Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision-Making

Sara Salimzadeh
Delft University of Technology
Delft, The Netherlands
s.salimzadeh@tudelft.nl

Gaole He
Delft University of Technology
Delft, The Netherlands
g.he@tudelft.nl

Ujwal Gadiraju
Delft University of Technology
Delft, The Netherlands
u.k.gadiraju@tudelft.nl

ABSTRACT

While existing literature has explored and revealed several insights pertaining to the role of human factors (*e.g.*, prior experience, domain knowledge) and attributes of AI systems (*e.g.*, accuracy, trustworthiness), there is a limited understanding around how the important task characteristics of complexity and uncertainty shape human decision-making and human-AI team performance. In this work, we aim to address this research and empirical gap by systematically exploring how task complexity and uncertainty influence human-AI decision-making. Task complexity refers to the load of information associated with a task, while task uncertainty refers to the level of unpredictability associated with the outcome of a task. We conducted a between-subjects user study ($N = 258$) in the context of a trip-planning task to investigate the impact of task complexity and uncertainty on human trust and reliance on AI systems. Our results revealed that task complexity and uncertainty have a significant impact on user reliance on AI systems. When presented with complex and uncertain tasks, users tended to rely more on AI systems while demonstrating lower levels of *appropriate reliance* compared to tasks that were less complex and uncertain. In contrast, we found that user trust in the AI systems was not influenced by task complexity and uncertainty. Our findings can help inform the future design of empirical studies exploring human-AI decision-making. Insights from this work can inform the design of AI systems and interventions that are better aligned with the challenges posed by complex and uncertain tasks. Finally, the lens of diagnostic versus prognostic tasks can inspire the operationalization of uncertainty in human-AI decision-making studies.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **User studies**.

ACM Reference Format:
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 CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3613904.3641905

This research has been supported by *ICAI AI for Fintech Research.*

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3641905

1 INTRODUCTION AND BACKGROUND

With the emergence of human-AI decision-making as a prominent paradigm across various domains, numerous investigations have been dedicated to understanding the factors that can impact trust and reliance on AI systems [84, 138, 142]. Such factors can be broadly classified into three primary categories: human-related factors [35, 95, 96], attributes of the AI systems [94, 98], and characteristics of the decision-making tasks [16, 56, 126]. Human factors such as prior experience [110, 119], cognitive biases [85, 102], and AI literacy [25], which can shape individuals' perceptions and interactions with AI systems. Attributes of the AI system include aspects such as predictions generated by AI [66, 76, 99], information about model predictions [11, 31, 93], as well as interventions that impact cognitive processes [17]. Furthermore, the level of trust and reliance on AI may differ across various domains and applications due to the attributes associated with decision tasks [42, 127].

The characteristics of tasks have been demonstrated to play a pivotal role in determining the level of reliance on AI systems, emphasizing the importance of methodically recognizing and comprehending these features in human-AI decision-making context. However, limited task characteristics have been systematically explored and their impact on human reliance on AI systems is not yet fully understood [68, 109]. Although a few studies have included multiple tasks with varying attributes [6, 14, 131], a systematic and empirical understanding of task features is notably absent from existing literature [68, 109]. Additionally, it remains unclear whether task attributes chosen in existing empirical studies have been appropriately considered, in a manner that is commensurate with the claims of the studies [42, 68, 75]. These limitations have the potential to undermine the credibility and generalizability of research findings, hindering our progress in developing effective strategies for human-AI decision-making [68, 109].

In this work, we propose empirically examining **task complexity** and **task uncertainty** as two essential objective task characteristics that are manipulable from the task's standpoint. *Task complexity* pertains to the characteristics of a task that contribute to an increased load of information [133], and it is distinct from task difficulty [100], which relates to an individual's perception of the task-based on their capabilities and previous experience [133]. It has been shown that task complexity is a crucial factor in determining both human performance and behaviour [3, 23, 83], as

well as the success of human-AI teams [9]. Additionally, prior work has demonstrated that individuals tend to rely more heavily on AI systems when confronted with more complex tasks [28] due to the challenges associated with analyzing large volumes of information [23]. In line with work by Parkes [100], Vasconcelos et al. [126], we operationalize task complexity as an objective task-related characteristic that can be measured based on the number of constraints involved in the task. On the other hand, the level of *task uncertainty* refers to the extent of unpredictability inherent in a given task [29]. We operationalize uncertainty in our study using **diagnostic** and **prognostic** tasks to capture different levels of uncertainty. *Diagnostic* tasks involve situations where participants are provided with detailed and comprehensive information about the task, (theoretically) enabling them to make accurate decisions. *Prognostic tasks*, on the other hand, involve situations where participants must make predictions about future events based on incomplete or limited information. By operationalizing uncertainty in this manner, we can effectively capture the diverse levels of uncertainty that arise from the inherent nature of a task and its connection to information availability. Intuitively, in prognostic tasks, users can benefit from using AI systems due to their ability to reduce uncertainties, particularly when choosing the optimal route for a future trip by considering anticipated weather and traffic conditions. Unlike planning immediate trips, this task entails a greater degree of uncertainty owing to future events' unpredictability.

Prior work has highlighted that appropriate trust and reliance play a critical role in achieving complementary human-AI team performance [58, 90, 139, 141]. Thus, it is essential to comprehend how task-related factors impact human trust and reliance on AI systems, as separate constructs [63, 90, 111], to foster successful collaboration between humans and AI. We thereby address the following research questions:

RQ1: How does task complexity influence user trust and reliance on an AI system?
RQ2: How does task uncertainty, characterized by *prognostic* versus *diagnostic* tasks, influence user trust and reliance on an AI system?
RQ3: How does task complexity interact with task uncertainty to shape user trust and reliance on an AI system?

To address these research questions, we selected the real-world scenario of *trip-planning* where both task complexity and uncertainty are prominent factors. In such scenarios, individuals are confronted with circumstances that necessitate a choice between relying on an imperfect AI system or exercising their own judgment. We conducted a 3 (*task complexity*) × 2 (*task uncertainty*) between-subjects study with 258 participants recruited from the Prolific crowdsourcing platform.

We found that users' reliance on the AI system varied depending on the level of *complexity* and *uncertainty* in the task. Individuals facing tasks characterized by `medium` complexity and uncertainty *i.e.,* prognostic tended to rely excessively on the AI system. However, their ability to differentiate accurate AI advice from misleading advice was compromised, leading to a relatively low appropriate reliance, a higher over-reliance on AI, and subsequently lower overall task performance. However, we observed a point of transition where participants started to increase their appropriate reliance on the AI system. This led to enhanced overall performance in prognostic tasks with high complexity, revealing a significant interaction between complexity and uncertainty.

2 RELATED WORK

2.1 Human-AI Collaborative Decision-Making

In recent years, the use of AI technologies has evolved to encompass more collaborative approaches that involve both humans and AI systems working together [5, 21, 22, 73, 129]. While fully automated decision-making by AI systems may not always be appropriate, certain tasks still require human judgment. For example, in high-stake scenarios such as in the medical [39, 61, 67, 97], legal [6, 81, 86, 131], and financial [27, 36, 43, 45, 46] domains, individuals tend to exhibit a preference for human decision-makers over AI systems. This preference could be motivated by ethical and legal concerns [68, 74, 104], as well as a desire for individual agency and accountability [54, 70, 81, 117]. Additionally, it may also stem from the limited trust [18, 19] surrounding AI systems, coupled with concerns about potential biases or errors in algorithms [77, 120], particularly when human lives or ethical considerations are at stake due to possible failures of AI systems [68, 74, 104].

The primary objective of integrating human and AI is to unite their respective strengths, resulting in enhanced decision outcomes through complementary capabilities [17, 51]. To this end, previous research has focused on identifying the factors that influence human-AI decision-making. Recent studies have explored variables that contribute to the fairness [31, 76, 124, 130] and trustworthiness [34, 48, 80, 139] of AI systems, as well as the impact of assigning different decision-making roles to humans and AI on the reliance on such systems [52, 103, 122, 144]. Prior work has also been dedicated to developing and evaluating interfaces [15, 30, 87, 89] and visualizations [43, 49, 134, 137, 140] aimed at improving human-AI collaboration.

2.2 Trust and Reliance on AI Systems

It is important to distinguish between trust and reliance, as they have different implications for the context of human-AI decision-making. Lee and See [71] proposed the following definition of trust, which we adopt for the scope of our work:

> *Trust is an attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability.*

Reliance, on the other hand, refers to the extent to which individuals rely on AI systems [71, 128]. When user decisions differ from AI advice, there are mainly three discernible patterns of reliance behavior [7, 112, 115], (i) appropriate reliance, switching to the AI advice when it is correct and overriding it when it is incorrect, (ii) over-reliance, excessively relying on AI advice even when it is incorrect, and (iii) under-reliance, not fully utilizing AI advice even when it is correct. While trust is an essential factor in determining the level of reliance on AI systems [55, 63, 71, 111], it is not always a guarantee. Prior studies have shown that individuals may not necessarily increase their reliance on AI systems even if they

trust them [62, 63, 90]. Instead, they might rely more on their own judgments despite acknowledging the capabilities of the AI system. This highlights that the *trusting behavior* of users can differ from their *trusting beliefs*. The evaluation of the system's trustworthiness by individuals to establish perceived trustworthiness significantly influences (subjective) trust and trusting behaviour (i.e., objective reliance) [113]. Even if a system is trustworthy, it does not automatically ensure accurate perceived trustworthiness [8, 113]. To align the perceived trustworthiness of AI systems with their actual value, it is essential to consider aspects like the availability and relevance of system information and the detection and utilization of this information by human decision-makers [113]. Trust in AI systems, namely perceived trustworthiness, can be evaluated through different methodologies, including subjective self-reported measures [26, 59, 63, 132] and relatively objective trust-related behavioral measures [51, 128, 138, 142], such as agreement and compliance.

Through a wide range of studies, researchers have consistently found that reliance on AI systems is influenced by various factors including human-related aspects [37, 51, 81, 101, 120], attributes of the AI systems [43, 79, 106, 107], and characteristics of the decision-making tasks [9, 12, 16, 46, 126]. Human factors encompass a variety of individual characteristics, including previous experience [95, 110], cognitive biases [85, 102], and AI knowledge [25]. For instance, cognitive [35, 68, 96] or meta-cognitive biases [51] have the potential to influence how individuals comprehend and appraise the outcomes generated by AI systems which in turn can affect their reliance on AI. In addition, the attributes of AI systems can enhance decision-making outcomes [68], which include aspects such as predictions generated by AI [66, 76, 99], information about AI predictions or AI systems themselves [13, 70, 118, 136], and interventions that impact cognitive processes [65, 99, 105]. For instance, various explanation methods have been explored to enhance the interpretability and transparency of AI algorithms, allowing humans to better understand AI advice [2, 50, 66]. Banovic et al. [8] discovered that reliance on AI systems is negatively affected when untrustworthy AI systems overstate their capabilities compared to trustworthy ones. This is primarily because users struggle to differentiate between the competence of trustworthy and untrustworthy AI systems, leading to deception and excessive reliance on the untrustworthy system. Moreover, the characteristics of the decision-making tasks can also significantly impact human reliance on AI systems [68, 109]. Hence, the level of reliance may differ across various domains and applications due to the attributes associated with decision-making tasks [42, 127]. For instance, in high-stake fields like healthcare or finance, individuals may exhibit distinct behaviours compared to low-stake areas such as entertainment [89, 136].

Recent research has revealed several challenges in fostering appropriate reliance on AI systems. Prior work has shown that depending on different factors [126, 143], users may blindly follow AI advice, leading to over-reliance [17], or underestimate the capabilities of AI, resulting in under-reliance [37, 131]. To overcome such challenges and improve performance-related outcomes, it is important to ensure that users can strike a balance between utilizing AI effectively while also considering the limitations of a given AI system. To this end, researchers and practitioners have explored the use of explanation methods [66, 94, 126], interventions such as tutorials [25, 84] and cognitive forcing functions [17] to foster appropriate reliance on AI systems with varying degrees of success.

Building on the body of literature, our study aims to enhance the comprehension of appropriate reliance on AI systems in human-AI decision-making by investigating how task complexity and uncertainty influence user trust and reliance. To this end, we conducted a between-subjects study in the context of trip-planning task. We measured the extent to which individuals rely on AI systems for decision-making in various conditions by leveraging a series of common metrics in the field.

2.3 Task Characteristics in Human-AI Decision-Making

Although much attention has been given to the effect of human and AI-related factors in shaping human reliance on AI, few studies have explored the influence of task characteristics. Lee [75] found that individuals exhibited lower trust in AI systems in tasks that involve human skills, such as work evaluation, compared to tasks that require more analytical skills. Additionally, Vasconcelos et al. [126] has also examined the concept of task difficulty by considering the cognitive load required. Their findings indicate that as tasks become more difficult, there is a tendency among users to rely excessively on AI advice, leading to over-reliance. A few studies have also explored the effect of task features on human-AI team performance. Bansal et al. [9] conducted a study where participants had to assess whether objects passing through a pipeline were defective or not. They manipulated the complexity by changing the number of the task features, such as color, shape, and size. They found that an excessive number of task features diminished the performance of human-AI teams significantly. Similarly, in a study by Poursabzi-Sangdeh et al. [105], participants were presented with varying numbers of features to predict apartment selling prices. The features included variables such as the number of rooms, area size, days on the market, distance to amenities, and building maintenance fees. They also found that participants struggle to distinguish AI errors in tasks with more features, leading to decreased performance. In contrast, Tolmeijer et al. [120] showed that the complexity of tasks did not significantly impact human-AI performance due to a learning effect. They conducted an experiment in which participants were tasked with finding a suitable house based on a set of constraints. The complexity of the tasks was manipulated, with some scenarios having three constraints (such as rent type, budget, and registration condition), while others had five constraints (including rental duration and proximity to amenities). Buçinca et al. [16] conducted a study examining the influence of proxy tasks, where participants were tasked to anticipate AI advice, compared to actual tasks where participants directly received AI advice. Their results indicate that participants' behavior in proxy tasks did not align with their behaviour in actual tasks, underscoring the importance of carefully designing experiments to draw valid conclusions. Additionally, high-stake [6, 45, 46, 97] tasks and low-stake [44, 46, 66] tasks have been studied individually in literature in relation to human reliance on AI systems.

Furthermore, there is a lack of comprehensive investigations into categorizing task attributes and their specific implications for

human-AI decision-making [109]. Lai et al. [68] proposed a framework that categorizes task characteristics in terms of their domain, required expertise, risk, and subjectivity. According to Lai et al. [69], tasks can also be differentiated based on whether they are emulating human intelligence, like object recognition [20], or based on discovered patterns in data such as recidivism prediction [86]. Some prior works have also provided a taxonomy of task types existing in the literature [1, 92]. However, these taxonomies often focus on general task types rather than specifically addressing the impact of these characteristics on human-AI decision-making. De-Arteaga et al. [29] introduced diagnostic and prognostic tasks in which there is clear grand-truth in diagnostic tasks, while prognostic tasks involve making predictions about future outcomes. They emphasized that the level of inherent uncertainty in predicting future outcomes is a crucial factor that can impact human reliance on AI systems. Inspired by this work, we operationalize task uncertainty in our study using the distinction between diagnostic and prognostic tasks.

In this paper, we aim to fill an empirical and research gap by examining the impact of task complexity and uncertainty, as important attributes in decision-making in real-world contexts. By providing application-grounded evaluation [32] with users relying on an AI system for assistance in practical tasks, our work is the first to explore task uncertainty and how task uncertainty interacts with task complexity in shaping human-AI decision-making.

3 HYPOTHESES AND TASK DESIGN

3.1 Hypotheses

The degree of task complexity is deemed one of the primary indicators for determining the success of Human-AI teams [3, 9, 23, 83]. Consequently, it can be anticipated that as tasks become more complex, their influence on human reliance on AI systems increases [23, 82, 105]. More complex tasks tend to require more cognitive effort [23], making individuals more likely to rely on AI systems for assistance. Moreover, as task complexity increases, the verifiability [40] and plausibility [57, 60] of AI advice tend to decrease. This can pose challenges for individuals in distinguishing misleading AI suggestions, leading to reduced levels of appropriate reliance on AI systems. Although there may not be a correlation between trust and reliance on AI systems [63, 71, 90, 114], prior work suggests a higher likelihood of individuals placing greater trust in AI systems for more complex tasks [53, 71].

When faced with prognostic tasks, individuals are likely to perceive them as more complex and unpredictable, thus increasing their reliance on AI systems for assistance. With the presence of uncertainty in a task, individuals may lack sufficient capability to verify the correctness of AI advice and therefore rely more heavily on the AI systems [29], leading to reduced appropriate reliance on AI systems. Previous research has also demonstrated the influence of uncertainty on trust formation in AI systems [121]. Considering highly complex and prognostic tasks, we hypothesize that individuals exhibit higher levels of trust and reliance on AI systems while showing a decrease in appropriate reliance. This could be due to the high cost of engaging cognitively in complex decision-making processes, leading to a greater reliance on AI systems for guidance [126]. Therefore, we formulate our hypotheses as shown in Table 1.

3.2 Trip-Planning Task

We chose trip-planning to as the scenario for our study due to two primary reasons. Firstly, trip-planning is a common real-world problem that individuals frequently encounter and seek assistance from AI systems to make decisions. Secondly, this task allows us to meaningfully manipulate complexity levels (e.g., the number of constraints) and uncertainty levels in our experimental conditions, thereby enhancing the ecological validity of our findings. In our study, participants are presented with a practical scenario where external assistance is potentially useful to successfully accomplish the task. We utilized an imperfect AI system with a 66.7% accuracy rate for trip-planning and manipulated its features accordingly (cf. section 4.1). This setup with the necessary complexity creates the desired sense of vulnerability and uncertainty, making it a suitable situation for analyzing human trust and reliance on AI systems [58, 71]. Note that while trip planning is a frequently encountered real-world task, the inclusion of time and budget limitations makes it unique, affecting how individuals rely on AI assistance.

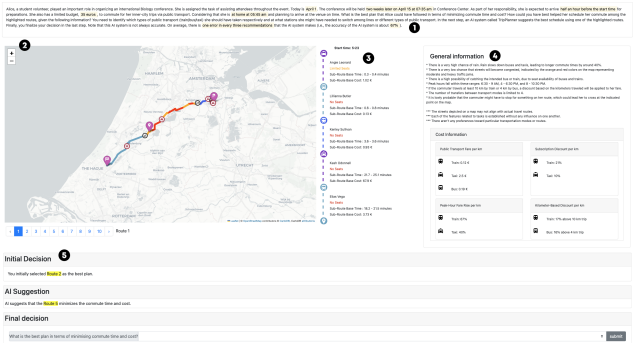


Figure 1: An overview of the trip-planning task interface that participants used including five components: (1) task scenario and description, (2) map, (3) route information, (4) general information, and (5) two-stage decision-making. Note that this screenshot is meant to convey a bird's-eye view of the interface. This interface is also dedicated to a highly complex scenario encompassing all constraints and the prognostic experimental condition with high uncertainty.

Planning a trip involves determining the most suitable route for travel, taking into account factors such as time limitations and budget constraints. Participants are tasked to select the trip that minimizes both travel time and expenses. Each task typically consists of multiple components that support participants in making well-informed decisions, as depicted in a bird's-eye view of the task interface in Figure 1.

Quality Control: To ensure the accuracy and reliability of the collected data in our study, we employed multiple methods. We initially offered instructional materials on the interface and task-related features, followed by a training session for participants that included both theoretical instruction and hands-on practice. Secondly, we evaluated participants' comprehension by administering

Table 1: Summary of Our Hypotheses.

Hypothesis	Description
H1a	Users demonstrate a lower level of appropriate reliance on AI systems for complex tasks compared to relatively less complex tasks.
H1b	Users trust AI systems to a greater extent in complex tasks compared to relatively less complex tasks.
H2a	Users demonstrate a lower level of appropriate reliance on AI systems in tasks with high levels of uncertainty compared to tasks with low levels of uncertainty.
H2b	Users trust AI systems to a greater extent in tasks with a high degree of uncertainty (`prognostic`) compared to tasks with lower levels of uncertainty (`diagnostic`).
H3	Users demonstrate a relatively low level of appropriate reliance on AI systems in tasks with relatively high complexity and uncertainty.

a quiz on task-related constraints. Individuals who scored below a certain threshold were excluded from the study to maintain the quality of data. Lastly, we incorporated four attention-check questions in the pre-questionnaire and post-questionnaire to screen out individuals who may not be fully engaged or attentive throughout the study. Detailed explanations of these methods are publicly available on our companion page.[1]

3.3 Design Considerations and Setups: Task Complexity vs. Task Uncertainty

Wood's seminal work [133] proposed that task complexity consists of three constructs: component, coordinative, and dynamic complexities. *Component complexity* relates to the number of features in a task, while *coordinative complexity* pertains to executing sequences or steps within the task. *Dynamic complexity* arises from changing world states requiring further considerations at the point of decision-making. We utilized component complexity to define task complexity and also adjusted the uncertainty as incomplete information in our setup. In dynamically complex tasks, decision-making must adapt as the situation changes, with all information accessible at each point. However, uncertain tasks involve incomplete information at the point of decision-making, setting them apart from dynamically complex tasks. Therefore, it is valid to consider these factors as separate dimensions although task uncertainty can increase task complexity.

3.3.1 ***Task Complexity:*** To operationalize task complexity in our experimental conditions, we manipulated the number of constraints that are given to participants. This approach has been used in previous studies to control the level of complexity for a given task [9, 105, 120]. We categorized the tasks into three levels of complexity: `low`, `medium`, and `high`. In low-complexity tasks, participants are presented with **four** features to consider while in medium-complexity tasks, **eight** features are provided. High-complexity tasks entail **twelve** different features that must be taken into account. This design choice is guided by prior neuroscience research by Miller [91], suggesting that human cognitive capacity for processing information is limited to around seven (± two) chunks of information at a time. Hence, we established five to nine task features as representative of a medium level of complexity based on this finding. Any number exceeding nine would classify as high complexity, while four or fewer would indicate low complexity [109].

3.3.2 ***Task Uncertainty:*** **Diagnostic** tasks entail circumstances where participants are given access to well-defined and comprehensive information about the current task, allowing them to make precise judgments [29]. **Prognostic** tasks, on the other hand, involve scenarios in which participants are presented with restricted or unclear data and need to generate predictions regarding future outcomes [29]. The necessity to anticipate uncertain results gives rise to increased uncertainty throughout the process of making decisions. To operationalize uncertainty in the contrasting experimental conditions pertaining to diagnostic and prognostic tasks, we employed various strategies.

For diagnostic tasks, participants are instructed to schedule a trip for the present moment within the narrative, while for prognostic tasks, participants are assigned to plan a trip that will take place two weeks later. Next, we customized the way task attributes are presented to align with the level of uncertainty. In situations involving diagnostic tasks, participants are given precise values for each constraint, eliminating any potential ambiguity. On the other hand, in prognostic tasks, a certain degree of uncertainty is introduced by offering participants ranges or estimates instead of exact values for each attribute. We also presented the probability of different outcomes for certain constraints. For example, we highlighted the high likelihood of encountering traffic congestion during the rush hour or the low chance of experiencing rain during the scheduled trip.

We created one task scenario for each task. In total, we generated 24 different scenarios, with four scenarios in each experimental condition that differed in terms of task complexity and uncertainty. The **full list of these task scenarios** and **all code for our implementation** is publicly accessible for the benefit of the research community and in the spirit of open science.[1]

3.3.3 ***Task Features:*** We designed task features to impart and define constraints in the decision-making tasks such that they do not affect each other and can be independently manipulated and measured. We communicated this independence explicitly and implicitly by ensuring that each feature is presented separately and does not rely on or interact with other features. All task features were inspired by considerations typical in real-world trip-planning contexts. In our research, we can classify task characteristics from two different viewpoints: each feature has the potential to influence

[1] https://osf.io/kt8m4/?view_only=c6930ba990c8412cb3948c2cf2b0a39c

either the overall duration of travel, the associated expenses, or both factors. Furthermore, each feature can be categorized as being either time-dependent or time-independent. Time-dependent features, such as traffic conditions and weather patterns, are prone to temporal changes based on external factors and their presentation differs when considering diagnostic tasks versus prognostic tasks. In tasks that have low complexity, we designed an equal distribution of time-dependent and time-independent features. However, for tasks with medium or high complexity, we increase the number of time-dependent features to enhance the degree of uncertainty that need to be considered in decision-making processes. Detailed explanations of all features are publicly available on our companion page.[1]

4 STUDY DESIGN

4.1 Experimental Conditions

Our study was approved by our institutional ethics board. We designed a between-subject study with a 3×2 factorial design. The three levels for task complexity were categorized as `low`, `medium`, and `high`, while the two distinct levels for uncertainty were `diagnostic` and `prognostic` tasks. We refer to these conditions as LowDiag, LowProg, MedDiag, MedProg, HighDiag, and HighProg. Participants were randomly assigned to one of the six experimental conditions while ensuring a balanced distribution of participants across the different task complexity and uncertainty levels. In each condition, participants were presented with three different task instances to complete with the assistance of an AI system. The three task instances were determined based on each condition's assigned complexity and uncertainty levels. Detailed explanations regarding the complexity and uncertainty levels are provided in section 3.3.

We fine-tuned the AI system to suggest routes that satisfy the given criteria with an accuracy of 66.7% across all experimental conditions. This level of accuracy was chosen since it is helpful if the system is relied on but still involves some risks. Hence, it calls for appropriate reliance instead of blindly following the AI system's advice. This design choice is motivated by prior work emphasizing the role of uncertainty in dictating the need to facilitate appropriate reliance [71]. This implies that within each batch of three task instances that a participant completes, to control for potential ordering effects, we ensure that incorrect advice is offered by the AI system once at random.

4.2 Measures

We leveraged a set of objective metrics to quantify participants' reliance on the AI system (cf. Table 2) [58, 88, 90, 113, 139, 141]. These metrics include Agreement Fraction, Switch Fraction [51, 138, 142], and Accuracy with Disagreement [51], Relative Positive AI Reliance, and Relative Positive Self-Reliance [112]. These parameters are commonly adapted in literature to capture the level of reliance within the human-AI interaction context. In addition to these measures of reliance, we also evaluated participants' decision-making accuracy, demonstrating the human-AI team performance [11, 108]. By measuring trust and reliance variables alongside human-AI team performance, we can gain a deeper understanding of whether performance outcomes result from under-reliance, appropriate reliance, or over-reliance on AI systems.

The subjective trust in the AI system was assessed using the Trust in Automation questionnaire (TiA) [63], which is a commonly employed and validated tool for measuring trust [78, 116, 120]. The questionnaire comprises multiple items that evaluate various aspects such as participants' perceptions regarding Reliability/Competence (TiA-R/c), Understanding/Predictability (TiA-U/P), Familiarity (TiA-Familiarity), Intention of Developers (TiA-IoD), the Propensity to Trust (TiA-PtT), and the overall level of trust placed in the AI system, Trust in Automation (TiA-Trust).

We collected information about participants' perceived numeracy skills as well as their affinity for technology in the pre-task questionnaire. To measure numeracy skills, we employed the Subjective Numeracy Scale [38], which is a self-report measure of perceived ability to perform various mathematical tasks and preference for the use of numerical information. Additionally, we administered the Affinity for Technology Interaction Scale (ATI) [41] to determine participants' level of comfort and familiarity with technology [120].

4.3 Participants

We first estimated the required sample size using G*Power software, considering a medium effect size of 0.25, a power of 0.90, and a significance level of 0.05, leading to a recommended minimum sample size of 210 participants, *i.e.,* 35 participants in each of our experimental condition. To obtain a sufficient sample for our study while accounting for potential exclusion, we enlisted the participation of 285 individuals using the Prolific crowdsourcing platform. To ensure the reliability of the data gathered, we applied inclusion criteria that were designed to select native English speakers with a minimum approval rate of 95% on the platform and at least 100 completed studies. A total of 27 participants who failed any attention-check questions or the quiz were excluded from participation in the study, resulting in a final sample size of 258 participants. On average, participants took approximately 25 minutes to complete the entire study. All participants were compensated at the fixed rate of 8 GBP per hour regardless of their performance in the study. Additionally, participants received bonus rewards amounting to 0.2 GBP for each accurate response they provided during the study period. Overall, participants earned an average of 8.44 GBP per hour, well over the wage considered to be '*good*' and recommended by the Prolific platform.

4.4 Procedure

The entire workflow of the study is illustrated in Figure 2. When participants entered the study, they were first provided with informed consent, a brief overview of the study's goals, and instructions on how to complete the tasks (step 1). If they consented to participate, they were directed to the pre-task questionnaire in step 2, where they were presented with a series of questions related to their numeracy skills and affinity for technology. Participants were then randomly assigned to one of the six different experimental conditions. According to the assigned condition, participants were presented with an interface tutorial and task tutorial that provided step-by-step instructions on how to navigate and complete the task followed by a training session on a sample task. The participants were given sufficient time to familiarize themselves with the sample

Table 2: An overview of the different metrics that we considered in our user study.

Metric Type	Metric Name	Value Type	Value Range
Performance	Accuracy	Continuous	[0,1]
Reliance	Switch Fraction	Continuous	[0,1]
	Agreement Fraction	Continuous	[0,1]
Appropriate Reliance [51, 112]	Accuracy-wid	Continuous	[0,1]
	RAIR	Continuous	[0,1]
	RSR	Continuous	[0,1]
Trust	TiA-ReliabilityCompetence	Likert	5-point, strong distrust to strong trust
	TiA-UnderstandingPredictability	Likert	5-point, strong distrust to strong trust
	TiA-Intention of Developers	Likert	5-point, strong distrust to strong trust
	TiA-Trust in Automation	Likert	5-point, strong distrust to strong trust
Covariates	Subjective Numeracy (SNS)	Likert	6-point: from low to high
	Affinity for Technology (ATI)	Likert	6-point: low to high
	TiA-Familiarity	Likert	5-point, strong distrust to strong trust
	TiA-Propensity to Trust (TiA-PtT)	Likert	5-point, strong distrust to strong trust

task and the interface. To ensure the understanding of the task, participants were required to answer a quiz related to the task features before proceeding to the main task. If participants did not pass the quiz, they were excluded from the study. Otherwise, they received immediate feedback on their quiz performance to ensure that participants proceeded to the main task with a complete understanding of the task and devoid of familiarity or comprehension-related biases. Participants were then asked to complete three trip-planning tasks. Each task instance consisted of a decision-making scenario, where participants had to analyze the information provided and make an AI-assisted decision. Lastly, participants were directed to fill out a post-task questionnaire to assess their perception of the task features and trust in the AI system.

Figure 2: Illustration of the procedure participants followed within our study.

5 RESULTS

5.1 Descriptive Statistics

5.1.1 Demographics. The resulting sample of 258 participants had an average age of 38 years old ($SD = 11.8$) and consisted of 39% females and 61% males. To account for potential confounding variables, we gathered information about the participants' subjective numeracy skill (SNS), affinity for technology (ATI), TiA-Familiarity, and TiA-Propensity to Trust (TiA-PtT). Participants reported a moderate level of perceived numeracy ($M = 4.28, SD = 0.80$) on the 6-point scale. Similarly, participants were found to have a moderate affinity for technology interaction ($M = 4.04, SD = 0.56$) measured on a 6-point scale, low familiarity ($M = 2.87, SD = 1.17$), and a moderate propensity to trust AI ($M = 3.72, SD = 0.49$) measured on a 5-point scale.

5.2 Hypothesis Tests

H1a. Impact of task complexity on appropriate reliance: To explore the main effect of complexity on appropriate reliance, we conducted a Kruskal–Wallis test, Table 3. Subsequently, we conducted Dunn's post-hoc test to determine which levels of complexity resulted in significant differences in appropriate reliance. We reported *adjusted p-values*, calculated using Bonferroni correction to account for the increased likelihood of falsely declaring statistical significance when conducting multiple tests. If the *adjusted p-value* for an individual hypothesis is less than the significance level (0.05), then the null hypothesis is rejected, indicating a statistically significant result [135]. We first report the influence of complexity on reliance, followed by our examination of appropriate reliance.

The observed significant difference in switch fraction between high and low-complexity tasks implies that task complexity does indeed exert an influence on **reliance**. In tasks with higher complexity levels, individuals tend to shift from relying on their own judgment to relying on the AI system. This can be attributed to a decrease in self-confidence regarding their decision-making abilities and, as a result, seeking guidance from the AI system.

Tasks of higher complexity tend to diminish the **appropriate reliance** on the AI system. Participants demonstrated significantly lower levels of *Accuracy-wid* in tasks with greater complexity compared to those with lower complexity. A similar trend is observed when examining *RSR*, wherein participants displayed significantly reduced levels of confidence in themselves during tasks with higher complexity than those with lower complexity. Consistent with these findings, participants exhibited a contrasting trend in displaying a significantly higher level of reliance on the AI system for tasks that were more complex compared to those of lower complexity, as indicated by higher *RAIR*. The rise in RAIR does not necessarily imply a higher appropriate reliance on the AI system. Rather, it suggests that individuals under-rely on the AI system in tasks with relatively lower complexity, and over-rely on the AI system in tasks with relatively higher complexity without being able to recognize when the advice may be inaccurate. This excessive reliance can ultimately have a negative impact on performance by reducing appropriate reliance levels.

Table 3: Kruskall-Wallis test for the main effect of task complexity on reliance. † indicates that the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.

Dependent Variable	*adjusted-p*	$M \pm SD$ (Low)	$M \pm SD$ (Medium)	$M \pm SD$ (High)	Post-hoc Results
Agreement Fraction	.8	0.62 ± 0.20	0.54 ± 0.25	0.55 ± 0.28	-
Switch Fraction	**.003**†	0.18 ± 0.32	0.26 ± 0.30	0.34 ± 0.36	Low < Medium < High
Accuracy	**<.001**†	0.79 ± 0.22	0.58 ± 0.27	0.61 ± 0.29	Low > Medium, High
Accuracy-wid	**.001**†	0.61 ± 0.40	0.40 ± 0.37	0.50 ± 0.35	Low > Medium, High
RAIR	**.001**†	0.22 ± 0.41	0.33 ± 0.41	0.43 ± 0.45	Low < Medium, High
RSR	**<.001**†	0.64 ± 0.48	0.34 ± 0.48	0.38 ± 0.49	Low > Medium, High

Furthermore, we found that the **accuracy** of participants is significantly lower in tasks with higher levels of complexity than those with lower complexity. This finding provides additional evidence to our previous findings regarding the influence of task complexity on appropriate reliance. Overall, these results **partially support** our hypothesis **H1a**.

H1b. Impact of task complexity on trust: We aimed to examine the main effect of task complexity on trust in the AI system. Therefore, we conducted a two-way ANCOVA to consider the potential confounding effects of the covariates, namely subjective numeracy skill, affinity for technology, TiA-Familiarity, and TiA-Propensity to Trust. We did not find a significant effect of task complexity on human trust in the AI system, leading us to **reject** our hypothesis **H1b**. However, this finding supports that the subjective nature of trust in the AI system does not always follow the objective measure of reliance on the AI system [90, 114].

H2a. Impact of task uncertainty on appropriate reliance: We investigated the main effect of task uncertainty on reliance by conducting the Kruskal–Wallis test, reported in Table 4. We found that task uncertainty significantly affects participants' **reliance** on the AI system. Participants showed significantly higher levels of switch fraction when faced with prognostic tasks, indicating their tendency to rely more on the AI system due to lower self-confidence. Our findings further suggest that individuals can accurately assess the level of uncertainty in a task and adjust their reliance on the AI system accordingly.

Furthermore, our findings revealed that the degree of uncertainty in a task significantly influenced participants' **appropriate reliance** on the AI system. We found that participants were more likely to appropriately rely on the AI system in diagnostic tasks, leading to higher accuracy rates, as indicated by higher *Accuracy-wid* compared to prognostic tasks. In line with this finding, we also observed that participants exhibited a slightly higher level of reliance on their own decision-making skills (*RSR*) when faced with diagnostic tasks. On the other hand, in prognostic tasks, participants showed significantly higher degree of reliance on the AI system as indicated by higher *RAIR*. This finding suggests that participants tend to rely heavily on the AI system in uncertain situations. However, this does not necessarily lead to appropriate reliance. It can be challenging for them to distinguish between accurate and inaccurate AI advice in prognostic tasks, resulting in lower appropriate reliance on the AI system and decreased accuracy levels. As a result, our findings **partially support** the hypothesis **H2a**.

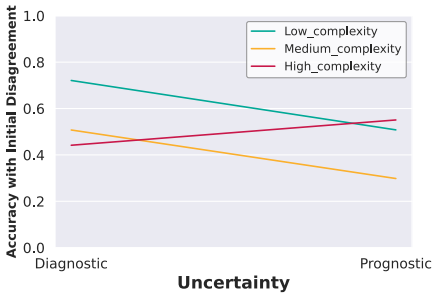
H2b. Impact of task uncertainty on trust: The main effect of task uncertainty on trust in the AI system was also examined in this study through the ANCOVA test. The results indicated that there was no significant main effect of task uncertainty on any trust subscales. These findings indicate that participants' trust in the AI system remains relatively stable regardless of the level of uncertainty in the task. Thus, we **reject** our hypothesis **H2b**.

H3. Interaction effect of task complexity and uncertainty: We conducted an ANOVA to investigate the interaction effect of task complexity and uncertainty on appropriate reliance and trust. We found a significant interaction effect between task complexity and uncertainty on *Accuracy-wid* as a measure of appropriate reliance. Figure 3a illustrates the interaction effect of task complexity and uncertainty on Accuracy-wid, focusing on different levels of complexity. We observed that the trend of Accuracy-wid is descending for tasks with low and medium complexity while increasing the level of uncertainty. However, for tasks with high complexity, the trend is the opposite, where Accuracy-wid increases with increasing uncertainty. Although we found earlier that participants have a lower *Accuracy-wid* for prognostic tasks, the interaction effect suggests that the impact of uncertainty on appropriate reliance depends on the level of task complexity. This finding suggests that participants tend to engage more cognitively in tasks they perceive as less complex, believing they can make accurate judgments. This trend is also observed in diagnostic tasks with high complexity. However, when faced with highly complex and prognostic tasks, participants are more likely to relinquish some cognitive control and rely heavily on the AI system. This could be attributed to their perception of the task's complexity exceeding their own capabilities. Participants may also view the AI advice as being more reliable and trustworthy, resulting in increased agreement and appropriate reliance. This finding is further supported by the significant interaction effect identified in *Accuracy*, Figure 4a, demonstrating that participants' ability to make accurate predictions increases when they are faced with prognostic tasks with high complexity, compared to prognostic tasks with medium and low complexity. Consequently, their level of accuracy aligns with that of the AI system due to their increased appropriate reliance. Figures 5a and 5b illustrate the Accuracy and Accuracy-wid for different levels of task complexity and uncertainty.

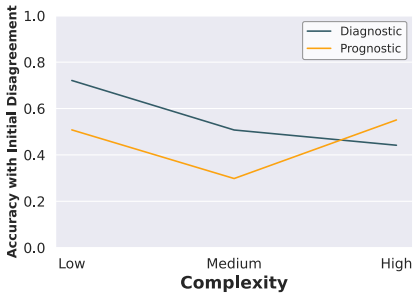
We can observe the interaction effect of complexity and uncertainty for diagnostic and prognostic tasks in Figure 3b. For diagnostic tasks, the trend *Accuracy-wid* is descending as the complexity of the task increases. However, for prognostic tasks, different effects

Table 4: Kruskall-Wallis test for the main effect of task uncertainty on reliance. † indicates the effect of the variable is significant in the comparisons shown in the 'Post-hoc Results' column.

Dependent Variable	*adjusted-p*	*M* ± *SD* (Diagnostic)	*M* ± *SD* (Prognostic)	Post-hoc Results
Agreement Fraction	**.01**†	0.60 ± 0.23	0.54 ± 0.26	Diagnostic > Prognostic
Switch Fraction	**.02**†	0.22 ± 0.32	0.31 ± 0.34	Diagnostic < Prognostic
Accuracy	**<.001**†	0.72 ± 0.30	0.60 ± 0.24	Diagnostic > Prognostic
Accuracy-wid	**.04**†	0.56 ± 0.43	0.45 ± 0.33	Diagnostic > Prognostic
RAIR	**.02**†	0.27 ± 0.42	0.38 ± 0.44	Diagnostic < Prognostic
RSR	.1	0.50 ± 0.50	0.40 ± 0.49	-

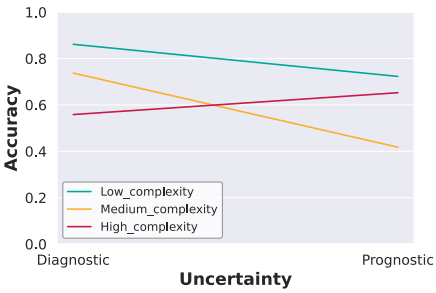


(a) Different task complexity levels across uncertainty levels

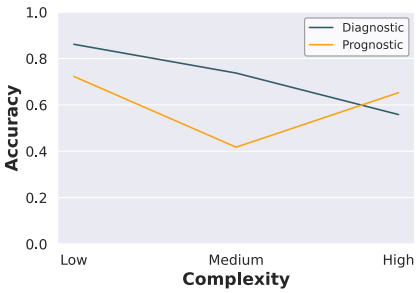


(b) Different task uncertainty level across complexity levels

Figure 3: Interaction effects between task complexity and uncertainty on the *Accuracy-wid* metric reflecting appropriate reliance.



(a) Different task complexity levels across uncertainty levels



(b) Different task uncertainty levels across complexity levels

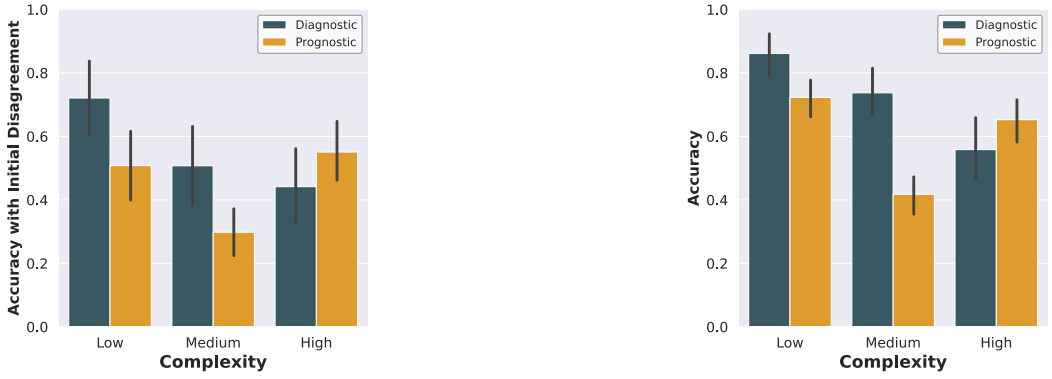
Figure 4: Interaction effect between complexity and uncertainty on *Accuracy* metric.

are observed. Participants tend to have lower *Accuracy-wid* as we increase the complexity from low to medium. In medium-complexity tasks, *Accuracy-wid* reaches its local minimum. So, as we further increase the complexity to high levels, *Accuracy-wid* starts to rise again, suggesting that participants rely more appropriately on the AI system, and their accuracy improves in highly complex prognostic tasks, aligning more closely with accuracy of the AI system (cf. Figure 4b). Furthermore, we can see that the appropriate reliance is always greater for diagnostic tasks compared to prognostic tasks, except for high complexity, where the values for prognostic tasks surpass those for diagnostic tasks, further supporting our findings. In summary, we found that the interaction effect between complexity and uncertainty in conditions with high complexity and uncertainty plays a significant role in human-AI decision-making. While the appropriate reliance drops as the complexity and uncertainty of a task increase, there is a turning point where participants start to rely more appropriately on the AI system, resulting in increased accuracy in prognostic tasks with high complexity. Thus, our findings **reject** hypothesis **H3**.

6 DISCUSSION

6.1 Key Findings

Our study examined the impact of task complexity and uncertainty on human-AI decision-making. The results of our study demonstrated that increasing the level of **complexity** and **uncertainty** in decision-making tasks led to significant differences in **users' reliance** on the AI system. In more complex and uncertain tasks, we found that users were often in initial disagreement with the advice provided by the AI system. However, they demonstrated a



(a) Mean of Accuracy-wid

(b) Mean of Accuracy

Figure 5: Mean of Accuracy-wid and Accuracy across different levels of task complexity and uncertainty.

heavy reliance on AI advice during the second stage of the decision-making process, leading to higher *Switch Fraction*. This can be attributed to the potential recognition that AI offers valuable insights for decision-making under complexity and uncertainty, coupled with a lack of confidence in their own judgment, corroborating what has been uncovered by other work in human-AI decision-making [23, 105]. Furthermore, the greater cognitive effort linked to complex tasks may also be a contributing factor. The cost of relying on the AI system would prove to be less compared to evaluating the reliability of the AI advice, thereby prompting individuals to lean towards following AI advice [126]. Additionally, users showed higher engagement and information-gathering behavior in prognostic scenarios, demonstrated by significantly more clicks on *route control buttons*, indicating greater inclination to explore different route options.

We also found that the **appropriate reliance** on the AI system varied significantly depending on **task complexity** and **uncertainty**. Users exhibited lower appropriate reliance on the AI system (lower *Accuracy-wid*), leading to lower accuracy in tasks with `medium` complexity or uncertainty compared to those with `low`. However, users demonstrated higher appropriate reliance on the AI system, resulting in improved accuracy in the experimental conditions with tasks with `high` complexity or uncertainty compared to those with `medium` complexity or uncertainty. Users perceived that tasks with higher complexity and uncertainty required greater effort and information processing, making them more willing to rely on the AI system. In such scenarios, their performance approaches AI accuracy, indicating the effectiveness of integrating AI in decision-making.

Our findings showed that individuals generally place significantly more **reliance** on the AI system when faced with tasks characterized by **high uncertainty**. However, in such prognostic tasks, their ability to **appropriately rely** on AI advice is lower compared to diagnostic tasks, subsequently affecting their overall performance. Tasks that involve inherent uncertainty are often those where humans tend to rely on AI systems for advice, such as loan approval [27, 34, 124], recidivism prediction [31, 47, 86], house price estimation [2, 13, 25], and student admission [13, 24]. Individuals may be more inclined to adhere to AI advice in these types of tasks. This could stem from the belief that AI systems possess advanced analytical abilities and have access to a greater amount of data [75]. On the other hand, when individuals are faced with tasks that have lower uncertainty, such as annotation and classification task [4, 77, 117], they tend to rely less on the AI advice and rely more on their expertise and judgment. Since the heavy reliance on AI systems in uncertain situations does not always lead to improved decision-making accuracy, several mechanisms have been proposed to optimize the combination of human and AI decisions to achieve the best outcomes and facilitate appropriate reliance on the AI system. These mechanisms include providing interpretable explanations for AI advice [21, 72, 123], using cognitive forcing functions[17, 47, 99], and incorporating feedback loops to enhance the interaction between humans and AI systems [10, 11, 139]. Despite implementing a two-stage decision-making process to encourage individuals to be cognitively involved in the procedure, as well as incorporating visual and textual explanations for increased transparency, our research emphasizes the necessity for additional exploration into strategies that can facilitate appropriate reliance on AI systems in contexts characterized by high levels of uncertainty.

The **complexity** of tasks plays a significant role in determining the degree of **reliance** on AI advice, consistent with the findings of [9, 105]. The more complex a task is, the more individuals may be inclined to rely on the AI system. We use the number of features or constraints as the measure of task complexity similar to previous studies [9, 105, 120]. Tasks with a larger number of constraints that need to be accounted for in decision-making are often more challenging for individuals to process, making them more likely to seek guidance from AI [63, 90, 111]. Our findings, which were based on objective measures, align with [126] study and suggest that users tend to rely more heavily on AI systems when faced with complex tasks that demand higher cognitive effort. This is further backed by [100] indicating that the complexity of a task can elevate its perceived difficulty, potentially resulting in greater reliance on AI systems. As shown by Salimzadeh et al. [109], the majority of tasks that have been studied in the context of decision-making are characterized by `low` and `medium` complexity. Prior studies that investigated tasks exceeding individual information processing capabilities (i.e., 9 constraints [91]) suggested employing visualization techniques to assist individuals in understanding the AI advice and the underlying decision-making process [43, 137, 140].

We used visual and textual techniques to support individuals in understanding the factors playing a role in shaping the given AI advice. However, in higher complexity scenarios, an individual still lacks cognitive engagement with the AI system and may be more likely to rely heavily on its advice. This is supported by the tendency of individuals to rapidly make their decision within approximately twenty seconds after receiving advice from AI, without carefully reassessing the provided information or exploring alternative route options. Although these visual and textual strategies have shown promise in improving decision-making outcomes in literature, they were not sufficient to mitigate over-reliance on AI advice in high complexity tasks.

According to the Trustworthiness Assessment Model (TrAM) [111], accurate perceived trustworthiness of AI systems is essential for establishing meaningful trust and reliance on AI systems. Factors such as relevance and availability of system information, as well as the ability of individuals to detect and utilize this information, play a crucial role in determining accurate perceived trustworthiness. In our study, we only presented relevant task features using visual and textual formats to participants. We utilized user behavior metrics and validation of participant perceptions through training and quizzes to ensure the detection of these features. However, we expected the complexity and uncertainty of tasks to impact the availability and utilization of system information, thus affecting perceived trustworthiness [98]. However, participant trust remained consistent regardless of task complexity or uncertainty, which was in contrast to what is suggested by the TrAM framework.

6.2 Implications of Our Work

6.2.1 Implications for Methodology and the HCI Community. The implications of methodology in HCI research pertain to the design and analysis of studies [125]. These implications specifically address data collection methods and the construction of new knowledge. Our work has important implications for the methods used to study human-AI decision-making, for increasing the external validity of empirical work and strengthening the understanding of the transferability of findings across different studies. It has been observed that task characteristics, such as complexity and uncertainty, are seldom examined or analyzed systematically in human-AI decision-making studies. While it may not be experimentally feasible to account for every facet of a task, our research emphasizes the significance of considering these factors when assessing human-AI collaboration. Future research should consider the incorporation of methodologies that take into account task-related features when evaluating human-AI decision-making. Our findings also contribute to the interpretation of human behaviour and reliance on AI systems through the lens of task complexity and uncertainty. Current studies often focus on generic decision-making scenarios or tasks with low to medium complexity, which may not fully reflect or represent the challenges and dynamics of the full range of real-world scenarios. This is particularly important in highly complex tasks coupled with high uncertainty, where humans tend to require, appreciate, and rely on advice from an AI system. Future research should consider the systematic identification and inclusion of task-specific characteristics in the design of studies in the realm of human-AI decision-making.

To initiate a systematic evaluation of task characteristics, we propose the lens of diagnostic and prognostic tasks as a framework for modeling uncertainty in decision-making, which can be used as a basis for designing experiments and gathering data on human-AI interactions. This approach acknowledges the inherent uncertainty in determining or estimating different constraints that influence decision outcomes. Additionally, it offers a relatively more precise representation of decision-makers' challenges. Incorporating this lens into research methodology would involve designing studies that specifically control the uncertainty inherent in diagnostic and prognostic tasks and exploring their impact on human-AI decision-making processes and outcomes. We also encourage researchers to consider highly complex tasks in their experiments to capture the challenges and nuances of decision-making in real-world scenarios. This can be achieved by developing scenarios or simulations that closely resemble complex decision-making situations in different domains. Our task details and all code for the interface are made publicly available to support future research in the community.[2]

Our study also highlights the need for further examination and development of techniques tailored specifically to support high-complexity and prognostic tasks in human-AI decision-making. Although many interventions have been developed for decision-making in various domains, there is still a need to focus on the unique challenges posed by high complexity and prognostic tasks. Such interventions could be targeted to offer users indicators that can help them accurately assess the reliability, plausibility, and verifiability of the AI advice. Consequently, these methods will promote appropriate reliance on the AI system in complex and uncertain decision-making scenarios. There is a heightened urgency in developing and creating these mechanisms to prevent potential deception arising from the complexity and uncertainty of tasks, which can make it challenging to detect untrustworthy AI systems [8]. By reducing the cost of verifiability and plausibility of such XAI techniques, decision-makers can gain a better understanding of the basis for AI advice based on their own expertise and judgment, potentially leading to improved performance and appropriate utilization of AI systems.

The decline in performance of human-AI teams when tackling tasks of medium complexity suggests that users may have faced challenges in accurately assessing their own abilities and the capabilities of AI systems, primarily by overestimating their own abilities [64]. This aligns with previous research findings, highlighting the need for interventions to assist users in evaluating their skills and appropriately adjusting their reliance on AI systems [25, 51, 69]. This may be particularly important in tasks with relatively moderate complexity which may lead to illusory self-assessments among some users, compared to tasks with evidently low or discernibly high complexity.

6.2.2 Implications for Theory. Theoretical implications focus on the understanding of task characteristics and their impact on human-AI decision-making. Based on our findings, it is evident that the complexity and uncertainty of tasks significantly influence how humans rely on AI systems. This study serves as the application-grounded evaluation [32] in the context of trip-planning, centering on the individuals the system intends to support in actual tasks. It

[2] https://osf.io/kt8m4/?view_only=c6930ba990c8412cb3948c2cf2b0a39c

empirically validates the commonly held belief that task complexity and uncertainty play a crucial role in determining human reliance on AI systems. While the primary objective of combining humans and AI is to achieve enhanced performance through collaboration, an over-reliance on AI can potentially impede the advantages offered by human judgment and decision-making abilities. Therefore, it is crucial for researchers to develop theoretical frameworks that can help identify and motivate the optimal balance between human and AI involvement in decision-making, taking into consideration task complexity and uncertainty.

Contrary to previous research suggesting that trust in AI systems increases with the complexity and uncertainty of tasks, our findings indicate that trust is orthogonal to these factors. These results suggest that trust is not the sole determinant of reliance on AI advice, and other factors such as task characteristics play a significant role. This also indicates the difference between human trustworthy beliefs and behavior toward AI systems, where trust may not always translate into increased reliance, highlighting the need to measure, calibrate, and understand factors beyond trust that influence human-AI decision-making.

6.3 Caveats and Limitations

According to the checklist of cognitive biases provided by Draws et al. [33], it is important to acknowledge that humans are prone to cognitive biases. In our task, we identify the familiarity bias and availability heuristic, which can cause individuals to exhibit an inclination towards decisions that align with their pre-existing beliefs or past experiences. Although we created artificial routes, individuals may still tend to prefer familiar or known options or prefer specific transport modes due to personal biases. Confirmation bias and overconfidence bias are other potential limitations, as individuals may be more likely to seek out and give more weight to information that confirms their preconceived notions or beliefs regarding AI capabilities and their decision-making abilities. We should also consider the self-interest bias, where individuals may prioritize their own monetary reward over objective decision-making criteria.

The findings discussed in this paper are not universally applicable to all decision-making tasks. Different tasks may have varying characteristics and contexts that can influence human-AI decision-making. Although this is a valid approach to operationalize uncertainty, it is important to acknowledge that there could be other approaches to capturing task uncertainty that were not explored in this study (e.g., missing data or conflicting information). Future research should consider exploring different operationalizations of task complexity and uncertainty to further understand their impact on human reliance on AI systems. It is worth noting that we asked participants in our study to consider that the traffic features were unrelated to each other and carried equal weights in determining the best route. This may not always be the case in real-world contexts. We also considered traffic conditions in both diagnostic and prognostic scenarios, although, in the real world, traffic conditions can change over time and at the time of decision-making, making them predominantly prognostic.

7 CONCLUSION AND FUTURE WORK

In this study, we explored how task complexity (**RQ1**) and task uncertainty (**RQ2**) and their interaction (**RQ3**) inform user trust and appropriate reliance on AI systems. To this end, we conducted a user study with 258 participants across six experimental conditions varying in three levels of task complexity (`low`, `medium`, and `high`) and two levels of task uncertainty (`diagnostic` and `prognostic`). We selected trip-planning as the decision-making task and evaluated participants' trust, reliance, and decision-making behaviors when interacting with an AI system. The study showed that task complexity and uncertainty significantly impact human reliance on AI systems. Participants tended to rely more on AI in tasks with higher complexity and uncertainty, with no significant differences in human trust across different levels of complexity and uncertainty.

Future studies should further explore the relationship between task complexity and uncertainty to better understand their interconnections in human-AI decision-making. Further research is needed across a range of domains and task types to fully understand the impact of task complexity and uncertainty. We encourage researchers to investigate the impact of other task characteristics, such as time pressure and information overload, on human-AI decision-making. Future work should also focus on understanding how to effectively present AI-generated predictions and explanations to enhance human understanding and decision-making, particularly in complex and uncertain situations. Given the increasing complexity and uncertainty of tasks, it becomes crucial to develop strategies that can help users evaluate the reliability and verifiability of AI advice in these scenarios.

REFERENCES

[1] Saranya A. and Subhashini R. 2023. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal* 7 (2023), 100230. https://doi.org/10.1016/j.dajour.2023.100230

[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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[3] Abdullah Almaatouq, Mohammed Alsobay, Ming Yin, and Duncan J. Watts. 2021. Task complexity moderates group synergy. *Proceedings of the National Academy of Sciences* 118, 36 (2021), e2101062118. https://doi.org/10.1073/pnas.2101062118 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2101062118

[4] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3377325.3377519

[5] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[6] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. https://doi.org/10.1145/3411764.3445736

[7] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T. Wolf, Evelyn Duesterwald, Casey Dugan, Werner Geyer, and Darrell Reimer. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 89 (apr 2021), 27 pages. https://doi.org/10.1145/3449163

[8] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (apr 2023), 17 pages. https://doi.org/10.1145/3579460

[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, 1 (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 Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 300, 9 pages. https://doi.org/10.1609/aaai.v33i01.33012429

[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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[12] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376718

[13] Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich. 2022. It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 248–266. https://doi.org/10.1145/3531146.3533090

[14] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. 2022. Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 10, 17 pages. https://doi.org/10.1145/3491102.3501965

[15] Nigel Bosch and Sidney K. D'Mello. 2022. Can Computers Outperform Humans in Detecting User Zone-Outs? Implications for Intelligent Interfaces. *ACM Trans. Comput.-Hum. Interact.* 29, 2, Article 10 (jan 2022), 33 pages. https://doi.org/10.1145/3481889

[16] 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* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[17] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[18] Jason Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2019. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* (2019). https://api.semanticscholar.org/CorpusID:210439660

[19] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of behavioral decision making* 33, 2 (2020), 220–239.

[20] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 258–262. https://doi.org/10.1145/3301275.3302289

[21] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[22] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[23] Siew H. Chan, Qian Song, and Lee J. Yao. 2015. The moderating roles of subjective (perceived) and objective task complexity in system use and performance. *Computers in Human Behavior* 51 (2015), 393–402. https://doi.org/10.1016/j.chb.2015.04.059

[24] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[25] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 148–161. https://doi.org/10.1145/3490099.3511121

[26] Shih-Yi Chien, Michael Lewis, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2018. The Effect of Culture on Trust in Automation: Reliability and Workload. *ACM Trans. Interact. Intell. Syst.* 8, 4, Article 29 (nov 2018), 31 pages. https://doi.org/10.1145/3230736

[27] 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* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. https://doi.org/10.1145/3397481.3450644

[28] Devleena Das and Sonia Chernova. 2020. Leveraging Rationales to Improve Human Task Performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 510–518. https://doi.org/10.1145/3377325.3377512

[29] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[30] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, and Qian Pan. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 392–401. https://doi.org/10.1145/3397481.3450698

[31] 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 275–285. https://doi.org/10.1145/3301275.3302310

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

[33] Tim Draws, Alisa Rieger, Oana Inel, Ujwal Gadiraju, and Nava Tintarev. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 9, 1 (Oct. 2021), 48–59. https://doi.org/10.1609/hcomp.v9i1.18939

[34] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 297–307. https://doi.org/10.1145/3377325.3377501

[35] Jessica Maria Echterhoff, Matin Yarmand, and Julian McAuley. 2022. AI-Moderated Decision-Making: Capturing and Balancing Anchoring Bias in Sequential Decision Tasks. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 161, 9 pages. https://doi.org/10.1145/3491102.3517443

[36] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For what it's worth: Humans overwrite their economic self-interest to avoid bargaining with AI systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[37] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of Algorithmic Decision Making on Human Behavior: Evidence from Ultimatum Bargaining. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 43–52. https://doi.org/10.1609/hcomp.v8i1.7462

[38] Angela Fagerlin, Brian J. Zikmund-Fisher, Peter A. Ubel, Aleksandra Jankovic, Holly A. Derry, and Dylan M. Smith. 2007. Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making* 27, 5 (September 2007), 672–680. https://doi.org/10.1177/0272989X07304449

[39] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference*

on Fairness, Accountability, and Transparency (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[40] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[41] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150 arXiv:https://doi.org/10.1080/10447318.2018.1456150

[42] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[43] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: Visual Counterfactual Explanations for Machine Learning Models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 531–535. https://doi.org/10.1145/3377325.3377536

[44] Kazjon Grace, Elanor Finch, Natalia Gulbransen-Diaz, and Hamish Henderson. 2022. Q-Chef: The Impact of Surprise-Eliciting Systems on Food-Related Decision-Making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 11, 14 pages. https://doi.org/10.1145/3491102.3501862

[45] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. https://doi.org/10.1145/3359152

[46] Ben Green and Yiling Chen. 2021. Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 418 (oct 2021), 33 pages. https://doi.org/10.1145/3479562

[47] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (nov 2019), 25 pages. https://doi.org/10.1145/3359280

[48] Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 537–548. https://doi.org/10.1145/3490099.3511111

[49] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300803

[50] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI'22)*.

[51] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. https://doi.org/10.1145/3544548.3581025

[52] Michael Hilb. 2020. Toward artificial governance? The role of artificial intelligence in shaping the future of corporate governance. *Journal of Management and Governance* 24 (2020), 851–870.

[53] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570 arXiv:https://doi.org/10.1177/0018720814547570 PMID: 25875432.

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

[55] Aya Hussein, Sondoss Elsawah, and Hussein A. Abbass. 2020. Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human Factors* 62, 8 (2020), 1237–1248. https://doi.org/10.1177/0018720819879273 arXiv:https://doi.org/10.1177/0018720819879273 PMID: 31590574.

[56] 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 (2022). https://doi.org/10.3390/app12031353

[57] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4198–4205. https://doi.org/10.18653/v1/2020.acl-main.386

[58] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[59] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 arXiv:https://doi.org/10.1207/S15327566IJCE0401_04

[60] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. 2023. Rethinking AI Explainability and Plausibility. *arXiv preprint arXiv:2303.17707* (2023).

[61] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439

[62] Alex Kirlik. 1993. Modeling Strategic Behavior in Human-Automation Interaction: Why an "Aid" Can (and Should) Go Unused. *Human Factors* 35, 2 (1993), 221–242. https://doi.org/10.1177/001872089303500203 arXiv:https://doi.org/10.1177/001872089303500203 PMID: 8349287.

[63] Moritz Körber. 2019. Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)*, Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.). Springer International Publishing, Cham, 13–30.

[64] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[65] 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* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2207676.2207678

[66] Johannes Kunkel, Tim Donkers, Lisa Michael, Catalin-Mihai Barbu, and Jürgen Ziegler. 2019. Let Me Explain: Impact of Personal and Impersonal Explanations on Trust in Recommender Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300717

[67] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 59–67. https://doi.org/10.1609/hcomp.v7i1.5280

[68] 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 *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. https://doi.org/10.1145/3593013.3594087

[69] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[70] 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* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[71] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[72] 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. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 156 (oct 2020), 27 pages. https://doi.org/10.1145/3415227

[73] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Interactive Hybrid Approach to Combine Machine and Human Intelligence for Personalized Rehabilitation Assessment. In

Proceedings of the ACM Conference on Health, Inference, and Learning (Toronto, Ontario, Canada) *(CHIL '20)*. Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/3368555.3384452

[74] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[75] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684 arXiv:https://doi.org/10.1177/2053951718756684

[76] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 182 (nov 2019), 26 pages. https://doi.org/10.1145/3359284

[77] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 72, 13 pages. https://doi.org/10.1145/3411764.3445522

[78] Mengyao Li, Brittany E. Holthausen, Rachel E. Stuck, and Bruce N. Walker. 2019. No Risk No Trust: Investigating Perceived Risk in Highly Automated Driving. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 177–185. https://doi.org/10.1145/3342197.3344525

[79] Qing Li, Sharon Chu, Nanjie Rao, and Mahsan Nourani. 2020. Understanding the Effects of Explanation Types and User Motivations on Recommender System Use. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (Oct. 2020), 83–91. https://doi.org/10.1609/hcomp.v8i1.7466

[80] Mengqi Liao, S. Shyam Sundar, and Joseph B. Walther. 2022. User Trust in Recommendation Systems: A Comparison of Content-Based, Collaborative and Demographic Filtering. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 486, 14 pages. https://doi.org/10.1145/3491102.3501936

[81] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 235, 17 pages. https://doi.org/10.1145/3411764.3445260

[82] Zhiyuan "Jerry" Lin, Jongbin Jung, Sharad Goel, and Jennifer Skeem. 2020. The limits of human predictions of recidivism. *Science advances* 6, 7 (2020), eaaz0652.

[83] Peng Liu and Zhizhong Li. 2012. Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics* 42, 6 (2012), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

[84] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[85] Zilin Ma and Krzysztof Z. Gajos. 2022. Not Just a Preference: Reducing Biased Decision-Making on Dating Websites. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 203, 14 pages. https://doi.org/10.1145/3491102.3517587

[86] Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. 2020. Do I Look Like a Criminal? Examining How Race Presentation Impacts Human Judgement of Recidivism. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376257

[87] Gonzalo Mendez, Luis Galárraga, and Katherine Chiluiza. 2021. Showing Academic Performance Predictions during Term Planning: Effects on Students' Decisions, Behaviors, and Preferences. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 22, 17 pages. https://doi.org/10.1145/3411764.3445718

[88] Stephanie M. Merritt, Deborah Lee, Jennifer L. Unnerstall, and Kelli Huber. 2015. Are Well-Calibrated Users Effective Users? Associations Between Calibration of Trust and Performance on an Automation-Aided Task. *Human Factors* 57, 1 (2015), 34–47. https://doi.org/10.1177/0018720814561675

[89] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2020. What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 173–182. https://doi.org/10.1145/3340631.3394844

[90] David Miller, Mishel Johns, Brian Mok, Nikhil Gowda, David Sirkin, Key Lee, and Wendy Ju. 2016. Behavioral Measurement of Trust in Automation: The Trust Fall. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 1 (2016), 1849–1853. https://doi.org/10.1177/1541931213601422 arXiv:https://doi.org/10.1177/1541931213601422

[91] George A. Miller. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review* 63, 2 (March 1956), 81–97. http://www.musanim.com/miller1956/

[92] 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. *ACM Comput. Surv.* 55, 13s, Article 295 (jul 2023), 42 pages. https://doi.org/10.1145/3583558

[93] Dong Nguyen. 2018. Comparing Automatic and Human Evaluation of Local Explanations for Text Classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 1069–1078. https://doi.org/10.18653/v1/N18-1097

[94] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, 1 (Oct. 2019), 97–105. https://doi.org/10.1609/hcomp.v7i1.5284

[95] Mahsan Nourani, Joanie T King, and Eric D. Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. In *AAAI Conference on Human Computation & Crowdsourcing*. https://api.semanticscholar.org/CorpusID:221186776

[96] 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* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[97] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. 2022. Understanding the Impact of Explanations on Advice-Taking: A User Study for AI-Based Clinical Decision Support Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 568, 9 pages. https://doi.org/10.1145/3491102.3502104

[98] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 35 (mar 2022), 33 pages. https://doi.org/10.1145/3495013

[99] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (nov 2019), 15 pages. https://doi.org/10.1145/3359204

[100] Alison Parkes. 2017. The effect of individual and task characteristics on decision aid reliance. *Behaviour & Information Technology* 36, 2 (2017), 165–177.

[101] Andisheh Partovi, Ingrid Zukerman, Kai Zhan, Nora Hamacher, and Jakob Hohwy. 2019. Relationship between Device Performance, Trust and User Behaviour in a Care-Taking Scenario. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 61–69. https://doi.org/10.1145/3320435.3320440

[102] Andi Peng, Besmira Nushi, Emre Kıcıman, Kori Inkpen Quinn, Siddharth Suri, and Ece Kamar. 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. In *AAAI Conference on Human Computation & Crowdsourcing*. https://api.semanticscholar.org/CorpusID:202541374

[103] Martin Petrin. 2019. Corporate Management in the Age of AI. *SSRN Electronic Journal* (01 2019). https://doi.org/10.2139/ssrn.3346722

[104] Samuele Lo Piano. 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Palgrave Communications* 7, 1 (2020), 1–7. https://EconPapers.repec.org/RePEc:pal:palcom:v:7:y:2020:i:1:d:10.1057_s41599-020-0501-9

[105] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[106] Maria Riveiro and Serge Thill. 2022. The Challenges of Providing Explanations of AI Systems When They Do Not Behave like Users Expect. In *Proceedings*

of the 30th ACM Conference on User Modeling, Adaptation and Personalization (Barcelona, Spain) *(UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 110–120. https://doi.org/10.1145/3503252.3531306

[107] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-Assisted Decision-Making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (Barcelona, Spain) *(UMAP '22)*. Association for Computing Machinery, New York, NY, USA, 223–233. https://doi.org/10.1145/3503252.3531311

[108] Y. Rong, T. Leemann, T. Nguyen, L. Fiedler, P. Qian, V. Unhelkar, T. Seidel, G. Kasneci, and E. Kasneci. 5555. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 01 (nov 5555), 1–20. https://doi.org/10.1109/TPAMI.2023.3331846

[109] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 215–227. https://doi.org/10.1145/3565472.3592959

[110] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[111] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI–Distinguishing Between Attitudinal and Behavioral Measures. *arXiv preprint arXiv:2203.12318* (2022).

[112] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916* (2022).

[113] Nadine Schlicker, Alarith Uhde, Kevin Baum, Martin C Hirsch, and Markus Langer. 2022. Calibrated Trust as a Result of Accurate Trustworthiness Assessment–Introducing the Trustworthiness Assessment Model. (2022).

[114] Anuschka Schmitt, Thiemo Wambsganss, Matthias Söllner, and Andreas Janson. 2021. Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. https://www.alexandria.unisg.ch/handle/20.500.14171/111308

[115] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making. *arXiv preprint arXiv:2304.08804* (2023).

[116] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence* 2, 8 (2020), 476–486.

[117] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[118] Alain D. Starke, Martijn C. Willemsen, and Chris Snijders. 2021. Using Explanations as Energy-Saving Frames: A User-Centric Recommender Study. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 229–237. https://doi.org/10.1145/3450614.3464477

[119] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, Textual or Hybrid: The Effect of User Expertise on Different Explanations. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. https://doi.org/10.1145/3397481.3450662

[120] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 77–87. https://doi.org/10.1145/3450613.3456817

[121] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. Rapid Trust Calibration through Interpretable and Uncertainty-Aware AI. *Patterns* 1, 4 (2020), 100049. https://doi.org/10.1016/j.patter.2020.100049

[122] A Trunk, H Birkel, and E Hartmann. 2020. On the current state of combining human and artificial intelligence for strategic organizational decision making. Bus. Res. 13 (3). , 875–919 pages.

[123] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 152, 17 pages. https://doi.org/10.1145/3411764.3445101

[124] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 245, 13 pages. https://doi.org/10.1145/3411764.3445365

[125] Niels van Berkel and Kasper Hornbæk. 2023. Implications of Human-Computer Interaction Research. *Interactions* 30, 4 (jun 2023), 50–55. https://doi.org/10.1145/3600103

[126] 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. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[127] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174014

[128] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 327 (oct 2021), 39 pages. https://doi.org/10.1145/3476068

[129] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 211 (nov 2019), 24 pages. https://doi.org/10.1145/3359313

[130] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376813

[131] 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* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[132] Lawrence R. Wheeless and Janis Grotz. 2006. The Measurement of Trust and Its Relationship to Self-Disclosure. *Human Communication Research* 3, 3 (03 2006), 250–257. https://doi.org/10.1111/j.1468-2958.1977.tb00523.x arXiv:https://academic.oup.com/hcr/article-pdf/3/3/250/22344414/jhumcom0250.pdf

[133] Robert E Wood. 1986. Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes* 37, 1 (1986), 60–82. https://doi.org/10.1016/0749-5978(86)90044-0

[134] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (Polo) Chau, and Diyi Yang. 2021. RECAST: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 181 (apr 2021), 26 pages. https://doi.org/10.1145/3449280

[135] S. Paul Wright. 1992. Adjusted P-Values for Simultaneous Inference. *Biometrics* 48, 4 (1992), 1005–1013. http://www.jstor.org/stable/2532694

[136] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. 2019. Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis. *ACM Trans. Comput.-Hum. Interact.* 26, 4, Article 24 (jun 2019), 27 pages. https://doi.org/10.1145/3319616

[137] 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* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 189–201. https://doi.org/10.1145/3377325.3377480

[138] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[139] 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* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[140] Rachael Zehrung, Astha Singhal, Michael Correll, and Leilani Battle. 2021. Vis Ex Machina: An Analysis of Trust in Human versus Algorithmically Generated Visualization Recommendations. In *Proceedings of the 2021 CHI Conference on*

Human Factors in Computing Systems (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 602, 12 pages. https://doi.org/10.1145/3411764.3445195

[141] Qiaoning Zhang, Matthew L Lee, and Scott Carter. 2022. You Complete Me: Human-AI Teams and Complementary Expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 114, 28 pages. https://doi.org/10.1145/3491102.3517791

[142] 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* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[143] Zelun Tony Zhang, Sven Tong, Yuanting Liu, and Andreas Butz. 2023. Is Overreliance on AI Provoked by Study Design?. In *IFIP Conference on Human-Computer Interaction*. Springer, 49–58.

[144] Chengbo Zheng, Yuheng Wu, Chuhan Shi, Shuai Ma, Jiehui Luo, and Xiaojuan Ma. 2023. Competent but Rigid: Identifying the Gap in Empowering AI to Participate Equally in Group Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 351, 19 pages. https://doi.org/10.1145/3544548.3581131