"The AI is uncertain, so am I. What now?": Navigating Shortcomings of Uncertainty Representations in Human-AI Collaboration with Capability-focused Guidance

ULRIKE SCHÄFER, Freie Universität Berlin, Germany
LARS SIPOS, Freie Universität Berlin, Germany
CLAUDIA MÜLLER-BIRN, Freie Universität Berlin, Germany

As AI becomes increasingly relevant, especially in high-stakes domains such as healthcare, it is important to investigate which approaches can improve human-AI collaboration and, if so, why. Current research focuses primarily on technically available approaches, such as explainable AI (XAI), often overlooking human needs. This study bridges this gap by adopting a well-established technical approach — model uncertainty representations — by considering users' familiarity with the format and numeracy skills. Despite being provided with uncertainty representations, users may still struggle to handle uncertain decisions. Thus, we introduce an educational approach that communicates the capabilities of humans and the AI system to users, supplementing the uncertainty representations. We conducted a pre-registered, between-subjects user study to determine whether these approaches resulted in improved human-AI team performance, mediated by the user's mental model of the AI. Our findings indicate that solely providing uncertainty representations does not improve team performance or the user's mental model in comparison to only AI recommendations shown. However, incorporating capability-focused guidance alongside uncertainty representations significantly enhances correct self-reliance and, to some extent, overall team performance. Our additional exploratory analyses suggest that factors such as task uncertainty, case difficulty, and case type, rather than numeracy skills, the need for cognition or familiarity, may influence team performance. We discuss these factors in detail, provide practical implications suggest directions for further research. Our research contributes to the CSCW discourse by demonstrating how technical approaches can be augmented with educational approaches to enhance human-AI collaboration in decision-making tasks.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Collaborative and social computing*; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: AI uncertainty, XAI, guidance, educational approach, human capabilities, AI capabilities, reliance, human-AI interaction, decision support tools, medical domain

ACM Reference Format:
Ulrike Schäfer, Lars Sipos, and Claudia Müller-Birn. 2025. "The AI is uncertain, so am I. What now?": Navigating Shortcomings of Uncertainty Representations in Human-AI Collaboration with Capability-focused Guidance. In *Proceedings of* . ACM, New York, NY, USA, 45 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' Contact Information: Ulrike Schäfer, Freie Universität Berlin, Berlin, Berlin, Germany, ulrike.schaefer@fu-berlin.de; Lars Sipos, Freie Universität Berlin, Berlin, Berlin, Germany, lars.sipos@fu-berlin.de; Claudia Müller-Birn, Freie Universität Berlin, Berlin, Berlin, Germany, clmb@inf.fu-berlin.de.

1 INTRODUCTION

Artificial Intelligence (AI) has gained widespread interest, especially in high-stakes domains such as healthcare [67, 96, 129]. This growing interest in AI highlights the importance of addressing the current challenges associated with human-AI collaboration [14, 53, 175], particularly the potential over-reliance on AI [21, 22, 147]. To address this problem, technical approaches have been developed [100], most notably *Explainable AI (XAI)* [11, 13, 21, 39, 45, 190]. XAI can be defined as an "Artificial Intelligence [...] that produces details or reasons to make its functioning clear or easy to understand" [13]. The goal of XAI may differ depending on the target audience; for example, for users of the AI system (e.g., medical doctors receiving diagnoses recommendations) providing XAI can help them to trust more appropriately and increase their scientific knowledge [11, 13, 21, 190]. A widespread technical approach entails providing information about the AI's uncertainty[1] for individual AI[2] recommendations and predictions to help users understand how reliable an individual prediction is [29, 82, 187], which we define as *uncertainty representations*. Within such representations, uncertainty is often expressed numerically in the form of confidence levels such as "The AI is X% confident in its suggestion", which can be calculated in many ways [11, 21, 22, 100, 102, 107]. Uncertainty representations can provide users valuable insights into the accuracy and reliability of AI recommendations [22, 29, 47, 143]. However, a recent study found that utilizing XAI approaches and uncertainty representations did not consistently improve *human-AI team performance*[3] [22, 29, 72, 102, 143, 172]. Recent research suggests that possible reasons for these inconclusive findings on the impact of uncertainty representations on improving human-AI collaboration may be due to a focus on technical availability [100], rather than on user needs, situational and human factors [47, 121, 155, 159]. Two key factors seem to influence the use of uncertainty representation, (1) users' numeracy skill limitations [29] and (2) unfamiliarity with the format [30, 110] (see Section 2.1). We, therefore, designed a **technical approach** combining visual and textual uncertainty representations, leveraging the strengths of both formats, which presents numerical values as frequencies [16, 148] and visual information as icon arrays [59, 163], leading to our first research question:

RQ1: Can combined uncertainty representations improve team performance? *(H1 (a-b))*

Our **technical approach** involves presenting combined uncertainty representations, derived by prior research (e.g., [11, 50, 107, 143]), in the form of the AI's uncertainty per prediction. However, when the AI system communicates uncertainty representations accompanying its recommendations, users may struggle to act based on them, especially in unusual decision-making cases with high uncertainty [140, 145]. Therefore, to complement our technical approach, we propose a human-centered **educational approach** that guides users directly in their decision-making process through audited information, which we define as information that has been validated, to educate users to collaborate more effectively with the AI system. Participants have explicitly expressed a need for information about AI's strengths and limitations [24], yet the articulation of human and AI capabilities to users remains understudied [42, 142, 153, 186]. Recent research suggests that educating users about AI capabilities can impact team performance [35, 142], so we investigate whether educating users about human and AI capabilities, i.e., *capability-focused guidance*, can enhance user interaction with uncertainty representations and improve team performance.

[1] There are a number of terms being used to describe uncertainty such as confidence, probability, scores, or accuracy (e.g., [11, 21, 50, 65, 93, 100, 102, 107, 179, 190]).
[2] In the following, we will only use the term "AI" instead of "AI system".
[3] In the following referred to as *team performance*.

RQ2: Can team performance be improved by providing capability-focused guidance in addition to combined uncertainty representations? *(H2 (a-b))*

By interacting with a system, users develop an understanding of its behavior and accuracy, which influences their interaction with it [9, 88, 95, 126], and this understanding in the form of an internal representation of the AI system can be thought of as their *mental model* [41]. A sound mental model can enable users to make informed decisions regarding when and to what extent they should rely on the AI's recommendations or its explanations [9, 23, 95]. Although prior research already acknowledged the role of users' mental models on team performance, they are often analyzed separately (e.g., [21, 27, 99]). Recent studies found that users' knowledge of the AI's capabilities influenced team performance [9, 78, 142, 166]; thus, we investigate whether the user's mental model of the AI's capabilities mediates the effect of our approaches on team performance. We want to capture whether users develop a correct mental model of whether the AI is more or less capable of making a correct decision for individual decision-making cases[4], which we refer to as *capability attribution* (see Section 2.3). We hypothesize that our approaches, combining uncertainty representations and capability-focused guidance, can improve human-AI collaboration by assisting users in creating more suitable mental models, leading to enhanced team performance. We therefore aim to address the following two research questions:

RQ3: Does the users' capability attribution mediate the effect of combined uncertainty representations on team performance? *(H3 (a-c))*

RQ4: Does the users' capability attribution mediate the effect of capability-focused guidance supplementing combined uncertainty representations on team performance? *(H4 (a-c))*

To investigate our research questions, we conducted a pre-registered, between-subject user study involving 144 participants, who were randomly assigned to one of three groups: Control Group **(CG)** (only showing AI's recommendation), Uncertainty Group **(UG)** (technical approach), and Capability-focused Guidance Group **(GG)** (combined technical and educational approach). In our crowdsourced experiment, we tested our hypotheses regarding the augmentation of a technical approach with an educational one and its impact on users' mental models and, ultimately, team performance and correct self-reliance. We utilize a simulated, high-stakes medical decision-making scenario, as uncertainty representations are commonly used in this domain [48, 131, 173]. This allows for designing tasks close to realistic settings (see Section 3.1). With our research, we want to make the following contributions to advancing the fields of human-AI collaboration in Computer-Supported Cooperative Work (CSCW) and HCI (Human-Computer Interaction):

(1) *Overcoming shortcomings of technical approaches:* We introduce an educational approach that enhances user understanding and utilization of uncertainty representations. Our user study demonstrates that integrating capability-focused guidance not only improves correct self-reliance but also enhances overall team performance to some extent.
(2) *Exploring factors influencing decision-making:* Our exploratory analysis sheds light on additional factors that impact team performance. We find that case difficulty, case type, and task uncertainty potentially affect human-AI collaboration, whereas numeracy, need for cognition, and familiarity with the format do not.

[4]We refer to a single situation where a decision has to be made as a *case*. For example, if images need to be classified, each image represents a case.

(3) *Practical design implications for an improved user interface:* Based on our findings, we provide practical recommendations for improving the user interface and interaction design to improve team performance.

2 RELATED WORK

Our research aims to improve the collaboration between humans and AI using technical and educational approaches. We begin by examining the existing challenges of technical approaches, in particular representing uncertainty. We then explore different formats of communicating uncertainty and propose a combined uncertainty representation. Next, we introduce educational approaches as a means to address the limitations of technical methods, focusing on the concept of capability-focused guidance. Since we aim to investigate the impact of our technical and educational approaches on users' mental models in human-AI collaboration, we discuss the relationship between users' mental models and performance, and whether our approaches may influence users' mental models.

2.1 Uncertainty representations in human-AI collaboration

This section summarizes recent challenges in technical approaches, focusing mainly on uncertainty representations, to improve human-AI collaboration. We discuss the common practice of representing the uncertainty of AI recommendations and propose the concept of a combined uncertainty representation.

2.1.1 Technical approaches and their challenges in human-AI collaboration. Recently, Lai et al. [100] provided an overview of *AI assistance elements*, i.e., UI elements assisting users in a human-AI decision-making task. They describe the display of AI predictions (e.g., AI that predicts whether an image shows skin cancer) as being a "natural form of assistance" that is widely used in the fields of CSCW and HCI in many forms [5]. Still, to overcome challenges associated in human-AI collaboration, such as humans' over-reliance on AI recommendations [21, 22, 147], more elaborate AI assistance elements were developed [100]. Most prominently, users are provided information about AI's predictions, ranging from showing uncertainty (e.g., [8, 11, 21, 22, 89, 190]) to more elaborate XAI (e.g., example-based explanations [17, 90]). In the following, we refer to *uncertainty representations* as providing information about the AI's confidence for individual AI recommendations and predictions (see Section 1). In this paper we focus on uncertainty representations as they are (1) extensively studied in CSCW and HCI to improve human-AI collaboration [29, 72, 82, 187], (2) a commonly used AI assistance element across a variety of domains and tasks [100][6] and (3) may have the potential to help users to decide to which degree the AI's recommendation should be considered [29, 44, 180].

However, recent meta-analyses and user studies in the field of CSCW have shown that uncertainty representations did not consistently lead to improved team performance [22, 29, 72, 102, 143, 156, 172]. For example, Bansal et al. [11] investigated the decision-making task of rating a review sentiment as positive or negative. Similar to existing studies [102], they found that showing individual AI's recommendations with uncertainty representations, i.e., as probabilities with a visualization, resulted in improved team performance in comparison to humans or AI performing the task independently. In contrast, other studies have shown uncertainty representations to not or only slightly improve team performance (e.g., income prediction task [190], medical diagnosis [22]).

One factor leading to these mixed results may be that "studies are often driven by technical availability such as new explanation techniques" [100], instead of focusing on user needs and human factors [47, 100, 155, 159]. Previous research

[5]For example, prediction [10, 17, 18, 33, 176, 188], classification of images or audio recommendations [22, 95, 185], continuous predictions [2, 109, 111], diverse domains such as law [63, 103], finance [38, 76], and cybersecurity [46].
[6]There are ranging from image classifications for medical diagnoses [22, 25, 29, 89], object detection in a military context [110], failure prediction in 3D printing [97] to sentiment detection in text [102] and more [81].

suggests two aspects that may hinder improving team performance when providing uncertainty representations: (1) users' *numeracy skills*[7] [16, 58], and (2) *familiarity* with the presentation format [30, 148]. First, in the field of human-AI collaboration, multiple studies pointed out, that users find it difficult to interpret numerical uncertainty representations [22, 179] and show "unwarranted faith in numbers" [47]. These findings are supported by the research of risk and uncertainty communication, which found that users and especially laypeople [29, 58, 163] may not have suitable numeracy skills [16, 191] to interpret uncertainty, e.g., in the format of probabilities [191]. Second, a users' familiarity with the format (e.g., different visualizations) of representing uncertainty may influence how well they understand and utilize this information. Studies investigating less familiar uncertainty representations have found that they can lead to mixed results [30, 110]. In the following section, we focus on different types of uncertainty representations to conclude which formats may lead to improved performance by addressing the challenges described above.

2.1.2 Designing uncertainty representations to improve team performance. Uncertainty is often expressed in the form of probabilities, some of which are embedded in text formats, such as "The AI is X% confident in its suggestion" %" [11, 21, 22, 100, 102, 107]. Studies have demonstrated that such representations can improve team performance [29, 102, 167] and may be preferred over visualizations [30, 164]. In contrast, other work suggests that a numerical format can lead to over-reliance (e.g., skin cancer diagnoses [29]), is difficult to interpret (e.g., in diagnosis suggestions [22]), not actionable or relatable (e.g., AI in healthcare [114]), and depends on numeracy skills [22, 47, 164]. Instead, a frequency format may be used to accommodate users' needs, such as "Out of every 100 cases, 5 are likely to be affected". While it has not been conclusively shown that the frequency format is superior [16], it can improve users' trust [190] and reliance [29]. Even though, the frequency format of communicating uncertainty seems to be promising [29, 190], such written representations carry the risk of misinterpretation due to insufficient numeracy skills of users [16, 29, 58]. Since users vary in terms of expertise, preference, and experience [16], complementing the written frequency representations with other visual formats may accommodate users' differing needs [81, 165, 169].

A multitude of studies focused on visualizing uncertainty [81]. Still, it was found that uncertainty visualizations led to mixed results on performance and other variables [30, 52, 110, 143]. Cassenti et al. [30]'s results indicated that in a decision-making scenario, unfamiliar visual representations were less preferred compared to a textual probability format, although the performance with the visualization was better compared to the textual frequency format [30]. Fernandes et al. [52] found that frequency-based quantile dot plots improved decision-making in a bus arrival prediction scenario more than cumulative distribution functions, with both visualizations outperforming textual uncertainty. In contrast, studies such as by Ling et al. [110], which used a more unconventional visualization in the form of a horizontal bar at the bottom of the user interface in a threat detection task, produced less conclusive performance results. These inconsistencies may be due to the different types of visualizations used, which varied in familiarity [30, 110]. In the field of risk communication, research demonstrated that icon arrays may improve accuracy, comprehension, and performance [6, 59], especially for low-numeracy lay users [146]. This may be due to the more intuitive, familiar format of sequential icon arrays that display discrete levels of information, showing a part-to-whole representation that may "invoke[s] automatic visual area processing and proportion judgments" [6].

Thus, we hypothesize that a combined uncertainty representation, i.e., a textual frequency format and an icon array visualization, could improve performance by combining the advantages of both formats, i.e., being familiar, preferred, appropriate for laypeople, and suitable for a diverse audience. Thus, our first research question is as follows:

[7] By low numeracy skills, i.e., innumeracy, we refer to Gigerenzer et al. [62]'s concept in the healthcare context for collective statistical illiteracy described as "widespread inability to understand the meaning of numbers." which affects novices and experts.

RQ1: Can combined uncertainty representations improve team performance? *(H1 (a-b))*

2.2 Guidance in human-AI collaboration

Although uncertainty representations provide users with more information to make an informed decision, without a suitable understanding of the AI's capabilities, this uncertainty information may not be actionable enough. In particular, uncertain cases (e.g., high uncertainty) could potentially lead to worse performance due to inappropriate reliance. In the next sections, we, therefore, introduce an educational approach, i.e., "capability-focused guidance", that may help overcome these challenges and guide users to understand their and the AI's capabilities.

2.2.1 Educational approaches in human-AI collaboration. Currently, research focuses on a narrow set of AI assistance elements that are technically available, such as new explanation techniques [100]. However, recent research on human-AI collaboration has shown that users need and want to be informed about the AI they interact with, especially in high-stakes domains such as healthcare [24, 142, 177]. Hence, utilizing "educational" approaches beyond technical approaches (see Section 2.1), such as providing users information about the AI systems' model and training data (e.g., using AI model cards [116]) or AI's capabilities (e.g., [142]), may equip users with needed knowledge to improve their collaboration with the AI [142]. In the following, we use the term *educational approaches*, which we define as educating users directly by providing audited information about the AI (e.g., AI's capabilities and reasoning) in order to enable them to collaborate more effectively with the AI system (based on Morana et al. [118][8]).

In a broader sense, we identified CSCW and HCI studies that educated users by using a training phase or tutorials [29, 31, 86, 95, 190]. However, these studies did not investigate if this information was understood or affected performance. Other studies have directly investigated whether educating users about certain aspects improves human-AI collaboration, such as providing information about the training data and model [43, 101], domain-specific guidelines [98] or AI's capabilities [142]. For example, Lai et al. [101] found that task-specific tutorials improved team performance at predicting text sentiment. Furthermore, Holstein et al. [78] showed participants information that the AI did not have access to a house prediction task[9], i.e., communicating AI limitations, with the intention of helping users gain an understanding of which cases the AI would be able to handle correctly. While human interaction behavior was impacted, performance was not necessarily affected [78]. Chiang and Yin [35] found that increasing awareness of the AI model's limitations can decrease over-reliance. Kawakami et al. [87] investigated AI-assisted decision-making in a child welfare agency and found that users' reliance on AI depends on their beliefs of the AI systems' capabilities and limitations. Pinski et al. [142] investigated the influence of AI knowledge in the form of an overview of general AI and human capabilities in an image classification task on a user's task delegation efficiency. It was shown that task delegation improved with AI knowledge, but also decreased the interest in using AI again [142]. Note that most of these studies represent low-stakes scenarios, e.g., house prediction [78], deceptive reviews detection [98], and simple object classification [142]. Nevertheless, they suggest that educational approaches may be able to improve human-AI collaboration in a variety of ways.

While more studies began to investigate educational approaches in a broader sense, only a few focused on educating users directly on the AI and their capabilities (e.g., [142]). Providing such information could focus the user's attention on the weaknesses of the AI system when needed and help them understand the context the AI has access to when

[8] "[T]he concept of supporting users with their decision-making, problem-solving, and task execution during system use by providing [...] information." [118]
[9] For example, AI has access to the year the house was built, but not the heating source.

deciding whether to rely on the AI or themselves. Therefore, in the next section, we focus on how to guide users with information about both their own and the AI's capabilities.

2.2.2 Designing capability-focused guidance to improve performance. Cassenti et al. [30] stated that the complementary capabilities of AI and humans suggest that human-AI collaboration is beneficial in complex decision scenarios [30]. Studies suggest that educational approaches, such as educating users about the strengths and limitations of the AI and themselves, can improve human-AI collaboration [24, 78, 142], see Section 2.2.1. For instance, previous research demonstrated that humans are less effective than AI at delegating tasks. This may be attributed to humans' limited capacity to accurately assess task difficulty and their own abilities [56, 153], which could be overcome by guiding users with knowledge of both the AI's and their own capabilities [60, 153]. For this, specific AI knowledge and literacy[10] is needed [15, 112, 142]. As discussed in Section 2.2.1, Pinski et al. [142] have already shown that participants enabled with AI knowledge "align their delegation decisions more closely with their assessment of how suitable a task is for humans or AI" leading to increased team performance, while uninformed participants exhibit less consistent behavior [142].

AI limits	**Human limits**
Unable to recognize patterns that were not trained [142].	Less experience with images of this classification task [142, 153].
Inflexible with untypical or low quality images [111, 142].	Stressed and distracted by surroundings [27, 29, 153].
AI strengths	**Human strengths**
Learned to recognize the object to classify on very large data sets [56, 142, 153].	Able to think out of the box and judge even highly distorted images [56, 142, 153].
Unaffected by disturbances [153].	Can adapt flexible based on new information [75, 142].

Table 1. Summary of relevant capabilities of humans and AI, with a focus on complementary skills fitting to the chosen task.

Building on the research of Pinski et al. [142] and others, we hypothesize that capability-focused guidance could enable users to make better use of uncertainty representations and AI recommendations. For example, if the AI system recommends option "A" over option "B" for a particularly difficult case, it is possible that the same recommendation would also be made for an easier case. This means that the user would receive the same recommendation for both cases, which may not reflect the nuances of the situation. If additional uncertainty representations are provided to the user, such as a confidence level of 53% or 68%, the user may be influenced to act differently, even if the AI's initial recommendation is incorrect in both cases. The user may develop a tendency to over-rely on the AI's suggestions, especially when faced with higher confidence levels, such as 68%. This highlights the potential risk of providing uncertainty information to users, as it can lead to inappropriate reliance and reduced critical thinking. Thus, we consolidated potential strengths and limitations of humans and AI suitable for our chosen task type and scenario based on literature (see Section 3.1) in Table 1[11], which can then be used to educate users and enable them to utilize the uncertainty representations.

Summarized, capability-focused guidance may help users interpret such uncertain cases by allowing them to infer why the AI is more or less certain. They can then make an informed decision about how much to consider the AI recommendation based on their newly gained knowledge about the AI's strengths and limitations. Thus, we

[10] Long and Magerko [112] define AI literacy as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool [...]" with one specific competency described as identifying AI's strengths and weaknesses and understanding for which problems AI might be appropriate to use [112].

[11] For more details, see OSF Material.

investigate whether educating users about the capabilities of the AI and themselves, in addition to providing uncertainty representations, can mitigate inappropriate over-reliance, leading to improved team performance:

> **RQ2:** Can team performance be improved by providing capability-focused guidance in addition to combined uncertainty representations? *(H2 (a-b))*

2.3 Mental models in human-AI collaboration

As described in Section 1, we assume that capability-focused guidance and uncertainty representations may help users improve their mental model. RQ1 and RQ2 examine team performance, which can be affected by the user's comprehension, specifically their mental model of the AI [9]. However, to gain a deeper understanding of users' mental models of the AI and the approaches used, we directly investigate selected mental model variables. Thus, in the next section, we describe the relationship between mental models and performance, followed by our mental model variable of interest, i.e., capability attribution.

2.3.1 The relationship between mental models and performance. During an interaction with an AI system, users develop an understanding of the AI's behavior and accuracy [9, 95, 126], which in turn influences users' interactions with the AI system [88]. This understanding can be described as a *mental model* of the AI. Mental models are incomplete, simplified representations that humans create to understand the real world [41]. They allow users to predict and understand the behavior of systems which with they interact [126, 127], including AI systems [95]. Kulesza et al. [95] found that users with a more sound mental model of the AI system "were significantly more likely to make the recommender operate to their satisfaction" [95]. Several studies in the field of CSCW and HCI have suggested that the more accurate the user's mental model, the more the user can appropriately rely on the AI system, leading to improved performance [9, 23, 142].

However, multiple factors have been found to influence users' mental models of an AI system and thus team performance, such as trust [190], confidence [190], observing system weaknesses [128], error boundaries [9], and perceived accuracy [92]. One recurring factor influencing users' mental models and, therefore, team performance is the information the users receive regarding the AI system and task (e.g., [11, 23, 49, 78, 142, 155, 190, 190]), for example, in the form of uncertainty representations [190] or educational information about an AI [78, 142, 166]. For instance, Bansal et al. [9] investigated how the AI's error boundary settings affected users' mental models of the AI's capabilities, and thus team performance [9]. The authors specify, that to support complementarity of humans and AI, i.e., having the human-AI team outperform the human and AI alone, users need to develop a sound mental model of the AI capabilities. As summarized in Section 2.2, providing users with information to enhance their knowledge about the AI's capabilities can improve human-AI performance. These findings are supported by research underlining the importance of users' awareness of both the AI and their own capabilities to be able to collaborate with the AI effectively, hence improving team performance by enhancing their mental model of the AI (e.g., [9, 24, 56, 60, 153]).

2.3.2 Investigating users' mental models of capability attribution. As described above in Section 2.3.1, users' awareness of both the AI and their own capabilities seems to be highly relevant when studying mental models and how they affect human-AI collaboration and performance. In the following, we refer to the user's mental model of whether the AI is more or less capable of making a correct recommendation for individual cases as *capability attribution* (see Section 3.6 for operationalization details). Besides our reasoning above, our concept of capability attribution is supported by the cognitive appraisal theory [54, 105, 142], which illustrates how humans cope with situations, including their perception

of the situation and their knowledge [37, 105]. In this context, "appraisal" entails humans' assessment of how suitable a task or case is based on the capabilities of the AI [54, 142]. As mentioned above, Pinski et al. [142] have already shown that users' assessment of how suitable a task is for the AI or themselves relates to users' delegation behavior and performance.

We hypothesize that providing uncertainty representations can improve the user's mental model of what leads to the AI being uncertain. This builds on the assumption that XAI as a retrospective approach serves to build more accurate mental models [49]. By providing uncertainty representations, we aim to enable users to build a mental model of the quality of each AI prediction [11, 32, 155, 190]. We believe that by experiencing and making decisions on multiple cases, users can develop an understanding of the AI's capabilities to make a correct recommendation based on the data provided (here, nail fungus images to assess, see Section 3.1). Thus, to extend our understanding of what improves human-AI collaboration in the CSCW and HCI field, we explore whether uncertainty representation can influence users' mental models, i.e., capability attribution, and thus team performance:

> **RQ3:** Does the users' capability attribution mediate the effect of combined uncertainty representations on team performance? *(H3 (a-c))*

Furthermore, Ahn et al. [4] found that people's background knowledge and beliefs about an underlying causal mechanism lead them to develop a causal explanation for a specific event. This supports our assumptions that based on users' experiences and knowledge (e.g., capability-focused guidance and shown uncertainty representations), users create mental models trying to develop a causal explanation why the AI makes a specific recommendation (e.g., capability attribution). Thus, we hypothesize that directly educating users about both their own and the AI system's capabilities may enable them to make use of the uncertainty representations and handle uncertain and difficult cases better (Section 2.2) due to an improved mental model of the AI's capabilities, leading to improved team performance.

> **RQ4:** Does the users' capability attribution mediate the effect of capability-focused guidance supplementing combined uncertainty representations on team performance? *(H4 (a-c))*

3 METHODS

In this study, we conducted a pre-registered [12] online experiment to understand if and how uncertainty representations and capability-focused guidance affect human-AI collaboration. The next section outlines the task and scenario, followed by the experimental design and study procedure. Then, the measures, sample, and analysis plan are described.

3.1 Task and scenario

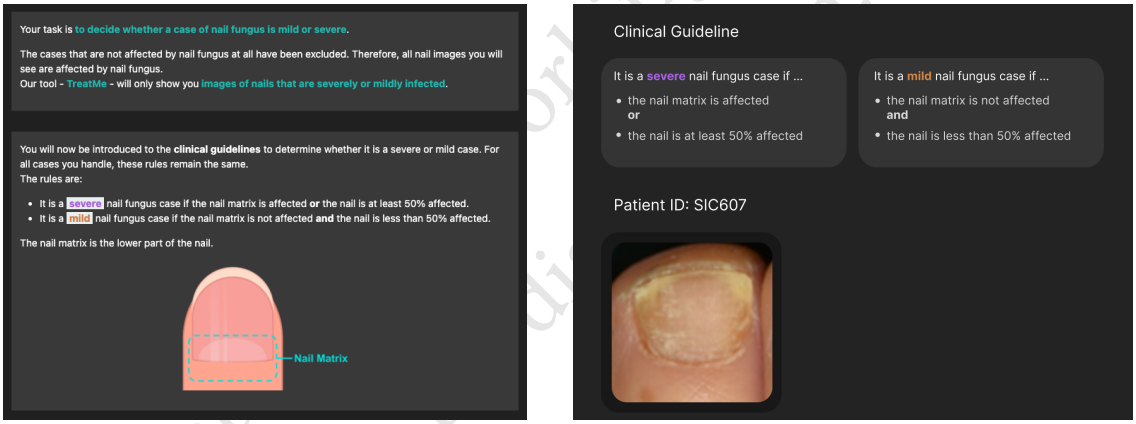
3.1.1 Scenario and high-stakes setting. Recently, Kumar et al. [96] conducted a systematic literature review that highlights the growing importance of AI in the field of disease diagnosis. We decided to utilize such a healthcare decision-making scenario, so that our findings contribute to recent research focused on improving human-AI collaboration in a high-stakes domain [24, 29, 89, 96, 104, 106, 108, 151, 171]. Furthermore, our intervention approaches suit computer vision tasks, which are a prominent focus in AI in healthcare [29, 48, 131, 162, 173].

Our chosen scenario deviates from studies using low-stakes scenarios (see [21, 78, 111, 142, 170]), which may be less relevant in realistic circumstances [96], as they may be perceived as less far-reaching and less consequential (e.g., do

[12] https://osf.io/mzay8/?view_only=96861621e73a4c5cae6c9d44c2185ef4

not have an impact on individuals or the society). Instead, we intentionally designed a task and scenario that could realistically be performed in the context of digitized healthcare [51, 83, 133, 158], where the use of AI is on the rise [96]. Specifically, we contextualized our experiment in a pharmaceutical decision-making scenario focused on classifying nail fungus (onychomycosis) images as mild or severe cases, i.e., we refer to each image to be assessed as a *case*. Nenoff et al. [123] emphasized that nail fungus is not merely a cosmetic problem, but also a serious health concern in the area of infectious diseases. For example, nail fungus can result in acute bacterial cellulitis, gram-negative toe web infection [123] and particularly affects patients with precondition such as diabetes [61, 61, 141]. Therefore, we deem this nail fungus scenario, along with the responsibility to make a decision that impacts the health and quality of life of potential patients, as high-stakes.

In comparison to other high-stakes scenarios (e.g., law, finance, professional [100]), this scenario has the advantage of being defined by official pharmaceutical guidelines [115, 122, 124], making it possible to realize this scenario in form of a fictional part-time job by guiding laypeople. After consulting our pharmaceutical expert and in light of increasingly outsourced task in the medical field to third party services or directly to AI [64, 66, 68, 182], we believe that this scenario reflects potential real life adaptations as digital healthcare services may also utilize guided laypeople for specific tasks in an online setting.



(a) Task description provided to the participants before receiving example cases with feedback.

(b) Example case showing a severely affected nail as the nail matrix, i.e., bottom half of the nail, and more than 50% is affected. Image credit: [69, 70]

Fig. 1. Rule introduction and example case showing a severely affected nail.

3.1.2 Task description and rule introduction. Our national official pharmaceutical guideline for onychomycosis states that local therapy is only possible if no contraindications exist and if the nail matrix is unaffected, no single nail is more than 40% affected, and/or maximally, 3 of the 10 toes are affected [115, 122, 124]. We deem it realistic, that such tasks can be compartmentalized. Therefore, in our experiment, participants only had to categorize one nail at a time as mild or severe based on the condition of the nail matrix and degree of affectedness. In order to make this task more realistic for the lay person in our part-time job scenario, we examined the current over-the-counter medications on the national market. These were mentioned as self-medication options in the pharmaceutical guideline [115]. One leaflet specified, that in addition to the nail matrix being unaffected, that no more than one half of a nail, i.e., 50%, should be

affected [13]. As the categorization of 50% seems clearer to communicate to laypeople, we used 50% of affectedness as an indicator of severity. Besides being introduced to the experiment and general scenario (see Section 3.3), the participants were familiarized with the rules and case interface as depicted in Figure 1[14]. Generally, our task can be described as human-centered and human-grounded, as we tested real humans with a simplified task [161].

3.1.3 Case preparation and data used. The images used were provided by Han et al. [70], who created a large data set of labeled healthy and disease affected nails (e.g., with nail fungus) [69, 70]. Since the pharmaceutical scenario involves deciding whether a nail fungus case is mild or severe, we further divided the images into mild and severe cases based on the pharmaceutical guidelines. Afterward, a pharmaceutical expert provided feedback on all pre-selected images, their classification as mild or severe and the difficulty of these cases. After three rounds of adjustments and feedback, we arrived at the final selection of cases. In addition, feedback from a pilot study of 12 laypeople and HCI experts, i.e., focus on design, decision-making and reflective informatics, was collected to test the technicalities of the survey and to check that the difficulty of the cases matched the intended difficulty to avoid floor or ceiling effects [168].

[13] In some leaflets no direct instructions regarding the decision between self medication and a needed doctor evaluation were mentioned (e.g., Ciclopirox Winthrop®, Ciclopoli®). In the leaflet for "Amorolfin-HCl 5% Acryl-Nagellack" with the active ingredient Amorolfin the instruction stated that for self-medication, the nail matrix has to be unaffected; no more than 2 nails should be affected, and affected nails shall only be affected in the upper half or on the edges of the nail, i.e., less than 50% and no nail matrix.

[14] The nail image shown is from Han [69] published under the CC BY 4.0 International License and adjusted by us (cropped).

3.2 Experimental design and groups

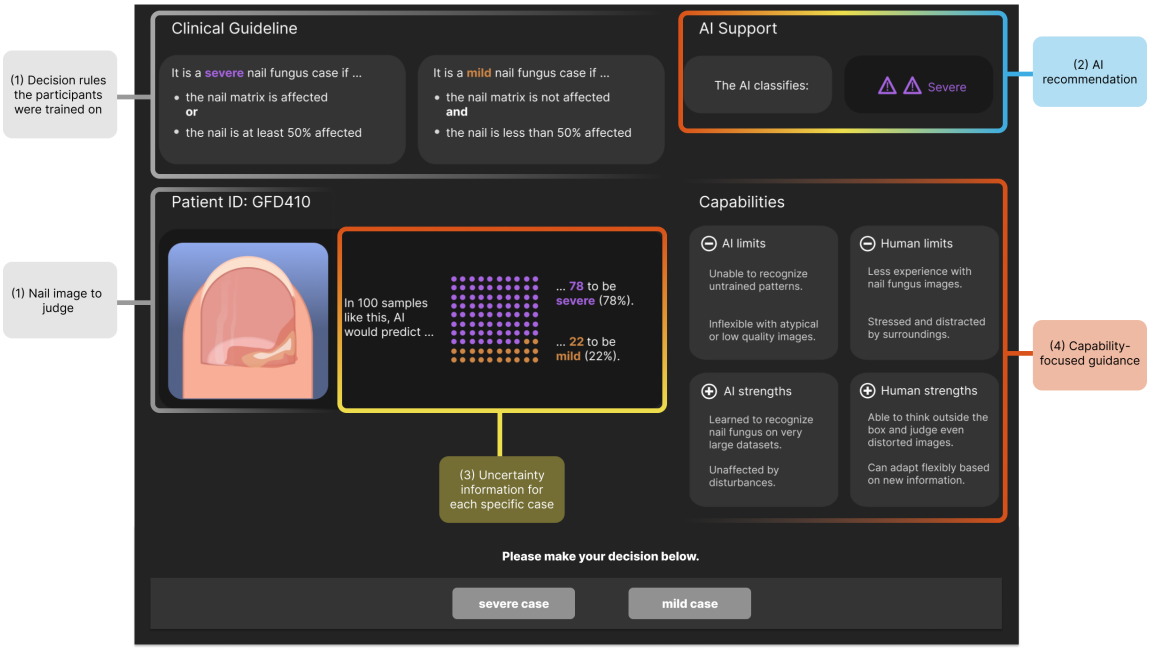


Fig. 2. **(1) Patient's case and guideline** In the human baseline the participants were shown the pharmaceutical guideline, the patient ID and nail image on the left, **(2) AI recommendation** After the human baseline all participants were shown the AI recommendation in the top right, **(3) Uncertainty representations** Participants in the UG and GG were shown uncertainty representations to the right of the nail image during the human-AI phase, based on Section 2.1, **(4) Guidance** Participants in the GG were shown capability-focused guidance on the bottom right during the human-AI phase, based on Section 2.2.

We are interested in how uncertainty representations and capability-focused guidance influence team performance and users' capability attributions. Therefore, we applied a between-subjects study design, consisting of three groups:

- Control Group **(CG)** : Only the AI recommendation is shown during the human-AI phase.
- Uncertainty Group **(UG)** : In addition to the AI recommendation, a combined uncertainty representation is provided. The AI's uncertainty is given in a percentage and frequency format based on Cao et al. [29] [15]. The certainty regarding it being a mild or severe case is given (e.g., "In 100 samples like this, AI would predict 20 to be severe (20%).") with a fixed denominator of 100 [30, 148, 163]. The given values are predefined based on the difficulty of the case [16]. Additionally, this frequency format is translated in a 10x10 dot icon array to visualize certainty, with purple colored dots [143] presenting severe and mild.
- Capability-focused Guidance Group **(GG)** : In addition to the AI recommendation and uncertainty representation, the participants were given information about their and the AI's capabilities, i.e., strengths and limitations, which were shown constantly during the task (see Section 2.2).

In Figure 2, we show an exemplary user interface, showing the interface elements that were provided for the CG.

[15] They used a calibrated frequency: "In 100 samples like this, AI would predict 72 to be benign, and 51 out of the 72 would actually be benign.".
[16] More ambiguous cases were presented as nearing 50%/50% uncertainty, whereas clear cases are more in the ranges of 80%/20%. The chosen cases were approved by a pharmacist.

3.3 Study procedure

The participants were informed about the eligibility criteria and consent information. The experiment was realized with LimeSurvey (version 6.4.0+231218).

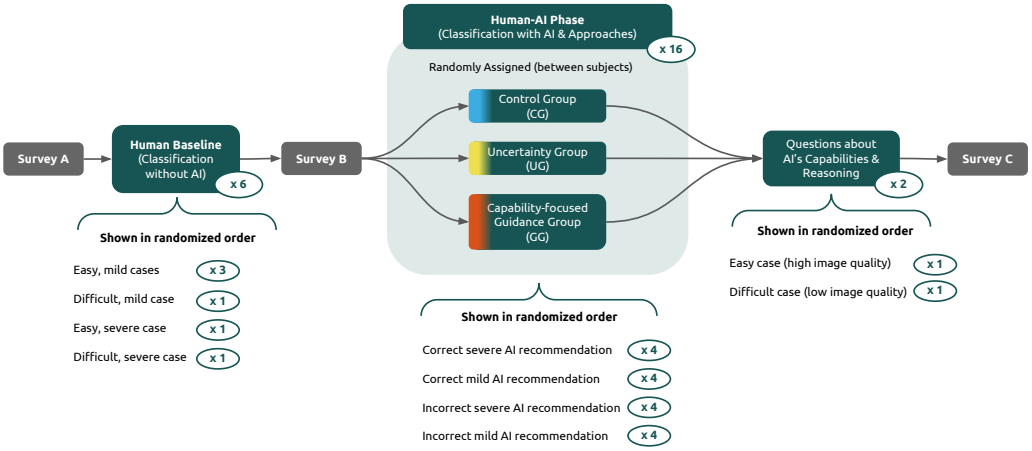


Fig. 3. Diagram of our experimental between-subjects study design. **(Survey A)** Participants are asked demographics questions. Next, in the **Human Baseline**, they assess nail fungus cases without AI assistant, followed by **Survey B** consisting of instructions and questions. During the **Human-AI Phase**, participants are shown AI's recommendations and uncertainty representations or guidance depending on which group they were randomly assigned to. Then, **Questions about AI's Capabilities and Reasoning** are asked. Lastly, participants complete **Survey C**.

As illustrated in Figure 3, the study begins with **Survey A**, in which the participants were asked demographic questions about their age [183], English language skills, education level [165], prior knowledge of AI, study design, or domain knowledge [78], followed by the first attention check [130, 135, 138] (see Appendix B). Previous studies have shown a lack of end-user training; thus, we explained the task and scenario in detail [189]. Afterward, all participants were introduced to the scenario, which simulates a part-time job at the fictional company *eHEALTH*, the decision-making task including the rules and the user interface (UI) (see Appendix A). Three cases were given to be assessed, and a solution was provided, followed by comprehension questions and a second attention check.

In the next section of the study, i.e., the **human baseline** (see Figure 3), the participants were asked to assess six nail fungus cases without the aid of the AI and reported their confidence in their recommendations, i.e., classification without AI. Of these cases, two were severely, and four were mildly affected by nail fungus. One of the mild cases and one of the severe cases were modified to be of lower image quality, which could later be considered more difficult for the AI to assess and are similar to cases that the AI misclassifies in the human-AI phase. These six cases were designed to familiarize participants with the task on their own, without feedback, to ensure that they had grasped the task well enough to be able to understand the new information in the intervention phase.

Subsequently, in **Survey B** (see Figure 3), all participants went through a third attention check and confidence questions. In addition, all participants were introduced to the new interface element, i.e., the AI recommendation. Participants of the UG and GG were introduced to the uncertainty representations and performed a comprehension check [21, 143]. Participants of the GG were introduced to the capability-focused guidance with additional comprehension checks. After the instructions, all participants assessed one test case to get used to the new interface.

In the **Human-AI Phase** (see Figure 3), all participants were shown 8 mild and 8 severe nail fungus images. Although each group viewed different UI elements, all participants were shown the same images and AI recommendations. Of the 16 cases shown, two were misclassified as mild, and two were misclassified as severe by the simulated AI[17]. Therefore, the AI recommendations were simulated with a 75% accuracy[18]. We decided to utilize 16 cases as a smaller number of trials, which seems more realistic for a flexible part-time job that may be performed asynchronously, for example, within a ticket system. Instead of using many cases, we put great effort into choosing appropriate cases to account for case difficulty and other potential effects based on feedback from our pilot study and pharmaceutical expert, see OSF Materials. Furthermore, prior research was able to identify the effects of uncertainty representations or guidance by utilizing 16 or fewer cases [29, 142]. To make the AI appear more realistic, we modified the images of the misclassified cases to be of lower image quality. The order of the cases shown was randomized.

Next, the participants were asked **Questions about AI's Capabilities & Reasoning** (see Figure 3). Two single nail image cases, one designed to be easy and one to be hard for the AI to assess, were separately presented only with the image and rules shown. The participants were asked how confident they were in their own and the AI's capabilities (needed for the mediator variable see Section 3.6) to state what they would choose and to predict the AI's recommendations. We decided against including such prediction questions directly into the 16 trials, as Buçinca et al. [20] highlighted the disadvantages of proxy tasks regarding validity. Finally, in **Survey C**, questions were asked to measure further variables of interest (see Section 3.4). A detailed study procedure can be found in Appendix D.

3.4 Measures

To investigate the influence of uncertainty representation and guidance on capability attribution and performance, we focused on three measures.

- *Overall performance*: Percentage of correctly classified cases in the human-AI phase [21, 117].
- *Correct Self-Reliance*: Percentage of disagreements with the AI in the human-AI phase when the AI made incorrect predictions [21, 157].
- *Capability attribution (Mental Model)*: Difference between the degree of capability participants assigned to the AI for an easy and for a hard for the AI to-assess case (see Section 3.6).

We decided to include only one mental model variable, i.e., capability attribution, directly in the mediator analysis for two main reasons: (1) previous literature provides a strong basis for the influence of users' mental models, in the form of their knowledge of the AI and its capabilities, on team performance (e.g., [9, 142] see Section 2.3), and (2) to avoid fishing for p-values [19]. Still, as using one mental model variable may provide limited information, we additionally collected participants' predictions of the AI recommendation for the hard and easy case shown after the main experiment [77], opinions on the self confidence per phase, perceived AI error tendency and accuracy, helpfulness of the UI elements, objective numeracy (Berlin Numeracy Test see [16, 40]), need for cognition [21, 36], subjective workload (NASA TLX see [71, 79, 150, 160]) and comprehension checks. The specific questions asked are described in Section 4. Additionally, the participants' age [183], prior knowledge and education [78] was collected.

[17] See the OSF Material for more details on the case design.
[18] Based on similar study designs with 73.3% [23], 75% [21], 87.96% [155], and 75% [190] accuracy.

3.5 Participants and sample size

A total sample of 144 participants was recruited via convenience sampling to participate in this online study via the crowdsourcing platform Prolific, which connects researchers with participants for various types of research (e.g., academic research, AI training, market research). We specifically chose Prolific because studies comparing Prolific to MTurk found that Prolific produced the highest quality research data in comparison to MTurk an others [5, 134, 136, 137, 178]. This could partially be explained by Prolific's strong vetting processes which are becoming more relevant due the increasing use of bots and AI [178][19].

Based on our G*Power a priori power analysis for a Wilcoxon-Mann-Whitney-Test for comparing two groups (as needed for the main hypotheses of interest), we decided on a sample size of 144, i.e., 48 individuals per group (d=0.62, $1-\beta$=0.9, α=0.05[20], $1-\beta$=0.9, n1=n2=48). In case the data fortunately fulfills the conditions for parametric testing, a t-test for independent group means will be conducted ($1-\beta$=0.91) by using a linear model regression, see Section 3.6. Due to the differences between tasks and scenarios used in the field of human-XAI collaboration, we extensively discussed the sample size[21]. Two main aspects led us to our final sample size: First, relevant studies in regards to uncertainty representations in the frequency format (N=50 [29]), and capability-focused guidance in form of AI knowledge as a moderator (n=55 [142], see Section 3.6) used similar sample sizes. Second, we looked at a recent meta-analysis summarizing AI-assisted and XAI-assisted performances including means and standard deviations [156]. We then focused on the studies with higher mean XAI-assisted performance[22], and used their averaged values for the power analysis above to derive the expected effect size (d=0.62)[23].

Each participant received $6.63 based on the rate of 9£/hour. Eligibility criteria included fluent English language proficiency, and failing less than two attention checks. All participants gave their informed consent prior to data collection. The study was approved by our institutional review board (IRB).

[19] For further information, please check https://www.prolific.com/.
[20] No Bonferroni adjustments were used, as this may lead to new challenges [139]. Instead, all tests that were conducted regarding the main hypotheses as well as the calculated p-value are directly reported, therefore the results can be interpreted uniquely depending on the reader's preferences.
[21] Sample size estimation focuses on RQ1-2 because the mediator hypotheses are very exploratory at the time of our data collection and we did not want to p-hack our results [73].
[22] As this is an exploratory approach, we decided on a mean difference greater that 1.
[23] (XAI: M=72.50, SD=9.30; AI: M=67.00, SD=8.43)

3.6 Analysis plan

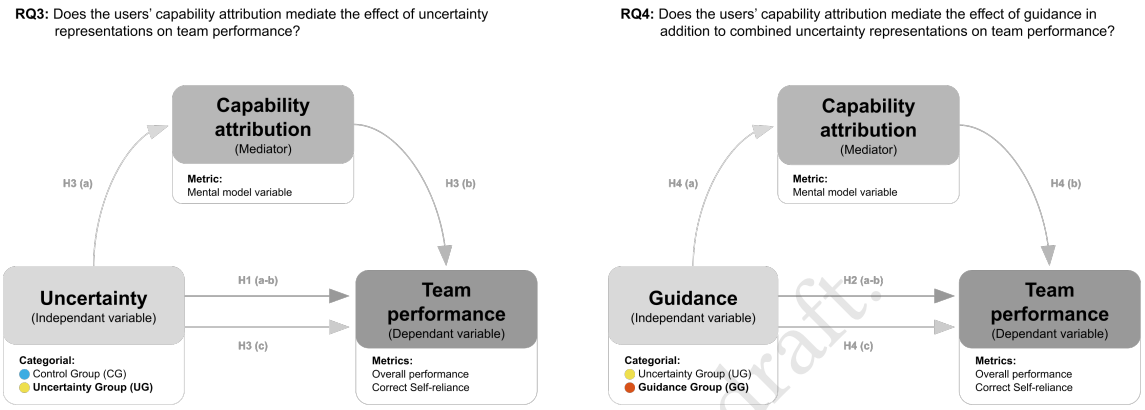


Fig. 4. Illustration of the mediator hypotheses H3 and H4. More details see Section 3.6.

Statistical analyses were conducted using R (version 4.3.2, R Core Team 2023) [144]. The hypotheses were tested by analysing overall performance, correct self-reliance and capability attribution. The post hoc power analyses were conducted with 1 - β = 0.9. To investigate the research questions RQ1 and RQ2, we formulated two hypotheses, each focusing on a performance measure of interest, i.e., overall performance or correct self-reliance, as defined in Section 3.4.

> **RQ1:** Can combined uncertainty representations improve team performance? *(H1 (a-b))*

H1 (a) Uncertainty representations lead to improved overall performance compared to none shown.
H1 (b) Uncertainty representations lead to improved correct self-reliance compared to none shown.

> **RQ2:** Can team performance be improved by providing capability-focused guidance in addition to combined uncertainty representations? *(H2 (a-b))*

H2 (a) Additional guidance leads to improved overall performance compared to only uncertainty representations.
H2 (b) Additional guidance leads to improved correct self-reliance compared to only uncertainty representations.

Overall performance was calculated by summing the correct decisions of each participant in the human-AI phase divided by the number of cases (= 16) which was then converted into percentages. The second performance measure used is *correct self-reliance*, i.e., only considering the cases where the AI's recommendation was incorrect. This value was also translated into percentages. Figure 4 illustrates the investigated relationship in form of an arrow pointing from "Uncertainty" or "Guidance" to the dependant variable "Team performance". To analyze the group differences for RQ1 and RQ2, we applied a linear model regression, dummy-coding the three groups with the CG as reference, see OSF Material. The hypotheses about the differences between the CG and UG (RQ1: H1 (a-b)) and UG and GG (RQ2: H2 (a-b)) were formulated as statistical hypotheses about the regression coefficients [55], whereby (a) refers to overall performance and (b) to correct self-reliance. In addition, adequate plots were analyzed, and the requirements were tested. As the requirements for this parametric test may not be met, additional non-parametric tests were performed.

In order to test the two research questions RQ3 and RQ4 regarding the mediator effect, the conditions according to Baron and Kenny [12] were analyzed, and a linear model regression was applied, see OSF Material. Each research question consists of three hypotheses (a-b). These hypotheses build upon each other. For example, if an effect in (a) is not found, (b) and (c) do not need to be tested, as it was already shown that no mediator effect was found based on the collected data. This mediator effect is first analyzed in regards to CG and UG (RQ3: H3 (a-c)) and second, for the groups UG and GG (RQ4: H4 (a-c)) and CG and GG. The following hypotheses were formulated:

> **RQ3:** Does the users' capability attribution mediate the effect of combined uncertainty representations on team performance? *(H3 (a-c))*

H3/4 (a) there is an effect of the group on capability attribution (directional: UG > CG),
H3 (b) there is an effect of capability attribution on the performance, and finally
H3 (c) the effect of the group (CG, UG) on the performance becomes smaller when capability attribution is added to the linear model.

> **RQ4:** Does the users' capability attribution mediate the effect of capability-focused guidance supplementing combined uncertainty representations on team performance? *(H4 (a-c))*

H3/4 (a) there is an effect of the group on capability attribution (directional: GG > UG),
H4 (b) there is an effect of capability attribution on the performance, and finally
H4 (c) the effect of the group (UG, GG) on the performance becomes smaller when capability attribution is added to the linear model.

We measured the user's mental model of interest, i.e., *capability attribution*, with subjective scales [23, 142, 190]. The use of qualitative response formats such as suitable scales could help explore statistical effects and reveal why some approaches lead to mixed results (e.g., [142]). Compared to open text formats and interviews, which can shed light on mental models and users' understanding [77, 90], a quantifiable scale format has the advantage that statistical inferences can be analyzed. This is necessary to conduct a mediator analysis. We calculated the mediator variable capability attribution based on the assessment of single cases after the main experiment without showing the AI recommendation. Previous studies collected responses for similar subjective measures for each case during the main experiment [29, 142, 166], which may lead to the proxy task effect [20]. In a more realistic setup, it is unlikely that such a question would be asked for each case. The participants were asked "How well do you think the AI is able to decide whether this is a mild or severe case of nail fungus?" for two images on a continuous scale ranging from 0 ("Not capable at all") to 100 ("Very capable"). Similar questions have been used in previous studies, which found that providing users with additional information improved human AI performance (e.g., as "task appraisal" [142], "instance-specific AI and self-efficacy" [166]). For example, Pinski et al. [142] asked whether "AI is suited to solve this exercise." to collect "AI-fit appraisal". In our experiment, the participants had to assess two nail images of moderate difficulty, stating how capable the AI is to classify them correctly. One image was of high quality, which should be optimal for the AI, another one was altered to be less clear and rotated similarly to the images that were wrongly classified during the previous trials. The capability attribution is the difference between the capability attribution of the easy minus the hard image to assess. For example, if a participant assigns the AI the value 85 (leaning to "Very capable") to assess the easy-to-assess image correctly and 45 to the hard-to-assess image (leaning toward "Not capable at all"), their value would be 40.

4 RESULTS

In the following, the demographics of the sample and the comprehension checks are analyzed. Afterward, the results of the hypotheses regarding uncertainty, guidance, and the assumed mediator effect are presented. At the end, exploratory findings are described.

Demographics	Characteristics	CG (n = 48)	UG (n = 48)	GG (n = 48)
Age	In years	29.0 (8.5)	28.1 (8.2)	28.4 (9.1)
English skills	Intermediate	4 / 48 (8.3%)	4 / 48 (8.3%)	0 / 48 (0%)
	Advanced	16 / 48 (33%)	16 / 48 (33%)	19 / 48 (40%)
	Fluent	28 / 48 (58%)	28 / 48 (58%)	29 / 48 (60%)
Education	High School or lower	5 / 48 (10%)	11 / 48 (23%)	10 / 48 (21%)
	Some graduate school or equivalent	8 / 48 (17%)	3 / 48 (6.3%)	6 / 48 (13%)
	Bachelor's Degree or equivalent	26 / 48 (54%)	21 / 48 (44%)	23 / 48 (48%)
	Master's Degree or equivalent	8 / 48 (17%)	9 / 48 (19%)	8 / 48 (17%)
	Others	1 / 48 (2.1%)	4 / 48 (8.3%)	1 / 48 (2.1%)
Prior knowledge	AI	20 / 48 (42%)	17 / 48 (35%)	24 / 48 (50%)
	Psychology	13 / 48 (27%)	7 / 48 (15%)	10 / 48 (21%)
	Medicine	4 / 48 (8.3%)	12 / 48 (25%)	6 / 48 (13%)
	Pharmacy	7 / 48 (15%)	4 / 48 (8.3%)	2 / 48 (4.2%)

Table 2. Demographic information of the sample based on subjective statements of the participants per group.

4.1 Demographics and Comprehension Checks

Table 2 presents relevant demographic information regarding age, English proficiency, education, and prior knowledge. There was no significant difference between the groups for age (one-way ANOVA: F(2,141)=0.15, p=0.86, η^2=0.002). On average, participants spent 28 minutes (GG: 29 min, UG: 30 min, CG: 25 min) to complete the study. The most common geographic region from which participants were recruited was Europe (65%), followed by Africa (25%), and others, including North America and Asia (10%), which were similarly distributed across the three groups.

To make sure that the participants understood the instructions, we asked if they understood the rules, to which they all replied "yes". In addition, it was checked whether the UG and the GG participants understood the uncertainty representations. Nearly all participants matched the correct icon array to a presented case with an uncertainty sentence. After the main experiment, they were asked a similar question asking what the icon array stood for, with the correct answer being "How many images like this the AI would classify as mild or severe". 67% of the UG and 75% of the GG chose the correct statement. The GG was also asked to answer a multiple-choice question about what the strengths of humans compared to AI are. This question was used to motivate the participants to read the provided capabilities thoroughly. The chosen human skills show that the participants leaned towards choosing the options compatible with the given capability information, see Table 3.

There was no significant difference between the groups for subjective workload (one-way ANOVA: F(2,141)=1.18, p=0.31, η^2=0.016), and need for cognition (one-way ANOVA: F(2,141)=1.02, p=0.34, η^2=0.015), see Table 3. As visualized in Appendix C, participants correctly solving the numeracy test in the CG and GG had a better overall performance, whereas the opposite is the case for the UG.

Variable	Characteristics	CG (n = 48)	UG (n = 48)	GG (n = 48)
Comprehension uncertainty representation	Correctly matching icon array to an uncertainty sentence	-	46 / 48 (95.8%)	48 / 48 (100%)
Perceived meaning of uncertainty icon array	"Similarity of the image to the training data"	-	0 / 48 (0%)	4 / 48 (8.3%)
	"How many images like this the AI would classify as mild or severe"	-	32 / 48 (67%)	36 / 48 (75%)
	"Amount of criteria met for a specific recommendation"	-	16 / 48 (33%)	8 / 48 (17%)
Comprehension of guidance (Selecting statement as human skill)	"Adapting to new contexts"	-	-	40 / 48 (83.3%)
	"Identifying poor quality images"	-	-	33 / 48 (68.8%)
	"Unaffected by street noise"	-	-	5 / 48 (10.4%)
	"Recognizing nail fungus from experience"	-	-	12 / 48 (25%)
	"Able to recognize new patterns"	-	-	25 / 48 (52%)
Berlin numeracy test	Correct answer	25 / 48 (52%)	19 / 48 (40%)	29 / 48 (60%)
Subjective Workload (NASA TLX)	Range from 0 "very low" to 100 "very high"	26.6 (13.5)	28.9 (13.8)	31.4 (17.8)
Need for cognition (NFC-10)	Range from 0 "very low" to 100 "very high"	60.3 (22.1)	62.8 (16.4)	66.0 (17.8)

Table 3. Amount of participants answering comprehension checks correctly, and their numeracy, subjective workload, and need for cognition score.

4.2 Effect of combined uncertainty representations on team performance

The performance for the three groups for the Human-AI phase is shown in Table 4 in the second column block. Even though the mean overall performance of the UG was higher, the one-sided t-tests on the linear model coefficients revealed that the UG (t(141)=1.50, p=0.068, $d_\Psi < 0.62$[24]) showed no significant in improvement of overall performance from the CG. Thus, the overall performance does not seem to be influenced by the uncertainty representation (H1 (a)). There was no significant difference between the UG and the CG in regards to correct self-reliance (t(141)=-0.09, p=0.535, $d_\Psi < 0.62$). More details are summarized in Table 5, and additional non-parametric analyses leading to the same findings are collected in Appendix E. **Thus, neither the overall performance (H1 (a)) nor the correct self-reliance (H1 (b)) seem to be improved by combined uncertainty representations.**

	CG	UG	GG	Order of means
Overall performance (cases: TN, TP, FP, FN)	78.5 (10.3)	81.4 (8.4)	82.7 (9.3)	AI <CG <UG <GG
Correct AI-reliance (cases: TN, TP)	86.6 (11.4)	90.6 (8.0)	88.0 (8.6)	CG <GG <UG
Correct self-reliance (cases: FP, FN)	54.2 (28.4)	53.6 (30.1)	66.7 (29.3)	UG <CG <GG
Capability attribution	10.5 (18.7)	11.6 (19.2)	10.1 (21.6)	GG <CG <UG

Table 4. Descriptive statistics of groups (CG, UG, and GG) of the human-AI phase in the format M (SD). The highest mean value per measure and group is marked gray for the human-AI phase. See Appendix F for differences of the human baseline.

[24] Always interpret effect size d_Ψ in terms of the linear combination of the model parameters. Cohen's d conventions do not apply in this case [91].

RQ1: Can combined uncertainty representations improve team performance?					
Hypotheses		Parametric testing (GLM: unpaired t-test)			
H1 (a)	Uncertainty representations UG lead to improved mean overall performance compared to no representations CG .	t(141)=1.50	p=0.068	$\widehat{d_\Psi}$ = 0.31	$1-\beta$ = 0.44
H1 (b)	Uncertainty representations UG lead to improved mean correct self-reliance compared to no representations CG .	t(141)=-0.09	p=0.535	$\widehat{d_\Psi}$ = -0.02	$1-\beta$ = 0.04

RQ2: Can team performance be improved by providing capability-focused guidance in addition to combined uncertainty representations?					
Hypotheses		Parametric testing (GLM: unpaired t-test)			
H2 (a)	Additional guidance GG leads to improved mean overall performance compared to only uncertainty representations UG .	t(141)=0.68	p=0.249	$\widehat{d_\Psi}$ = 0.14	$1-\beta$ = 0.17
H2 (b)	Additional guidance GG leads to improved mean correct self-reliance compared to only uncertainty representations UG .	t(141)=2.18	**p=0.015**	$\widehat{d_\Psi}$ = 0.44	$1-\beta$ = 0.7
Additional	Guidance in addition to uncertainty representations GG leads to improved mean overall performance compared to no representations CG .	t(141)=2.18	**p=0.016**	$\widehat{d_\Psi}$ = 0.44	$1-\beta$ = 0.7
Additional	Guidance in addition to uncertainty representations GG leads to improved mean correct self-reliance compared to no representations CG .	t(141)=2.09	**p=0.019**	$\widehat{d_\Psi}$ = 0.43	$1-\beta$ = 0.67

Table 5. Hypotheses testing for the research question 1 and 2, and Hypotheses H1 (a - b) and H2 (a - b). As the residuals are not optimally normal distributed and observations were tied, non-parametric tests were conducted in addition, see Appendix E.

4.3 Effect of capability-focused guidance on team performance

The performance for the three groups is shown in Table 4 in the second column block. The one-sided t-test on the linear model coefficients revealed that the GG performed significantly better than the CG (t(141)=2.18, p=0.016, $\widehat{d_\Psi}$ = 0.44, $1-\beta$ = 0.7), but the GG showed no significant difference from the UG (t(141)=0.68, p=0.249, d_Ψ < 0.62). **Even though the hypothesis that capability-focused guidance leads to improved overall performance in comparison to only uncertainty representations (H2 (a)) must be rejected, the combination of capability-focused guidance and uncertainty representations led to significantly better overall performance in comparison to the CG.**

Regarding, the variable correct self-reliance, there was a significant difference between the GG and the CG (t(141)=2.09, p=0.019, $\widehat{d_\Psi}$ = 0.43, $1-\beta$ = 0.67) and the GG showed a significant difference from the UG (t(141)=2.18, p=0.015, $\widehat{d_\Psi}$ = 0.44, $1-\beta$ = 0.7). Non-parametric testing supporting these results and more details are summarized in Appendix E. **This confirms the hypothesis that additional capability-focused guidance leads to improved correct self-reliance in comparison to only uncertainty representations given (H2 (b)) and in comparison to neither uncertainty representations nor capability-focused guidance given.**

4.4 Capability attribution as a mediator between the approaches and performance

The proportion of variance explained did not significantly differ from zero (one-way ANOVA: F(2,141)=0.076, p=0.93, R^2=0.01; Kruskal-Wallis rank sum test: χ^2=0.06, p=0.97), therefore no significant difference between the groups regarding capability attribution was found. Additionally, capability attribution showed no significant correlation (Pearson's product-moment correlation) with overall performance (r=-0.08, p=0.37) or correct self-reliance (r=0.004, p=0.96). A linear model with capability attribution as criterion and the groups as predictors was fitted, see OSF Material. No significant difference of capability attribution between the UG and the CG for H3 (a) (t(141)=0.28, p=0.39, d_Ψ < 0.62) or the UG and the GG for H4 (a) (t(141)=-0.38, p=0.65, d_Ψ < 0.62) was found. Therefore, adding capability attribution to the linear model for analysing the group effects did not result in significant changes in the group differences regarding the performance (H3 (b-c), H4 (b-c)). **Consequently, it cannot be confirmed that the mental model variable capability attribution mediates the effect of uncertainty representations (H3) or capability-focused guidance (H4) on team performance.**

4.5 Exploratory data analyses

In addition to hypothesis testing, we conducted exploratory data analyses. First, we explored if human-AI teams outperformed the AI. Second, additional mental model variables are investigated. Third, we looked at performance differences based on case difficulty and case type.

4.5.1 Comparing human-AI team performance with human and AI performance. As Vaccaro et al. [172] noted in their meta-analysis, it is of great interest to compare the performance of the human-AI teams with the human baseline and the AI performance. Although our focus was on comparing interventions, we exploratively tested whether the human-AI teams outperformed the AI, i.e., simulated accuracy of 75%, or humans, i.e., participants' performance in the human baseline without AI assistance consisting of 6 trials[25] (see Section 3.3). We compared the human-AI team performance with the simulated AI accuracy of 75% by using a one-sided, one-sample Wilcoxon signed rank test for possibly tied observations and hypothesized, that team performance is greater than the AI alone per group. We found all groups to significantly outperform the performance of 75% of the AI[26] (CG: V=590, p<0.01; UG: V=687, p<0.001; GG=803, p<0.001). As the mean performance in the human baseline for all groups was higher than the overall team performance[27], indicating an effect in the opposite direction, no hypothesis testing was needed. Human-AI team performance did not outperform the human baseline performance.

[25] Calculated as percentage of correctly classified cases in the human baseline.
[26] The R package "exactRankTests" was used to consider tied observations.
[27] M(SD) of the human baseline - CG: 78.8(15.3), UG: 85.4(13.1), GG: 83.3(14.2). Group differences were tested, see Appendix F.

Variable	Description	CG	UG	GG
Capability attribution to the AI	"How well do you think the AI is able to decide whether this is a mild or severe case of nail fungus?" (0 "Not capable at all" to 100 "Very capable") Calculation: $\Delta = CA_{\text{easy}} - CA_{\text{hard}}$	10.5(18.7)	11.6(19.2)	10.1(21.6)
	CA_{easy}: Capability attribution the user attributes to the AI for an easy for the AI to-assess case (mild case). Question see above.	86.4(13.5)	87.5(14.3)	83.6(14.0)
	CA_{hard}: Capability attribution the user attributes to the AI for a hard for the AI to-assess case (severe case). Question see above.	75.9(18.5)	75.9(18.0)	73.4(17.7)
Capability attribution to oneself	Capability attribution users attribute to themselves for the easy to judge case (mild case).	87.4(15.6)	86.7(16.5)	86.6(13.0)
	Capability attribution user attribute to themselves for the hard to judge case (severe case).	78.8(15.6)	76.6(16.5)	81.3(13.0)
Perceived AI accuracy	"How often do you think the AI's recommendation is accurate?" (0 "Never" to 100 ""All the time")	73.7(13.5)	73.7(14.2)	71.8(14.4)
Perceived AI error tendency	"What mistakes do you think the AI was prone to?" (0 "incorrectly recommending mild" to 100 "incorrectly recommending severe")	51.6(13.5)	55.8(14.2)	56.8(14.4)
Self tendency	"If you were unsure whether the case was mild or severe, which decision would you lean toward?" (0 "Mild" to 100 ""Severe")	47.9(29.3)	52.8(23.5)	46.5(22.4)
Perceived similar reasoning	"How similar do you think is the AI's reasoning to your own reasoning?" (0 "Very different" to 100 "Very similar")	59.5(23.6)	61.8(21.3)	62(17.2)
Self confidence	"How confident were you in the recommendations you made to the patients?". (0 "Not confident at all" to 100 "Very confident")			
	Question above again after the human baseline.	71.5(21.0)	67.5(19.2)	71.0(21.4)
	Question above again after the human-AI phase.	75.7(19.5)	74.2(16.4)	76.3(13.5)
Helpfulness	All items were ranked from 0 "Not helpful at all" to 100 "Very helpful" ...			
	... the nail image.	89.6(16.3)	86.8(16.0)	92.4(10.4)
	... the AI's recommendation.	70.1(23.1)	71.0(19.8)	70.8 (19.5)
	... the instructions and rules.	93.8(9.9)	86.9(15.9)	84.7(17.1)
	... the AI's confidence sentences.	-	67.3(18.3)	67.6(22.3)
	... the AI's confidence visualization.	-	71.0(22.0)	70.0(23.2)
	... the listed strengths and limits of the AI and humans.	-	-	57.3 (27.2)

Table 6. Descriptive summary of collected mental model variables per group. The format is in M(SD) as all measures are metric.

4.5.2 Additional mental model variables. In addition to capability attribution, we collected additional factors which may give an insight into the users' mental models of the AI system. Their descriptive values and variable descriptions are

summarized in Table 6. In this section, we will highlight interesting findings. First, on average, all groups underestimated AI accuracy (75%), i.e., *perceived AI accuracy*. In addition, the participants were asked to state their *perceived AI error tendency*. All participants reported that the AI was generally slightly more likely to recommend incorrectly classifying mild cases as severe as the other way around. This bias was most pronounced in the GG, followed by the UG, and lastly the CG. Furthermore, we asked a question to get an insight into the perceived similarity of the AI's reasoning to one own's, i.e., *perceived similar reasoning*, which was on average leaning towards "very similar". Moreover, participants rated the helpfulness of the UI elements. All groups rated the nail image and rules as most helpful, followed by the AI recommendation. The UG and the GG rated the uncertainty representations as similarly helpful as the AI recommendation. The lowest value was attributed to the capability-focused guidance by the GG.

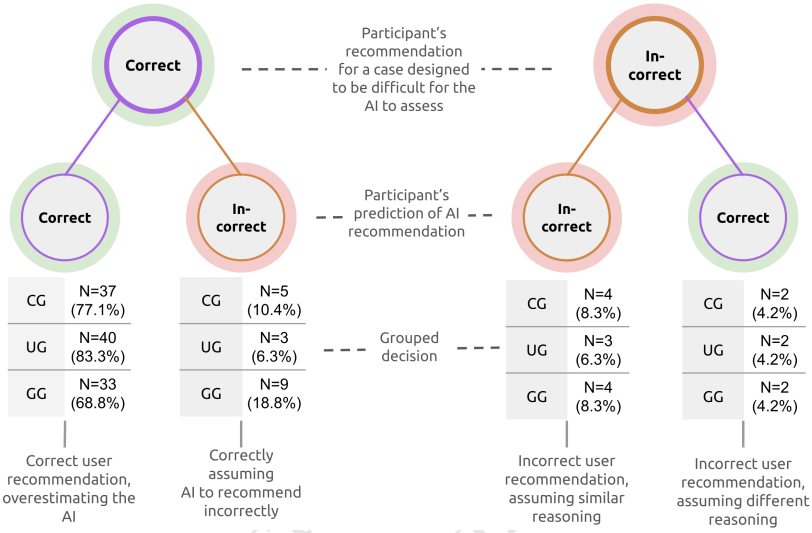


Fig. 5. Amount of participants' recommendations and predictions of the AI's recommendation for a severe case with a low-quality image. The color purple represents deciding for "severe", and orange for "mild".

Additionally, we asked the participants to *predict what the AI would recommend* and what they would choose for the two cases on which the capability attribution was calculated. One case showed a rotated, altered image of a severe case, which should be more difficult for the AI to assess correctly. The other one showed a clear image of a mild case, which should be easier for the AI to assess correctly. The proportions of recommendations and predictions made per group are summarized in Appendix G. Despite the UG having the highest capability attribution, see Table 4 last row, in the UG only 6% correctly recommended "severe" and predicted the AI to state "mild" for the hard-to-assess case, in the GG 19% and in the CG 10%, see Figure 5.

4.5.3 The effect of the intervention on correct decision-making depending on case difficulty and case type. Figure 6 and Figure 7 show the percentage of the average number of correct human decisions depending on the case type, intervention group, and case difficulty. The simulated uncertainty value in the UG and the GG matched the difficulty of the cases (see OSF Material). In addition, we distinguish between case types for which the AI correctly made a mild (True Negative/TN) or severe (True Positive/TP) recommendation, or incorrectly made a mild (False Negative/FN) or severe (False Positive/FP) recommendation.

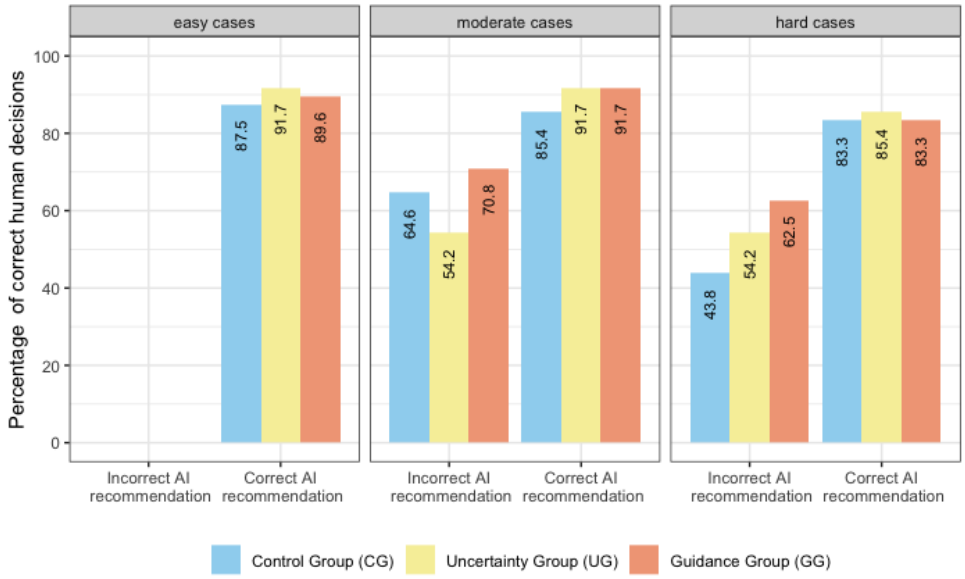


Fig. 6. Percentage of the average number of correct human decisions (in %) per group depending on the case type (correct, incorrect) and the case difficulty (easy, moderate, and hard cases).

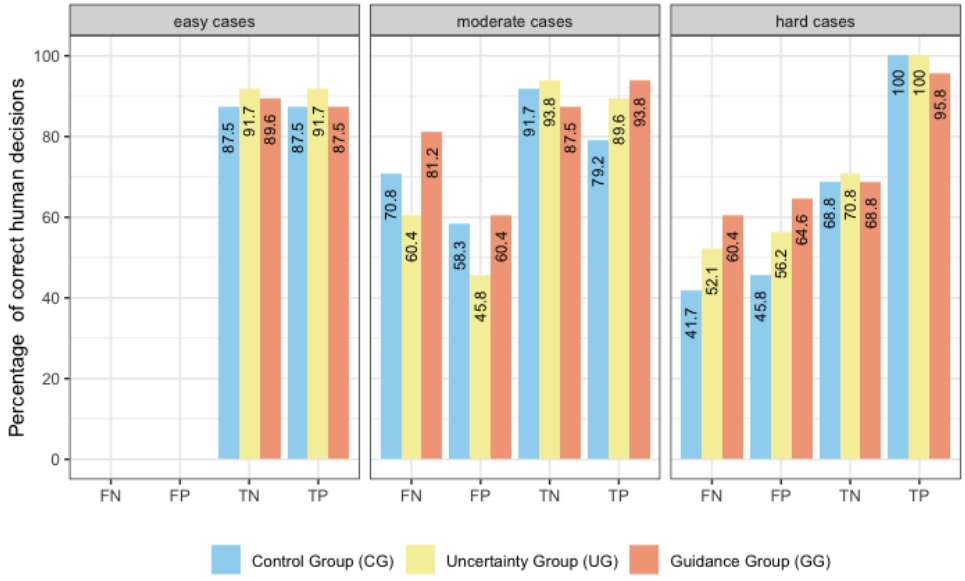


Fig. 7. Percentage of the average number of correct human decisions (in %) per group depending on the case type (FN, FP, TN, TP) and the case difficulty (easy, moderate, and hard cases).

Overall, participants performed best on easy cases and better on moderate cases than on hard cases, indicating that the task design worked as intended. In Figure 6 we see, that for hard cases where the AI's uncertainty indicated very low confidence and provided a wrong recommendation, the correct self-reliance of the GG was higher than of the UG, and the CG's correct self-reliance was the lowest. Therefore, it seems that wrong AI recommendations were handled better with uncertainty than without for hard cases, and best with guidance. Cases with correct AI recommendations, i.e., correct AI-reliance, seem to be handled similarly between the groups. Interestingly, all groups accepted the correct severe recommendation (TP) for the hard cases more than the moderate cases, and in turn, cases with correct mild recommendations (TN) were less followed for hard than moderate cases, see Figure 7. For hard cases, the number of participants correcting the false severe recommendation (FP) was higher across all groups in comparison to correcting the false mild case. When summarizing the correct and incorrect AI recommendations, it becomes apparent that the GG performed better overall, see Figure 6. Interestingly, the UG performed better than the CG for hard, but not for moderate cases.

5 DISCUSSION

Our study investigated how uncertainty representations with additional capability-focused guidance affect users' mental models of the AI and, thus, team performance. Our between-subjects user study yielded three main findings. In the following section, we present and interpret our key findings. Then, we provide practical implications derived based on our findings. Finally, we discuss the generalizability of our results and limitations.

5.1 Key findings

5.1.1 Technical approaches alone did not improve performance. Prior research has indicated that a greater focus on user needs in the design of technical approaches could address current challenges, such as over-reliance or a lack of understandability, and thereby improve team performance. We, therefore, combined uncertainty representations for which research has shown that they are likely to be suitable for laypeople, taking into account familiarity with the format [30] and low numeracy skills [16, 29, 146]. **However, regarding RQ1, we found that uncertainty representations alone did not significantly improve team performance.** Thus, our results support previous findings indicating that uncertainty information does not consistently improve team performance [172].

> **#K1 Key finding:** Combined uncertainty representations did not increase human-AI team performance. Thus, even uncertainty representations designed focusing on human factors do not consistently improve team performance.

Previous studies have identified human factors, such as users' numeracy skills, that influence human-AI collaboration and decision making. **Surprisingly, these factors, i.e., user's numeracy skills, format familiarity, need for cognition, and cognitive overload, did not explain why uncertainty representations did not improve team performance in our experiment.** We measured *objective numeracy* [40], as numeracy skills might influence how well the uncertainty representations are understood. More participants in the uncertainty group answered the numeracy test incorrectly than in the control group, but the participants who answered incorrectly surprisingly performed better overall. Furthermore, because *unfamiliarity* was assumed to hinder the utilization of uncertainty representations [30, 110], we measured familiarity indirectly with comprehension checks. Of the groups provided with uncertainty representations, most participants answered the comprehension questions correctly (see Section 4.1). In addition, prior research has found that the stable personality trait *need for cognition*, i.e., a users' drive to engage in demanding mental activities, may affect how well approaches improve human-AI collaboration [21, 36]. A lower need for cognition in the uncertainty

group could have explained the lack of improvement in team performance. Yet, we found that the uncertainty group had a higher average need for cognition than the control group. Furthermore, stress-inducing factors are known to affect decision-making [140]. For example, *cognitive overload* is known to affect human-AI collaboration [23, 85, 181], which we measured in the form of subjective workload (NASA TLX) [150]. We found that the NASA TLX score to be higher with uncertainty representations than without and the highest in the guidance group, albeit not significantly. Still, the guidance group with the highest subjective workload did outperform the other groups.

> **#K2 Key finding:** Known factors such as users' numeracy skills, familiarity with the format, need for cognition, and cognitive overload do not explain our findings.

Instead, we identified another stress-inducing factor, described by Phillips-Wren and Adya [140] as "task uncertainty", which may explain our results. "Task uncertainty" is defined as "an inadequate availability of knowledge about a situation requiring action or resolution", which can cause indecisiveness and disrupt rational decision-making [140]. Similar to technically unreliable explanations reducing trust [34], the awareness of the AI's uncertainty may have led to indecisiveness and worsened performance [140] in the uncertainty group. **We assume that users may have tended to over-rely due to the perceived "uncertain" decision-making situation, especially for cases of high uncertainty.** These findings align with previous research finding that unreliable predictions can lead to increased self-doubt [32] and over-reliance [174], and that uncertainty can lead to indecisiveness and impaired decision-making [143, 145]. Since additional guidance improved team performance (see Section 5.1.2), we assume that only showing uncertainty representations may not be actionable as hypothesized in Section 2.2. Nevertheless, we discuss general study limitations in Section 5.3, which may also contribute to the non-significance of our results.

> **#K3 Key finding:** User indecisiveness due to *task uncertainty* may explain why uncertainty representations did not improve team performance, highlighting the necessity to prioritize user needs over technological availability.

5.1.2 Additional educational approaches led to improved performance. Based on previous studies, we hypothesized that educating users about their own and the AI's capabilities could lead to improved team performance [142]. Our educational approach was designed to help users utilize the uncertainty representations and infer why the AI is more or less certain, especially when the AI communicates high uncertainty. **With respect to RQ2, we found that capability-focused guidance, in addition to uncertainty representations, can reduce over-reliance.** Overall performance was also improved by capability-focused guidance, although only significantly compared to the control group, which may be due to our study limitations (see Section 5.3.3). Our findings support previous research using educational approaches focusing on different metrics such as task delegation [142] and performance [32, 98, 101]. Thus, educational approaches seem to have a positive effect on human-AI collaboration in both low-stakes and high-stakes scenarios.

> **#K4 Key finding:** Capability-focused guidance in addition to uncertainty representations can improve team performance, i.e., correct self-reliance. Hence, educational approaches appear to address the shortcomings of technical approaches in high-stakes scenarios.

5.1.3 Complex relationship of users' mental models on performance. In addition to performance measures, we investigated whether the approaches used influenced the user's mental model, as this may, in turn, affect performance. **The mediator hypotheses RQ3 and RQ4, assuming that capability attribution mediates the effect of the uncertainty representations and additional capability-focused guidance on team performance, could not be**

confirmed. Our findings are in contrast with previous research suggesting that AI knowledge can align delegation behavior with task appraisal, i.e., measured by the question "AI is suited to solve this exercise", and increase performance [142]. Therefore, it is crucial to contextualize our results within the framework of our scenario and task design (see Section 3.1). Compared to previous work [142], we did not focus on task delegation, performed a mediator analysis, and used a high-stakes scenario. We specifically designed two cases representing images that an AI would have more or less difficulty predicting to measure capability attribution after the experiment and to avoid the proxy task effect [20]. Future studies may utilize more trials after the main human-AI phase to still avoid the proxy task effect and get a richer understanding of users' capability attribution. Furthermore, we utilized a high-stakes human-AI collaboration scenario, which may have affected users' capability attribution differently. Pinski et al. [142] only focused on classifying general objects, such as animals, without specific decision rules. Our scenario may be of higher task complexity and more uncertain, which can lead to over-reliance [152] and indecisiveness (see Section 5.1.1). Thus, our results suggest that while some findings may be transferable from low-stakes to more complex, high-stakes scenarios (see Section 5.1.2), others, such as users' mental models, may not be transferable and thus require further investigation.

> **#K5 Key finding:** Our results did not suggest a mediating role of the user's mental model of the AI's capabilities in a high-stakes decision-making scenario.

Capability attribution was the lowest (most incorrect) when capability-focused guidance was provided; however, participants who received guidance were better at predicting whether the AI recommendation would be incorrect for a given case (see Section 4.5.2). This aligns with the observation that providing capability-focused guidance increased correct self-reliance and raises the question of whether our educational approach led users to recognize which cases the AI was capable of handling without them being aware of their knowledge of the AI's capabilities. This assumption is supported by the fact that participants rated the provided capability-focused guidance the least helpful compared to other information, even though it improved appropriate reliance (see Table 6). Cassenti et al. [30] found a similar tension, with participants preferring the approach they performed best with the least. Thus, further research is needed to explore this tension and to capture users' mental models (see Section 5.2.3).

> **#K6 Key finding:** Users mental model of the AI's capabilities and what they assume the AI predicts does not seem to be aligned.

5.1.4 Influence of case difficulty and case type on team performance. We identified additional factors that influence human-AI collaboration, namely applied rules, case type[28], and case difficulty[29] (see Section 4.5.3). We noticed that the team performance for the hard, severe cases (TP) was the highest at nearly 100% across all groups. To understand why, we looked at the specific case shown and noticed that this particular nail fungus case (TP) had an affected nail matrix. We assume that this rule[30] may have been easier to apply for the participants than the first rule regarding the percentage of the nail affected. Furthermore, we found that for cases with incorrect AI recommendations, providing additional capability-focused guidance led to the highest team performance across all case difficulty levels. Interestingly, only providing uncertainty representations outperformed the control group for hard, but not for moderate cases (see Figure 6). **As assumed, without guidance, users seemed to interpret uncertainty to their disadvantage,**

[28] FP = False positive, FN = false negative, TP = True positive, and TN = true negative.
[29] We differentiate between task difficulty and case difficulty. Task difficulty refers to the type of task, i.e., depending on the goal, task type, scenarios, and more. In our experiment, case difficulty refers to a specific instance, i.e., if users have to assess three patient images each of those images is referred to as a case. The first case, for example, might be straightforward (easy case difficulty), whereas the second case might be more complex (hard case difficulty).
[30] "It is a severe nail fungus case if the nail matrix is affected."

specifically for moderate cases[31]**.** For example, if users saw a moderate case with a less confident uncertainty value (e.g., 67%), they may still have perceived that value as high, leading them to follow the AI's recommendation. However, when the AI showed a low uncertainty percentage (e.g., 54%) for hard cases, users seemed to give the recommendation appropriately less weight. Previous research has already suggested that case difficulty [28, 154] and the value of uncertainty percentages [11, 22] may influence human-AI collaboration. For example, Bussone et al. [22] used a healthcare diagnosis scenario and found that users' needs for additional explanations appeared to increase when the uncertainty value was low or unexpected and that users slightly over-relied when it was high [22]. Furthermore, studies suggest that participants have an uncertainty threshold, below or above, which they may distrust or trust the AI (e.g., for sentiment classification of reviews [11], clinical decision-making [25]).

> **#K7 Key finding:** Case type and case difficulty influenced team performance, highlighting the shortcomings of displaying less confident (mid-range) uncertainty values.

5.2 Practical design implications and future research

5.2.1 Adaptive approaches based on uncertainty values and human factors. Our main findings suggest that even when known factors are taken into account, uncertainty representations may not improve team performance (**#K1**, **#K2**), but may hinder rational decision-making and lead to indecisiveness (**#K3**). Rassin et al. [145] state that individual differences play a role in whether a user experiences a decision as difficult to assess. Hence, a one-size-fits-all approach may not be appropriate. We recommend utilizing measurements such as the Indecisiveness Scale (IS) [145]. Based on the user's general indecisiveness, it could be decided whether to show uncertainty presentations at all or only over specified thresholds. In addition, human factors such as numeracy (**#K2**) may be relevant to the design of individualized interfaces, although we did not find them to influence our results. We suggest to study them further with a direct focus (e.g., with quota sampling and a balanced study design) rather than as an add-on to determine their relevance.

> **#P1 Practical implication:** Individualizing whether to show uncertainty representations based on users' indecisiveness could improve team performance. These and other human factors should be primarily investigated in order to utilize them similarly.

Key finding **#K7** shows that case difficulty and specific uncertainty values influence users' reliance. This finding is supported by prior research suggesting that users may have uncertainty thresholds for trusting the AI [11, 25]. Showing low confidence values (near 50%) may indicate more clearly to not rely on the AI and high confidence (near 100%) to trust the AI. Whereas mid-range values may be less actionable. Thus, we recommend dynamically displaying uncertainty representations depending on the value. For example, the uncertainty representations could only be shown if the value is not in the mid-range, i.e., very high (near 100%) or low (near 50%), not in between. Future studies are needed to test whether this interactive approach improves team performance, and if **#P1** and **#P2** can be combined.

> **#P2 Practical implication:** An adaptive approach of not showing uncertainty representations when the certainty is neither clearly low or high may increase team performance.

5.2.2 Integration and adaptation of capability-focused guidance for a wider range of use. We found capability-focused guidance to overcome the challenges of uncertainty representations (**#K4**). As uncertainty representations are used

[31] We matched uncertainty values to case difficulty, showing that the easier the case, the better the performance (see Section 4.5.3).

across many domains and in various formats such as predicting and classifying diagnoses [89, 106], the military [110] and more [100], we highly recommend to utilize and further study educating users directly about the AI system they interact with. Other decisions-making scenarios may require adjustments. For example, we derived our capability-based guidance specifically for an image classification scenario (see Section 2.2.2), which is relevant for human-AI collaboration scenarios in healthcare (e.g., [7, 24, 29, 89]). For other data types and scenarios (e.g., recidivism [63] or diabetes prediction [3, 125]), guiding information has to be redesigned to include limitations and strengths of such AI systems. For more complex scenarios, i.e., generative AI [57], more research is needed to understand human and AI capabilities. However, studies already indicate that users need transparency in the reasoning of generative AI [184][32].

> **#P3 Practical implication:** Depending on the AI system and task, we recommend adapting our approach by formulating task-specific AI and user capabilities. We focused on image classification; future research may focus on text-based data or even GenAI.

Furthermore, Abdel-Karim et al. [1] studied whether AI-based assistance can induce reflection. They found that higher uncertainty in users' initial judgment and perceiving a conflict with the AI recommendation can lead to deeper reflections. We expect that in situations where the uncertainty value is ambiguous, users may experience a tension that causes them to reflect. Therefore, we suggest that when uncertainty values are in a range less actionable for users (e.g., mid-range, see **#K3**), leading them to be more indecisive, guidance may be needed the most. Thus, we suggest using additional reflection prompts, i.e., pop-up alerts to remind users to consider the additional information (e.g., capability-focused guidance). Further studies are needed to examine the feasibility of combining this with the above recommendations (**#P1, #P2**).

> **#P4 Practical implication:** We recommend providing users with information about both their own and the AI's capabilities. Specifically, we suggest prompting users to consider this guiding information depending on the uncertainty value.

5.2.3 Investigating and utilizing users' mental model of AI's capabilities. Our key findings **#K5** and **#K6** highlight the challenges of investigating users' mental models of the AI's capabilities. Based on our findings, we recommend exploring new approaches to study users' mental models, i.e., asking users to sort cases of various difficulties (e.g., nail fungus images) based on how reliably the AI would handle them correctly. This could lead to more nuanced knowledge about the user's mental model, as it can combine attributing capability and predicting what the AI would recommend. Afterward, users could be informed about what the AI would predict for these cases and with which uncertainty value. Therefore, this approach could be used to train users or to assess and study their mental model after a first interaction.

> **#P5 Practical implication:** Exploring new methods to investigate users' mental models such as sorting cases (e.g., nail fungus images) based on how reliable the AI can handle them correctly may lead to more nuanced insights.

Furthermore, we found case characteristics such as the decision rules relevant to a specific case to influence team performance (**#K7**). We recommend future studies to report in detail how cases are designed and whether their design may impact the findings to avoid systematic errors. Especially in controlled, Wizard-of-Oz experiments or studies using a limited number of trials, this may help to explain findings. For example, we suggest observing whether the AI's

[32] Yan et al. [184] investigated how users interact with GenAI in qualitative research (thematic analysis).

recommendations correlate with specific decision rules for the chosen cases in the experiment. Such a correlation could affect how users interact with the AI and thus affect the results. In practice, it could help to understand whether the rules actually correlate with incorrect AI recommendations, as this knowledge could then be taught to users to create a more efficient heuristic of AI behavior.

> **#P6 Practical implication:** We recommend future studies document and assess the characteristics of the individual cases presented, as they may affect the results. For example, by informing users whether the AI's incorrect recommendations may be related to specific decision rules, users could create more efficient heuristics.

5.3 Generalizability of our results & Limitations

5.3.1 Generalizability of our high-stakes scenario. Our results may not be as applicable to low-stakes scenarios or specific user groups as they are to similar scenarios in the prominent research area of AI-assisted digitized healthcare [96]. Still, the difficulties with uncertainty representations (see Section 2.1) and the advantages of educational approaches (see Section 2.2) seem to occur across domains. Since higher task complexity and uncertainty may lead to more over-reliance [152], it may be particularly important to examine high-stakes scenarios. We assume that our findings are more generalizable to other high-stakes domains in comparison to research focusing on low-stakes domains (e.g., [78, 142]). For tasks similar to ours, i.e., with communicable rules (e.g., [29]), we deem it realistic to divide them into sub-tasks if many decision rules need to be considered (see Section 3.1). In addition, for high-stakes scenarios of higher complexity, we recommend extensive training and consider focusing on experts, as investigating laypeople may not be realistic. Nonetheless, our findings have value for such high-stakes scenarios as we utilized a consequential scenario[33] (see Section 3.1), creating a more stressful setting [140], an aspect that may affect high-stakes scenarios more than low-stakes ones. In addition, our exploration of task uncertainty and case difficulty seem to be relevant for human-AI decision-making scenarios in general [140, 145]. However, to test the generalizability of our results to other scenarios, further research is needed.

5.3.2 Generalizability and limitations of the controlled crowdsourcing approach. We utilized a simulated task in an online experiment, i.e., a controlled environment, with the crowdsourcing platform Prolific. Our results may not apply to settings where experts make decisions in-person influenced by various situational factors. However, given the rising use of AI-assistance in digital healthcare (see Section 3.1), where task outsourcing and the use of third party services[34] is becoming more relevant [64, 66, 68, 182], we deemed a crowdsourcing setting with Prolific as realistic. Since only registered users of Prolific could participate, who may have certain human characteristics, selection bias may have occured [74]. However, the selection bias in recruiting interested individuals via Prolific may be analogous to the selection bias in recruiting part-time workers, which reflects our scenario[35]. In addition, given that our sample represents users who are more likely to be recruited as part-time workers in the context of our scenario, our results may differ from samples that only include participants from the US or UK. We have discussed this cultural factor, which does not seem to be highly relevant for our controlled scenario, but should be considered if different measurements are of interest (see Appendix H). Instead, we measured factors known to be relevant to human-AI decision-making

[33] In contrast, the decisions in our scenario would have a relevant impact on individuals because they affect the health of patients.
[34] For example, labelyourdata or upwork.
[35] Prolific provides access to a diverse pool of potential part-time workers and is used for industry-specific work such as AI training. Also, Prolific's fair payment policies map more closely to potential part-time workers' payments; see Prolific advantages.

(see Section 3.4). Summarized, our findings are most suitable for controlled decision-making settings similar to ours (Section 3.1); further research is needed to investigate if our findings translate to other settings (Section 5.3.3).

5.3.3 Further limitations. We decided to carefully craft our cases using a simulated AI, enabling us to investigate influences such as case difficulty. Still, future research is needed to investigate whether less controlled, non-simulated, more complex, and unpredictable scenarios also profit from educational approaches. For example, our findings may not be directly transferable to complex in-person settings with disturbances such as the experimenter effect [149]. In addition, we only tested one combined uncertainty representation. Different visualizations (e.g., bar plots) may lead to different results [16, 81]. Furthermore, we chose to focus on two individual cases to measure capability attribution after the main human-AI phase to create a more realistic experience, which may have led to our non-significant findings (see Section 5.1.3). In addition, recent research indicates the importance of comparing a more extensive human baseline and AI accuracy with team performance [172]. As we did not design our experiment to focus on this comparison, we used a limited number of trials in the human baseline and found that the human-AI team only outperformed the AI, not the human baseline (Section 4.5.1). Finally, we recognize that our sample size may limit the generalizability of our findings. A larger sample size (e.g., [26]) or utilizing within-between design (e.g., [119]) may be able to find differences, especially in regard to the non-significant findings for uncertainty representations and the mediator effect.

6 CONCLUSION

Especially in high-stakes domains such as healthcare, it is crucial to determine if and why certain approaches can improve human-AI collaboration, as over-reliance on AI and misinterpretation of explanations remain major concerns. Thus, we conducted an empirical online user study using a simulated high-stakes healthcare scenario to investigate different approaches. To address human needs more effectively, we introduced a combination of textual and visual uncertainty representations. Additionally, we supplemented this technical approach with capability-focused guidance to educate users. Our findings show that uncertainty representations alone did not improve team performance. However, the inclusion of capability-focused guidance on top of uncertainty representations had a positive impact. Although we expected our two approaches to enhance users' understanding of the AI's capabilities, i.e., the user's mental model, our analysis did not reveal a significant effect. We examined additional factors that may have influenced our results and found task uncertainty, case difficulty, and case type to be relevant. Based on these findings, we derived practical implications and discussed how and to what extent our findings can be used in the broader CSCW field. Most notably, we found that additional educational approaches seem to be necessary to help users overcome indecisiveness, especially when uncertainty values are neither low nor high.

As AI becomes more prevalent in high-stakes domains like healthcare, we believe that the demand for understandable and human-centered approaches becomes increasingly critical. Our research takes us closer to harnessing the full potential of AI as a decision-making aid, ensuring that these technologies enhance human capabilities in an effective and actionable manner.

REFERENCES

[1] Benjamin M Abdel-Karim, Nicolas Pfeuffer, K Valerie Carl, and Oliver Hinz. 2023. How AI-Based Systems Can Induce Reflections: The Case of AI-Augmented Diagnostic Work. *Management Information Systems Quarterly* 47, 4 (2023), 1395–1424.

[2] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y. Lim. 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*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376615

[3] Shamim Ahmed, M. Shamim Kaiser, Mohammad Shahadat Hossain, and Karl Andersson. 2024. A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. *IEEE Access* (2024), 1–1. https://doi.org/10.1109/ACCESS.2024.3422319

[4] Woo-kyoung Ahn, Charles W. Kalish, Douglas L. Medin, and Susan A. Gelman. 1995. The Role of Covariation versus Mechanism Information in Causal Attribution. *Cognition* 54, 3 (March 1995), 299–352. https://doi.org/10.1016/0010-0277(94)00640-7

[5] Derek A. Albert and Daniel Smilek. 2023. Comparing Attentional Disengagement between Prolific and MTurk Samples. *Scientific Reports* 13, 1 (Nov. 2023), 20574. https://doi.org/10.1038/s41598-023-46048-5

[6] Jessica S. Ancker, Yalini Senathirajah, Rita Kukafka, and Justin B. Starren. 2006. Design Features of Graphs in Health Risk Communication: A Systematic Review. *Journal of the American Medical Informatics Association* 13, 6 (Nov. 2006), 608–618. https://doi.org/10.1197/jamia.M2115

[7] Giulia Anichini, Chiara Natali, and Federico Cabitza. 2024. Invisible to Machines: Designing AI That Supports Vision Work in Radiology. *Computer Supported Cooperative Work (CSCW)* (May 2024). https://doi.org/10.1007/s10606-024-09491-0

[8] Syed Z. Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. Investigating User Confidence for Uncertainty Presentation in Predictive Decision Making. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. ACM, Parkville VIC Australia, 352–360. https://doi.org/10.1145/2838739.2838753

[9] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[11] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445717

[12] Reuben M Baron and David A Kenny. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* 51, 6 (1986), 1173.

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

[14] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[15] Nicholas Berente, Bin Gu, Jan Recker, and Radhika Santhanam. 2021. Managing Artificial Intelligence. *MIS quarterly* 45, 3 (2021), 1433–1450.

[16] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Virtual Event USA, 401–413. https://doi.org/10.1145/3461702.3462571

[17] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–14. https://doi.org/10.1145/3173574.3173951

[18] Or Biran and Kathleen R McKeown. 2017. Human-Centric Justification of Machine Learning Predictions.. In *IJCAI*, Vol. 2017. 1461–1467.

[19] J.Robert Boston, Thomas E. Rudy, and John A. Kubinski. 1991. Multiple Statistical Comparisons: Fishing with the Right Bait. *Journal of Critical Care* 6, 4 (Dec. 1991), 211–220. https://doi.org/10.1016/0883-9441(91)90021-K

[20] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 454–464. https://doi.org/10.1145/3377325.3377498

[21] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[22] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*. IEEE, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[23] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–21. https://doi.org/10.1145/3579612

[24] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–24. https://doi.org/10.1145/3359206

[25] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-Based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–20. https://doi.org/10.1145/3544548.3580682

[26] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. 2022. Modeling Adoption of Intelligent Agents in Medical Imaging. *International Journal of Human-Computer Studies* 168 (Dec. 2022), 102922. https://doi.org/10.1016/j.ijhcs.2022.102922

[27] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How Time Pressure in Different Phases of Decision-Making Influences Human-AI Collaboration. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–26. https://doi.org/10.1145/3610068

[28] Shiye Cao and Chien-Ming Huang. 2022. Understanding User Reliance on AI in Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–23. https://doi.org/10.1145/3555572

[29] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–32. https://doi.org/10.1145/3637318

[30] Daniel N. Cassenti, Lance M. Kaplan, and Aayushi Roy. 2023. Representing Uncertainty Information from AI for Human Understanding. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 67, 1 (Oct. 2023), 21695067231193649. https://doi.org/10.1177/21695067231193649

[31] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make VQA models more predictable to a human? *arXiv preprint arXiv:1810.12366* (2018).

[32] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–32. https://doi.org/10.1145/3610219

[33] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300789

[34] Teodor Chiaburu, Frank Haußer, and Felix Bießmann. 2024. Uncertainty in XAI: Human Perception and Modeling Approaches. *Machine Learning and Knowledge Extraction* 6, 2 (May 2024), 1170–1192. https://doi.org/10.3390/make6020055

[35] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. ACM, Virtual Event United Kingdom, 120–129. https://doi.org/10.1145/3447535.3462487

[36] Francesca Chiesi, Kinga Morsanyi, Maria Anna Donati, and Caterina Primi. 2018. Applying Item Response Theory to Develop a Shortened Version of the Need for Cognition Scale. *Advances in Cognitive Psychology* 14, 3 (Sept. 2018), 75–86. https://doi.org/10.5709/acp-0240-z

[37] Yi-Te Chiu, Yu-Qian Zhu, and Jacqueline Corbett. 2021. In the Hearts and Minds of Employees: A Model of Pre-Adoptive Appraisal toward Artificial Intelligence in Organizations. *International Journal of Information Management* 60 (Oct. 2021), 102379. https://doi.org/10.1016/j.ijinfomgt.2021.102379

[38] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 307–317. https://doi.org/10.1145/3397481.3450644

[39] Miruna A Clinciu and Helen F Hastie. 2019. A survey of explainable AI terminology. In *1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence 2019*. Association for Computational Linguistics, 8–13.

[40] Edward T. Cokely, Mirta Galesic, Eric Schulz, Saima Ghazal, and Rocio Garcia-Retamero. 2012. Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making* 7, 1 (Jan. 2012), 25–47. https://doi.org/10.1017/S1930297500001819

[41] Alan F Collins, Philip Levy, Peter E Morris, and Mary M Smyth. 1994. *Cognition in Action*. Psychology Press.

[42] Dominik Dellermann, Philipp Ebel, Matthias Söllner, and Jan Marco Leimeister. 2019. Hybrid Intelligence. *Business & Information Systems Engineering* 61, 5 (Oct. 2019), 637–643. https://doi.org/10.1007/s12599-019-00595-2

[43] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 275–285. https://doi.org/10.1145/3301275.3302310

[44] Xiao Dong and Caroline C. Hayes. 2012. Uncertainty Visualizations: Helping Decision Makers Become More Aware of Uncertainty and Its Implications. *Journal of Cognitive Engineering and Decision Making* 6, 1 (March 2012), 30–56. https://doi.org/10.1177/1555343411432338

[45] D Doran. 2017. What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794* (2017).

[46] Kate Ehrlich, Susanna E. Kirk, John Patterson, Jamie C. Rasmussen, Steven I. Ross, and Daniel M. Gruen. 2011. Taking Advice from Intelligent Systems: The Double-Edged Sword of Explanations. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*. ACM, Palo Alto CA USA, 125–134. https://doi.org/10.1145/1943403.1943424

[47] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. (2021). https://doi.org/10.48550/ARXIV.2107.13509

[48] Eyad Elyan, Pattaramon Vuttipittayamongkol, Pamela Johnston, Kyle Martin, Kyle McPherson, Carlos Francisco Moreno-García, Chrisina Jayne, and Md. Mostafa Kamal Sarker. 2022. Computer Vision and Machine Learning for Medical Image Analysis: Recent Advances, Challenges, and Way Forward. *Artificial Intelligence Surgery* (2022). https://doi.org/10.20517/ais.2021.15

[49] Mica R. Endsley. 2023. Supporting Human-AI Teams:Transparency, Explainability, and Situation Awareness. *Computers in Human Behavior* 140 (March 2023), 107574. https://doi.org/10.1016/j.chb.2022.107574

[50] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me?: Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 229–239. https://doi.org/10.1145/3301275.3302265

[51] Jefferson Gomes Fernandes. 2022. Artificial Intelligence in Telemedicine. In *Artificial Intelligence in Medicine*, Niklas Lidströmer and Hutan Ashrafian (Eds.). Springer International Publishing, Cham, 1219–1227. https://doi.org/10.1007/978-3-030-64573-1_93

[52] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173718

[53] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena. 2021. *An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*. Springer International Publishing, Cham, 19–39. https://doi.org/10.1007/978-3-030-81907-1_3

[54] Susan Folkman. 2020. Stress: appraisal and coping. In *Encyclopedia of behavioral medicine*. Springer, 2177–2179.

[55] John Fox. [n. d.]. *Applied regression analysis and generalized linear models* (3 ed.).

[56] Andreas Fügener, Jörn Grahl, Alok Gupta, and Wolfgang Ketter. 2022. Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research* 33, 2 (June 2022), 678–696. https://doi.org/10.1287/isre.2021.1079

[57] Fiona Fui-Hoon Nah, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. Generative AI and ChatGPT: Applications, Challenges, and AI-human Collaboration. *Journal of Information Technology Case and Application Research* 25, 3 (July 2023), 277–304. https://doi.org/10.1080/15228053.2023.2233814

[58] Mirta Galesic. 2010. Statistical Numeracy for Health: A Cross-cultural Comparison With Probabilistic National Samples. *Archives of Internal Medicine* 170, 5 (March 2010), 462. https://doi.org/10.1001/archinternmed.2009.481

[59] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. 2009. Using Icon Arrays to Communicate Medical Risks: Overcoming Low Numeracy. *Health Psychology* 28, 2 (2009), 210–216. https://doi.org/10.1037/a0014474

[60] Baocheng Geng and Pramod K. Varshney. 2022. Human-Machine Collaboration for Smart Decision Making: Current Trends and Future Opportunities. In *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, Atlanta, GA, USA, 61–67. https://doi.org/10.1109/CIC56439.2022.00019

[61] Mahmoud Ghannoum and Nancy Isham. 2014. Fungal nail infections (onychomycosis): a never-ending story? *PLoS pathogens* 10, 6 (2014), e1004105.

[62] Gerd Gigerenzer, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, and Steven Woloshin. 2007. Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest* 8, 2 (Nov. 2007), 53–96. https://doi.org/10.1111/j.1539-6053.2008.00033.x

[63] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–25. https://doi.org/10.1145/3359280

[64] Jonathan Guo and Bin Li. 2018. The Application of Medical Artificial Intelligence Technology in Rural Areas of Developing Countries. *Health Equity* 2, 1 (Aug. 2018), 174–181. https://doi.org/10.1089/heq.2018.0037

[65] Shunan Guo, Fan Du, Sana Malik, Eunyee Koh, Sungchul Kim, Zhicheng Liu, Donghyun Kim, Hongyuan Zha, and Nan Cao. 2019. Visualizing Uncertainty and Alternatives in Event Sequence Predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. https://doi.org/10.1145/3290605.3300803

[66] Amar Gupta, Raj K Goyal, Keith A Joiner, and Sanjay Saini. 2008. Outsourcing in the healthcare industry: Information technology, intellectual property, and allied aspects. *Information Resources Management Journal (IRMJ)* 21, 1 (2008), 1–26.

[67] Pavel Hamet and Johanne Tremblay. 2017. Artificial Intelligence in Medicine. *Metabolism* 69 (April 2017), S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

[68] Pavel Hamet and Johanne Tremblay. 2017. Artificial Intelligence in Medicine. *Metabolism* 69 (April 2017), S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

[69] Seung Seog Han. 2017. Model Onychomycosis Training Datasets (JPG Thumbnails) and Validation Datasets (JPG Images). , 4619093790 Bytes pages. https://doi.org/10.6084/M9.FIGSHARE.5398573.V2

[70] Seung Seog Han, Gyeong Hun Park, Woohyung Lim, Myoung Shin Kim, Jung Im Na, Ilwoo Park, and Sung Eun Chang. 2018. Deep Neural Networks Show an Equivalent and Often Superior Performance to Dermatologists in Onychomycosis Diagnosis: Automatic Construction of Onychomycosis Datasets by Region-Based Convolutional Deep Neural Network. *PLOS ONE* 13, 1 (Jan. 2018), e0191493. https://doi.org/10.1371/journal.pone.0191493

[71] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (Oct. 2006), 904–908. https://doi.org/10.1177/154193120605000909

[72] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–29. https://doi.org/10.1145/3610067

[73] Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. The Extent and Consequences of P-Hacking in Science. *PLOS Biology* 13, 3 (March 2015), e1002106. https://doi.org/10.1371/journal.pbio.1002106

[74] James J Heckman. 1990. Selection bias and self-selection. In *Econometrics*. Springer, 201–224.

[75] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review.

[76] Otman Hijazi, Kawtar Tikito, and Khadija Ouazzani-Touhami. 2023. A Systematic Review on Artificial Intelligence Models Applied to Prediction in Finance. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, Las Vegas, NV, USA, 0183–0188. https://doi.org/10.1109/CCWC57344.2023.10099222

[77] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for Explainable AI: Explanation Goodness, User Satisfaction, Mental Models, Curiosity, Trust, and Human-AI Performance. *Frontiers in Computer Science* 5 (Feb. 2023), 1096257. https://doi.org/10.3389/fcomp.2023.1096257

[78] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–20. https://doi.org/10.1145/3579628

[79] Peter Hoonakker, Pascale Carayon, Ayse P. Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M. Walker. 2011. Measuring Workload of ICU Nurses with a Questionnaire Survey: The NASA Task Load Index (TLX). *IIE Transactions on Healthcare Systems Engineering* 1, 2 (April 2011), 131–143. https://doi.org/10.1080/19488300.2011.609524

[80] Torsten Hothorn and Kurt Hornik. 2002. Exact Nonparametric Inference in R. In *Compstat*, Wolfgang Härdle and Bernd Rönz (Eds.). Physica-Verlag HD, Heidelberg, 355–360. https://doi.org/10.1007/978-3-642-57489-4_52

[81] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

[82] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. 2022. A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks. *Applied Sciences* 12, 3 (Jan. 2022), 1353. https://doi.org/10.3390/app12031353

[83] Ying-Chun Jheng, Chung-Lan Kao, Aliaksandr A. Yarmishyn, Yu-Bai Chou, Chih-Chien Hsu, Tai-Chi Lin, Hou-Kai Hu, Ta-Kai Ho, Po-Yin Chen, Zih-Kai Kao, Shih-Jen Chen, and De-Kuang Hwang. 2020. The Era of Artificial Intelligence–Based Individualized Telemedicine Is Coming. *Journal of the Chinese Medical Association* 83, 11 (Nov. 2020), 981–983. https://doi.org/10.1097/JCMA.0000000000000374

[84] Julian D. Karch. 2021. Psychologists Should Use Brunner-Munzel's Instead of Mann-Whitney's *U* Test as the Default Nonparametric Procedure. *Advances in Methods and Practices in Psychological Science* 4, 2 (April 2021), 251524592199960. https://doi.org/10.1177/2515245921999602

[85] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Léger. 2022. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience* 16 (June 2022), 883385. https://doi.org/10.3389/fnins.2022.883385

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

[87] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–18. https://doi.org/10.1145/3491102.3517439

[88] Markelle Kelly, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. 2023. Capturing Humans' Mental Models of AI: An Item Response Theory Approach. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1723–1734. https://doi.org/10.1145/3593013.3594111

[89] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P. Langlotz, Robyn L. Ball, Thomas J. Montine, Brock A. Martin, Gerald J. Berry, Michael G. Ozawa, Florette K. Hazard, Ryanne A. Brown, Simon B. Chen, Mona Wood, Libby S. Allard, Lourdes Ylagan, Andrew Y. Ng, and Jeanne Shen. 2020. Impact of a Deep Learning Assistant on the Histopathologic Classification of Liver Cancer. *npj Digital Medicine* 3, 1 (Feb. 2020), 23. https://doi.org/10.1038/s41746-020-0232-8

[90] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2022. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. (2022). https://doi.org/10.48550/ARXIV.2210.03735

[91] Rex B. Kline. 2013. *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences (2nd Ed.).* American Psychological Association, Washington. https://doi.org/10.1037/14136-000

[92] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 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*. ACM, Glasgow Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300641

[93] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 5686–5697. https://doi.org/10.1145/2858036.2858529

[94] Patrick M Kreiser, Louis D Marino, Pat Dickson, and K Mark Weaver. 2010. Cultural influences on entrepreneurial orientation: The impact of national culture on risk taking and proactiveness in SMEs. *Entrepreneurship theory and practice* 34, 5 (2010), 959–984.

[95] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Austin Texas USA, 1–10. https://doi.org/10.1145/2207676.2207678

[96] Yogesh Kumar, Apeksha Koul, Ruchi Singla, and Muhammad Fazal Ijaz. 2023. Artificial Intelligence in Disease Diagnosis: A Systematic Literature Review, Synthesizing Framework and Future Research Agenda. *Journal of ambient intelligence and humanized computing* 14, 7 (2023), 8459–8486.

[97] Nahyun Kwon, Tong Steven Sun, Yuyang Gao, Liang Zhao, Xu Wang, Jeeeun Kim, and Sungsoo Ray Hong. 2024. 3DPFIX: Improving Remote Novices' 3D Printing Troubleshooting through Human-AI Collaboration Design. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–33. https://doi.org/10.1145/3637288

[98] Vivian Lai. 2022. *Empowering Humans in Human-AI Decision Making.* Ph. D. Dissertation. University of Colorado, Colorado.

[99] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–18. https://doi.org/10.1145/3491102.3501999

[100] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[101] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why Is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376873

[102] Vivian Lai and Chenhao Tan. 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*. ACM, Atlanta GA USA, 29–38. https://doi.org/10.1145/3287560.3287590

[103] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[104] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[105] Richard S Lazarus. 1984. *Stress, appraisal, and coping*. Vol. 464. Springer.

[106] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez I Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 1–27. https://doi.org/10.1145/3415227

[107] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez I Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445472

[108] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez I Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445472

[109] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the Impact of Automated Suggestions on Decision Making: Domain Experts Mediate Model Errors but Take Less Initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445522

[110] Shihong Ling, Yutong Zhang, and Na Du. 2024. More Is Not Always Better: Impacts of AI-Generated Confidence and Explanations in Human–Automation Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (March 2024), 00187208241234810. https://doi.org/10.1177/00187208241234810

[111] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 1–45. https://doi.org/10.1145/3479552

[112] Duri Long and Brian Magerko. 2020. What Is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–16. https://doi.org/10.1145/3313831.3376727

[113] Maris G. Martinsons and Robert M. Davison. 2007. Strategic Decision Making and Support Systems: Comparing American, Japanese and Chinese Management. *Decision Support Systems* 43, 1 (Feb. 2007), 284–300. https://doi.org/10.1016/j.dss.2006.10.005

[114] Stina Matthiesen, Søren Zöga Diederichsen, Mikkel Klitzing Hartmann Hansen, Christina Villumsen, Mats Christian Højbjerg Lassen, Peter Karl Jacobsen, Niels Risum, Bo Gregers Winkel, Berit T Philbert, Jesper Hastrup Svendsen, and Tariq Osman Andersen. 2021. Clinician Preimplementation Perspectives of a Decision-Support Tool for the Prediction of Cardiac Arrhythmia Based on Machine Learning: Near-Live Feasibility and Qualitative Study. *JMIR Human Factors* 8, 4 (Nov. 2021), e26964. https://doi.org/10.2196/26964

[115] Peter Mayser and Niehaus Niehaus. 2023. Leitlinien-Update Onychomykose. *Evidence for Self-Medication International Review Journal* (2023). https://doi.org/10.52778/efsm.23.0002

[116] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 220–229. https://doi.org/10.1145/3287560.3287596

[117] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Dec. 2021), 1–45. https://doi.org/10.1145/3387166

[118] Stefan Morana, Silvia Schacht, Ansgar Scherp, and Alexander Maedche. 2017. A Review of the Nature and Effects of Guidance Design Features. *Decision Support Systems* 97 (May 2017), 31–42. https://doi.org/10.1016/j.dss.2017.03.003

[119] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–39. https://doi.org/10.1145/3641022

[120] Ralf Müller, Konrad Spang, and Sinan Ozcan. 2009. Cultural differences in decision making in project teams. *International journal of managing projects in business* 2, 1 (2009), 70–93.

[121] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *Comput. Surveys* 55, 13s (Dec. 2023), 1–42. https://doi.org/10.1145/3583558

[122] Pietro Nenoff. 2022. S1-Leitlinie Onychomykose (AWMF-Register-Nr. 013-003).

[123] P. Nenoff, G. Ginter-Hanselmayer, and H.-J. Tietz. 2012. Onychomykose – ein Update: Teil 1 – Prävalenz, Epidemiologie, disponierende Faktoren und Differenzialdiagnose. *Der Hautarzt* 63, 1 (Jan. 2012), 30–38. https://doi.org/10.1007/s00105-011-2251-5

[124] Pietro Nenoff, Dieter Reinel, Peter Mayser, Dietrich Abeck, Guntram Bezold, Philipp P Bosshard, Jochen Brasch, Georg Daeschlein, Isaak Effendy, Gabriele Ginter-Hanselmayer, Yvonne Gräser, Gudrun Hamm, Ulrich Hengge, Uta-Christina Hipler, Peter Höger, Alexandra Kargl, Annette Kolb-Mäurer, Constanze Krüger, Bartosz Malisiewicz, Johannes Mayer, Hagen Ott, Uwe Paasch, Martin Schaller, Silke Uhrlaß, and Miriam Zidane. 2023. S1-Leitlinie Onychomykose. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft* 21, 6 (June 2023), 678–694. https://doi.org/10.1111/ddg.14988_g

[125] Akihiro Nomura, Masahiro Noguchi, Mitsuhiro Kometani, Kenji Furukawa, and Takashi Yoneda. 2021. Artificial Intelligence in Current Diabetes Management and Prediction. *Current Diabetes Reports* 21, 12 (Dec. 2021), 61. https://doi.org/10.1007/s11892-021-01423-2

[126] Donald A Norman. 1988. *The Psychology of Everyday Things.* Basic books.

[127] Donald A Norman. 2014. Some Observations on Mental Models. In *Mental Models.* Psychology Press, 15–22.

[128] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces.* ACM, College Station TX USA, 340–350. https://doi.org/10.1145/3397481.3450639

[129] David B Olawade, Aanuoluwapo C David-Olawade, Ojima Z Wada, Akinsola J Asaolu, Temitope Adereni, and Jonathan Ling. 2024. Artificial intelligence in healthcare delivery: Prospects and pitfalls. *Journal of Medicine, Surgery, and Public Health* (2024), 100108.

[130] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology* 45, 4 (July 2009), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

[131] S. O'Sullivan, M. Janssen, Andreas Holzinger, Nathalie Nevejans, O. Eminaga, C. P. Meyer, and Arkadiusz Miernik. 2022. Explainable Artificial Intelligence (XAI): Closing the Gap between Image Analysis and Navigation in Complex Invasive Diagnostic Procedures. *World Journal of Urology* 40, 5 (May 2022), 1125–1134. https://doi.org/10.1007/s00345-022-03930-7

[132] A Ben Oumlil and Joseph L Balloun. 2017. Cultural variations and ethical business decision making: a study of individualistic and collective cultures. *Journal of Business & Industrial Marketing* 32, 7 (2017), 889–900.

[133] Danica Mitch M Pacis, Edwin D C Subido, and Nilo T Bugtai. 2018. Trends in Telemedicine Utilizing Artificial Intelligence. In *2ND BIOMEDICAL ENGINEERING'S RECENT PROGRESS IN BIOMATERIALS, DRUGS DEVELOPMENT, AND MEDICAL DEVICES: Proceedings of the International Symposium of Biomedical Engineering (ISBE) 2017.* Bali, Indonesia, 040009. https://doi.org/10.1063/1.5023979

[134] Stefan Palan and Christian Schitter. 2018. Prolific.Ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* 17 (March 2018), 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

[135] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (Aug. 2010), 411–419. https://doi.org/10.1017/S1930297500002205

[136] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology* 70 (May 2017), 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

[137] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2021. Data Quality of Platforms and Panels for Online Behavioral Research. *Behavior Research Methods* 54, 4 (Sept. 2021), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

[138] Weiping Pei, Arthur Mayer, Kaylynn Tu, and Chuan Yue. 2020. Attention Please: Your Attention Check Questions in Survey Studies Can Be Automatically Answered. In *Proceedings of The Web Conference 2020.* ACM, Taipei Taiwan, 1182–1193. https://doi.org/10.1145/3366423.3380195

[139] Thomas V Perneger. 1998. What's Wrong with Bonferroni Adjustments. *BMJ* 316, 7139 (April 1998), 1236–1238. https://doi.org/10.1136/bmj.316.7139.1236

[140] Gloria Phillips-Wren and Monica Adya. 2020. Decision Making under Stress: The Role of Information Overload, Time Pressure, Complexity, and Uncertainty. *Journal of Decision Systems* 29, sup1 (Aug. 2020), 213–225. https://doi.org/10.1080/12460125.2020.1768680

[141] Gérald E. Piérard and Claudine Piérard-Franchimont. 2005. The Nail under Fungal Siege in Patients with Type II Diabetes Mellitus. *Mycoses* 48, 5 (Sept. 2005), 339–342. https://doi.org/10.1111/j.1439-0507.2005.01140.x

[142] Marc Pinski, Martin Adam, and Alexander Benlian. 2023. AI Knowledge: Improving AI Delegation through Human Enablement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3580794

[143] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 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.* ACM, Sydney NSW Australia, 379–396. https://doi.org/10.1145/3581641.3584033

[144] R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[145] Eric Rassin, Peter Muris, Ingmar Franken, Maartje Smit, and Maggie Wong. 2007. Measuring General Indecisiveness. *Journal of Psychopathology and Behavioral Assessment* 29, 1 (Jan. 2007), 60–67. https://doi.org/10.1007/s10862-006-9023-z

[146] Gabriel Recchia, Alice C E Lawrence, and Alexandra L J Freeman. 2022. Investigating the Presentation of Uncertainty in an Icon Array: A Randomized Trial. *PEC Innovation* 1 (Dec. 2022), 100003. https://doi.org/10.1016/j.pecinn.2021.100003

[147] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, GI Genius CADx Study Group, Giulio Antonelli, Halim Awadie, Sebastian Bernhofer, Sabela Carballal, Mário Dinis-Ribeiro, Agnès Fernández-Clotett, Glòria Fernández Esparrach, Ian Gralnek, Yuta Higasa, Taku Hirabayashi, Tatsuki Hirai, Mineo Iwatate, Miki Kawano, Markus Mader, Andreas Maieron, Sebastian Mattes, Tastuya Nakai, Ingrid Ordas, Raquel Ortigão, Oswaldo Ortiz Zúñiga, Maria Pellisé, Cláudia Pinto, Florian Riedl, Ariadna Sánchez, Emanuel Steiner, Yukari Tanaka, and Andrea Cherubini. 2022. Experimental Evidence of Effective Human–AI Collaboration in Medical Decision-Making. *Scientific Reports* 12, 1 (Sept. 2022), 14952. https://doi.org/10.1038/s41598-022-18751-2

[148] Valerie F Reyna and Charles J Brainerd. 2008. Numeracy, Ratio Bias, and Denominator Neglect in Judgments of Risk and Probability. *Learning and Individual Differences* 18, 1 (Jan. 2008), 89–107. https://doi.org/10.1016/j.lindif.2007.03.011

[149] Robert Rosenthal. 1976. Experimenter effects in behavioral research. (1976).

[150] Susana Rubio, Eva Díaz, Jesús Martín, and José M Puente. 2004. Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology* 53, 1 (Jan. 2004), 61–86. https://doi.org/10.1111/j.1464-0597.2004.00161.x

[151] Patrik Sabol, Peter Sinčák, Pitoyo Hartono, Pavel Kočan, Zuzana Benetinová, Alžbeta Blichárová, L'udmila Verbóová, Erika Štammová, Antónia Sabolová-Fabianová, and Anna Jašková. 2020. Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images. *Journal of biomedical informatics* 109 (2020), 103523.

[152] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

[153] Alan Samuel, Jocelyn Cranefield, and Yi-Te Chiu. 2023. AI to Human:"Help Me to Help You Collaborate More Effectively"-a Literature Review from a Human Capability Perspective. (2023).

[154] Beau G. Schelble, Christopher Flathmann, Nathan J. McNeese, Thomas O'Neill, Richard Pak, and Moses Namara. 2023. Investigating the Effects of Perceived Teammate Artificiality on Human Performance and Cognition. *International Journal of Human–Computer Interaction* 39, 13 (Aug. 2023), 2686–2701. https://doi.org/10.1080/10447318.2022.2085191

[155] Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. 2023. Towards Effective Human-AI Decision-Making: The Role of Human Learning in Appropriate Reliance on AI Advice. (2023). https://doi.org/10.48550/ARXIV.2310.02108

[156] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, Oxford United Kingdom, 617–626. https://doi.org/10.1145/3514094.3534128

[157] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 410–422. https://doi.org/10.1145/3581641.3584066

[158] Karthik Seetharam, Nobuyuki Kagiyama, and Partho P Sengupta. 2019. Application of Mobile Health, Telemedicine and Artificial Intelligence to Echocardiography. *Echo Research & Practice* 6, 2 (June 2019), R41–R52. https://doi.org/10.1530/ERP-18-0081

[159] Lars Sipos, Ulrike Schäfer, Katrin Glinka, and Claudia Müller-Birn. 2023. Identifying Explanation Needs of End-users: Applying and Extending the XAI Question Bank. In *Mensch Und Computer 2023*. ACM, Rapperswil Switzerland, 492–497. https://doi.org/10.1145/3603555.3608551

[160] Phil So. 2022. NASA Task Load Index.

[161] Kacper Sokol and Julia E Vogt. 2024. What Does Evaluation of Explainable Artificial Intelligence Actually Tell Us? A Case for Compositional and Contextual Validation of XAI Building Blocks. *arXiv preprint arXiv:2403.12730* (2024). arXiv:2403.12730

[162] Hyewon Song, Anh-Duc Nguyen, Myoungsik Gong, and Sanghoon Lee. 2016. A Review of Computer Vision Methods for Purpose on Computer-Aided Diagnosis. *Journal of International Society for Simulation Surgery* 3, 1 (June 2016), 1–8. https://doi.org/10.18204/JISSIS.2016.3.1.001

[163] David Spiegelhalter. 2017. Risk and Uncertainty Communication. *Annual Review of Statistics and Its Application* 4, 1 (March 2017), 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148

[164] David Spiegelhalter, Mike Pearson, and Ian Short. 2011. Visualizing Uncertainty About the Future. *Science* 333, 6048 (Sept. 2011), 1393–1400. https://doi.org/10.1126/science.1191181

[165] Elias Spinn. 2023. Probability Expressions in AI Decision Support: Impacts on Human+ AI Team Performance. (2023).

[166] Philipp Spitzer, Joshua Holstein, Patrick Hemmer, Michael Vössing, Niklas Kühl, Dominik Martin, and Gerhard Satzger. 2024. On the Effect of Contextual Information on Human Delegation Behavior in Human-AI Collaboration. *arXiv preprint arXiv:2401.04729* (2024). arXiv:2401.04729

[167] Anna Taudien, Andreas Fügener, Alok Gupta, and Wolfgang Ketter. 2022. Calibrating Users' Mental Models for Delegation to AI. (2022).

[168] Tish Holub Taylor. 2024. *Ceiling Effect*. SAGE Publications, Inc., Thousand Oaks. https://doi.org/10.4135/9781412961288

[169] Karl Halvor Teigen. 2023. Dimensions of Uncertainty Communication: What Is Conveyed by Verbal Terms and Numeric Ranges. *Current Psychology* 42, 33 (Nov. 2023), 29122–29137. https://doi.org/10.1007/s12144-022-03985-0

[170] Ian Thomas, Song Young Oh, and Danielle Albers Szafir. 2024. Assessing User Trust in Active Learning Systems: Insights from Query Policy and Uncertainty Visualization. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. ACM, Greenville SC USA, 772–786. https://doi.org/10.1145/3640543.3645207

[171] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–17. https://doi.org/10.1145/3411764.3445101

[172] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. 2024. When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nature Human Behaviour* (Oct. 2024). https://doi.org/10.1038/s41562-024-02024-1

[173] Bas H.M. Van Der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, and Max A. Viergever. 2022. Explainable Artificial Intelligence (XAI) in Deep Learning-Based Medical Image Analysis. *Medical Image Analysis* 79 (July 2022), 102470. https://doi.org/10.1016/j.media.2022.102470

[174] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–38. https://doi.org/10.1145/3579605

[175] Sahil Verma and Julia Rubin. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, Gothenburg Sweden, 1–7. https://doi.org/10.1145/3194770.3194776

[176] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 318–328. https://doi.org/10.1145/3397481.3450650

[177] Quentin Waymel, Sammy Badr, Xavier Demondion, Anne Cotten, and Thibaut Jacques. 2019. Impact of the Rise of Artificial Intelligence in Radiology: What Do Radiologists Think? *Diagnostic and Interventional Imaging* 100, 6 (June 2019), 327–336. https://doi.org/10.1016/j.diii.2019.03.015

[178] Margaret A. Webb and June P. Tangney. 2024. Too Good to Be True: Bots and Bad Data From Mechanical Turk. *Perspectives on Psychological Science* 19, 6 (Nov. 2024), 887–890. https://doi.org/10.1177/17456916221120027

[179] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. (2019). https://doi.org/10.48550/ARXIV.1907.03324

[180] Odette Wegwarth, Gert G. Wagner, Claudia Spies, and Ralph Hertwig. 2020. Assessment of German Public Attitudes Toward Health Communications With Varying Degrees of Scientific Uncertainty Regarding COVID-19. *JAMA Network Open* 3, 12 (Dec. 2020), e2032335. https://doi.org/10.1001/jamanetworkopen.2020.32335

[181] Christopher D. Wickens. 2008. Multiple Resources and Mental Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (June 2008), 449–455. https://doi.org/10.1518/001872008X288394

[182] Bethany Jill Williams, David Bottoms, and Darren Treanor. 2017. Future-Proofing Pathology: The Case for Clinical Adoption of Digital Pathology. *Journal of Clinical Pathology* 70, 12 (Dec. 2017), 1010–1018. https://doi.org/10.1136/jclinpath-2017-204644

[183] Honglian Xiang, Jia Zhou, and Bingjun Xie. 2023. AI Tools for Debunking Online Spam Reviews? Trust of Younger and Older Adults in AI Detection Criteria. *Behaviour & Information Technology* 42, 5 (April 2023), 478–497. https://doi.org/10.1080/0144929X.2021.2024252

[184] Lixiang Yan, Vanessa Echeverria, Gloria Milena Fernandez-Nieto, Yueqiao Jin, Zachari Swiecki, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2024. Human-AI Collaboration in Thematic Analysis Using ChatGPT: A User Study and Design Recommendations. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–7. https://doi.org/10.1145/3613905.3650732

[185] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. 2020. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. ACM, Cagliari Italy, 189–201. https://doi.org/10.1145/3377325.3377480

[186] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376301

[187] Jie Yu, Dingyan Wang, and Mingyue Zheng. 2022. Uncertainty Quantification: Can We Trust Artificial Intelligence in Drug Discovery? *iScience* 25, 8 (Aug. 2022), 104814. https://doi.org/10.1016/j.isci.2022.104814

[188] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate?: An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 460–468. https://doi.org/10.1145/3301275.3302277

[189] Hubert D Zając, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Transactions on Computer-Human Interaction* 30, 2 (April 2023), 1–39. https://doi.org/10.1145/3582430

[190] Yunfeng Zhang, Q Vera Liao, and Rachel K E Bellamy. 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*. ACM, Barcelona Spain, 295–305. https://doi.org/10.1145/3351095.3372852

[191] Brian J. Zikmund-Fisher, Dylan M. Smith, Peter A. Ubel, and Angela Fagerlin. 2007. Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations. *Medical Decision Making* 27, 5 (Sept. 2007), 663–671. https://doi.org/10.1177/0272989X07303824

A SCENARIO AND TASK DESCRIPTIONS

A.1 Scenario description

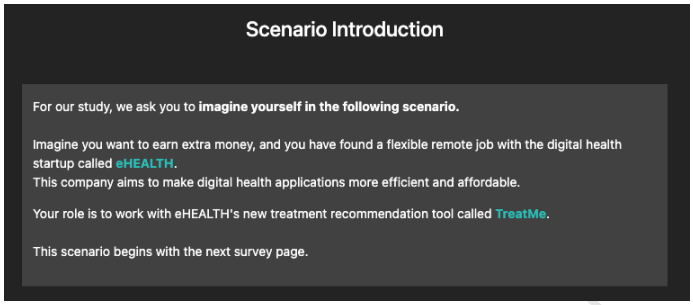


Fig. 8. The participants were informed in the consent form, that they will be introduced to a scenario. Before the task instruction all participants were given this scenario introduction.

A.2 Task instructions

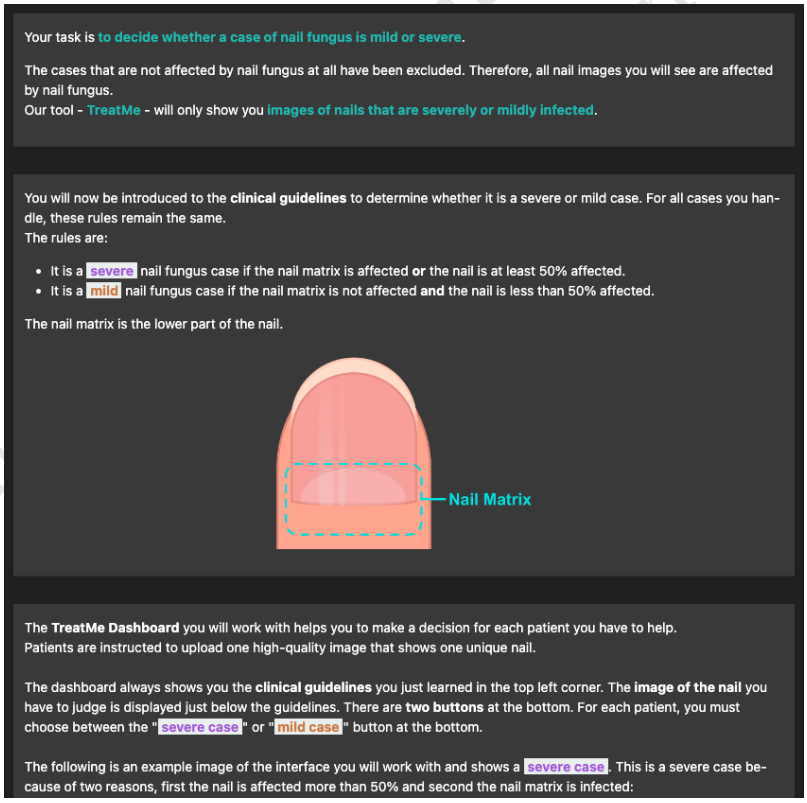


Fig. 9. Task instruction for the general decision-making task. Not depicted in the screenshot is the described example case of the whole interface, which was shown to the participants below.

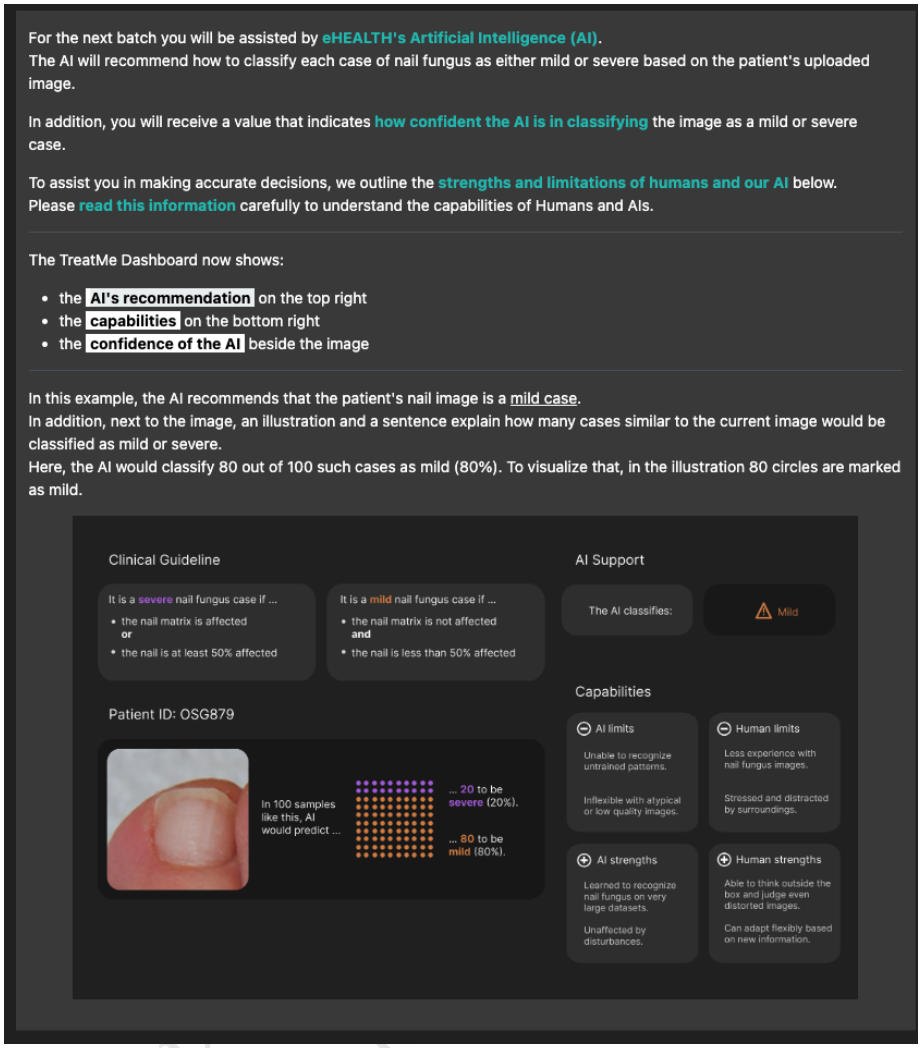


Fig. 10. Task instruction for the GG before the human-AI phase. The UG did not receive the introduction to the capabilities. The CG did not receive the introduction to the capabilities nor the uncertainty representation introduction. Before the human-AI phase participants were introduced to the AI. Participants of the UG and GG were additionally introduced to the uncertainty representation including a comprehension check. The GG was also introduced to the AI's and their capabilities including a comprehension check. In the following only the instruction of the GG is shown as they include all the information the other groups got. Afterward, comprehension check for the UG and GG were asked. The nail image shown is from Han [69] published under the CC BY 4.0 International License and adjusted by us (cropped).

B ATTENTION CHECKS

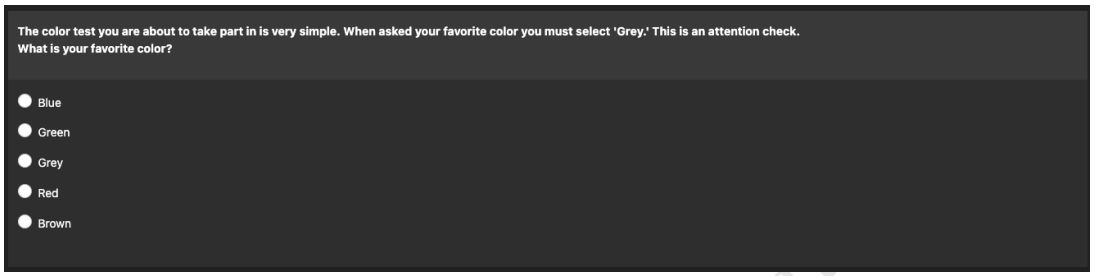


Fig. 11. Attention check 1 is an instructional manipulation check based on [130] and was asked after the demographic questionnaire.

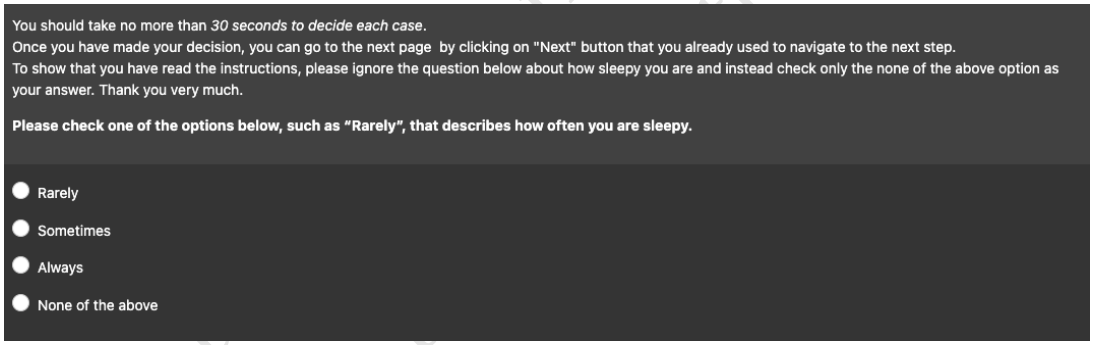


Fig. 12. Attention check 2 is an instructional manipulation check based on [130] and was asked after the instruction before the baseline phase.

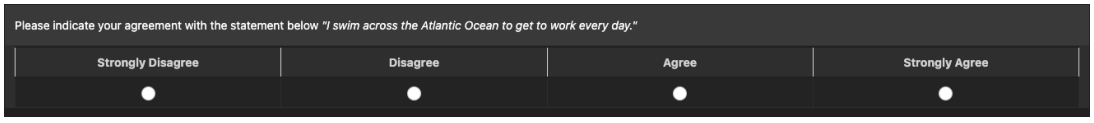


Fig. 13. Attention check 3 is a nonsensical item based on [135] was asked after the baseline phase.

C NUMERACY AND OVERALL PERFORMANCE

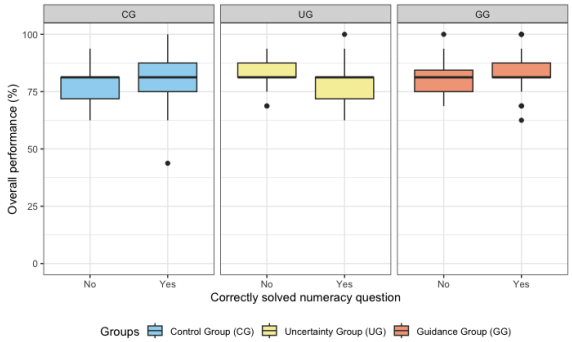


Fig. 14. Overall performance (%) per group and correctly or incorrectly solved numeracy question.

D STUDY PROCEDURE

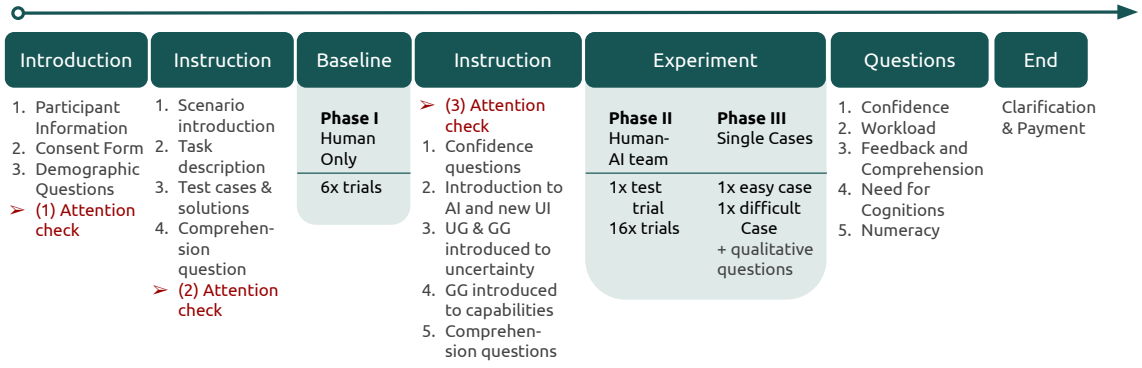


Fig. 15. Visualization of the study procedure.

E DETAILED HYPOTHESES TESTING INCLUDING NON-PARAMETRIC TESTINGS

RQ1: Can combined uncertainty representations improve team performance?					
Hypotheses		Parametric testing (GLM: unpaired t-test)	Non-parametric testing (Wilcoxon rank sum test)	Levene's Test for Homogeneity of Variance (center = median)	Brunner-Munzel Test for stochastic equality
H1 a)	Uncertainty representations (UG) lead to improved mean overall performance compared to no representations (CG).	t(141)=1.50 p=0.07 $\widehat{d_\Psi}$ = 0.31 $1-\beta$ = 0.44	W=1331.5 p=0.09	F=0.31 p=0.74	B(93.4)=-1.36 p=0.09 $\widehat{p}^{\text{“}}$=0.42
H1 b)	Uncertainty representations (UG) lead to improved mean overall performance compared to no representations (CG).	t(141)=-0.09 p=0.53 $\widehat{d_\Psi}$ = -0.02 $1-\beta$ = 0.04	W=1161.5 p=0.47	F=0.37 p=0.67	B(93.4)=-0.07 p=0.47 $\widehat{p}^{\text{“}}$=0.49
RQ2: Can providing capability-focused guidance in addition to uncertainty representations improve team performance?					
Hypotheses		Parametric testing (GLM: unpaired t-test)	Non-parametric testing (Wilcoxon rank sum test)	Levene's Test for Homogeneity of Variance (center = median)	Brunner-Munzel Test for stochastic equality
H2 a)	Guidance (GG) leads to improved mean overall performance compared to only uncertainty representations (UG).	t(141)=0.68 p=0.25 $\widehat{d_\Psi}$ = 0.14 $1-\beta$ = 0.17	W=1235 p=0.27	F=0.31 p=0.74	B(93.2)=-0.62 p=0.27 $\widehat{p}^{\text{“}}$=0.46
H2 b)	Guidance (GG) leads to improved mean over-reliance performance compared to only uncertainty representations (UG).	**t(141)=2.18** **p=0.02** $\widehat{d_\Psi}$ = **0.44** $1-\beta$ = **0.7**	**W=1425.5** **p=0.02**	F=0.37 p=0.67	**B(93.6)=-2.14** **p=0.02** $\widehat{p}^{\text{“}}$=**0.38**
Additional	Guidance in addition to uncertainty representations (GG) leads to improved mean overall performance compared to no no representations (CG).	**t(141)=2.18** **p=0.02** $\widehat{d_\Psi}$ = **0.44** $1-\beta$ = **0.7**	**W=1400.5** **p=0.03**	F=0.31 p=0.74	**B(94)=-1.92** **p=0.03** $\widehat{p}^{\text{“}}$=**0.39**
Additional	Guidance in addition to uncertainty representations (GG) leads to improved mean over-reliance performance compared to no representations (CG).	**t(141)=2.09** **p=0.02** $\widehat{d_\Psi}$ = **0.43** $1-\beta$ = **0.67**	**W=1437.5** **p=0.02**	F=0.37 p=0.67)	**B(93.5)=-2.25** **p=0.01** $\widehat{p}^{\text{“}}$=**0.37**

Table 7. Hypotheses testing for the research question 1 and 2 and Hypotheses H1 (a-b) and H2 (a-b). As the residuals are not optimally normal distributed, a non-parametric Wilcoxon-Mann-Whitney U-test (Wilcoxon rank sum test) was conducted with the R package "exactRankTests" to consider tied observations [80]. We provide exact p-values that are consistent in significance with the asymptotic Wilcoxon rank sum test. To be able to interpret the median difference, the variances have to be homogen, which is the case for all hypotheses. As the shape of correct self-reliance for the GG is different from the other groups, the results need to be interpreted with caution. In addition, we calculated the Brunner-Munzel Test (Generalized Wilcoxon Test) including the effect size (stochastic superiority: $\widehat{p}^{\text{“}}$=P(X <Y)+0.5*P(X = Y)), which tests for stochastic equality capturing the group difference by comparing the individual performances [84].

F GROUP DIFFERENCES BETWEEN THE HUMAN BASELINE

There was no significant difference between the groups for the human baseline performance (M(SD) - CG: 78.8(15.3), UG: 85.4(13.1), GG: 83.3(14.2); one-way ANOVA: F(2,141)=2.70, p=0.07, η^2=0.04). The only significant two-sided t-test was between the UG and the CG, revealing that the UG had greater overall human baseline performance (t(141)=2.27, p=0.024, $\widehat{d_\Psi}$ = 0.46, $1-\beta$ = 0.62).

G PARTICIPANTS' RECOMMENDATION AND PREDICTION OF THE AI'S RECOMMENDATION

		CG		UG		GG	
		M	**S**	M	**S**	M	**S**
Hard case (correct: severe)	Own recommendation	6	**42**	5	**43**	6	**42**
	AI recommendation prediction	9	**39**	6	**42**	13	**35**

		CG		UG		GG	
		M	S	**M**	S	**M**	S
Easy case (correct: mild)	Own recommendation	**48**	0	**48**	0	**47**	1
	AI recommendation prediction	**46**	2	**48**	0	**47**	1

Table 8. Participants' recommendation and prediction of the AI's recommendation for a more difficult case due to a low-quality image (top) and an easier case due to a high quality image (bottom). Bold indicates the correct decision.

H INFLUENCE OF PARTICIPANTS' GEOGRAPHIC LOCATION

Studies have found differences in decision-making between cultures especially in regard to business, strategic and social decisions for example between collective and individualistic cultures or risk taking [94, 113, 120, 132]. Most of these findings (e.g., business decisions) do not seem to limit our findings. Especially since the frequencies of the geographic locations of our participants are similar between the groups. If other measurements are of interest in future research such as decision-making speed or social and strategic aspects (e.g., accepting changes in decisions, negotiation styles) [120, 132], cultural backgrounds should be considered. In our case, we collected demographic data relevant for the decision-making tasks, such as language fluency, level of schooling, prior knowledge [78], age [183] and comprehension checks, so that future studies referencing our work may assess whether they represent the population of interest.

Received 02 July 2024; revised XX X 2024; accepted XX X 2024