How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System

He, Gaole; Buijsman, Stefan; Gadiraju, Ujwal

DOI
10.1145/3610067

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the ACM on Human-Computer Interaction

How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System

GAOLE HE* and STEFAN BUIJSMAN*, Delft University of Technology, Netherlands
UJWAL GADIRAJU, Delft University of Technology, Netherlands

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

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

Additional Key Words and Phrases: human-subjects experiment, analogy, trust, reliance

ACM Reference Format:
Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How Stated Accuracy of an AI System and Analogies to Explain Accuracy Affect Human Reliance on the System. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 276 (October 2023), 29 pages. https://doi.org/10.1145/3610067

1 INTRODUCTION

It is becoming more and more common for humans to make decisions supported by machine learning algorithms. Whether it is in financial risk assessment [23, 37], medical diagnosis [15, 33] or in public employment services [10], such collaborative, socio-technical systems (*i.e.,* a decision procedure where humans and AI are jointly involved in making the decision) are ubiquitous. And while initial hopes were that such a combination would lead to better decisions [34], it has proved

*Both authors contributed equally to this research.

Authors' addresses: Gaole He, g.he@tudelft.nl; Stefan Buijsman, s.n.r.buijsman@tudelft.nl, Delft University of Technology, Mekelweg 5, Delft, Netherlands, 2628 CD; Ujwal Gadiraju, Delft University of Technology, Mekelweg 5, Delft, Netherlands, 2628 CD, u.k.gadiraju@tudelft.nl.

2573-0142/2023/10-ART276
https://doi.org/10.1145/3610067

tough to mitigate unexpected reliance (*i.e.,* under-reliance and over-reliance) on the AI system. In this paper, *appropriate reliance* is defined as the tendency for users to rely on the system in situations where it is accurate (or more precisely, where it is more accurate than humans) and not to rely on it when the system is inaccurate (or, ideally, whenever it is wrong). This follows the conceptualization of appropriate system reliance established in the Human-AI interaction, collaboration, and teaming fields over the last few years [8, 30, 39, 41, 55]. Users in the real world, however, find it difficult to determine their own accuracy in difficult tasks as well as the system's accuracy (in individual cases). That in turn means they have a hard time deciding when an AI system is more accurate than they are. This tension has been shown to result in both under-reliance [14, 40] and over-reliance [8] of users on AI systems, often leading to detrimental outcomes.

There are several complementary approaches to facilitating appropriate system reliance, such as research in explainable AI attempting to elucidate the reasons for model output [28, 67]. Such tools can help, especially if users are actively made to reflect on explanations using cognitive forcing interventions [7]. Another approach, and one which is explored further in this paper, is to give users information on the confidence and overall accuracy of the system. Papenmeier et al. [51], Yin et al. [65] found that users adjust their reliance on AI systems based on the reported system accuracy. However, even after seeing the high stated accuracy, users do not rely on the system as often as the accuracy warrants (*e.g.* adopting system advice 80% of the time while system accuracy is 95%, resulting in an inferior overall performance than the theoretical potential). We explore if this under-reliance among users is a result of their potentially limited understanding of the system accuracy measure. We do not hold the position that reliance on AI systems is universally good. On the contrary, preventing over-reliance on AI systems is just as important. However, a fundamental pre-requisite to designing and facilitating human-AI interactions that can effectively support humans in a given task, is to advance our current understanding of how users rely on AI systems. An unanswered question in this context pertains to why users tend to under-rely on AI systems despite their relatively high stated accuracy. Perhaps users do not properly calibrate their reliance on the AI system because they have trouble identifying the right accuracy level when presented only with an overall accuracy value.

We use analogies to counter such lack of understanding of global accuracy measures, which is to our knowledge the first attempt of its kind to elucidate system measures. An analogy can be interpreted as a structural mapping of a target domain which is to be clarified (in this case, overall system accuracy) onto a source domain which the recipient of the analogy is more familiar with [25, 32]. As a simple example, one might elucidate how hard a task is by saying 'it is as hard as finding a needle in a haystack'. As the recipient is likely to know that finding a needle will be difficult in this case, the inference on the target domain can be made that the relevant task will also be difficult. While such simple examples may not make a convincing case for the use of analogies, there is strong empirical evidence that more specific analogies can help people to individuate and identify risk levels, as discussed further in Section 2.

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

RQ1: How does the understanding of stated system accuracy affect reliance of users on the AI system?

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

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

Our results highlight that a limited understanding of the system accuracy measure is not the reason why users rely on AI systems lesser than warranted by the relatively higher system accuracy. Instead, it is likely that users' overestimation of their own ability to solve the given task drives their under-reliance on the system. This interpretation is supported by various findings in prior work [9, 31, 38, 43]. We outline this as a direction for further study. Empirical studies that explore why and how humans tend to rely on AI systems play a vital role in furthering our understanding of how we can build better human-AI interactions in a variety of tasks, scenarios, and domains. It is in this context that our work makes important contributions by (a) advancing our understanding of user under-reliance on AI systems, (b) exploring the effectiveness of analogies as an instrument to explain measures like stated system accuracy, and (c) investigating whether an improved understanding of global AI system measures can lead to more appropriate reliance.

In addition, although we considered several potentially important user factors (such as numeracy level and familiarity with and trust in the analogy domain), most of them did not significantly impact user reliance behaviors. Only users' general propensity to trust automated systems emerged as an important user factor which contributes to both subjective trust and objective reliance. Based on the results from our empirical study, we synthesized and discussed favorable conditions for the use of analogies and pointed out promising future directions for further research exploring user reliance on AI systems. Our findings contribute to the growing body of literature on human-AI decision making and further our understanding of under-reliance on AI systems.

2 RELATED WORK

This paper contributes to the growing literature on user reliance on AI systems by focusing on how users might be helped to calibrate their reliance by analogies that clarify stated accuracy measures. Our goal is to explore whether a limited understanding of stated accuracy is to blame for under-reliance on an AI system (within the scope of **RQ1**) and whether improving this understanding can lead to more appropriate reliance (within the scope of **RQ2**). As such, the research combines three strands of literature: the general literature on user reliance of AI systems (2.1). The more specific literature on how that reliance is affected by stated accuracy measures (2.2) and finally the literature on analogies, which have been shown to benefit risk perception (2.3).

On the one hand, the research focuses on the use of accuracy scores to engender (appropriate) reliance on AI systems. As merely stating the accuracy has been found to be insufficient for reaching

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

appropriate reliance, the contribution of this paper is to explore whether that is due to a limited grasp of the implications of the accuracy scores. Another area of research that is therefore relevant for this paper is the literature on analogies in risk perception, where the use of analogies to elucidate percentages in a similar setting has been investigated. That gives us a basis to postulate that analogies improve this understanding.

2.1 Reliance on AI Systems

There is a wide range of factors that affects how users rely on AI systems. For example, Dietvorst *et al.* [11] and Dzindolet *et al.* [13] found that users stop relying on a system after seeing it make a mistake. Meanwhile, Yeomans *et al.* [63] found that people did not rely on system advice in a highly subjective domain – namely a task to predict which jokes others will find funny – even if the system performed better than they did. At the same time, Dietvorst *et al.* [12] saw that participants are more willing to rely on systems if they are able to alter the final decision somewhat, rather than having to follow the exact prediction. Such prior research has generally found that it is hard to get users to rely on a system appropriately. Inspired by the design of these studies, in our study we used a two-stage decision making process that allows users to alter their final decision after seeing the AI advice (see Section 3.1).

Different solutions for this challenge have been examined. We investigate the option of presenting users with accuracy measures (2.2), but the other major option is to provide users with explanations of the system output (XAI). In a risk assessment task (for a loan approval and a pretrial domain), Green *et al.* [27] looked at whether explanations or feedback per decision help users calibrate their reliance, but found mostly null effects. They show that people are unable to evaluate their own accuracy at risk assessments, do not calibrate their reliance based on observed accuracy and only had a positive effect from explanations on the loan approval task. And whereas Green *et al.* [27] found some positive effects of explanations, Zhang *et al.* [66] failed to find similar appropriate reliance when users were given (feature importance) explanations. However, they did observe an improvement in reliance when presenting confidence scores for the system, with users switching more often to (*i.e.,* relying on) AI predictions with high confidence scores than to those with lower confidence scores or none at all. This is in line with the proposal of Bhatt *et al.* [4] to use uncertainty measures to help users rely appropriately on AI systems. Yet the addition of confidence scores in the study by Green *et al.* did not improve the accuracy of participants using the AI system.

One complicating factor here is the interplay between subjective trust and objective reliance. In this paper, we consider that subjective trust influences objective reliance. And indeed Lu *et al.* [41] found similar patterns for both objective reliance and subjective trust when feedback on model performance is limited. Both trust and reliance are significantly affected by the level of agreement between people and a model on decision making tasks that people have high confidence in. However, other conflicting results have also been found. Through an extensive user study, Buçinca *et al.* [6] pointed out that "when using actual decision making tasks, subjective results do not predict objective performance results," which reveals a gap between the subjective trust attitude of users and their objective reliance behavior. Similarly, a gap between stated trust and actual reliance was reported by Schmitt *et al.* [56], and Bansal *et al.* [1] observed that explanations can promote blind trust rather than lead to appropriate reliance on AI systems. We thus hold that subjective trust *can* promote objective reliance, but keep in mind that subjective trust measures can give an overly optimistic image of reliance and therefore focus on objective reliance.

2.2 Reliance and System Accuracy

Though research specifically on stated accuracy is sparse, prior experiments do show that the stated accuracy of a system has an effect on the degree to which people rely on the system. Yin *et al.* [64]

first reported a significant effect of stated accuracy on reliance and further expanded on this in [65]. Here, in a task where users had to predict if someone wanted to see his or her date a second time, they compared reliance on the system across conditions with different stated accuracies (and included a control with no stated accuracy). They observed significant differences in the fraction of cases in which users agreed with the system and in the fraction of cases in which users changed their initial decision so that their final decision agreed with the system advice. However, they found that participants struggle to calibrate their reliance. When there was no stated accuracy, users agreed in about 75% of the final decisions with the system. For decisions with an initial disagreement between users and the system, users switched to agree with the system in 30% of cases. This did not change for a stated accuracy of 60% or 70% and only increased for a stated accuracy of 90 and 95%. However, the effect of the stated accuracy is not as high as it should be: for 90% and 95%, users only agreed with the system in 80% of cases. Finally, the effect of stated accuracy was canceled out by the effects of observed accuracy when these were presented to users midway through the study.

This relevance of observed accuracy has further been underscored by Papenmeier *et al.* [51], who found that the effect of varying observed accuracy on reliance was stronger than the effect of explanations of system outputs (either no, low-fidelity, or high-fidelity explanation). So, system accuracy has been shown to be relevant for calibrating reliance, and therefore the extent to which users understand what this system accuracy means. Recent work by Nourani *et al.* has shown that users do not rely on what they do not understand [48]. It is this lack of understanding that we hope to alleviate through the use of analogies.

2.3 Analogies in Risk Perception

There is a long-standing use of analogies to explain statistical concepts [42, 45] and medical risk levels [21, 22]. What emerges from this is that it can be difficult to get analogies to deliver benefits, as the meta-study by Sopory *et al.* [57] on the effect of metaphor's persuasive effects underlines. Analogies, as they intricately depend on how they are perceived by the recipient, can be hard to calibrate to the audience. If successful, however, they can have clear cognitive benefits. Sopory *et al.* [57] found that when they are novel, have a familiar source domain (*i.e.,* the 'needle in a haystack' part in 'x is as difficult as finding a needle in a haystack') and are used early in the message then they are used optimally and have a clear effect on persuasiveness. A later meta-study by Van *et al.* [61] confirms this, finding that metaphorical messages are, when using a familiar source domain, more effective than literal messages.

Such effects can be found in the existing literature on risk perception too. Barilli *et al.* [2] tested the use of analogies to improve the risk perception between a 1 in 100 chance and a 1 in 900 chance. While adding analogies does not make these risks more discriminable, they do lower the overall risk perception on a 7-point scale (from 3.5 to 2.5 for 1 in 100, from 3.1 to 2.1 for 1 in 900). The lack of effects here has, however, been hypothesized to be due to the choice of analogies: stated analogies were about the odds of drawing a red ball out of a jar, something which we do not encounter or deal with on a regular basis. More familiar analogies studied by Galesic *et al.* [22], such as 'as a flu vaccine is to flu' or 'as a car alarm is to theft', did show a clear effect of analogies. Performance on difficult medical problems was improved for people with high numeracy skills and performance on easy problems was improved for people with low numeracy skills. Numeracy here means the ease and skill with which participants work with numbers. Their interpretation of the finding, therefore, was that analogies help when problems are not too difficult and performance is not at ceiling. Interestingly for the current study, Galesic *et al.* [22] also looked at what makes analogies helpful and again ranked familiarity with the source domain highly.

The effect of numeracy level on findings has, moreover, been collaborated in other studies. Pighin *et al.* [52] found that high-numeracy participants do improve on discrimination of risk levels after seeing analogies. Participants with low numeracy showed no improvement in the discrimination between a 1 in 5390, 1 in 770 and 1 in 110 risk on a 7-point Likert scale. Similarly, with a more visual analogy in the form of a risk ladder, Keller *et al.* [35] found the visualisation to suffice for high-numeracy participants in discriminating between different risk levels. Low-numeracy participants only managed to do so after also seeing analogies with the number of cigarettes one would smoke a day. So, here too, familiarity with the source domain is likely to have been high, to support understanding of the risk levels.

To sum up, analogies have been found to be effective tools to improve risk perception and performance on related medical problems, though a number of relevant factors have emerged that interact with the effectiveness. These have informed our hypotheses 3 and 4. Numeracy level is important, as also underlined by a recent overview study [24], and especially low numeracy individuals can use help in understanding the meaning of percentages. This finding supports our motivation to look into the possibility that participants fail to calibrate reliance to accuracy scores because they might not fully understand the presented information. Aside from numeracy, familiarity with the source domain used to explain the percentages is an important factor for the success of analogies. Hence, we have used a range of analogies in our study that vary with respect to familiarity and included a question in the post-task questionnaire to measure user's (subjective) familiarity with the source domain.

3 TASK AND HYPOTHESIS

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

3.1 Loan Prediction Task

The basis for our experimental setup is a task where participants have to decide whether to accept or reject a loan application using the publicly available loan prediction dataset.[2] This task was chosen as a realistic scenario for human-AI collaboration, where there is a clear risk and a benefit to the adoption of AI advice. As such, it fits in with the risk perception research where analogies were pioneered. It has also been adopted by existing research in behavioral economics [3] and human-AI collaboration [27].

Participants thus made decisions on whether to grant a loan or not based on twelve features such as income, the absence or presence of a credit history and the loan amount. This simulates a realistic scenario where participants interact with an AI system and may rely on it due to the complexity in simultaneously considering multiple features for successful decision making, but also due to a relatively high stated accuracy of the AI system. Furthermore, we consider this to be a suitable task to test the influence of user numeracy level, as almost all the presented information is in numerical format. The task interface is shown in Figure 1.

Task Selection. Participants were presented with twelve such cases, of which two were example cases and ten trial cases. These cases were selected by first training a linear regression model on the full dataset. The two example cases were the top-*1* most confident correct cases for approval and rejection (with respect to the linear regression model). The ten trial cases used in the actual experimental task were: two high confidence correct predictions, two medium confidence correct predictions, two borderline correct predictions, two borderline wrong predictions and the two least confident wrong predictions (again, with respect to the linear regression model). Cases were evenly

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

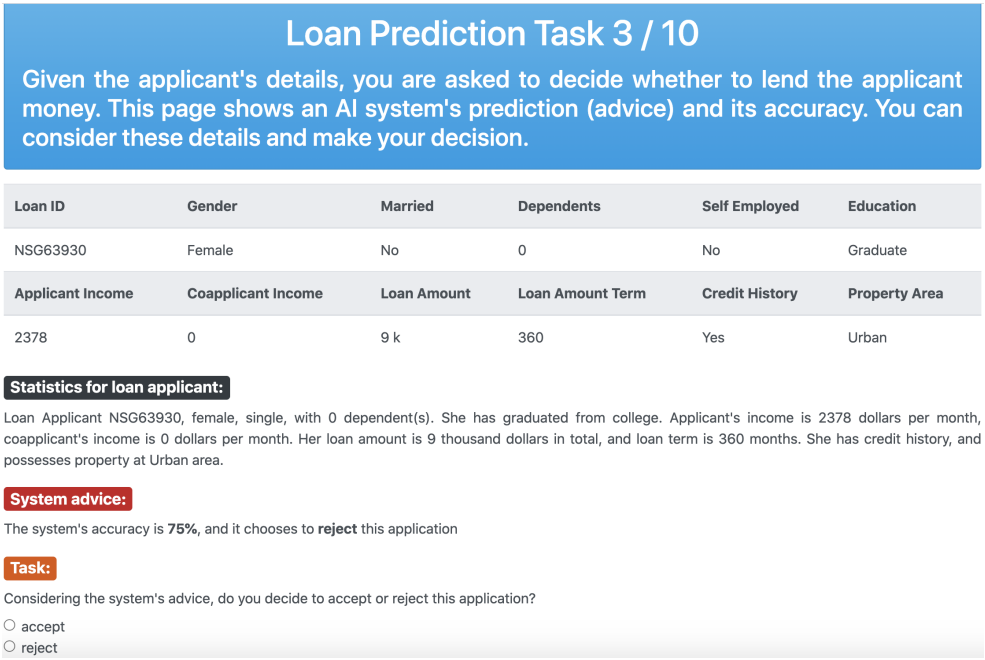

Fig. 1. Illustration of the interface that participants used to complete the loan prediction task.

split between those where the loan should be approved and those where the loan should be rejected and the order of the trial cases was randomized to prevent order effects [50].

Two-stage Decision Making. In trial cases, participants of all conditions were first presented with the applicant information corresponding to the case and then asked to make a decision whether to accept or reject the loan application (see screenshot in Figure 1). This first time, they were not presented with the systems' prediction, or with any additional information. After making an initial choice they saw the same case again, but now additionally saw the systems' prediction and (depending on the experimental condition) also the system accuracy and analogy. Participants were then asked to make a final decision. This setup of an initial unaided decision and the presentation of system advice in order to make a second and final choice is similar to the update condition in [27], and in line with findings that people first make a decision on their own and only then decide whether to incorporate system advice [26]. It also fits with the research of Dietvorst *et al.* [12] on trust in two-stage decision making.

3.2 Hypotheses

Our study was designed to answer questions about the effectiveness of well-understood stated accuracy on reliance, and the use of analogies to improve user understanding of the accuracy level. As stated accuracy has been found to be effective in improving (appropriate) reliance [65], we expect to observe the same effect here:

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

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

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

In particular, we expect that this effect will depend on how familiar users are with the target (the stated accuracy) and source (*e.g.,* train punctuality) domain of the analogy, as discussed in Section 2. Thus, we further hypothesize that the numeracy level of users, *i.e.,* how familiar they are with quantitative measures, shapes the usefulness of analogies. Participants with a high numeracy level might understand the task and stated accuracy well enough already for analogies to offer little improvement, whereas participants with low numeracy might have a lack of understanding of these numbers that is alleviated by the analogy. As the role of analogies is to make this target domain (accuracy of the system) easier to understand by creating a structural mapping onto a source domain that the user is potentially more familiar with, we also formulate a hypothesis around the familiarity with the source domain:

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

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

4 STUDY DESIGN

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

4.1 Experimental Conditions

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

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

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

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

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

(1) Vaccine efficacy: 'the system is 75% accurate, which is about as reliable as the AstraZeneca vaccine is for protecting against covid' (which is about 70% effective against the then-current Delta variant and somewhat more effective against earlier variants [53]). [4]
(2) Accuracy of weather predictions: 'the system is 75% accurate, which is about as reliable as the five-day weather prediction' (which is also typically around 75% accurate). [5]
(3) Train punctuality: 'the system is 75% accurate, which is about as reliable as the French trains are on punctuality' (which is 75% as listed in the 7th Rail Market Monitoring Report of the European Commission).

4.2 Measures And Variables

As mentioned, we use analogies to investigate whether a lack of appropriate reliance is due to a lack of understanding of global accuracy measures. It is important for this investigation to note the difference between (objective) reliance, which is the focus of our study, and (subjective) trust. We follow Lee *et al.* [39] in postulating that "trust in automation guides reliance when the complexity of the automation makes a complete understanding impractical and when the situation demands adaptive behavior that procedures cannot guide." Thus, we operationalize trust as a subjective user attitude, and reliance as objective user behavior that can be influenced by trust. As such, subjective trust can help us illuminate the effects we see on objective reliance [58].

To answer **H1** and **H2** we measure the reliance of participants on the system via two metrics: the agreement fraction and the switch fraction. These look at the degree to which participants are in agreement with system advice, and how often they adopt system advice in cases of initial disagreement. They are commonly used in the literature, for example in [65, 66]. In addition, we consider the overall accuracy and the accuracy under initial disagreement (*i.e.,* accuracy-wid) to measure participants' performance and appropriate reliance respectively. Since cases without initial disagreement do not clearly signal reliance on the system we restrict the scope of the appropriate reliance measure to accurately understand how participants handle divergent system advice. Following Max *et al.* [55], we adopted the relative positive AI reliance (RAIR) and relative positive self-reliance (RSR) metrics to measure appropriate reliance. When the AI system provides correct advice and the user makes a wrong initial decision, there are two possible reliance patterns: positive AI reliance (users switch to AI advice), negative self-reliance (users do not follow correct AI advice). When the AI system provides wrong advice and the user makes a correct initial decision, there are two other possible reliance patterns: positive self-reliance (users insist on their own initial decision) and negative AI reliance (users switch to another option). These measures are computed as follows:

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

[4] https://www.nature.com/articles/d41586-021-02261-8
[5] https://spectrumlocalnews.com/tx/austin/weather/2020/10/08/wisconsin-weather-blog-meteorologist-wrong-rudd

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

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

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

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

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

To answer **H3** , we measured the numeracy level of the participants in our study. To do so we used the Subjective Numeracy Scale [16, 68], which has been widely validated as a measure for numeracy level in risk perception literature. We chose this subjective scale as opposed to an objective measure (asking participants to answer a number of quantitative questions) since prior work by Zikmund-Fisher *et al.* revealed that participants find objective tests stressful and unenjoyable [68]. Furthermore, the subjective scale has also been shown to correlate with the helpfulness of analogies in increasing risk perception [35], motivating our hypotheses.

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

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

- The Trust in Automation (TiA) (post-task) questionnaire [36], a validated instrument to measure (subjective) trust [58] consisting of 6 subscales: *Reliability/Competence* (TiA-R/C), *Understanding/Predictability* (TiA-U/P), *Propensity to Trust* (TiA-PtT), *Familiarity* (TiA-Familiarity), *Intention of Developers* (TiA-IoD), and *Trust in Automation* (TiA-Trust). Thus, we consider possible effects of trust on reliance, in accordance with Lee *et al.* [39].
- The Affinity for Technology Interaction Scale (ATI) [18], administered in the pre-task questionnaire. Thus, we account for the effect of participants' affinity with technology on their reliance on systems [58].

Table 1 presents an overview of all the variables considered in our study.

4.3 Participants

Sample Size Estimation. Before recruiting participants, we computed the required sample size in a power analysis for a Between-Subjects ANOVA using G*Power [17]. To correct for testing multiple hypotheses, we applied a Bonferroni correction so that the significance threshold decreased to $\frac{0.05}{4} = 0.0125$. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.0125$ (*i.e.,* due to testing multiple hypotheses), a statistical power of $(1 - \beta) = 0.9$, and that we will investigate 3 different experimental conditions/groups. This resulted in a required sample size of 273 participants. We thereby recruited 316 participants from the crowdsourcing platform Prolific[6], in order to accommodate potential exclusion.

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

Table 1. The different variables considered in our experimental study. "DV" represents a dependent variable.

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

Compensation. All participants were rewarded with £1.5, amounting to an hourly wage of £7.5 deemed to be "*good*" payment by the platform (estimated completion time was 12 minutes). We rewarded participants with extra bonuses of £0.1 for every correct decision in the 10 trial cases. By incentivizing participants to reach a correct decision, we operationalize the concomitant "vulnerability" discussed by Lee and See[39] as a contextual requirement to encourage appropriate system reliance.

Filter Criteria. All participants were proficient English-speakers above the age of 18 and they had an approval rate of at least 90% on the Prolific platform. We excluded participants from our analysis if they failed at least one attention check (2 participants), or represented an outlier in terms of the amount of time they spent on our study. Outliers were participants (33 in total) who spent less than 7 minutes on the entire study. The resulting sample of 281 participants had an average age of 27 (SD = 8.64) and a gender distribution (70.1% female, 28.5% male, 1.4% other).

4.4 Procedure

The full procedure that participants followed in our study is illustrated in Figure 2. All participants first read the same basic instructions on the loan prediction task. Next, participants were asked to complete a pre-task questionnaire to measure their numeracy level and affinity for technology interaction.

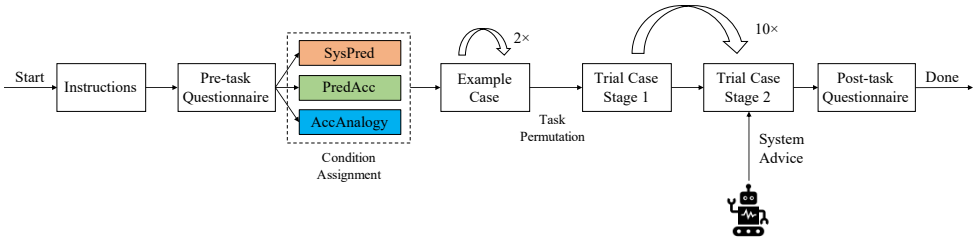

Fig. 2. Illustration of the procedure that participants followed within our study.

Participants were then randomly assigned to one of three different experimental conditions, that differed in whether or not the system's prediction was supplemented with its accuracy and an analogy to explain the accuracy. After assignment, the participants were trained with two

example cases before 10 trial cases. Selection of these cases is described in section 3.1. Finally, a post-task questionnaire was administered, using the 6 subscales of the TiA questionnaire discussed in section 4.2. Participants in the **AccAnalogy** condition were additionally asked for their familiarity with the source domain and the perceived helpfulness of the analogy they were presented with. To further ensure reliability of responses gathered in the questionnaires and the loan decisions, we added five attention check questions spread out at random through the different stages of the procedure [20].

4.5 Pilot Study

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

This validated our task selection strategy, and suggested that the task is relatively difficult for humans to complete accurately, and decision support from an AI system would be realistic and meaningful. A 75% accuracy of the system is, then, a level which is helpful if the system is relied on, but still involves some risks and so calls for *appropriate* reliance, as opposed to blindly following the system advice. Note that this design choice is motivated by Lee and See's work which emphasizes the role of uncertainty in dictating the need to facilitate appropriate reliance [39]. Had we set the accuracy at 90 or 95%, the situation would have been less clearly one of uncertainty for participants following the system advice.

5 RESULTS

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

5.1 Descriptive Statistics

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

Distribution of Covariates and Reliance Behavior. The covariates' distribution is as follows: *numeracy level* ($M = 4.48$, $SD = 0.78$, 6-point Likert scale, *1: low*, *6: high*), *ATI* ($M = 3.82$, $SD = 0.78$, 6-point Likert scale, *1: low, 6: high*), *familiarity with analogy domain* ($M = 3.36$, $SD = 1.52$, 5-point Likert scale, *1: unfamiliar*, *5: very familiar*), *TiA-Propensity to Trust* ($M = 2.79$, $SD = 0.60$, 5-point Likert scale, *1: tend to distrust*, *5: tend to trust*), and *TiA-Familiarity* ($M = 2.38$, $SD = 0.98$, 5-point Likert scale, *1: unfamiliar*, *5: very familiar*). This is illustrated in the boxplots in Figure 3.

Overall, all participants had at least one initial disagreement with system advice and 83.6% participants switched at least one decision after viewing the system's advice. On average, the initial decision was the same as the final decision in 77.6% of all decisions. A small portion of participants

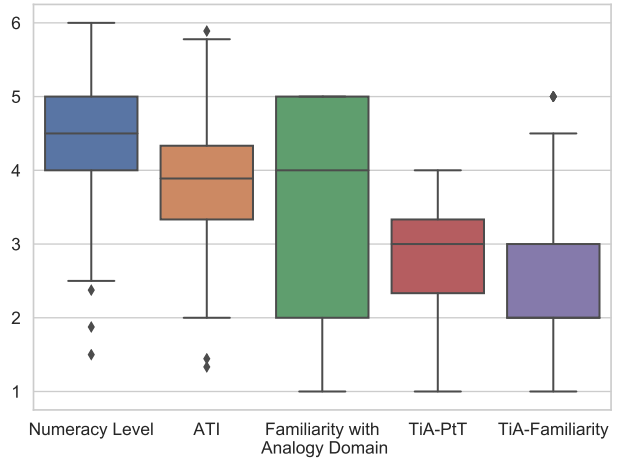

Fig. 3. Box plot illustrating the distribution of the different covariates considered in our study. Among these covariates, *numeracy level* and *ATI* were measured on a 6-point scale, while others were measured on a 5-point scale.

(0.5% across all conditions) changed their mind despite an initial agreement with the system, to reach a final decision different from both their initial decision and the system advice.

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

Table 2 shows the accuracy and error analysis for each of the 10 loan prediction tasks. In all tasks, we observe that the average accuracy of task and participants' error cause is highly correlated to its difficulty level (determined as described in Section 4.4). On relatively easy tasks, participants achieved high accuracy, and the errors in such cases are mainly caused by adopting incorrect system advice. In contrast, participants achieved a low accuracy on hard tasks, and demonstrated a reluctance to rely on the AI system which achieved superior performance on hard tasks. On average, however, we see that the mistakes made by participants are evenly split between cases where they should have relied on the system (49.3%) and cases where they should have disagreed with the system (50.7%).

5.2 Hypothesis Tests

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

Effect on Objective Reliance. To analyze the main effect of system accuracy (**H1**) and analogies (**H2**) on reliance, we conducted a Kruskal-Wallis H-test by considering the *experimental condition* as independent variable. The results showed no significant effects of *experimental condition* on reliance measures. The only effect that was significant was one of *experimental condition* on *participant accuracy*; $H(2) = 11.42$, $p = 0.003$. Participants in the **`AccAnalogy`** condition perform worse on *participant accuracy* ($M = 0.48$, $SD = 0.14$) than those in the **`SysPred`** condition ($M = 0.54$, $SD = 0.15$) and the **`PredAcc`** condition ($M = 0.55$, $SD = 0.14$). Post-hoc Mann-Whitney tests using a Bonferroni-adjusted alpha level of 0.0125 ($\frac{0.05}{4}$) were used to compare all pairs of conditions.

Table 2. Participant performance on loan prediction tasks. Observed errors are split into two cases: 'Error-reliance' refers to the fraction of errors that were a result of participants agreeing with the system when it was wrong. 'Error-non-reliance' refers to the fraction of errors that were a result of participants disagreeing with the system when it was in fact correct. The difficulty levels are from 1 (*very easy*) to 5 (*very hard*), obtained by leveraging the predictions from a linear regression model. 'Accuracy', 'Error-reliance' and 'Error-non-reliance' are reported in percent (%).

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

The difference in *participant accuracy* between **SysPred** condition and **PredAcc** condition was not significant; $U(N_{\texttt{SysPred}} = 87, N_{\texttt{PredAcc}} = 92) = 3682, p = 0.345$.

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

Effect on Subjective Trust. The impact of subjective trust was analyzed using an *Analysis of Covariance* (ANCOVA) with the *experimental condition* as between-subjects factor and *numeracy level*, *ATI*, *TiA-Familiarity* and *TiA-Propensity to Trust* as covariates. This allows us to explore the main effects of system accuracy (**H1**) and analogy-based explanation (**H2**) on subjective trust as measured by the relevant four subscales of the TiA. We decided to conduct AN(C)OVAs despite the anticipation that our data may not be normally distributed because these analyses have been shown to be robust to Likert-type ordinal data [47]. Table 3 shows the ANCOVA results pertaining to the four trust-related dependent variables.

Table 3. ANCOVA test results for **H1** and **H2** on trust-related dependent variables. "†" indicates the effect of variable is significant at the level of 0.0125.

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

As can be seen, there is no effect on any of the four subjective trust subscales by experimental condition. This suggests that the reduced accuracy in the analogy group (considered broadly)

is not due to a lack of subjective trust in the system. Subjective trust in the particular system participants was presented with did correlate significantly with their familiarity with similar systems (*TiA-Familiarity*) and their general propensity to trust automated systems (*TiA-PtT*), as one would expect. Likewise, general affinity to technology (*ATI*) had a significant effect on subjective feeling of understanding the system (*TiA-U/P*) and trusting the intentions of the designers (*TiA-IoD*). This strengthens our confidence that we did succeed in measuring subjective trust in the system, as it depends on other subjective measures in the way one would expect. In a further Spearman rank-order test we observed that *TiA-PtT* significantly affects reliance and accuracy. Namely, there is a significant positive correlation between *TiA-PtT* and the reliance-based measures: *agreement fraction*, $r(279) = 0.277$, $p = 0.000$; *switch fraction*, $r(279) = 0.271$, $p = 0.000$; *accuracy-wid*, $r(279) = 0.191$, $p = 0.001$; *participant accuracy*, $r(279) = 0.203$, $p = 0.001$; *RAIR*, $r(279) = 0.266, p = 0.000$; *RSR*, $r(279) = -0.177, p = 0.003$. This confirms our postulated link between subjective trust and objective reliance and so our null findings on objective reliance *w.r.t.* the experimental conditions can be partially explained by the observed lack of improvement in subjective trust. However, this fails to explain why the accuracy decreased in the analogy condition. We discuss this further while assessing the results for **H4**, where we examine the different analogy domains in detail.

5.2.2 *H3: Numeracy level.*

To verify **H3**, we calculated Spearman rank-order correlation coefficients for *numeracy level* and dependent variables on the different experimental conditions and the sub-groups of the **AccAnalogy** condition. As can be seen in Table 4, we found that *numeracy level* does not significantly correlate with reliance measures when considering all participants in the **AccAnalogy** condition. Nor does it significantly correlate with reliance measures when focusing on participants in any of the three subgroups. We thus find no evidence in support of **H3**.

Table 4. Spearman rank-order correlation coefficient for numeracy level on reliance.

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

We carried out an exploratory analysis to examine the overall effect of numeracy level on reliance. To do so, we split the participants in all conditions into three groups: those with high (top 25%), medium (25-75%) and low (bottom 25%) numeracy. We conducted Kruskal-Wallis H-test with *numeracy group* and all dependent variables. The results indicate that there is no statistically significant difference between the three groups with different numeracy levels in terms of either reliance or subjective trust measures (see Table 5).

However, as shown in Table 5, participants in the low numeracy group did exhibit a higher agreement fraction and as a result had a higher accuracy in the task. Meanwhile, in cases with an initial disagreement between user decision and system advice, participants in the medium numeracy group achieved higher appropriate reliance and switch fraction than other two groups. Oddly enough, low numeracy participants report virtually the same subjective understanding of the system as high numeracy participants, but lower subjective trust on the other measures. Though these results were not statistically significant, they potentially suggest that participants with lower numeracy might have felt the need to rely more on the system as they were less comfortable with the numerical task.

Table 5. Mean of dependent variables on different numeracy groups. "p" refers to the p-value for Kruskal-Wallis H-test results between three groups.

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

5.2.3 ***H4**: Familiarity with analogy domains.*

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

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

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

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

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

5.3 Participant Perception of Analogy-based Explanations

Finally, we analyzed the written responses of participants to the prompts "*Why did you find the analogy to be helpful or not helpful?*", and "*Please share any comments, remarks or suggestions regarding the use of analogies to explain the accuracy of the system.*" Authors of this paper manually coded all participants' responses about the analogy-based explanations into the mutually exclusive categories of — `positive` ($N = 32$), `negative` ($N = 57$), `neutral` ($N = 4$), or `not reported` ($N = 9$). Using a random sample of the responses from participants, authors agreed on the categories for coding. We do not report inter-rater reliability, as disagreement between the authors was resolved through detailed discussions and critical reflection [44]. Example excerpts of the feedback received from participants are presented in Table 6. Using the thematic analysis software, ATLAS.ti,[7] we conducted a thematic analysis and selected the top-3 topics mentioned by users across three analogy domains (shown in Table 7).

Table 6. Excerpts from participants' responses to open questions regarding the analogy-based explanations.

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

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

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

[7] https://atlasti.com

Table 7. Resulting main themes from the thematic analysis of participants' responses to the open questions pertaining to analogy-based explanations across domains.

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

- 3 participants (9.4%) felt the analogy-based explanation improved their risk perception.

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

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

- 8 participants (14.0%) reported that they did not trust the corresponding analogy domain and therefore found the analogies to be less helpful.
- 5 participants (8.8%) found that the analogy was irrelevant to the task at hand and therefore less helpful.

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

We believe that this is due to the explanation that well-understood accuracy highlighted the fact that the system can be wrong, thereby making users more aware of the risk (for example, the second comment in Table 6), and leading to a slight change in decision making that led to lower accuracy. As discussed in Section 5.2.1, we found that accuracy decreased in the **`AccAnalogy`** condition, but subjective trust did not. If analogies indeed improved risk perception, as prior work [11, 13] have shown in other contexts, then participants may have viewed relying on the system as riskier than making their own decisions. We discuss this further in the next section, in light of the earlier findings on reliance when users are presented with information on system accuracy.

6 FOLLOW-UP STUDY: THE INFLUENCE OF DIFFERING USER TRUST IN ANALOGY DOMAINS

To further understand the impact of users' trust in the analogy domains on their appropriate reliance, we conducted a within-subjects study in which each participant worked with AI systems where their stated accuracy was explained using analogies from three different analogy domains. This study was approved by the human research ethics committee of our institution.[8]

6.1 Experimental Setup

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

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

Procedure. We followed a similar procedure as in the main study (see Section 4.4). The main difference is that we did not separate participants into different experimental conditions. Instead, we separately assessed the user factors in each analogy domain before participants worked on one group of tasks explained with a single analogy domain.

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

Measures. We consider all covariates and reliance-based measures in the main study (see Section 4.2). However, we calculated the reliance-based measures according to each analogy domain. In addition, we assessed familiarity, trust, and confidence with the relevant analogy domain before each block of 4 tasks using that analogy domain. This was done using the following questions on a 6-point Likert scale:

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

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

Participants. Before recruiting participants, we computed the required sample size in a power analysis for a Within-Subjects ANOVA using G*Power [17]. We specified the default effect size $f = 0.25$ (*i.e.,* indicating a moderate effect), a significance threshold $\alpha = 0.025$ (*i.e.,* due to testing multiple hypotheses, **H3** and **H4**), a statistical power of $(1 - \beta) = 0.95$. This resulted in a required sample size of 245 participants.

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

6.2 Results and Analysis

Domain-specific User Factor Distribution. The distribution of analogy-specific user factors is visualized in Figure 4. Most participants reported a low *Familiarity* with the punctuality of French trains ($M = 1.70$, $SD = 1.14$). In comparison, most participants were familiar with the five-day weather forecast ($M = 5.08$, $SD = 0.94$) and AstraZeneca vaccine ($M = 4.65$, $SD = 1.25$). *Trust* was similar for all analogy domains, with the punctuality of French trains scoring lowest ($M = 3.57$, $SD = 0.99$), the weather report scoring slightly higher ($M = 3.85$, $SD = 1.04$) and the AstraZeneca vaccine getting the highest trust scores ($M = 4.36$, $SD = 1.33$). As for *Confidence*, this too was lowest for the French train punctuality ($M = 2.77$, $SD = 1.48$). Both the weather report ($M = 3.79$, $SD = 1.03$) and AstraZeneca vaccine ($M = 4.00$, $SD = 1.26$) scored higher on *Confidence*. As can be seen, standard deviations indicate that there were individual differences in how participants perceived these different analogies, while the aggregate results also show that the choice of analogy has an overall impact. Mann-Whitney tests using a Bonferroni-adjusted alpha

level of 0.025 ($\frac{0.05}{2}$) were used to compare all pairs of analogy domains. Our results indicate that: (1) participants showed a significantly higher *Familiarity*, *Trust*, and *Confidence* in the five-day weather report accuracy and the AstraZeneca vaccine effectiveness than the French train punctuality; (2) comparing the weather report and the AstraZeneca vaccine domains, we found that although participants reported a significantly higher *Familiarity* with the five-day weather report accuracy, they showed a significantly higher *Trust* and *Confidence* in the AstraZeneca vaccine effectiveness.

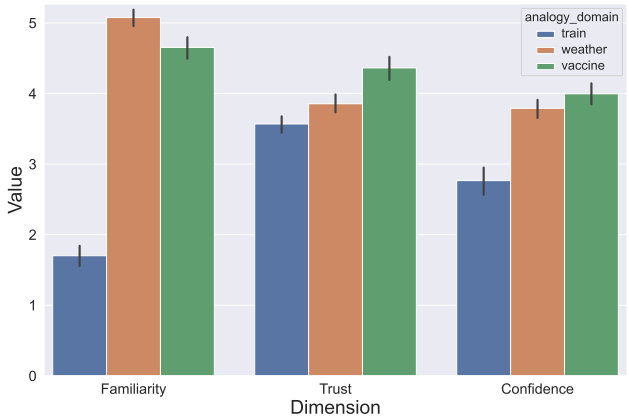

Fig. 4. Bar plot illustrating the distribution of the different user factors considered in our study. All user factors were measured on a 6-point scale.

Main Effect of Domain-specific User Factors. To analyze whether these differences had an effect on performance, we conducted Friedman tests for reliance-based measures across the different analogy domains. The results show that no significant difference exists between the reliance-based measures across the three analogy domains: *Agreement Fraction*, $\chi^2 = 0.19$, $p = 0.91$; *Switch Fraction*, $\chi^2 = 0.41$, $p = 0.81$; *Accuracy-wid*, $\chi^2 = 1.28$, $p = 0.53$; *Participant Accuracy*, $\chi^2 = 1.37$, $p = 0.50$; *RAIR*, $\chi^2 = 0.62$, $p = 0.73$; *RSR*, $\chi^2 = 2.89$, $p = 0.24$. While participants show relatively lower *Familiarity*, *Trust*, and *Confidence* on French train punctuality, no significant difference exists in the reliance-based measures. This indicates that, although participants perceive the three analogy domains differently, their reliance on the system is not affected by these differences in perception. Thus, we are reassured that our findings in the first study were not biased due to individual differences.

Table 8. Spearman rank-order correlation coefficient for user characteristics on reliance. "†" indicates the effect of variable is significant at the level of 0.025.

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

Correlation Analysis for User Factors on Reliance. For further insights about all user factors on user reliance behaviors, we calculated Spearman rank-order correlation coefficients for reliance-based dependent variables across all groups of tasks. As can be seen in Table 8, we found that

participants' trust, familiarity, and confidence with the analogies do not significantly affect reliance on the system. This further confirms our finding that differences in the perception of analogies do not affect reliance. Only participants' general *Propensity to Trust* shows a significant positive correlation with *Agreement Fraction*, *Swith Fraction*, and *Participant Accuracy*. This also aligns with our findings in main study (see Table 3) where the subjective trust in the AI system correlated significantly with their general *Propensity to Trust*. We also observed a positive correlation between users' Confidence and the RAIR they demonstrated, which indicates that users who have more confidence in the AI system, tend to more appropriately rely on the AI system.

7 DISCUSSION

7.1 Key Findings

Our analysis of the responses to the analogies suggests that the problem is not one of a lack of understanding of what the stated accuracy measure means. Nor was the decline in reliance observed in the analogy case the result of a reduction in subjective trust. As discussed, there were no significant effects on the various TiA subscales, even though these subscales correlated as expected with other subjective measures. In fact, the cases where participants were familiar with the analogies led to a significantly higher subjective understanding of the system, though here too there was no translation into higher reliance. We thus see a significant decline in accuracy that does not seem to be explainable in terms of a decline in subjective trust. According to the results discussed in Section 5.2.2, participants who reported a higher numeracy level tended to rely less on the AI system and achieved worse appropriate reliance and team performance (*i.e.,* accuracy). Therefore, we argue it is likely that participants overestimated their skills to deal with numeracy and loan prediction task, and did so more in the **`AccAnalogy`** condition. Combined with existing findings that analogies help improve risk perception in dealing with numeracy, the reduced reliance on AI system may be caused by the risk perception brought by analogies. The only unexpected effect is that it improved risk perception to their detriment: making users think that relying on the relatively accurate AI system was riskier than trusting their own answer. User comments such as the second and fourth in Table 6 match this interpretation of the results. For example, "The weather can be unpredictable, and so even the experts cannot be 100% sure at all times. The analogy helped to determine whether I should take the system's advice 100% or not".

Positioning in Existing Work. Our findings may seem at first to contrast with the findings of Yin *et al.* [65], where the authors found a significant effect of stated accuracy on reliance. We did not find this to be the case in our study using the loan prediction task. When aiming to better explain the stated accuracy measure through the aid of analogies, we even saw a reduction in reliance. How do these contrasting findings fit together? We consider the crucial difference to their study [65] to be that the observed effect of stated accuracy on reliance was only found for very high stated accuracy levels (90 and 95%) and even then users only agreed with the system in 80% of cases (up from 75% with no/lower stated accuracy). Our study intentionally did not consider these high accuracy levels, to avoid inducing system reliance simply due to the near certain promise of making the right decision when relying on the system (and thus acquiring the monetary reward). At 75% accuracy, though significantly better than human performance, users (especially those with high self-reported numeracy level) were reluctant to rely on the AI system. And indeed, for stated accuracies around 75% Yin *et al.* also did not find an improvement in reliance. In fact, even for a stated accuracy of 50% the observed agreement fraction was around 80% – they did not find effective calibration of reliance, especially for lower levels of stated accuracy.

This explanation of the findings is also in line with the findings of Yin *et al.*, where participants started to rely more on the system after they were given an overview of their own performance and

that of the system midway through the task (where generally the system performed better) [65]. This also aligns with the observed effect of *Propensity to Trust* and *Numeracy Level* in our study where the AI system shows superior performance than human performance. Participants who reported higher numeracy levels tended to rely less on the AI system — potentially due to thinking they can do better than the AI system with a 75% accuracy. Their reduced reliance and accuracy can be caused by the illusion of their own competence with numeracy and this task [31]. In contrast, participants who showed a higher propensity to trust tended to treat the AI system advice as more trustworthy, and relied more on the AI system.

Potential Cause — Dunning-Kruger Effect. Prior work in human behavior and psychology that have studied poor task performance have observed participants' overestimation of their own performance as an important reason. These studies attribute the overestimation to a cognitive bias called the Dunning-Kruger effect [38, 43]. The Dunning–Kruger effect describes a tendency for incompetent individuals to overestimate their ability, and has been replicated across several tasks in different domains including crowd work [19]. While we cannot entirely attribute the under-reliance of participants on the AI system in our study to the overestimation of their skills on the loan prediction task, there is a substantial amount of support for this plausible explanation in existing literature [31, 54].

Numeracy Levels Did Not Play a Role. Following on from overestimation of one's skills as the potential cause for under-reliance on the AI system, our results suggest that this occurs regardless of the numeracy level of participants. Having said that, we did observe that participants with low numeracy levels exhibited a higher reliance, *i.e.,* agree with and switch towards system advice more often (see Table 5), though this effect is not significant. Furthermore, participants with lower numeracy levels tend to have lower Trust in Automation scores, which is significant for the Intention of Developers measure (cf. Tables 5). As these findings are statistically insignificant, we refrain from drawing conclusions from them. At most, we think that should it turn out that findings regarding numeracy are significant in later studies then they make intuitive sense. Low-numeracy participants might rely more on a system not because of higher subjective trust, but rather due to a struggle with the range of numerical information they have to deal with. Hence, they report lower subjective trust but display higher objective reliance.

7.2 Caveats and Limitations

Observations on Single Accuracy Level. While it is informative to observe a lack of calibration to the stated accuracy level of 75%, our study is limited due to the restriction to a single accuracy level. As discussed above, the research of [65] only found an effect for higher accuracy levels when participants were not given feedback on their own performance, so perhaps the lack of findings regarding analogies is partly a result of our chosen accuracy level. That being said, participants would have been significantly better off relying more on the AI system, so even with a single accuracy level the question of how to get users to rely appropriately on such a system remains a valuable and important one. Thus, the findings of our study are important even though a single accuracy level was used.

Limitations of Analogy Domains. Furthermore, while the analogies we chose differed on the main feature of familiarity (with participants generally being unfamiliar with French trains and familiar with weather reports and covid vaccines), and all had a relevant structural mapping from accuracy in the AI domain and reliability in the various analogy domains, none were very close to the AI domain. Thus, it may be that participants' knowledge of the analogy domains was hard to apply in the AI domain. Alternatively, they might have preferred analogies closer to the task

domain (loan predictions), to clarify the meaning of accuracy in that context. That being said, participants who were familiar with the presented analogy domains did rate their understanding of the system higher and found the analogies to be helpful. According to the results in the follow-up study, we also found that the differences in perception of analogies (on *Familiarity*, *Trust*, and *Confidence*) did not show a significant impact on reliance-based measures. We, therefore, do not consider the choice of analogies to be the reason behind the significant decrease in user reliance on the AI system in the **`AccAnalogy`** condition.

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

Analogies can benefit users in understanding something that is not easy to digest [29, 30]. So in tasks with input data which is easy to comprehend (*e.g.,* visual input), our findings may not apply. Furthermore, as reported by Nourani *et al.* [49], the domain knowledge (expertise) plays an important role in facilitating reliance. In the presence of such potentially dominant factors, which appear to have a significant impact on trust formation and reliance behavior of users, our findings may not hold. In short, if users do not lack in their understanding (e.g., of measures like the AI system accuracy) analogies may be of little help, and explanations may not be needed in the first place.

Consideration of Task Type. The loan prediction task has been widely used to study human-AI decision making where there is a clear risk associated with the decision and a potential benefit in adopting AI advice [5, 9, 27, 60]. This task also follows the scenario-based exploration of end-user interpretability of AI systems championed by prior work [59]. However, the external validity beyond this scenario and domain (i.e., in other human-AI decision making tasks) and type of data (i.e., other than numerical data) cannot be ascertained. Future work could explore the effectiveness of analogy-based explanations, and consider alternative XAI methods altogether, in different scenarios [46].

7.3 Implications and Future Work

Based on our findings, we reason that an overestimation of users' skills in the task may explain their under-reliance on the AI system. Future work should further explore the effects of providing feedback to users on their performance. For whereas Green *et al.* [27] found that feedback on single decisions was of little use, Yin *et al.* [65] found feedback of average user accuracy to be a good motivator for increased reliance on system advice (though note, again, that reliance in their study was not optimal either). The question is whether and how this increased reliance can be calibrated properly to the system accuracy. Note that it is not the aim of our work to treat reliance on AI systems as universally desirable. However, to design and facilitate optimal team performance in human-AI decision making, it is pivotal to understand why users fail to achieve the theoretically possible higher accuracy — particularly when aided by a relatively more accurate AI system — and why users tend to demonstrate under-reliance. This is the spirit in which we explored the RQs in our work.

Regarding the use of analogy-based explanations, a complementary direction would be to consider the use of analogies to elucidate other general features of algorithms (*e.g.,* their decreased reliability when applied on outlier data, as such explanations have helped for appropriate reliance [8]), or to use analogies to explain more technical measures such as confidence scores and Shapley

values. These instance-level measures may be harder to interpret than the global accuracy measure explored in our work, and allow for a more dynamic presentation of analogies. If users lack enough expertise to comprehend these instance-level measures, then we believe that analogies can be helpful. Analogies may fit how humans actually reason, as Wang *et al.* note in their discussion of analogical reasoning [62] and we have observed some subjective effects from the use of analogies for stated accuracy. For that reason, they might be useful in explaining other parts of AI systems. An interesting finding from our work in this context, is that an improved risk perception can lead to under-reliance on AI systems and perhaps result in sub-optimal final decisions. Thus, more work is required to understand how to balance these two — promote criticality with which users rely on AI systems to prevent over-reliance on the one hand, and encourage reliance on AI systems when the advice is accurate to decrease under-reliance on the other hand. The ultimate aim should be to support users in their decision making, while fostering a better understanding of the AI system and promoting appropriate reliance of users on the system.

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

8 CONCLUSIONS

The two main research questions for this paper were: 'How does the understanding of stated system accuracy affect reliance of users on the AI system?' and 'How does explaining stated system accuracy using analogies affect the reliance of users on the AI system?'. As we have discussed, the conclusion to draw from our experiment is that users are no better at calibrating their reliance on the system when they better understand system accuracy. In fact, analogies made users less accurate, presumably because they became more aware of the risk that the system makes mistakes. A lack of understanding of the accuracy level is not the reason users fail to rely on the system appropriately. Thus, the limited understanding of stated accuracy is not to blame for under-reliance. This tallies with our finding that numeracy level, a factor one would expect to be relevant for a task filled with numerical information, had no significant effects on system reliance or accuracy.

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

of *Propensity to Trust* on reliance. Further research is needed here, but it is striking that the level of understanding of the presented numerical information has little bearing on user reliance.

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

ACKNOWLEDGMENTS

This work was partially supported by the Delft Design@Scale AI Lab, the 4TU.CEE UNCAGE project, and the Convergence Flagship "ProtectMe" project. We made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-1806, EINF-3888 and EINF-5571. We finally thank all participants from Prolific and experts from our department.

REFERENCES

[1] 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*. 1–16.

[2] Elisa Barilli, Lucia Savadori, Stefania Pighin, Sara Bonalumi, Augusto Ferrari, Maurizio Ferrari, and Laura Cremonesi. 2010. From chance to choice: The use of a verbal analogy in the communication of risk. *Health, Risk & Society* 12, 6 (2010), 546–559.

[3] Marianne Bertrand, Sendhil Mullainathan, and Eldar Shafir. 2006. Behavioral economics and marketing in aid of decision making among the poor. *Journal of Public Policy & Marketing* 25, 1 (2006), 8–23.

[4] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*. ACM, 401–413.

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*. 1–14.

[6] 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*. 454–464.

[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[8] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. 120–129.

[9] 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*. 307–317.

[10] Sam Desiere, Kristine Langenbucher, and Ludo Struyven. 2019. Statistical profiling in public employment services: An international comparison. (2019).

[11] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[12] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.

[13] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. The perceived utility of human and automated aids in a visual detection task. *Human Factors* 44, 1 (2002), 79–94.
[14] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the aaai conference on human computation and crowdsourcing*, Vol. 8. 43–52.
[15] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.
[16] 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 (2007), 672–680.
[17] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.
[18] 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.
[19] Ujwal Gadiraju, Besnik Fetahu, Ricardo Kawase, Patrick Siehndel, and Stefan Dietze. 2017. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 4 (2017), 1–26.
[20] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1631–1640.
[21] Mirta Galesic and Rocio Garcia-Retamero. 2011. Communicating consequences of risky behaviors: Life expectancy versus risk of disease. *Patient education and counseling* 82, 1 (2011), 30–35.
[22] Mirta Galesic and Rocio Garcia-Retamero. 2013. Using analogies to communicate information about health risks. *Applied Cognitive Psychology* 27, 1 (2013), 33–42.
[23] Jorge Galindo and Pablo Tamayo. 2000. Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics* 15, 1 (2000), 107–143.
[24] Rocio Garcia-Retamero, Agata Sobkow, Dafina Petrova, Dunia Garrido, and Jakub Traczyk. 2019. Numeracy and risk literacy: What have we learned so far? *The Spanish journal of psychology* 22 (2019).
[25] Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7, 2 (1983), 155–170.
[26] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.
[27] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[28] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.
[29] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. 2022. It Is Like Finding a Polar Bear in the Savannah! Concept-Level AI Explanations with Analogical Inference from Commonsense Knowledge. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 89–101.
[30] 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)*.
[31] 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*. 1–18.
[32] Douglas R Hofstadter and Emmanuel Sander. 2013. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic Books.
[33] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci. 2017. Risk stratification of lung nodules using 3D CNN-based multi-task learning. In *International conference on information processing in medical imaging*. Springer, 249–260.
[34] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.
[35] Carmen Keller, Michael Siegrist, and Vivianne Visschers. 2009. Effect of risk ladder format on risk perception in high-and low-numerate individuals. *Risk Analysis: An International Journal* 29, 9 (2009), 1255–1264.
[36] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.
[37] Gang Kou, Yi Peng, and Guoxun Wang. 2014. Evaluation of clustering algorithms for financial risk analysis using MCDM methods. *Information Sciences* 275 (2014), 1–12.

[38] 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.
[39] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.
[40] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[41] 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*. 1–16.
[42] Michael A Martin. 2003. "It's Like… You Know": The Use of Analogies and Heuristics in Teaching Introductory Statistical Methods. *Journal of Statistics Education* 11, 2 (2003).
[43] Matan Mazor and Stephen M Fleming. 2021. The Dunning-Kruger effect revisited. *Nature Human Behaviour* 5, 6 (2021), 677–678.
[44] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[45] Eric Mintz and Truls Østbye. 1992. Teaching statistics to health professionals: the legal analogy. *Medical Teacher* 14, 4 (1992), 371–374.
[46] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).
[47] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15, 5 (2010), 625–632.
[48] 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. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.
[49] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
[50] 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*. 340–350.
[51] Andrea Papenmeier, Gwenn Englebienne, and Christin Seifert. 2019. How model accuracy and explanation fidelity influence user trust in AI. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019*.
[52] Stefania Pighin, Lucia Savadori, Elisa Barilli, Rino Rumiati, Sara Bonalumi, Maurizio Ferrari, and Laura Cremonesi. 2013. Using comparison scenarios to improve prenatal risk communication. *Medical Decision Making* 33, 1 (2013), 48–58.
[53] Katharine Sanderson. 2021. COVID vaccines protect against Delta, but their effectiveness wanes. *Nature* (2021).
[54] 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*. 240–251.
[55] 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. In *ACM Conference on Human Factors in Computing Systems (CHI'22), Workshop on Trust and Reliance in AI-Human Teams (trAIt)*.
[56] 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. In *International Conference on Information Systems (ICIS)*.
[57] Pradeep Sopory and James Price Dillard. 2002. The persuasive effects of metaphor: A meta-analysis. *Human communication research* 28, 3 (2002), 382–419.
[58] 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*. 77–87.
[59] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).
[60] 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*. 1–13.

[61] Stephanie K Van Stee. 2018. Meta-analysis of the persuasive effects of metaphorical vs. literal messages. *Communication Studies* 69, 5 (2018), 545–566.
[62] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
[63] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. 2019. Making sense of recommendations. *Journal of Behavioral Decision Making* 32, 4 (2019), 403–414.
[64] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Does stated accuracy affect trust in machine learning algorithms. In *Proceedings of ICML2018 Workshop on Human Interpretability in Machine Learning (WHI 2018)*, Vol. 7.
[65] 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*. 1–12.
[66] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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*. 295–305.
[67] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.
[68] Brian J Zikmund-Fisher, Dylan M Smith, Peter A Ubel, and Angela Fagerlin. 2007. Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations. *Medical Decision Making* 27 (2007), 663–671.

Received July 2022; revised January 2023; accepted March 2023