Evaluating the Impact of Uncertainty Visualization on Model Reliance

Jieqiong Zhao,Yixuan Wang,Michelle Mancenido,Erin K. Chiou,Ross Maciejewski

Abstract—Machine learning models have gained traction as decision support tools for tasks that require processing copious amounts of data. However, to achieve the primary benefits of automating this part of decision making, people must be able to trust the machine learning model's outputs. In order to enhance people's trust and promote appropriate reliance on the model, visualization techniques such as interactive model steering, performance analysis, model comparison, and uncertainty visualization have been proposed. In this study, we tested the effects of two uncertainty visualization techniques in a college admissions forecasting task, under two task difficulty levels, using Amazon's Mechanical Turk platform. Results show that (1) people's reliance on the model depends on the task difficulty and level of machine uncertainty and (2) ordinal forms of expressing model uncertainty are more likely to calibrate model usage behavior. These outcomes emphasize that reliance on decision support tools can depend on the cognitive accessibility of the visualization technique and perceptions of model performance and task difficulty.

Index Terms—Uncertainty, model reliance, trust, human-machine collaborations

1 Introduction

Artificial intelligence and machine learning (AI/ML) models have gained traction as decision support tools in tasks that are highly repetitive, laborious, and/or require processing copious amounts of data. Despite the benefits of automation, some application domains that are security-critical and/or high stakes are not yet amenable to full automation. Examples include automated baggage screening systems in airport security checkpoints [1] and models that predict the chances of parolee recidivism [2]. In these domains, human experts are tasked to corroborate information from the model and, in some cases, to intervene and/or to provide a final system decision. This is due to the well-established fact that, no AI/ML model is perfectly accurate.

A prerequisite for optimizing the performance of human-AI/ML systems is the appropriate calibration of the human's trust in the model's outputs i.e., the human should be able to recognize when to rely on model predictions and when to spot flaws in the model's judgment [3], [4]. Aligning a human's perceived trust with the model's actual capabilities (i.e., performance, trustworthiness) and limitations (i.e., uncertainties) in varied operating environments remains one of the open questions in the design of AI/ML-enabled systems. Failure to appropriately calibrate trust results in the over-reliance or under-reliance on the model's outputs [5], both of which diminish the benefits of a joint human-automation system. Trust calibration can be achieved through various approaches, like providing information about how decisions are generated by AI/ML models (e.g., interpretability) or providing information to qualify weaknesses in the AI/ML model (e.g., uncertainty). In this paper, we mainly study the latter.

Trust calibration in AI/ML decision support systems has been posed as a problem of efficient and effective communication of the AI/ML's limitations or imperfections [6]. In practice, one potential method for revealing a model's imperfections to a human counterpart is through uncertainty awareness. As Tomsett, et al. [7] put it,"while (model) interpretability makes clear what the system knows, uncertainty awareness reveals what the system doesn't know." Uncertainty awareness may assist decision makers and auditors in forming an appropriate representation of the AI/ML's limitations so that suitable corrections or adjustments could be made [8], [9]. The estimated uncertainty of a prediction by an AI/ML model can potentially alert human decision makers to how that prediction will perform, with higher uncertainty indicating a higher potential for poorer performance. Traditionally, the uncertainty of a prediction is mathematically presented in a probabilistic form. For people to perceive uncertainty information quickly, effective visualizations are required to convey the probability scores, often represented as distributions of a prediction target within a range. The domain of uncertainty quantification is expansive and we refer to the comprehensive discussion in Hüllermeier and Waegeman [10]. Meanwhile, rigorous quantitative evaluation methods to determine the effectiveness of uncertainty visualizations in decision making have been gradually established [11].

To date, few studies have quantitatively measured the impact of communicating AI/ML uncertainty on model adoption or rejection, and by proxy trust calibration, in the use of decision support systems. In addition to the method of revealing the ground truth and model predictions [12], Dietvorst et al. [13] communicated a model's uncertainty through the outright disclosure of the model's average error rate by a text description (i.e., "the model has an average error rate of x"). These studies tracked how often participants selected the model outputs over human judgments in the presence of model uncertainty. The overwhelming

- *Jieqiong Zhao, Yixuan Wang, Michelle Mancenido, Erin K. Chiou, Ross Maciejewski are with Arizona State University.*
E-mail: {Jieqiong.Zhao, ywan1290, mickey.mancenido, Erin.Chiou, rmacieje}@asu.edu.

Manuscript received MM DD, 2021; revised MM DD, 2021.

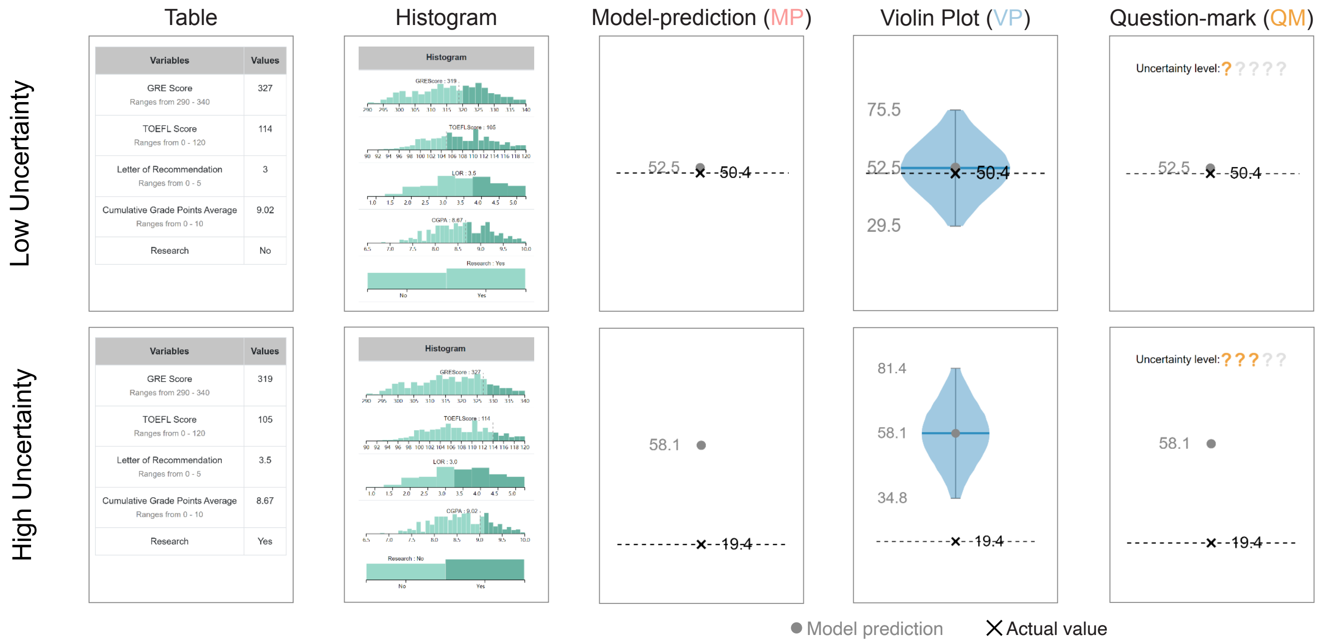


Fig. 1. Our experiment used five visual representations that could be grouped into three categories: (a) Table and Histogram, which do not have model prediction values, (b) Model-prediction, which consists solely of model prediction values, (c) Violin Plot and Question-mark, which include both model prediction values and uncertainty. Two instances of predictions displaying varying levels of model uncertainty are displayed across the rows. The actual values are only presented in the model performance demonstration stage, not in the prediction task.

consensus is that *without appropriate interventions*, people are heavily biased to adopt imperfect human judgments over imperfect models [12], [14]. The question of which interventions work to mitigate this bias, such that people make the appropriate choices, returns to the importance of trust calibration and aligning AI/ML capabilities to people's understanding and use of those capabilities. Dietvorst et al. [13] showed that people will use imperfect models *if* they have the ability to modify the model's outputs. In those studies [12], [13], [14], the AI/ML models were regarded as fully automated decision aids, and participants had only to make a single decision — accept model predictions for all cases (fully automated) or discard model predictions but use their own predictions instead for all case (fully manual). The single decision measures overall reliance on fully automated models. However, with modern visual analytics systems, people are able to make dynamic decisions depending on the performance of individual predictions with varying levels of uncertainty, and the difficulty of tasks, which have not been quantified in previous studies.

In this paper, our major aim is to explore different uncertainty visualization techniques as a possible intervention for aligning human decisions with a model's actual capabilities. In essence, if a model has uncertainty about a prediction, does an uncertainty visualization unilaterally prompt human aversion, or, are there uncertainty visualizations that would encourage human adoption of the model's outputs? Theoretical groundwork for this question has been established in uncertainty visualization for statistical models [15], [16], [17], uncertainty-aware black-box systems [7], and algorithm aversion [12], [14]. However, these works are disparate and do not specifically establish an association between visual representations of uncertainty and the adoption of model outputs in the presence of high or low uncertainty. Here, we note that our study focuses only on the impact of uncertainty representation on model adoption. For the model chosen in our experiments, low uncertainty typically indicates that the model is reliable. However, this is not necessarily true for all models, and future work addressing specific decision contexts is needed to confirm if such representations can result in inappropriate reliance, which is considered a useful, albeit indirect, measure of miscalibrated trust [18].

Our study examines two designs for uncertainty visualization in comparison with a baseline case in which only the model prediction is present. To test these designs, we devised a forecasting task for study participants in Amazon Mechanical Turk (MTurk). The test bed resembles the regression prediction task used in Dietvorst et al. [12], in which participants were tasked to predict an applicant's chance for admission into a graduate school program with or without observing model performance by presenting the model prediction result in plain text. In our study, we build upon Dietvorst et al.'s work [12] and explore the impacts of visualization on model reliance in a Multiple Linear Regression model. Here, the regression model acted as an AI/ML agent. We chose this test bed because of the relative transparency and simplicity of quantifying uncertainty in linear regression. Additionally, the available data set allowed us to control for what we refer to as model uncertainty, which we closely associate with task difficulty. In this experiment, we employed static visualization exclusively, in order to reduce potential confounds (e.g., animation in a hypothetical outcome plot), as our first step in investigating the impact of uncertainty visualization on model reliance. We considered two forms of representations to convey model uncertainty, ordinal vs. distributional, and adopted Question-mark glyph and Violin Plot (shown in Figure 1) as their representatives, respectively. We chose these two straightforward representations to prevent participants from being overwhelmed with excessive information when performing regression tasks with given data attributes. In the results from our experiment:

1) We show that when a decision task has low model uncertainty, people tend to adopt model predictions and tend to use model information when predictions include uncertainty visualization; however, when a decision task has higher model uncertainty, the uncertainty visualization has no significant effect on model adoption (Section 5.1 Q1).
2) We compare two designs of uncertainty visualization namely, the Violin Plot and a Question-mark glyph, and show which visualization between the two better aligns human decisions and behaviors with the model's uncertainty level, with better alignment suggesting better trust calibration (Section 5.1 Q2).
3) Finally, we show which uncertainty visualization design results in more positive perceptions of model trustworthiness, confidence, security, and reliability which are proxy measures of the decision maker's trust (Section 5.1 Q3).

2 Related Work

Based on previous work addressing trust in automation, reliance behaviors with decision support systems, uncertainty-aware AI/ML models, and data visualization, this study explores the human decision dynamics of forecasting tasks when supported by visualizations of model uncertainty.

2.1 Trust Calibration and Model Reliance

Trust in decision support systems has received much attention, particularly with recent developments in ML-enabled systems that are more difficult to manually inspect. A large body of literature has argued that it is critical for people to calibrate their trust to a machine's actual capabilities, if the appropriate human behaviors while using these machines are to be achieved and sustained, for better overall performance from the human-machine system [4], [5], [19]. Due to the difficulty of directly measuring trust, which is a social-psychological construct, many researchers resort to using behavioral measures, such as reliance, as a proxy for trust, even though trust can only partly predict reliance. However, such behavioral measures are often the primary outcomes of interest when it comes to the study of trust. During the decision-making process in a human-AI/ML teaming environment, without appropriate trust calibration, humans can either overly rely on predictions made by AI/ML models (known as automation bias [20]) or disregard predictions made by AI/ML models (known as algorithm aversion [12]). Automation bias and algorithm aversion are two opposite tendencies where people fail to align their trust with a model's actual capabilities. Automation bias has been defined as the tendency to follow the automated model suggestion "as a heuristic replacement for vigilant information seeking and processing" [21]. It has been well-studied in a variety of domains such as aviation [22], [23], health care [24], [25], [26], and process control [27], [28], [29]. These studies found that automation bias can occur even when the automated model provided wrong advice. This phenomenon has been verified by various human-computer interaction (HCI) studies that show people may over-rely on automated models and follow their incorrect suggestions, even when they would make a better decision on their own [30], [31], [32]. Logg et al. [33] further investigated automation bias, and found that lay people are more prone to rely on algorithmic suggestions while experts are more likely to be affected by algorithm aversion.

Previous studies reveal potential reasons why people have algorithm aversion, namely they prefer human predictions over model predictions [34], [35]. First, algorithmic models are inherently distrusted in many application scenarios [36], [37], [38]. However, in these scenarios, if people would adopt model predictions appropriately, the overall human-AI/ML collaborative performance could improve, especially for predictions with low model prediction uncertainty. In algorithm aversion studies conducted by Dietvorst et al. [12], people discarded model predictions if errors were observed. Even when informed that human forecasters performed worse than the model, people were still inclined to select the human forecaster. If people were provided with the opportunities to adjust predictions made by the model, then model adoption rate increases significantly [13]. The increase indicates that modifying a model's predictions can increase human trust and model adoption; yet, overall performance was worse after human adjustments.

It is notable that these initial studies applied models as black-boxes and evaluated peoples' reliance on models by displaying model predictions with ground truth. Later, Yang et al. [39] provided visual explanation of a classifier to support appropriate trust through model transparency. In this paper, our goal is to test the effects of visualizing uncertainty on model usage, with model usage representing reliance and a proxy measure of trust the model.

2.2 Uncertainty Awareness in AI/ML Models

Showing uncertainty information can aid in understanding the causes of potential errors, variations, and biases [16], [40], [41] in the data, model outputs, and visual mapping. For instance, AI systems with uncertainty-aware designs can inform decision-makers about the known unknowns [7], which helps decision-makers adopt different strategies as needed. There are two types of uncertainty, aleatoric uncertainty and epistemic uncertainty [10], [42]. Aleatoric uncertainty indicates the inherent randomness of a system or a model, such as a flip of a coin. Epistemic uncertainty denotes uncertainty caused by a lack of knowledge, which can be mitigated by observing additional data. In this paper, we are referring to epistemic uncertainty which is possible to mitigate through visualization.

Visualizing uncertainty can be used to calibrate trust to a model's actual capabilities and enhance appropriate model reliance, and some studies have been conducted to evaluate the impact of uncertainty on reliance decisions. One study conducted by Cai and Lin [9] showed that presenting confidence levels of a system improved trust calibration in dynamic scenarios (e.g., autonomous driving). Furthermore, a study conducted by Conway et al. [43] discovered that the actual capability of a decision aid was inadequately perceived if multiple cues with uncertainty were given. In our study, we use visual representations to show the uncertainty of a regression model to confirm associations between model use and uncertainty visualization.

2.3 Uncertainty Visualization

Visualizations may have significant differences in their ability to communicate uncertainty across different tasks, with controlled experiments showing that even researchers cannot interpret confidence intervals correctly for statistical inference [44], [45]. However, more recent work has begun to explore the effectiveness of various uncertainty visualizations on a per task bases. Distributional visualization with frequency information (e.g., quantile dot plots and hypothetical outcome plots) have been found more effective for communicating uncertainty [46], [47], [48], in particular, work by Padilla et al. [49] found that humans can better reason with direct and indirect uncertainty when presented as quantile dot plots. Several studies on bar charts revealed that bar charts with error bars were not efficient for inferring statistical information including sample means [50] and error distributions [51], while color density encoded horizontal bars were found to be effective in expressing temporal uncertainty [52], [53]. The dynamic variance of data distribution (data uncertainty) in progressive visualization was evaluated by Procopio et al. [54]. Kale et al. [55] investigated the interpretation of effect size using eight types of uncertainty visualizations. They found that the variance of distributions is a critical factor in their experiment to impact decision making. Guo et al. [56] introduced an alternative-aware uncertainty visualization to enhance humans' confidence in selecting between two options, in which they discovered that the level of uncertainty plays a critical role. Although these works applied different approaches to obtain the amount of uncertainty, it is evident that enhancing uncertainty awareness by visualizations is critical. Thus, in our experiment, we also considered the level of uncertainty to inform the selection of samples in our prediction task (details in Section 4.2)

Our study adopted a Question-mark representation (QM) and a Violin Plot (VP) to show the uncertainty of a model prediction. Compared to frequency-based formats (e.g., quantile dot plots, HOP), the QM ordinal representation – a discretized presentation – was simplified for laypeople using the number of golden question marks to inform rough uncertainty levels without precise numbers regarding the upper and lower bounds of uncertainty. Furthermore, using the number of golden markers to indicate a rating score is widely adopted in popular websites and mobile apps. These two uncertainty representations were selected due to their different complexity in expressing uncertainty: continuous probability density distribution vs. an ordinal value. We evaluated these uncertainty representations' (QM and VP) impact on model usage (adoption or rejection of the ML model prediction) compared with observing solely the model prediction (MP). We considered not only uncertainty representations but also the experimental factors and procedure, task complexity, and completion time as a whole. We prefer a within-subjects experimental design to maximize the signal-to-noise ratios because each participant serves as their own control. Therefore, we only adopted two uncertainty visualizations in our experiments to ensure the experiment could be completed within a reasonable time and avoid fatigue among participants.

3 Testbed and MTurk Platform

The current study investigated how different visualizations of uncertainty information, including no visualization of uncertainty, impacted adoption behavior of model predictions among a general population. Participants were asked to inspect visualizations that provided varying levels of detail about a data set and a model's predictions based on that same data set. Participants then chose whether or not to adopt the model's prediction. The study was conducted with Amazon Mechanical Turk (MTurk) participants who were instructed to examine the information presented for each task and decide whether to use the model's prediction or to provide their own prediction. Therefore, model adoption in this study served as a proxy for reliance. In this section, we describe in detail the data set, the prediction model, and the prediction task.

3.1 Dataset

The testbed was designed based on a publicly available data set of graduate student admissions.[1] The data set consists of five hundred unique applicants with seven independent predictors for admission including cumulative GPA, GRE score, TOEFL score, university rating, statement of purpose, letter of recommendation, and research experience. To simplify the prediction task for our general participants, we pre-processed the data by transforming the target outcome variable (likelihood of admission) and performing stepwise regression. For the target variable, the original values (likelihood of admission) were converted into percentiles to eliminate any contextual information (i.e., the range of likelihood of admission). In this case, the percentile score for a given applicant was equivalent to the percentage of applicants whose likelihood of admission was restrictively lower. For example, an applicant with a percentile score of 50 means that 50% of the applicant pool had a lower chance of admission than the applicant.

3.2 AI/ML Agent

Multiple linear regression models are among the simplest AI/ML tools for supervised learning where a set of predictors and a target variable make up the input-output space. Using a full set of the graduate admissions data ($N = 498$, excluding two applicants whose percentile scores equal zero), we performed ordinary least squares estimation and stepwise regression (JMP® version 15) on the original predictors, with the percentiles as the target variable. Stepwise regression is a step-by-step iterative process of choosing the best predictors out of a large set of predictors. The forecasting tasks for participants were deliberately selected according to the level of uncertainty (more details in Section 4.2). The final regression model that served as the AI/ML agent for the experiments included five predictors

[1] https://www.kaggle.com/mohansacharya/graduate-admissions

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3251950

TABLE 1
An example of a graduate school applicant with the 5 variables used in the prediction task.

Variable	Value
GRE Score [290 - 340]	305
TOEFL Score [0 - 120]	112
Letter of Recommendation [0 - 5]	3.5
Cumulative Grade Points [0 - 10]	8.65
Research	No

(Table 1) and one active interaction effect between TOEFL score and research experience ($R^2 = 0.837$, $R^2_{adj} = 0.837$, $RMSE = 11.66$). To reduce task complexity, only the active predictors were shown to participants. Often, an active predictor has two characteristics: (1) the prediction performance of a single predictor exceeds a given threshold, and; (2) not highly correlated with other active predictors. Auxiliary model information, such as performance indicators and regression coefficients, were not disclosed. This task has been crafted to resemble tasks in other studies on algorithm aversion (i.e., [12]), in which participants were asked to predict admissions for MBA applicants. We extended these prior studies by adding visualizations, including distributions of variables, model prediction values and uncertainty.

3.3 MTurk Platform

Participants were instructed to assume the role of a graduate school admissions officer, whose main task was to predict the percentile score of an applicant. To motivate performance, participants were informed that lower error rates would lead to higher bonuses, which were calculated based on the **a**verage **a**bsolute **e**rror (AAE) in Equation 1 (k denotes the number of cases evaluated). AAE is easier for participants as it is a linear calculation. Participants only observe AAE information both for their own predictions and the model's predictions. The mapping between AAEs and bonuses is shown in Table 2.

$$AAE = \frac{1}{k}\sum_{i=1}^{k} |y_{predicted(i)} - y_{actual(i)}| \qquad (1)$$

The bonus structure was such that participants received a one dollar bonus if, on average, their final predictions (a selection of either a model prediction value or their own prediction value) deviated from the actual value within 5 percentile points. For every additional 5 percentile deviation, the bonus decreased by 20 cents. If the participant's AAE was larger than 25 percentile points, there was no bonus. Similar incentive structures were used in previous work on algorithmic aversion [12].

TABLE 2
Mapping of AAE to bonuses.

AAE	5	10	15	20	25
Bonus	$ 1.00	$ 0.80	$ 0.60	$ 0.40	$ 0.20

4 Methods

The goal of the MTurk experiments is to test the effects of different uncertainty visualizations on the adoption or rejection of model outputs by general participants. We conducted two independent experiments (Figure 2) with the same test bed, conditions, and protocols but with different participant groups and tasks. The *task* in this study was a forecasting instance (i.e., one set of predictors for one applicant) presented to participants for their prediction of the applicant's acceptance percentile. For brevity, we will refer to participants from the first study as the *study group* and participants from the second study as the *confirmatory group*. The second, confirmatory study, serves to corroborate results from the first study to test whether the results are consistent across participant groups and tasks. The concept of confirmatory validation runs has been applied in design of experiments [57] and response surface methodology [58] to check if certain singles found in a prior experiment can be generalized to a reasonable range of conditions in follow-up experiments. We only report results that are statistically significant for both groups in Section 5. In this section, we describe our research questions and associated metrics (also known as response or target variables); experimental factors, conditions, and design (Section 4.2); participant recruitment and demographics (Section 4.3); the experimental protocol (Section 4.4); and finally, the method for statistical analysis (Section 4.5).

4.1 Research Questions

Q1. *Does uncertainty visualization help to align appropriate human decisions in a forecasting task, given a certain level of model uncertainty i.e., low and high uncertainty?* For low uncertainty tasks, regression models are expected to perform better than a human forecaster, so the optimal decision should be to adopt model predictions. Conversely, the appropriate behavior would be to reject model predictions under high model uncertainty. Thus, the response variable is whether a participant, in a specific task, adopted the model prediction or opted for their own prediction. Additionally, participants who opted for their own prediction were asked "Did you use model information or not?", another dichotomous measure. Here, we note that uncertainty does not always indicate a model is performing well; however, for this experimental setup, low uncertainty is aligned with good performance.

Q2. *Which of the two uncertainty visualizations explored in this study (Violin Plot or Question-mark glyph) supports better decisions given the level of model uncertainty (i.e., supports trust calibration better)?* The response variables are the same as Q1.

Q3. *Which of the two uncertainty visualizations explored in this study (Violin Plot or Question-mark glyph) has higher perceived trustworthiness, reliability, and confidence in predictions?* At the end of the experiment, participants were administered a 37-item post-task questionnaire to measure perceptions of the model and visualization tools. The majority of questionnaire items solicited Likert-type responses. For each task, in addition to the same behavioral measures collected in Q1 and Q2,

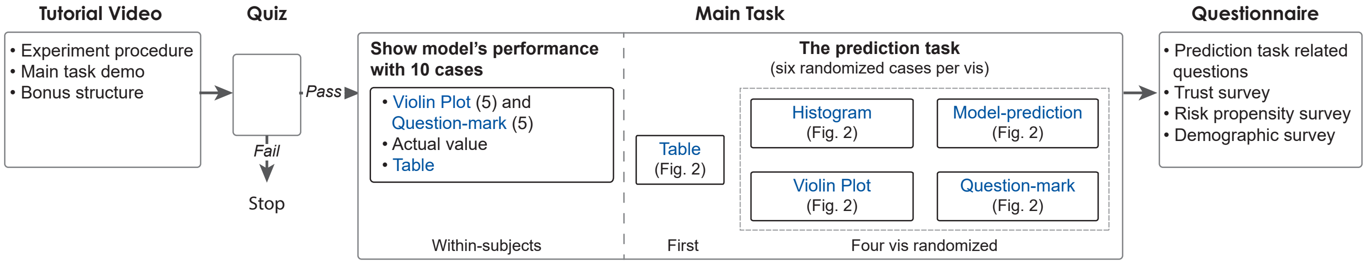


Fig. 2. Experimental procedure of the study. Participants completed the same forecasting task for six applicants across five representations (within-subjects quasi-split plot design). For the prediction task, the raw data Table was presented first, and the order of the Histogram, Model-prediction (MP), Violin Plot (VP), and Question-mark (QM) representations were randomized.

participants were also asked to elaborate their reason for adopting or rejecting model predictions in a free-response field.

Auxiliary information, such as the AAE (Equation 1), task time, and demographic details, were also analyzed.

4.2 Experimental Design

Five visual representations of the data (Figure 1) were implemented in this study, including: two representations of the raw data (Table, Histogram), a representation of the model prediction value only (Model-prediction), and two representations of uncertainty (Violin Plot, Question-mark glyph) along with the model prediction. The Table representation replicates Dietvorst et al.'s experiments [12], [13], which uses a table to list the variables and their values in text for a prediction task. The Histogram representation uses visual aids to further convey the percentile information of each variable. We used the Table and Histogram as comparative representations to benchmark the effects of other representations with model prediction, uncertainty information, or both. In addition, we also used Table and Histogram as an attention check to confirm that participants were engaged in the task since these two representations do not include model predictions. That is, if AAE under Table and Histogram conditions were significantly higher than the three other representations that contained model predictions, we would conclude that a participant was likely making uninformed or random guess decisions.

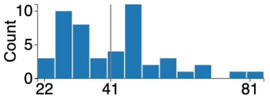
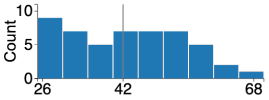
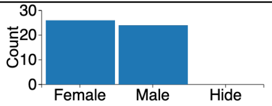
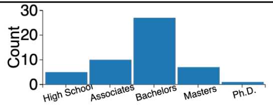
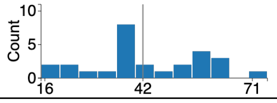
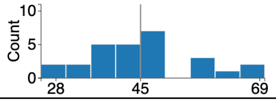
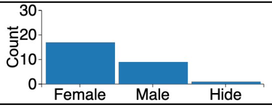
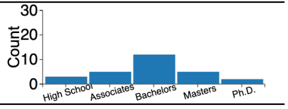
The amount of uncertainty accompanying a prediction from linear regression models is referred to as the *prediction variance*. It is a function of the uncertainty in parameter estimates, also known as epistemic uncertainty – the conditional variance around an observation (given the specific settings of the predictors). The model's measure of uncertainty may contain epistemic or systematic uncertainty, as there are input variables that could affect the response that were not included in the model. In this case, the model fit and diagnostics indicate the lack of epistemic uncertainty. Thus, to determine what forecasting tasks would comprise the high and low levels of model uncertainty, we considered two criteria: (1) the prediction variance of an individual observation [59]; (2) the set deletion statistics, DFFITS and DFBETA, of an individual observation, which points to the "rarity" of a specific data point in relation to the subspace of predictors (x) and actual observations y [60]. While set deletion statistics such as DFFITS and DFBETA are traditionally used to measure the influence of an observation in regression models, we used both as guidelines to detect edge cases. Thus, data points which had high prediction variances and DFFITS and DFBETA statistics relative to the rest of the data comprised the "high model uncertainty" category. These are data points which are unusual in both the x and y values; intuitively, both the model and human agents would find these cases difficult to predict. This distinction between high and low levels of model uncertainty is the basis for designing two types of Violin Plots and Question-mark glyphs, one that reflects high model uncertainty (difficult cases) and one that reflects low model uncertainty (theoretically easy cases for the model and participants).

10 experimental conditions (5 levels of representation × 2 levels of model uncertainty) were tested in the study. We selected 3 prediction instances for each level of model uncertainty, with each prediction instance presented through the 5 representations. As we opted for a within-subjects experimental design, every participant had to predict 6 unique instances (3 high, 3 low) across all 5 representations, resulting in 30 prediction cases in total. Participants were not informed that they were making predictions for the same cases, differing only in the representation or visualization method. To alleviate the threat of maturation and test-retest bias, we first randomized the order of 5 representations, with the exception of the Table always presented first and served as the baseline condition. Then, within each representation, we further randomized the order of prediction instances. Participants were *not* informed of the ground truth after each prediction.

The experimental design follows a factorial structure (each factor level completely crossed with levels of the other) with restricted randomization (where representations were randomized within a case, but not across cases) and repeated measures. We opted for this study design for several reasons. First, the uniformity of cases across representations deterred possible noise resulting from task-to-task differences. Thus, if there were differences in responses among experimental conditions, they would most likely be due to the factors of interest. Second, a within-subjects (repeated measures) design was used for a similar reason i.e., individual differences among participants would most likely add variation that we were not interested in estimating. Finally, we only used 6 unique applicants because of concerns about experimental task fatigue among participants.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3251950

TABLE 3
Detailed information about participants in the study group and the confirmatory group, including number of participants, study completion time, age, gender, and educational background. For completion time and age, the min, max, and median values are highlighted in the Histograms.

Group	N	Completion Time (minutes)	Age	Gender	Education
Study Group	50				
Conf. Group	27				

Our experiment setup was that each participant was tasked to complete 30 predictions, which could be accomplished within a reasonable duration.

To identify the consistency across two samples, we used the same approach for the confirmatory experiment, with two key differences: a different set of prediction cases and a new group of MTurk participants. Similar criteria for establishing baselines of model uncertainty level were employed as in the first study. Due to the change of the set of prediction cases, the model prediction values and corresponding uncertainty (prediction variance) intervals are different, which may impact participants' behavior of model adoption and rejection. It implies that the final prediction performance and adoption behavior may rely on the performance of the model. Our findings only reported results consistent between both study and confirmatory groups.

4.3 Participants

Participants recruited through MTurk were filtered based on: (1) a HIT approval rate $\geq 98\%$, (2) number of approved HITs $\geq$ 5000, and (3) location, i.e., US only. Participants were required to pass a quiz following a tutorial video. They cannot proceed with the experiment if they do not pass the quiz. In total, 162 MTurk workers started the quiz, while 14 workers did not complete it, and 56 workers failed. The 34.57% failure rate is somewhat common compared to other online experiments conducted on MTurk with higher failure rates [61], [62]. Ninety two (92) MTurk workers passed the quiz but only 77 workers successfully submitted a HIT. Ultimately, 50 participants were recruited for the study group, while 27 participants were recruited for the confirmatory group. These sample sizes were projected as adequate for the statistical models used for data analysis (see Section 4.5). Participants were compensated with a base payment of three dollars and a maximum bonus of one dollar. More details about the participant samples in the study and confirmatory groups are shown in Table 3.

4.4 Protocol

The experimental procedure, reviewed and approved by the university's Institutional Review Board (IRB), is shown in Figure 2. Participants were firstly asked to read an information sheet and to watch a video tutorial, which introduced the prediction task, independent variables, prediction target, study interface, and the bonus structure. Then, brief explanations of the interpretation of percentiles, uncertainty ranges, and absolute error were provided. To assist the interpretation of the amount of uncertainty, in the tutorial video, we verbally stated that a larger prediction interval in a Violin Plot or more golden question marks in a Question-mark representation indicates a higher amount of uncertainty. Meanwhile, high vs. low uncertainty of Violin Plots and Question-mark representations were presented in a matrix format similar to Figure 1. After watching the video, a short quiz was administered to participants to ensure that they completed task training and be able to distinguish, for each representation, when a higher amount of uncertainty is shown. In detail, the quiz consisted of five questions, with the first three questions being general questions about the prediction task, prediction target, and bonus structure, while the last two assessed the participants' understanding of uncertainty information presented in the Violin Plots and Question-mark representations respectively by aligning high uncertainty and low uncertainty representations side-by-side.[2]

The main task then proceeded in two stages (separated by the dashed line in Figure 2), the model performance demonstration and the prediction task. In the first stage, participants observed the model forecasts of 10 randomly selected applicants. For each applicant, participants were shown different representations of an applicant's information (see Table 1 for an example), and the applicant's actual percentile score, i.e., ground truth (illustrated in Figure 1). Showing the randomly selected cases that included the ground truth information was designed to indirectly inform participants of model performance. At the end of the first stage, participants were shown the summary of model performance, which included actual and predicted values, absolute error for each applicant, and the model's average absolute error (e.g., *"The average absolute error for the model is 8.08"*). In general, the model performance demonstration stage can be regarded as another form of training. In this stage, uncertainty information and ground truth are both available to participants to help establish the association between model performance and prediction uncertainty. Moreover, participants are expected to gauge their reliance on the model while uncertainty information is offered.

The 30 prediction cases were successively shown, with the order of experimental conditions following the randomization structure discussed in Section 4.2. For conditions that showed the model predictions (Model-prediction, Violin Plot, Question-mark), participants were shown the model prediction and then were asked to either adopt the model

[2] Quiz questions are provided in Appendix B page 3

prediction or use their own. Participants were subsequently asked to provide reasons to justify their decisions. In cases where participants decided to use their own predictions, they were asked if the model was helpful in formulating their forecasts. Finally, after each prediction case, participants subjectively rated their self-confidence in the final prediction on a 7-point Likert scale.

Participants also completed several post-task questionnaires, including task-related perception survey, trust survey for three representations that had model predictions (randomized 12 trust items proposed by Jian et al. [63]), risk propensity survey (6 items proposed by [64]), and demographic survey. The study concluded with a report showing a participant's bonus, AAE, and absolute errors for individual cases.

4.5 Statistical Analysis

Processing and analysis of data from the experiments were performed using the statistical computing software SAS® 9.4 and JMP® Pro 15. Due to the categorical nature of the majority of the response variables and the presence of restrictions on randomization and repeated measures, we used generalized linear mixed models (GLMM), a flexible family of models that accommodates both fixed and random effects and non-normally distributed responses belonging to the exponential family of distributions. The reader is referred to [59] for more information about GLMM's. A crude version of the GLMM formulated for task-level responses (i.e., responses that were analyzed per prediction task, per participant):

$$G(E(y_{ij})) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij} x_{2ij} + u_{0j} \quad (2)$$

where $G(\cdot)$ is a transformation function for the expected value of the response y [59]; β_1, β_2, and β_3 are the coefficients associated with the visual representations (x_1), uncertainty level (x_2), and the interaction between the two ($x_1 x_2$); β_0, depending on the model, could be a vector or scalar of intercepts; and, u_0, random intercepts, addressing the repeated measures structure and constrained randomization by imposing a nesting between subjects and tasks for the residuals. The full SAS® scripts are included in the supplementary materials.

When the response is scaled as binary ($0, 1$; yes, no), the resulting model is a logistic regression model [59] with correlated residuals. When the response is scaled as ordinal (e.g., Likert scale), the model is simply the proportional odds model [59] with correlated residuals. For continuous, numeric responses (e.g., AAE's), we assume Gaussian-distributed errors, resulting in a mixed effects model.

5 Results

In this section, we present results from quantitative and qualitative (free responses) data collected from participants. Unless otherwise stated, special cases of GLMMs described in Section 4.5 were used in the analysis of quantitative data. We included a summary table in Appendix A listing the details of response variables and corresponding statistical models and tests.

5.1 Quantitative Analysis

Average Absolute Error (AAE) The objective of including Table and Histogram as raw data representations was to check the reliability of final predictions and to ensure that participants were thoroughly engaged in the task. This metric also provides some insights on the effectiveness of uncertainty visualization in improving prediction performance among subjects. Type III Tests of Fixed Effects (F-tests) showed statistical significance at $\alpha = 0.05$ ($p < 0.01$) in AAE's for the interaction effect between *Visual Representation* and *Level of Model Uncertainty*, suggesting that differences in AAE's across the representations were not constant across the two levels of models uncertainty. This result was consistent for both the study group ($F(4, 1486) = 5.33$, $\text{p-value} = 0.\,0003$) and confirmatory group ($F(4, 796) = 3.59$, $\text{p-value} = 0.\,004$) in Appendix A Table II.

This result has several implications. First, it implies that there are differences in AAE's among the representation groups and that the magnitude of these differences are dependent on the level of model uncertainty. As expected, representations with model predictions had significantly lower AAE's than Table and Histogram ($d_1 = 9.82, d_2 = 9.12, p < 0.0001$)[3]; these differences were lower in magnitude for high uncertainty tasks ($d_1 = 5.00, d_2 = 4.48, p < 0.0001$) (Appendix A Table III), a consequence of the active interaction effect. Secondly, it also shows that while there are significant differences between representations with model predictions and those without, there was no significant difference between the AAE's of the two uncertainty visualizations and *Model Prediction Only* (Table 4). Results from both study group and confirmatory group are consistent and similar in magnitude.

TABLE 4
Differences of average AAEs between *Model Prediction Only* and two uncertainty visualizations combined (Uncertainty Viz).

Uncertainty Level	Diff. of Avg. AAEs	p-value
Low Uncertainty	$d_1 = 1.47$	$p = 0.14$
	$d_2 = 1.61$	$p = 0.23$
High Uncertainty	$d_1 = 1.51$	$p = 0.13$
	$d_2 = -2.09$	$p = 0.12$

d_i = *Model prediction only* − *Average of Uncertainty Viz*, where $i = 1$ is the *study group* and $i = 2$ is *confirmatory group*

Q1. Trust Calibration To determine if behaviors were aligned with the model's level of uncertainty, we compared the percentage of trials in which participants adopted the model prediction values (chose "Model prediction") with the percentage of trials in which participants provided their own predictions (chose "Own prediction"). Our expectation is that for low uncertainty tasks, uncertainty visualization would significantly drive participants to pick the model's prediction over their own; the reverse is hypothesized for high uncertainty tasks. Looking at the descriptive statistics in the graphs of Figures 3, it is evident that participants increased adoption of model predictions in the presence of low model uncertainty and uncertainty visualization. For both study and confirmatory groups, the proportion

[3] d_i: difference in average AAEs between representation groups where $i = 1$ is the *study group* and $i = 2$ is *confirmatory group*

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3251950

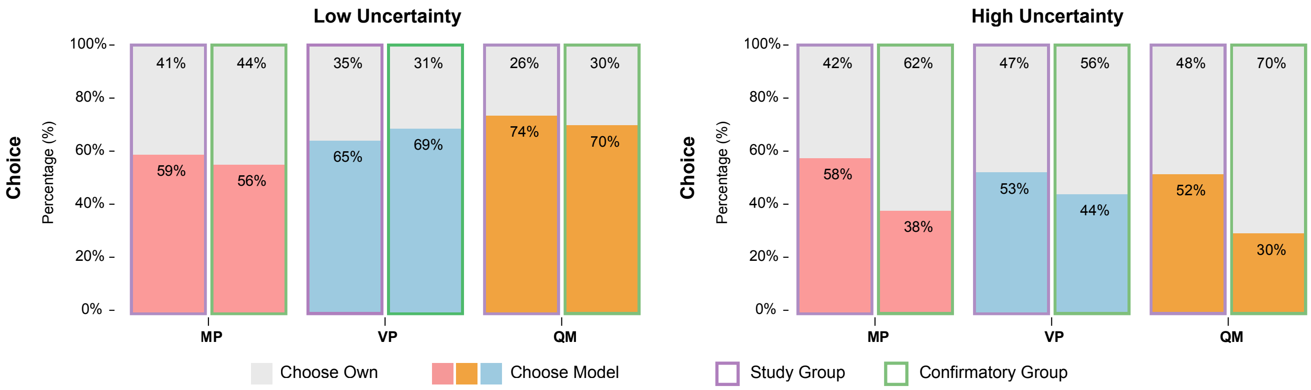


Fig. 3. The percentages of trials in which participants adopted the model prediction values for different levels of model uncertainty, with low uncertainty tasks on the left and high uncertainty tasks on the right. The study group is marked by purple borders, while the confirmatory group is marked by green borders. The colors correspond to different visual representations.

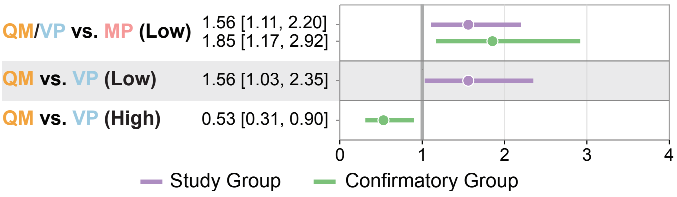


Fig. 4. 95% CI interval plots of OR between two representations for a specific uncertainty level (annotated inside parentheses). The threshold at OR equals 1.0 is highlighted.

of times participants selected model predictions was seemingly higher with uncertainty visualization compared to when only the model prediction was present. For high uncertainty tasks, appropriate reliance was observed for the study group, as participants gravitated towards their own predictions in the presence of either uncertainty visualization. But, for the confirmatory group, only the Question-mark representation increased rejection of model outputs.

Whether these signals are legitimate (and not just due to noise) remains a question. Hence, we conducted formal statistical tests by fitting a logistic regression model with correlated errors. Type III Tests of Fixed Effects yielded similar results as in analysis of AAE, where the interaction effect between *Visual Representation* and *Level of Model Uncertainty* has a significant impact (study group: $F(2, 596) = 5.20$, p-value $= 0.006$; confirmation group: $F(2, 320) = 3.59$, p-value $= 0.029$; shown in Appendix A Table IV). This provides some evidence that there are different adoption/rejection behaviors across visual representations, depending on whether the task is easy (low uncertainty) or difficult (high uncertainty).

Subsequently, **O**dds **R**atios (OR)[4] were estimated and tested ($H_0 : OR = 1$ vs. $H_1 : OR > 1$) to provide insights on the order of magnitude of effects from uncertainty visualization and level of uncertainty. Figure 4 shows the estimated odd ratios and corresponding 95% confidence intervals. For low uncertainty tasks, an odds ratio significantly greater than 1.0 implies increased model adoption when uncertainty visualization is present, while for high uncertainty tasks, an odds ratio less than 1.0 implies increased model rejection with uncertainty visualization.

Figure 4 confirms that for low uncertainty tasks, uncertainty visualization generally promotes the adoption of model predictions i.e., the chance of choosing model predictions over one's own is 1.56 (study group) or 1.85 (confirmatory) times more with uncertainty visualization than without. Both odds ratios are statistically different from 1.0. This result implies that anywhere from $60 - 65\%$ of users will adopt model predictions in the presence of uncertainty visualization when the model is more certain about a prediction, compared to only $35-40\%$ in its absence.

In the case of high model uncertainty, the estimated odds ratios were not significantly different from 1.0 for both groups of participants. A straightforward interpretation of this result is that uncertainty visualization has potentially no impact for difficult cases. However, the differences in proportions between Model-prediction and Question-mark glyph in Figure 3 suggest that there may be potential for the right type of uncertainty visualization to deter model adoption behavior when the model is less certain.

The aggregated pattern of model adoption behaviors is analyzed using OR. We further investigated individual participants' behavior over time. To do so, we applied the Bernoulli **cu**mulative **sum** (CUSUM) chart [65], which is a type of control chart designed for binary variables (e.g., model adoption = 1; model rejection = 0). Control charts with upper and lower control limits are often used to monitor the occurrence of a particular event, and data points out of control limits indicate systematic changes from the baseline rate to an alternative rate that exceeded control limits. In our case, we aim to detect if systematic changes (i.e., significant increased or decreased model adoption) exist in 18 trials of the prediction task. Table and Histogram conditions do not include model predictions, so these two conditions are excluded from the temporal analysis. Referring to the OR and control limits setting applied in [66], we used the average percentage of model adoption of all participants as the baseline rate (i.e., 56%), increasing or decreasing the chance of model adoption by 15%-25% as the upper and lower limit respectively (i.e., OR+=2, OR-=0.5). After generating Bernoulli CUSUM charts of each participant, we acquired three categories of participants :

[4] *Here, the odds ratio is defined as the odds of* Model Prediction Only *vs. the odds of* Model Prediction with Uncertainty Visualization *(averaged). The odds is defined as the chance of choosing the model prediction vs. formulating their (participants) own.*

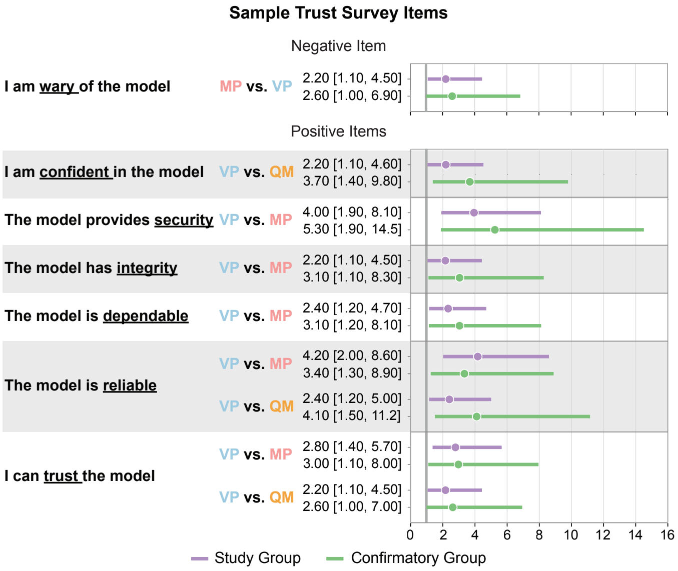


Fig. 5. A sample of subjective ratings of trust survey items (based on the scales of trust developed by Jian et al. [63]) from participants in both study group and confirmatory group. The trust survey items have significant differences consistent in both groups are included exclusively. The 95% CI interval plots show the odds ratios between two representations.

(1) do not have distinct temporal patterns over time (70 out of 77 participants), (2) have distinct temporal patterns over time (one out of 77 participants), (3) chose to adopt model prediction results for all 18 trials (six out of 77 participants). Based on the results, we can conclude that most participants had no obvious temporal dependencies, such as increasing the adoption of the model over time. The participant who exceeded the lower control limit adopted model prediction twice. To avoid lacking degrees of freedom in analyzing other ordinal variables, we decided to keep all data.

Q2. Uncertainty Viz Designs and Trust Calibration Now that it has been established that uncertainty visualization has the potential to calibrate trust as measured by reliance behaviors depending on model uncertainty, we compare the two uncertainty viz designs tested in the study. The Question-mark representation shows more propensity to affect reliance behaviors, an observation evident in Figure 4 where for the study group with low uncertainty tasks, the chance of adopting model outputs is 1.56 times more than the Violin Plot. However, this finding was not corroborated in the confirmatory group, in which the odds ratio was not significant at $\alpha = 0.05$. However, for high uncertainty tasks, the confirmatory group posted increased rejection of model outputs for the QM representation ($OR = 1/0.53 \approx 2.0$), which was not observed in the study group. Though tests of significance of odds ratios yielded inconsistent results across the two study groups, directions of estimates are congruous – QM consistently outperformed VP in terms of calibrating trust and affecting reliance behaviors.

Q3. Perceptions In addition to reliance behaviors, measures of perception toward the visual representations were collected through a post-task questionnaire. Unless otherwise mentioned, survey responses were on a 7-point Likert scale, with higher values being more desirable in this case. Due to the ordinal scale of the response data, we used a proportional odds model with *Task* as a random effect in the analysis of survey responses.

Trust After completing the forecasting tasks, participants were asked to provide their ratings on the 12-item trust questionnaire from Jian et al. [63]. To eliminate maturation and fatigue biases, the questionnaire items were presented to participants in a randomized fashion [67]. Every visual representation with model prediction (*MP*, *VP*, *QM*) was evaluated. Similar to *Q1* and *Q2*, odds ratios were estimated and tested to compare the relative impact of uncertainty visualization on trust perceptions. Figure 5 lists the questionnaire items that had *ORs*[5] which were statistically different from 1.0. Among the 7 positively worded items in the survey, only the statement *"I am familiar with the model (a regression model)"* did not post any differences. Among the negatively worded trust items, only (*"I am wary of the model"*) posted a significant difference, favoring uncertainty visualization VP over MP.

Interestingly, the Violin Plot was consistently perceived as more superior with respect to the trust questionnaire items. In comparison to the frequency based representation, VP seemed to inspire more confidence and trust in the model, as well as perceptions of increased security, integrity, dependability, and reliability by at least a factor of 2 for both the study group and confirmatory group.

Difficulty and Confidence in Prediction Participants also rated the overall level of difficulty[6] of the forecasting task when using a specific representation. Both the study group and confirmatory group revealed that the uncertainty visualizations *VP*, *QM* were easier for prediction tasks than MP by at least a factor of 2 (study group: $OR = 3.57\ [1.74,\ 7.32]$, p-value $= 0.0005$; confirmation group: $OR = 2.88\ [1.10,\ 7.54]$, p-value $= 0.0309$; shown in Appendix A Table V). However, participants did not perceive any differences in difficulty between VP and QM.

Additionally, confidence ratings on a 7-point Likert scale (the higher, the better) were also collected to gain insights about participants' level of confidence in their predictions when a specific representation was present. Results show a similar pattern where VP and QM raised perceptions of confidence in predictions over MP by a factor of at least 3 (study group: $OR = 3.29\ [1.64,\ 6.58]$, p-value $= 0.0008$; confirmation group: $OR = 3.07\ [1.19,\ 7.94]$, p-value $= 0.0201$; shown in Appendix A Table VI). Similarly, there were no difference in perceptions between VP and QM.

5.2 Qualitative Feedback

Participants from both study and confirmatory groups elaborated their reasons for selecting or rejecting model predictions at the end of each task. We combined, categorized, and analyzed the responses according to two groups, (1) reasons for choosing model predictions (Table 5) and (2) reasons for choosing their own predictions (Table 6). We looked at the top 3 reasons (based on the cumulative frequency of responses being $70-80\%$) and performed separate statistical

[5] Here, OR is defined as the ratio of the odds of being in a higher category (e.g., 5, 6, 7) over a lower category (e.g., 1, 2, 3) between two representations. For positively worded questions, $OR > 1$ suggests that the first representation is more desirable. For negatively worded questions, $OR > 1$ suggests that the second is desirable.

[6] Measured on a 5-point Likert scale (1=Very Difficult, 5=Very Easy)

TABLE 5
The categories of reasons for choosing **model** predictions (adoption).

	Reason	Pct.(%)	Accum.
1	I agree with the model prediction./The model prediction seems reasonable to me.	50.32	50.32
2	I trust the model./The model is reliable.	14.90	65.22
3	Due to the low **uncertainty**.	12.23	77.45
	My prediction is close to the model prediction and the model prediction is more accurate.	7.77	85.22
	I am not confident/certain about my prediction.	4.71	89.94
	Due to the high **uncertainty**.	1.27	91.21
	Others	8.03	99.24
	Invalid	0.76	100.00

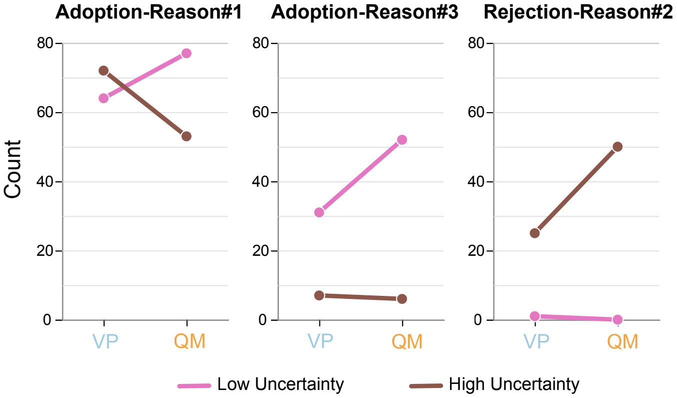


Fig. 6. Significant results of a homogeneity test between Uncertainty Level and Uncertainty Visualization Type for top three reasons of both adoption and rejection behavior (listed in Table 5 and 6 respectively). For adoption reason #1 (left), there is an active interaction effect between Uncertainty Level and Uncertainty Viz Type. For adoption reason #3 (middle), participants used "low uncertainty" to justify their adoption behavior more frequently for QM than for VP. For rejection reason #2 (right), participants used "high uncertainty" to justify their adoption behavior more frequently for QM than for VP.

analyses on the top 3 for each behavior category (adoption, rejection). The goal of the analysis is to investigate differences in the observed frequencies of a top 3 cited reason across the uncertainty visualization types (*QM*, *VP*) and uncertainty levels (low, high).

First, we performed a test of homogeneity (a.k.a, 2×2 χ^2 test of independence) between Uncertainty Level and Uncertainty Viz Type. This was done for each top 3 reason and behavior type (adoption, rejection), resulting in 6 separate and independent tests. Out of the 6, only the first cited reason for adoption ("I agree with the model prediction./The model prediction seems reasonable for me") turned out to be significant at the 5% level ($\chi^2 = 3.95$, $p < 0.05$; shown in Appendix A Table VII). This result implies that Uncertainty Level and Uncertainty Viz Type are involved in an active interaction effect, further suggesting that differences in frequency between the two uncertainty viz types are dependent on the Uncertainty Level. Unfortunately, we did not have enough degrees of freedom (i.e., samples) to explore the magnitude of this interaction effect using GLMM's. Further examination of the raw data, however, shows that for high uncertainty tasks, participants tended to cite that model predictions are reasonable with a higher frequency for VP than for QM. For tasks with low model uncertainty, the opposite is true – participants tended to cite this reason less for QM than for VP (see Figure 6 left). This presents some evidence that QM calibrates trust better than VP due to the reason stated i.e., if the model is more uncertain, QM facilitated this understanding more than VP by having a lower number of participants citing that the model prediction is reasonable.

TABLE 6
The categories of reasons for choosing **own** predictions (rejection).

	Reason	Pct.(%)	Accum.
1	I think that the model prediction value is not good enough/not realistic.	63.42	63.42
2	Due to the high **uncertainty**.	12.75	76.17
3	I partially agree with the model prediction or **uncertainty** value, but I want to use my own prediction.	7.38	83.56
	I am more confident about my own prediction.	4.36	87.92
	My prediction is close to the model prediction, but I trust myself.	0.67	88.59
	Others	11.41	100.00

For the other reasons cited, we performed GLMM with Uncertainty Level and Uncertainty Viz Type (no interaction effect due to lack of degrees of freedom to fit interaction effects). Because our response is count data (frequency), we used the Poisson distribution with a natural logarithmic transform on the mean, resulting in the Poisson regression model. Of the 5 reasons remaining, only the third reason for adoption ("Due to the low uncertainty") and second reason for rejection ("Due to the high uncertainty") resulted in statistically significant effects (Appendix A Table VIII and Table IX). What we found by investigating the raw data is that for high uncertainty tasks, there were approximately the same number of participants who cited Reason #3 for adoption. For low uncertainty tasks, however, more participants cited this reason for QM than for VP. The same trend is observed for Reason #2 for rejection.

What these results tell us is that QM seems to be better than VP at calibrating trust with a model's uncertainty level. Note that the cited reasons were acquired in verbatim. Participants seem to observe, address, and take action on a model's uncertainty level with the Question-mark glyph than the Violin Plot.

5.3 Summary of Key Results

- Uncertainty visualization does not improve overall prediction performance (Section 5.1 AAE, Table 4).
- Uncertainty visualization encourages adoption of model predictions for low uncertainty (easy) tasks. For high uncertainty tasks, the results are not conclusive (Section 5.1 Q1, Figure 3).
- As an uncertainty visualization tool, the question-mark representation shows more promise in calibrating trust based on model uncertainty. For low uncertainty tasks,

there is a higher chance that a person will adopt model predictions while for high uncertainty tasks, rejection of model outputs is more likely (Section 5.1 Q2, Figure 4).

- A violin plot is better rated as promoting trust-related perceptions of the model in measures of "confident", "reliable", and "trust", but not in measures of "security", "integrity", and "dependable" compared with the question-mark representation on the positively-worded items of the trust questionnaires (Section 5.1 Q3 Trust, Figure 5).
- Uncertainty visualization, in a forecasting task, improves ease of use of model outputs and increases confidence in predictions. However, neither uncertainty visualization (the violin plot or the question-mark representation) is superior to the other (Section 5.1 Q3 Difficulty and Confidence in Prediction, Appendix A Table V and Table VI).
- Based on the reasons participants cited for adopting or rejecting model outputs, the question-mark representation seems to align their reliance behaviors better with the model's level of uncertainty compared to the Violin Plot (Section 5.2, Figure 6).

6 Discussion and Limitations

The analysis of AAE between two uncertainty visualizations (VP and QM) shows no statistical significance, which indicates we do not have enough evidence to conclude that uncertainty visualization reliably improves overall prediction performance. This is consistent with prior work showing that individual human predictions tend to be inferior to AI/ML model predictions [31], [32], [68], [69]. Although uncertainty visualization can impact reliance on the model (adoption or rejection of the model's predictions), the ultimate prediction results as measured by the average absolute error are not enhanced. Likewise, Dietvorst et al. [13] discovered that human adjustments of model outputs could worsen the prediction results. The recruited MTurk participants could be considered to be a non-expert sample rather than domain or model experts within our particular forecasting task. With a non-expert sample, participants may be less sensitive to the magnitude of prediction errors.

There is a disagreement between potentially more effective uncertainty visualization for trust calibration (as measured by reliance behaviors) and perceived trustworthiness. The QM seems to promote better trust calibration (reliance behaviors relative to the model's uncertainty); however, the VP would improve people's self-reported trust perceptions in the model. It is interesting that QM somewhat improves trust calibration, based upon statistical analysis on choice (Section 5.1 Q2) and categorization on supporting reasons (Section 5.2), but participants would prefer VP that shows a bell curve and the prediction range. Although participants usually consider a model to be more trustworthy when provided with more details regarding uncertainty information (i.e., VP), a general ordinal representation (i.e., QM) is generally sufficient to indicate the relative amount of uncertainty in practice. This aligns with current literature that a simple ordinal representation could be more accessible to general audiences than probability distributions regarding uncertainty, as it can be challenging to interpret uncertainty information correctly [70], [71]. Perceptually, it may be demanding for the general population to build a mental map between statistical instances and the magnitude of uncertainty.

There are several limitations in our experimental design. First, we constructed the experimental prediction task referring to prediction tasks utilized by Dietvorst et al. [12], [13], which asked participants to predict the performance of MBA students in percentiles given numerical or categorical attributes. We believe that using only numeric and categorical attributes to represent a candidate's performance and asking a participant to predict the relative percentile ranking is a valid prediction task. However, it may not be identical to a practical scenario where more detailed information about a candidate is often provided in text documents, including letters of recommendation and statements of purpose.

Formulation of (ecologically and contextually) valid tasks in human-AI/ML decision support systems is, in our opinion, an open area for research i.e., no standard guidelines appear in the literature on how to address this when there are many possible tasks to choose from. In our study, we defined explicit guidelines for choosing potential instances for inclusion using statistical criteria (e.g., DFFITS and DFBETA; prediction variance). In the current experiment setting, model uncertainty does not equivalent to model accuracy. This is primarily the reason why we included a task complexity (complex for both humans and AI/ML models) experimental factor to make the study design more ecologically valid. It is true that model accuracy cannot be assessed locally in real-world scenarios, which is why the evaluation of model uncertainty is of paramount importance. Also, note that this is why we used participants from the general population in lieu of experts who would exhibit gut hunches, domain knowledge, etc. Moreover, we designed and picked the tasks intentionally so that those that are "low uncertainty" are easy cases to evaluate while the high uncertainty ones emulate perturbations or outliers (e.g., students with low GPAs but high standardized scores and with research experience). Because this is a white-box model, we understand the mechanisms that produce the predictions. Therefore, we were able to manipulate both uncertainty and accuracy because the exogenous variables are well-understood. For black-box models, we understand that this would not always be the case.

Our study adopted limited types of static uncertainty visualizations for conveying the model's uncertainty to the subject. Numerous uncertainty visualizations exist, and this study explores only two. Future work should explore the impact of uncertainty representations on model reliance within a broader scope, including variations of uncertainty visualizations and direct and indirect inform of uncertainty in order to further explore more generalities of trust calibration when presented with uncertainty visualization. In addition to exploring a wider range of uncertainty visualizations, we can study how varying model prediction performance can affect reliance behavior. We can utilize similar experimental settings (e.g., testbed, procedure, protocol) to investigate additional experimental factors, such as varying model performance, by including other AI/ML models or even possibly adding in a "nonsense" model as a more extreme control condition in our future work.

Finally, trust calibration is a complex topic. For the regression model developed, the model performed better in the low uncertainty conditions than in the high uncertainty conditions. However, low uncertainty does not always mean that the model is reliable. While a high-accuracy model implies it might be advisable for a decision maker to adopt the model recommendation under a low uncertainty condition, there are cases in which the model can still be wrong about its recommendation, in which case the decision maker should reject this recommendation. Our study does not explore this condition. However, our results indicate that participants calibrated their trust relative to the uncertainty, and such a calibration could be harmful in different model configurations. Future work is needed to further investigate model reliance and trust calibration under various specialized populations and decision domain conditions, to tease apart best practices in which visualization can be used effectively to establish appropriate trust in those cases.

7 Conclusions and Future Work

To investigate the impact of uncertainty visualization on trust and reliance on model predictions, a study was conducted on MTurk exploring two experimental factors: the type of visual representation for conveying information and the level of model uncertainty. For each prediction trial, participants were first asked to inspect a model prediction value (with or without uncertainty information) and then decide to use the model prediction value or provide their own prediction. The most obvious findings to emerge from this study are that, for low uncertainty tasks, proper visualization of model uncertainty can enhance an appropriate adoption of model predictions. This was especially true for an ordinal question-mark representation of uncertainty, which potentially led to more appropriate adoption of model predictions for low uncertainty tasks. Nevertheless, the Violin Plot which provides a statistical distribution with upper and lower limits was rated as the more satisfying and trustworthy visualization. The higher rating reflects that participants trusted the model more when Violin Plot was presented and could partially explain why participants did not reject model predictions for high uncertainty tasks with Violin Plot. Participants often used different strategies when they encountered difficult cases, exhibiting different behaviors even when the same reason was offered in the free response field of each task. For instance, “Due to high uncertainty” is a shared reason that was cited for both adoption (Table 5) and rejection (Table 6) behavior. We surmise this is due to different inner intentions of participants, which further experiments should be conducted to capture participants’ actual intentions for adoption or rejection behavior using focused interviews. Taken together, despite the ambiguity for high uncertainty tasks, these results suggest that it is promising to calibrate people’s trust and reliance on model predictions when proper uncertainty visualization is presented. Different from previous studies that looked at people’s overall reliance on black-box model outputs, in this study we provided a visualization of uncertainty and measured model adoption for each individual decision case to better understand reliance behaviors given different levels of model uncertainty.

A future study could assess the long-term effects of model reliance based on workload. With the increase of workload, people may not have adequate time and effort to inspect every case. It would be interesting to explore when people intend to rely on automated decision aids entirely in the future – and to what extent visualization play a part in that. Moreover, it would be beneficial to identify long term reliance measures and design appropriate alert systems to ping human counterparts at a proper frequency to inspect model prediction results. Besides uncertainty, additional studies could be conducted to investigate other factors influencing model reliance and human-machine trust, such as the explainability of a predictive model, people’s familiarity with the model, performance of a model, and application scenarios (low-stake vs. high-stake). The targeted users of the current study are without specializations in predictive models and domain knowledge. It would be interesting to assess if the impact of these visualizations would differ for and between expert populations.

Supplemental Materials

All supplemental materials can be found on OSF at https://osf.io/mjrh9/?view_only=d8bdea8d469841b3913df59ecec9e612, released under a CC BY 4.0 license. They include (1) CSV files containing raw data collected from participants in the study and confirmatory groups, (2) additional statistical test results in Appendix A, (3) screenshots of experiment interface in Appendix B, (4) post-task questionnaires in Appendix C, and (5) SAS code implementation of GLMM.

Acknowledgments

This work was supported by the U.S. Department of Homeland Security under Grant Award 2017-ST-061-QA0001 and 17STQAC00001-03-03, and the National Science Foundation Program on Fairness in AI in collaboration with Amazon under award No. 1939725. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

[1] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, “I trust it, but i don’t know why: Effects of implicit attitudes toward automation on trust in an automated system,” *Human Factors*, vol. 55, no. 3, pp. 520–534, Jun 2013.

[2] G. J. Stahler, J. Mennis, S. Belenko, W. N. Welsh, M. L. Hiller, and G. Zajac, “Predicting recidivism for released state prison offenders: Examining the influence of individual and neighborhood characteristics and spatial contagion on the likelihood of reincarceration,” *Criminal Justice and Behavior*, vol. 40, no. 6, pp. 690–711, Feb 2013.

[3] B. M. Muir, “Trust between humans and machines, and the design of decision aids,” *International Journal of Man-Machine Studies*, vol. 27, no. 5, pp. 527–539, Nov 1987.

[4] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human Factors*, vol. 46, no. 1, pp. 50–80, Jan 2004.

[5] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” *Human Factors*, vol. 39, no. 2, pp. 230–253, Jun 1997.

[6] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman, "A design methodology for trust cue calibration in cognitive agents," in *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, Jun 2014, pp. 251–262.

[7] R. Tomsett, A. Preece, D. Braines, F. Cerutti, S. Chakraborty, M. Srivastava, G. Pearson, and L. Kaplan, "Rapid trust calibration through interpretable and uncertainty-aware AI," *Patterns*, vol. 1, no. 4, pp. 1–9, Jul 2020.

[8] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 240–249, Jan 2016.

[9] H. Cai and Y. Lin, "Tuning trust using cognitive cues for better human-machine collaboration," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 54, no. 28, pp. 2437–2441, Sep 2010.

[10] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, no. 3, pp. 457–506, Mar 2021.

[11] E. Dimara and J. Stasko, "A critical reflection on visualization research: Where do decision making tasks hide?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1128–1138, Jan 2022.

[12] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, Feb 2015.

[13] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Management Science*, vol. 64, no. 3, pp. 1155–1170, Mar 2018.

[14] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[15] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson, "Visualizing summary statistics and uncertainty," *Computer Graphics Forum*, vol. 29, no. 3, pp. 823–832, Aug 2010.

[16] K. Brodlie, R. Allendes Osorio, and A. Lopes, "A review of uncertainty in data visualization," in *Expanding the Frontiers of Visual Analytics and Visualization*, J. Dill, R. Earnshaw, D. Kasik, J. Vince, and P. C. Wong, Eds., 2012, pp. 81–109.

[17] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, "Uncertainty displays using quantile dotplots or CDFs improve transit decision-making," in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2018, paper 144, pp. 1–12.

[18] S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, and T. H. Shaw, "Measurement of trust in automation: A narrative review and reference guide," *Frontiers in Psychology*, vol. 12, pp. 1–23, Oct 2021.

[19] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5-6, pp. 527–539, Nov 1987.

[20] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, Oct 2010.

[21] K. L. Mosier and L. J. Skitka, "Human decision makers and automated decision aids: Made for each other?" in *Automation and Human Performance: Theory and Applications*. CRC Press, 2018, pp. 201–220.

[22] K. L. Mosier, E. A. Palmer, and A. Degani, "Electronic checklists: Implications for decision making," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 36, no. 1, pp. 7–11, Oct 1992.

[23] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, "Automation bias: Decision making and performance in high-tech cockpits," in *Decision Making in Aviation*. London, UK: Routledge, 2017, pp. 271–288.

[24] T. L. Tsai, D. B. Fridsma, and G. Gatti, "Computer decision support as a source of interpretation error: The case of electrocardiograms," *Journal of the American Medical Informatics Association*, vol. 10, no. 5, pp. 478–483, Sep 2003.

[25] E. Alberdi, A. Povyakalo, L. Strigini, and P. Ayton, "Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography," *Academic Radiology*, vol. 11, no. 8, pp. 909–918, 2004.

[26] E. Alberdi, A. A. Povyakalo, L. Strigini, P. Ayton, and R. Given-Wilson, "CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, no. 1, pp. 115–122, Jun 2008.

[27] J. E. Bahner, M. F. Elepfandt, and D. Manzey, "Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, no. 19, pp. 1330–1334, Sep 2008.

[28] D. Manzey, J. Reichenbach, and L. Onnasch, "Human performance consequences of automated decision aids," *Journal of Cognitive Engineering and Decision Making*, vol. 6, no. 1, pp. 57–87, Mar 2012.

[29] C. D. Wickens, B. A. Clegg, A. Z. Vieane, and A. L. Sebok, "Complacency and automation bias in the use of imperfect automation," *Human Factors*, vol. 57, no. 5, pp. 728–739, Aug 2015.

[30] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *Proceedings of the International Conference on Healthcare Informatics*, Oct 2015, pp. 160–169.

[31] V. Lai and C. Tan, "On human predictions with explanations and predictions of machine learning models: A case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, 2019, p. 29–38.

[32] M. Jacobs, M. F. Pradier, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, and K. Z. Gajos, "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection," *Translational Psychiatry*, vol. 11, no. 1, pp. 1–9, Jun 2021.

[33] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103, Feb 2019.

[34] C. Eroglu and K. L. Croxton, "Biases in judgmental adjustments of statistical forecasts: The role of individual differences," *International Journal of Forecasting*, vol. 26, no. 1, pp. 116–133, Jan 2010.

[35] M. Lawrence, P. Goodwin, M. O'Connor, and D. Önkal, "Judgmental forecasting: A review of progress over the last 25 years," *International Journal of Forecasting*, vol. 22, no. 3, pp. 493–518, 2006.

[36] R. Fildes and P. Goodwin, "Against your better judgment? how organizations can improve their use of management judgment in forecasting," *Interfaces*, vol. 37, no. 6, pp. 570–576, Dec 2007.

[37] S. Highhouse, "Stubborn reliance on intuition and subjectivity in employee selection," *Industrial and Organizational Psychology*, vol. 1, no. 3, pp. 333–342, Sep 2008.

[38] S. I. Vrieze and W. M. Grove, "Survey on the use of clinical and mechanical prediction methods in clinical psychology," *Professional Psychology: Research and Practice*, vol. 40, no. 5, pp. 525–531, 2009.

[39] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt, "How do visual explanations foster end users' appropriate trust in machine learning?" in *Proceedings of the International Conference on Intelligent User Interfaces*, Mar 2020, pp. 189–201.

[40] G.-P. Bonneau, H.-C. Hege, C. R. Johnson, M. M. Oliveira, K. Potter, P. Rheingans, and T. Schultz, "Overview and state-of-the-art of uncertainty visualization," in *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, C. D. Hansen, M. Chen, C. R. Johnson, A. E. Kaufman, and H. Hagen, Eds., Jan 2014, vol. 37, pp. 3–27.

[41] C. D. Correa, Y.-H. Chan, and K.-L. Ma, "A framework for uncertainty-aware visual analytics," in *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, Oct 2009, pp. 51–58.

[42] G. R. Fox and G. Ülkümen, *Perspectives on Thinking, Judging, and Decision Making*. Oslo: Universitetsforlaget, 2011, ch. 1, pp. 21–35.

[43] D. Conway, F. Chen, K. Yu, J. Zhou, and R. Morris, "Misplaced trust: A bias in human-machine trust attribution – in contradiction to learning theory," in *Proceedings of the ACM Conference Extended Abstracts on Human Factors in Computing Systems*, vol. 94, no. 6, May 2016, pp. 3035–3041.

[44] S. Bella, F. Fidler, J. Williams, and G. Cumming, "Researchers misunderstand confidence intervals and standard error bars," *Psychological Methods*, vol. 10, no. 4, pp. 389–396, 2005.

[45] R. Hoekstra, R. D. Morey, J. N. Rouder, and E. J. Wagenmakers, "Robust misinterpretation of confidence intervals," *Psychonomic Bulletin and Review*, vol. 21, no. 5, pp. 1157–1164, 2014.

[46] J. Hullman, P. Resnick, and E. Adar, "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering," *PLoS ONE*, vol. 10, no. 11, pp. 1–25, Nov 2016.

[47] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, “When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2016, pp. 5092–5103.

[48] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, “Uncertainty displays using quantile dotplots or CDFs improve transit decision-making,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, Apr 2018, paper 144, pp. 1–12.

[49] L. M. K. Padilla, M. Powell, M. Kay, and J. Hullman, “Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations,” *Frontiers in Psychology*, vol. 11, pp. 579 267:1–23, Jan 2021.

[50] G. E. Newman and B. J. Scholl, “Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias,” *Psychonomic Bulletin & Review*, vol. 19, no. 4, pp. 601–607, 2012.

[51] M. Correll and M. Gleicher, “Error bars considered harmful: Exploring alternate encodings for mean and error,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2142–2151, Dec 2014.

[52] T. Gschwandtnei, M. Bögl, P. Federico, and S. Miksch, “Visual encodings of temporal uncertainty: A comparative user study,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 539–548, Jan 2016.

[53] F. Du, C. Plaisant, N. Spring, and B. Shneiderman, “EventAction: Visual analytics for temporal event sequence recommendation,” in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, Oct 2016, pp. 61–70.

[54] M. Procopio, A. Mosca, C. Scheidegger, E. Wu, and R. Chang, “Impact of cognitive biases on progressive visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 9, pp. 3093–3112, Sep 2022.

[55] A. Kale, M. Kay, and J. Hullman, “Visual reasoning strategies for effect size judgments and decisions,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 272–282, Feb 2021.

[56] S. Guo, F. Du, S. Malik, E. Koh, S. Kim, Z. Liu, D. Kim, H. Zha, and N. Cao, “Visualizing uncertainty and alternatives in event sequence predictions,” in *Proceedings of the ACM Conference on Human Factors in Computing Systems*, May 2019, paper 573, pp. 1–12.

[57] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 4th ed. Hoboken, NJ, USA: John Wiley & Sons, 2016.

[58] N. T. Stevens and C. M. Anderson-Cook, “Design and analysis of confirmation experiments,” *Journal of Quality Technology*, vol. 51, no. 2, pp. 109–124, Apr 2019.

[59] W. W. Stroup, *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC press, 2012.

[60] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[61] M. Mancenido, P. Salehi, E. Chiou, A. Mosallanezhad, A. Shah, and M. Cohen, “Challenges of data collection on mturk: A human-ai joint face-matching task,” *Proceedings of the IISE Annual Conference*, pp. 175–180, 2021.

[62] E. Peer, D. Rothschild, A. Gordon, Z. Evernden, and E. Damer, “Data quality of platforms and panels for online behavioral research,” *Behavior Research Methods*, vol. 54, no. 4, p. 1643–1662, Aug 2022.

[63] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, “Foundations for an empirically determined scale of trust in automated systems,” *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, Jun 2000.

[64] D. C. Zhang, S. Highhouse, and C. D. Nye, “Development and validation of the general risk propensity scale (grips),” *Journal of Behavioral Decision Making*, vol. 32, no. 2, pp. 152–167, Apr 2019.

[65] M. R. R. Jr. and Z. G. Stoumbos, “A cusum chart for monitoring a proportion when inspecting continuously,” *Journal of Quality Technology*, vol. 31, no. 1, pp. 87–108, Feb 1999.

[66] J. Neuburger, K. Walker, C. Sherlaw-Johnson, J. van der Meulen, and D. A. Cromwell, “Comparison of control charts for monitoring clinical performance using binary data,” *BMJ Quality & Safety*, vol. 26, no. 11, pp. 919–928, Nov 2017.

[67] R. S. Gutzwiller, E. K. Chiou, S. D. Craig, C. M. Lewis, G. J. Lematta, and C.-P. Hsiung, “Positive bias in the ‘trust in automated systems survey’? an examination of the Jian et al. (2000) scale,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 217–221, Nov 2019.

[68] A. Bussone, S. Stumpf, and D. O’Sullivan, “The role of explanations on trust and reliance in clinical decision support systems,” in *Proceedings of the International Conference on Healthcare Informatics*, Oct 2015, pp. 160–169.

[69] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, and E. Ragan, “Machine learning explanations to prevent overtrust in fake news detection,” *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 15, no. 1, pp. 421–431, May 2021.

[70] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, “In pursuit of error: A survey of uncertainty visualization evaluation,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 903–913, Jan 2019.

[71] J. Hullman, “Why authors don’t visualize uncertainty,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 130–139, Jan 2020.

Jieqiong Zhao is a postdoctoral research associate in the School of Computing and Augmented Intelligence at Arizona State University. She received her Ph.D. degree in Electrical and Computer Engineering from Purdue University in 2020, M.S. degree in Computer Science from Tufts University in 2013. Her broad research interests include visual analytics, information visualization, human-computer interaction, and applied artificial intelligence and machine learning.

Yixuan Wang is a Ph.D. candidate at Arizona State University in the School of Computing and Augmented Intelligence. She received her masters degree in Computer Systems Engineering in 2020 from Northeastern University, MA, USA. Her research interests include visual analytics, machine learning, and human-computer interaction.

Michelle Mancenido is an assistant professor at Arizona State University in the School of Mathematical and Natural Sciences. Her research focuses on the design and analysis of statistical experiments in engineering, scientific, and industrial applications. Her expertise is in optimal experimental designs, statistical modeling for chemical and mixture experiments, and sensory experiments. She is an advocate of well-designed experiments as the key to robust scientific conclusions, user-centric product design and efficient industrial processes.

Erin K. Chiou is an assistant professor of Human Systems Engineering in the Polytechnic School at Arizona State University, and director of the Automation Design Advancing People and Technology (ADAPT) Laboratory. Her research interests include human-agent interaction, trust in technology, sociotechnical systems design, and safety critical work environments. She has served on several committees for the National Academies as an expert on human factors science and human systems integration. She received her PhD in industrial and systems engineering from the University of Wisconsin-Madison.

Ross Maciejewski is a professor and director of the School of Computing and Augmented Intelligence at Arizona State University and director of the Center for Accelerating Operational Efficiency (CAOE) – a Department of Homeland Security Center of Excellence. His primary research interests are in the areas of geographical visualization and visual analytics focusing on homeland security, public health, dietary analysis, social media, criminal incident reports, and the food-energy-water nexus. He has served on the organizing committees for the IEEE Conference on Visual Analytics Science and Technology and the IEEE/VGTC EuroVis Conference, and he currently serves as the co-chair of the Visualization Executive Committee (VEC).