). *Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?* arXiv Preprint arXiv:2005.01831

Hemmer, P., Schemmer, M., Kühl, N., Vössing, M., & Satzger, G. (2024). *Complementarity in human-AI collaboration: Concept, sources, and evidence*. arXiv Preprint arXiv:2404.00029.

Hemmer, P., Schemmer, M., Vössing, M., & Kühl, N. (2021). Human-AI complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, *78*, 118.

Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023). 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* (pp. 453–463). ACM.

Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). *Generating counterfactual explanations with natural language*. arXiv Preprint arXiv:1806.09809.

Herm, L.-V. (2023). *Impact of explainable AI on cognitive load: Insights from an empirical study*. arXiv Preprint arXiv:2304.08861

Herz, P. J., & Schultz, J. J. Jr. (1999). The role of procedural and declarative knowledge in performing accounting tasks. *Behavioral Research in Accounting*, *11*, 1.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Horvatić, D., & Lipic, T. (2021). Human-centric AI: The symbiosis of human and artificial intelligence. *Entropy*, *23*(3), 332. https://doi.org/10.3390/e23030332

Huang, F., Kwak, H., Park, K., & An, J. (2024). *ChatGPT rates natural language explanation quality like humans: But on which scales?* arXiv Preprint arXiv:2403.17368.

Huang, Y., Song, J., Wang, Z., Zhao, S., Chen, H., Juefei-Xu, F., & Ma, L. (2023). *Look before you leap: An exploratory study of uncertainty measurement for large language models*. arXiv Preprint arXiv:2307.10236.

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (2021). Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence. In *Information systems and neuroscience: NeuroIS Retreat 2021* (pp. 237–246). Springer.

Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L. H., & Quinn, G. (2023). Advancing human-AI complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, *30*(5), 1–29. https://doi.org/10.1145/3534561

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420.

Kauffmann, J., Ruff, L., Montavon, G., & Müller, K.-R. (2020). *The clever Hans effect in anomaly detection*. arXiv Preprint arXiv:2006.10609.

Kayser, M., Menzat, B., Emde, C., Bercean, B., Novak, A., Espinosa, A., Papiez, B. W., Gaube, S., Lukasiewicz, T., & Camburu, O.-M. (2024). *Fool me once? Contrasting textual and visual explanations in a clinical decision-support setting*. arXiv Preprint arXiv:2410.12284.

Kendeou, P., Smith, E. R., & O'Brien, E. J. 2013. Updating during reading comprehension: Why causality matters. *Journal of Experimental Psychology. Learning, Memory, and cognition*, *39*(3), 854–865. https://doi.org/10.1037/a0029468

Kim, E., Choe, K., Yoo, M., Chowdhury, S. S., & Seo, J. 2025. *Beyond tools: Understanding how heavy users integrate LLMs into everyday tasks and decision-making*. arXiv Preprint arXiv:2502.15395.

Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM.

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). 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*. (pp. 1–14). ACM.

Lai, V., & Tan, C. (2019). 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* (pp. 29–38). ACM.

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. (pp. 79–85). ACM.

Laux, J. (2024). Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI & Society*, *39*(6), 2853–2866. https://doi.org/10.1007/s00146-023-01777-z

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, *13*(5), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(1), 19.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Ma, S., Chen, C., Chu, Q., & Mao, J. (2024). *Leveraging large language models for relevance judgments in legal case retrieval*. arXiv Preprint arXiv:2403.18405

Madsen, A., Chandar, S., & Reddy, S. (2024). *Are self-explanations from Large Language Models faithful?* [Paper presentation]. Findings of the Association for Computational Linguistics ACL, Bangkok, Thailand (pp. 295–337). https://doi.org/10.18653/v1/2024.findings-acl.19

Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity 2021*(1), 6634811.

McCormick, R. (1997). Conceptual and procedural knowledge. *International Journal of Technology and Design Education*, *7*(1–2), 141–159. https://doi.org/10.1023/A:1008819912213

Meske, C., Brenne, J., & Uenal, E., Oelcer, S., & Doganguen, A. (2025). *From explainable to explanatory artificial intelligence: Toward a new paradigm for human-centered explanations through generative AI*. arXiv Preprint arXiv:2508.06352

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*(2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). 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. https://doi.org/10.1145/3387166

Molinet, B., Marro, S., Cabrio, E., & Villata, S. (2024). Explanatory argumentation in natural language for correct and incorrect medical diagnoses. *Journal of Biomedical Semantics*, *15*(1), 8. https://doi.org/10.1186/s13326-024-00306-1

Morrison, K., Spitzer, P., Turri, V., Feng, M., Kühl, N., & Perer, A. (2024). The impact of imperfect XAI on human-AI decision-making. In *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–39.

Mozannar, H., Lee, J., Wei, D., Sattigeri, P., Das, S., & Sontag, D. (2023). Effective human-AI teams via learned natural language rules and onboarding. *Advances in Neural Information Processing Systems*, *36*, 30466–30498.

Mukhtorov, D., Rakhmonova, M., Muksimova, S., & Cho, Y.-I. (2023). Endoscopic image classification based on explainable deep learning. *Sensors*, *23*(6), 3176. https://doi.org/10.3390/s23063176

Nahdi, D. S., & Jatisunda, M. G. (2020). *Conceptual understanding and procedural knowledge: A case study on learning mathematics of fractional material in elementary school* [Paper presentation]. Journal of Physics: Conference Series (Vol. 1477, pp. 042037). IOP Publishing.

Nonaka, I. (2009). The knowledge-creating company. In *The economic impact of knowledge* (pp. 175–187). Routledge.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). *How model accuracy and explanation fidelity influence user trust*. arXiv Preprint arXiv:1907.12652

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., & Banovic, N. (2023). Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 379–396). ACM.

Qin, H. X., Zhu, G., Fan, M., & Hui, P. (2025). Toward personalizable AI node graph creative writing support: Insights on preferences for generative AI features and information presentation across story writing processes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1–30). ACM.

Randl, K., Pavlopoulos, J., Henriksson, A., & Lindgren, T. (2024). Evaluating the reliability of self-explanations in large language models. In *International Conference on Discovery Science* (pp. 36–51). Springer.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial intelligence,* 32(1).

Riefle, L., Hemmer, P., Benz, C., Vössing, M., & Pries, J. (2022). *On the influence of cognitive styles on users' understanding of explanations.* arXiv Preprint arXiv:2210.02123

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

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1–85. https://doi.org/10.1214/21-SS133

Sadeghi, M., Pöttgen, D., Ebel, P., & Vogelsang, A. (2024). Explaining the unexplainable: The impact of misleading explanations on trust in unreliable predictions for hardly assessable tasks. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (pp. 36–46). ACM.

Schemmer, M., Bartos, A., Spitzer, P., Hemmer, P., Kühl, N., Liebschner, J., & Satzger, G. (2023). *Towards effective human-AI decision-making: The role of human learning in appropriate reliance on AI advice*. arXiv Preprint arXiv:2310.02108

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 410–422). ACM.

Schemmer, M., Kühl, N., & Satzger, G. (2021). *Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence*. arXiv Preprint arXiv:2109.13827

Schemmer, M., Kühl, N., Benz, C., & Satzger, G. (2022). *On the influence of explainable AI on automation bias.* arXiv Preprint arXiv:2204.08859

Schmitt, V., Villa-Arenas, L.-F., Feldhus, N., Meyer, J., Spang, R. P., & Möller, S. (2024). The role of explainability in collaborative human-AI disinformation detection. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2157–2174). ACM.

Schoeffer, J., De-Arteaga, M., & Kuehl, N. (2024). Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–18). ACM.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., & Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, *2*(8), 476–486. https://doi.org/10.1038/s42256-020-0212-3

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management*, *57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174

Senoner, J., Schallmoser, S., & Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). *Explainable AI improves task performance in human-AI collaboration.* arXiv Preprint arXiv:2406.08271

Shi, L., Ma, C., Liang, W., Ma, W., & Vosoughi, S. (2024). *Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms*. arXiv Preprint arXiv:2406.07791

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, *39*(7), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Singh, C., Priya Inala, J., Galley, M., Caruana, R., & Gao, J. (2024). *Rethinking interpretability in the era of large language models.* arXiv Preprint arXiv:2402.01761

Soon, C., & Goh, S. (2018). Fake news, false information and more: Countering human biases. *Institute of Policy Studies (IPS) Working Papers* (Vol. 31). Institute of Policy Studies.

Speith, T. (2022). A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2239–2250). ACM.

Spitzer, P., Holstein, J., Hemmer, P., Vössing, M., Kühl, N., Martin, D., & Satzger, G. (2024). *On the effect of contextual information on human delegation behavior in human-AI collaboration.* arXiv Preprint arXiv:2401.04729

Spitzer, P., Kühl, N., Goutier, M., Kaschura, M., & Satzger, G. (2024). *Transferring domain knowledge with (X) AI-based learning systems*. https://doi.org/10.48550/arXiv.2406.01329

Spitzer, P., Kühl, N., Heinz, D., & Satzger, G. (2023). ML-based teaching systems: a conceptual framework. In *Proceedings of the ACM on Human-Computer Interaction* (pp. 1–25). ACM.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2495–2507). ACM. https://doi.org/10.1145/3630106.3659051

Subramanian, H. V., Canfield, C., & Shank, D. B. (2024). Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review. *Artificial Intelligence in Medicine*, *149*(2024), 102780. https://doi.org/10.1016/j.artmed.2024.102780

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022). Trust in human-AI interaction: Scoping out models, measures, and methods. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–7). ACM.

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*(2021), 103404. https://doi.org/10.1016/j.artint.2020.103404

Vasconcelos, H., Bansal, G., Fourney, A., Liao, Q. V., & Vaughan, J. W. (2023). *Generation probabilities are not enough: Exploring the effectiveness of uncertainty highlighting in AI-powered code completions.* arXiv Preprint arXiv:2302.07248

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. In *Proceedings of the ACM on Human-Computer Interaction* (pp. 1–38). ACM.

Vicente, L., & Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, *13*(1), 15737. https://doi.org/10.1038/s41598-023-42384-8

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, *31*, 841.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–15). ACM.

Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, *144*, 107714. https://doi.org/10.1016/j.chb.2023.107714

Xu, D., Fan, S., & Kankanhalli, M. (2023). Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9291–9298). ACM.

Xu, Z., Tao, D., Zhang, Y., Wu, J., & Tsoi, A. C. (2014). *Architectural style classification using multinomial latent logistic regression* [Paper presentation]. Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, (pp. 600–615). Springer.

Yeung, A., Joshi, S., Williams, J. J., & Rudzicz, F. (2020). *Sequential explanations with mental model-based policies.* arXiv Preprint arXiv:2007.09028

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 Chi Conference on Human Factors in Computing Systems* (pp. 1–12). ACM.

Zhang, Y., Liao, Q. V., & Ke Bellamy, R. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). ACM.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). 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* (pp. 1–20). ACM.

Zytek, A., Pidò, S., & Veeramachaneni, K. (2024). *LLMs for XAI: Future directions for explaining explanations.* arXiv Preprint arXiv:2405.06064

About the authors

Philipp Spitzer is a postdoctoral researcher leading the Applied AI in Services Lab in the research group Digital Service Innovation at the Karlsruhe Institute of Technology. His research is on human-computer interaction, specifically on the interplay of knowledge and human-AI interaction, focusing on designing robust and effective collaborative systems.

Joshua Holstein is a PhD student at the Karlsruhe Institute of Technology. His work primarily focuses on enhancing human-AI collaboration by supporting the development of humans' mental models of AI systems and their underlying data, examining how these affect reliance behavior and team performance.

Katelyn Morrison is a PhD student at the Human-Computer Interaction Institute at Carnegie Mellon University and focuses her research on bridging technical machine learning approaches and human-centered methods to design and evaluate explainable AI (XAI) systems that enhance human-AI collaborations in critical decision-making contexts.

Kenneth Holstein is an Assistant Professor in the Human-Computer Interaction Institute at Carnegie Mellon University, where he is directing the CMU CoALA Lab. His research focuses broadly on participatory and expertise-driven approaches to AI design, development, and evaluation, with a particular interest in AI's impacts on human workers.

Gerhard Satzger is a professor for Digital Service Innovation at the Karlsruhe Institute of Technology. His research centers around conceiving and developing digital services and corresponding business models—with a particular focus on human-centric design for services as well as on services that extract value from data via Artificial Intelligence.

Niklas Kühl is a Professor of Information Systems and Human-Centric Artificial Intelligence at the University of Bayreuth and Group Leader at the Fraunhofer FIT. He focuses on AI design, human–AI collaboration, and data-driven decision-making. His work bridges computer science, business, and social impact, emphasizing explainability, fairness, and responsible AI.