When Does Uncertainty Matter?: Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making

Anonymous authors
Paper under double-blind review

Abstract

As machine learning (ML) models are increasingly being employed to assist human decision makers, it becomes critical to provide these decision makers with relevant inputs which can help them decide if and how to incorporate model predictions into their decision making. For instance, communicating the uncertainty associated with model predictions could potentially be helpful in this regard. In this work, we carry out user studies to systematically assess how people with differing levels of expertise respond to different types of predictive uncertainty i.e., posterior predictive distributions with different shapes and variances, in the context of ML assisted decision making. Our results demonstrate that showing uncertainty information leads to smaller disagreements with the machine, regardless of the type of uncertainty, but that these effects are sensitive to expertise in both ML and the domain. This suggests that posterior predictive distributions can potentially serve as useful decision aids which should be used with caution and take into account the type of uncertainty distribution and the expertise of the human.

1 Introduction

As machine learning models are increasingly being deployed in critical domains such as healthcare and criminal justice, there has been a growing emphasis on the need for human interpretable models and predictions which can enable decision makers to decide if and how much to rely on these predictions. In order to interpret model predictions, the following two complimentary approaches have been proposed in literature. The first paradigm involves building inherently simpler models such as decision trees/lists/sets (Letham et al., 2015; Rudin, 2019; Lakkaraju et al., 2016b), point systems that can be memorized (Ustun & Rudin, 2016) or generalized additive models in which the impact of each feature on the model's prediction is explicitly given (Caruana et al., 2015b). However, complex models such as deep neural networks and random forests seem to achieve higher accuracy compared to these inherently simpler models in several real world settings (Ribeiro et al., 2016). Therefore, an alternate approach of constructing post hoc explanations to interpret these complex models has been proposed in literature (Ribeiro et al., 2016; Lundberg & Lee, 2017). Both simple models and post hoc explanations are attempts at creating more interpretable decisions aids.

In addition to the aforementioned simple models and explanations, other auxiliary information such as uncertainty associated with predictions could also potentially serve as useful decision aids. While there are reasons to believe that conveying information pertaining to predictive uncertainty might help decision makers in figuring out how to incorporate model predictions into their decision making, there is little empirical work that systematically explores the validity of this hypothesis (Bhatt et al., 2021). While some prior research has examined how predictive uncertainty in various contexts ranging from weather forecasting to public transit scheduling can be communicated to the public (Greis et al., 2016; Roulston et al., 2006; Fernandes et al., 2018), these works largely focus on the cases where uncertainty can be expressed using a low variance normal distribution. However, in reality, predictive uncertainties could be much more complex to reason about – e.g., posterior predictive distributions which are bimodal or normal with high variance. In such cases, interpreting predictive uncertainty and incorporating it in decision making is non-trivial, and is heavily driven by the end user's domain expertise and familiarity with topics such as probability and machine learning.

To illustrate, let us consider the following example: if a machine learning model predicts that Alice should sell her car for $\$20,000$, should she place an advertisement for that amount or rely on her background knowledge to price the car herself? If she knew the model was highly certain, the rational choice may be to defer to its prediction, but if it is highly uncertain, she may instead prefer her own. However this picture is complicated by both Alice's own level of expertise, and the form of the uncertainty associated with the prediction by the model. For example, what if Alice is an expert used car appraiser? Or what if the model is highly certain that the car will sell for either $\$17,000$ or $\$23,000$? A rational choice for Alice might look different between these cases, and it is unclear whether the human decision maker will even follow a rational strategy. Therefore, it is important to systematically study how different types of predictive uncertainty (e.g., posterior predictive distributions with different shapes and variances) and different contexts (namely decision makers with varying degrees of domain expertise and familiarity with machine learning) impact decision making under different contexts.

In this paper, we explore how decision making is impacted when decision makers (end users) are shown estimates of predictive uncertainty under several settings. More specifically, we consider posterior predictive distributions as our measure of uncertainty, and study how posterior predictive distributions with various shapes and variances impact decision making in cases where decision makers have varying degrees of expertise in machine learning and the problem domain. While the goal is to understand how uncertainty affects the decision making process as a whole, in this work, we focus on systematically measuring how closely participants agree with the predictions of the model based on the factors described above. This allows us to separate the impact of uncertainty from the quality of the machine learning model. To the best of our knowledge, this work makes one of the first attempts at systematically exploring the aforementioned research questions.

To find answers to the above questions, we conduct a user study with 95 participants where each participant is asked to predict monthly rental prices of apartments in a particular city. Participants were mainly students and researchers spanning various fields across multiple universities. Our results demonstrate: 1) Showing uncertainty leads to smaller disagreements with the machine than not showing uncertainty-regardless of the type of uncertainty. 2) Any model prediction regardless of uncertainty information has the potential to equalize differences between people with and without domain expertise. 3) People respond most to predictions with accompanying uncertainty information when that information is Gaussian and has low variance, and this effect is particularly marked for people with ML expertise. This suggests that posterior predictive distributions can potentially serve as useful decision aids which should be used with caution and take into account the type of uncertainty distribution and the expertise of the human.

2 Related Work

ML Assisted Decision Making: Machine learning is increasingly being employed to assist human decision makers. For example, Kleinberg et al. (2017) examine ML assisted decisions in the context of the the criminal justice system, and Giannini et al. (2019) studies how ML can assist in a clinical care context. Consequently, there has been a lot of work in how decision makers make use of the output of machine learning models (Lai et al., 2021). In some cases, users may trust models even when the model's predictions are inaccurate (Poursabzi-Sangdeh et al., 2021). In other cases, users will ignore a model's prediction even when it is expected to perform better than humans (Yin et al., 2019). More generally, Skitka et al. (1999) found that people made more errors when aided by highly accurate automation, suggesting the existence of an automation bias that affects people's decision making with automated aids. We explore how showing uncertainty information for a specific prediction alongside a machine learning prediction can affect the downstream decisions made by a human decision maker.

Model Interpretability: Prior research has suggested that model interpretability can be extremely helpful in ML assisted decision making (Doshi-Velez & Kim, 2017). Different classes of interpretable models have been proposed (Wang & Rudin, 2015; Zeng et al., 2017; Letham et al., 2015; Bien & Tibshirani, 2009; Lakkaraju et al., 2016a; Lou et al., 2012; Caruana et al., 2015a) and studied in the context of human decision making (Lage et al., 2019; Poursabzi-Sangdeh et al., 2021). While these simpler interpretable models are often easier for users to interact with, complex models such as deep neural networks and random forests are often shown to achieve higher accuracy (Ribeiro et al., 2016). This motivated the development of post-hoc explanation methods, including perturbation-based local explanation methods (Ribeiro et al., 2016; 2018; Lundberg & Lee, 2017; Slack et al., 2021), gradient-based local explanation methods (Selvaraju et al., 2017; Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017), global explanation methods (Bastani et al., 2017; Lakkaraju et al., 2019), and other methods (Koh & Liang, 2017).

However, these post-hoc techniques have been shown to have flaws. Rudin (2019) argued that post hoc explanations are not reliable, as these explanations are not necessarily faithful to the underlying models and present correlations rather than information about the original computation. There has also been recent work on exploring vulnerabilities of black box explanations (Adebayo et al., 2018; Slack et al., 2020; Lakkaraju & Bastani, 2020b; Rudin, 2019; Dombrowski et al., 2019). Moreover, there is a growing literature on evaluating the effectiveness of post-hoc explanations for ML assisted decision making (Doshi-Velez & Kim, 2017; Kaur et al., 2020; Bhatt et al., 2020; Hong et al., 2020; Lakkaraju & Bastani, 2020a; Poursabzi-Sangdeh et al., 2021; Buçinca et al., 2020; Krishna et al., 2022), which has highlighted a number of challenges and subtleties. For instance, while prior studies have shown that decision makers can perform better when provided with ML predictions and corresponding explanations in various tasks such as medical diagnosis (Cai et al., 2019; Lundberg et al., 2018), data annotation (Schmidt & Biessmann, 2019), and deception detection (Lai & Tan, 2019), Bansal et al. (2021) argued that such improvements may be due to decision makers simply following the recommendation of the highly accurate ML models rather than placing so-called appropriate reliance on the models (Lee & See, 2004). Indeed, Bansal et al. (2021) found that decision makers were more likely to follow the model's prediction when provided with explanations, regardless of the model's correctness.

We focus on studying the impact of predictive uncertainty on user behavior since it is often more easily attainable than a faithful and interpretable explanation. Based on insights from the explainable AI literature (Buçinca et al., 2020), we focus on the impact of predictive uncertainty on user behavior in an actual decision-making task rather on than proxy tasks or subjective measures such as user trust or preference.

Predictive Uncertainty as Auxiliary Input: Bhatt et al. (2021) outline key considerations for conveying uncertainty for ML assisted decision making. Many works have empirically studied how to best communicate uncertainty to people making decisions ranging from reacting to weather forecasts to catching public transportation (Correll & Gleicher, 2014; Kay et al., 2016; Greis et al., 2016; Roulston et al., 2006; Fernandes et al., 2018; Kirschenbaum et al., 2014; Leffrang & Müller, 2021; Koval & Jansen, 2022). For example, Greis et al. (2016) finds that people earned more when shown uncertainty estimates in a farming game that depends on weather forecasts. They also found that the performance was best when showing a simpler uncertainty visualization than the full distribution. Roulston et al. (2006) finds that showing participants error bars in addition to a point estimate can help them make a larger profit in a game where they must decide when to salt roads based on a snow forecast. They also find, however, that showing explicit probabilities does not lead to further improvements. Fernandes et al. (2018) explores the question of which types of uncertainty visualizations help users make the 'best' decisions according to a specified payoff function when catching a bus. They explore different uncertainty visualizations and find that all of them allow people to improve over time, and their preferred method–quantile dotplots, resulted in the most consistent performance. Kirschenbaum et al. (2014) studied the impact of spatial vs. tabular uncertainty representation on submarine detection and found that spatial uncertainty raised the performance of non-experts almost to that of experts. Motivated in part by a preprint of our work, Leffrang & Müller (2021) studied the impact of showing 95% confidence interval plots and ensemble displays in time series forecasts of the number of hospital beds occupied by COVID-19 patients. They found that users were less willing to follow the model's prediction when shown more salient visualizations of uncertainty, and called for further studies evaluating the impact of conveying uncertainty visualizations. Our work differs from previous research in that we systematically study how decision makers are impacted when shown posterior predictive distributions under several settings. We also consider expert decision makers–unlike most of these works–who may be able to work with more complex information about the probability distribution (Spiegelhalter et al., 2011).

Finally, some work has been conducted on how people make decisions with machine learning models that present uncertainty information (Zhou et al., 2017; Arshad et al., 2015). For instance, Zhou et al. (2017)

conducted a user study to evaluate how uncertainty intervals around a point estimate affects users' predictive decision making. They found that showing users such uncertainty intervals increased trust, but only under conditions with low cognitive load. We study the impact of showing uncertainty information on downstream decisions, rather than relying on trust as a proxy for team performance. We also explore different features of the uncertainty distribution that may affect human decisions.

3 Goals and study design

3.1 Goals

In this project, we seek to understand how people make decisions when provided with the uncertainty of a machine learning model as an input to their decision process. The uncertainty of a machine learning model is auxiliary information provided by many machine learning models that can be used as an input to human decision making in addition to the model's prediction. Specifically, we explore how different factors related to the model's uncertainty distribution impact ML-assisted decision making. We measure this through user agreement with the model's prediction before and after being shown the uncertainty information, which allows us to isolate the important factor of agreement with the machine prediction from whether the prediction is correct. We also study how ML-assisted decision making differs depending on user expertise in the domain or in machine learning. We describe each of these factors in more detail below.

We operationalize the notion of predictive uncertainty in a Bayesian setting where predictions may be given as point estimates or as entire posterior predictive distributions. We show uncertainty information as a visualization of the model's full posterior predictive distribution, in addition to its prediction marked at the mean of the posterior predictive.

We study the effect of 2 factors related to the uncertainty distribution: the shape of the uncertainty distribution, and its variance. Different shapes of distributions have different properties that may both affect how much people are willing to take them into account when making decisions, and what the optimal way to take them into account may be. We study 3 shapes of posterior predictives: a normal distribution, a skewed distribution, and a bimodal distribution. A normal distribution is commonly studied in past work (e.g. Fernandes et al. (2018)), and has the property that its mean is equal to its mode. This is not true however for the bimodal distribution, which has 2 modes and very little probability on its mean, or the skewed distribution which has a mode to one side of the mean and a long tail of reasonably high probability outcomes. Whether people will still update their estimates to reflect the machine learning model's prediction when presented with these distributions with more complicated properties is an open question.

We also explore the effect of variance on people's behavior. We wish to answer the question of how much uncertainty affects people's decisions when that uncertainty is larger and encompasses more plausible predictions vs. when it is smaller and more targeted. Are people more likely to agree with a machine learning prediction when the uncertainty around the prediction is low? That is, do they correctly interpret and respond to uncertainty? And how does this compare to only showing a point estimate? We study this in the context of the normal distribution.

Moreover, we explore how decision differs between experts (in either the domain or in machine learning) and non-experts. Will people with differing levels of machine learning expertise feel more or less compelled to follow the machine's prediction regardless of the provided uncertainty? Will they respond more or less strongly to the presented uncertainty estimates. Orthogonally, will people who know more about the domain we study feel more confident in their prediction and therefore place comparatively less weight on the machine's prediction? We study the relationship between these notions of expertise and decision making with various forms of uncertainty.

Finally, we study the correlation between subjective measures of trust and perceived usefulness of the machine learning model and participant agreement with the model. In summary, our main research questions are:

- Does showing predictive uncertainty affect how closely people follow model predictions?

- Does the effect of showing predictive uncertainty depend on the type of uncertainty–either the shape or the variance of the distribution?
- Do participant with more expertise, either in the domain or in machine learning, more closely follow model predictions?

3.2 Study design

In this section, we describe the details of the user study that we carried out to answer the research questions outlined in the previous section.

3.2.1 Domain/Model

Our user study was based on predicting the monthly rental prices of apartments in City A[1]. Participants were provided with the number of square feet and number of bedrooms of each apartment.

To set realistic apartment prices, we collected details such as the monthly rental listing price, square footage, and number of bedrooms of 75 apartments in City A from Zillow.com in November 2019. A linear regression model was fit to this data to estimate the monthly rental listing price using the square footage and number of bedrooms as features. This model was then used to set the point estimates for each of the apartments in the study.

The posteriors were not generated from the model, but were instead hand generated to embody the properties we aim to study. These are described in more detail below.

All participants in the study were shown the same 10 hypothetical apartments. Specifically, the number of bedrooms and square feet for the 10 hypothetical apartments were respectively (1, 500), (1, 800), (2, 500), (2, 800), (2, 1100), (3, 800), (3, 1100), (3, 1400), (4, 1400), and (4, 1700). Three of these (namely, (2, 500), (3, 1100), (4, 1700)) were used as practice examples for participants to familiarize themselves with the study (see Section 3.2.3 for more details).

3.2.2 Conditions

We consider 5 conditions in our study each of which corresponds to different properties of the posterior predictive distribution shown to the participant:

No Uncertainty: Participants were only given access to the model's point estimate of the predicted rental price for each apartment.

Normal, Low Variance: Participants were shown a normally distributed posterior with a standard deviation (SD) of 80 for each apartment.

Normal, High Variance: Participants were shown a normally distributed posterior with a SD of 250 for each apartment.

Skew: Participants were shown a right-skewed posterior − generated from shifted and scaled beta distribution − with a SD of 250 for each apartment.

Bimodal: Participants were shown bimodal posterior − generated from a mixture of two normal distributions with equal weights − with a SD of 250 for each apartment.

Participants were randomly assigned with equal probability to one of the 5 conditions. The same set of apartments was shown in each condition, although the order of the apartments was randomized. The mean of the distribution for each apartment was held fixed across conditions, and the distribution that the participant had been randomly assigned to was scaled to have that mean. The mean was marked on the visualization as the model's prediction, and it corresponds to the point prediction given in the no uncertainty case.

[1] City name will be revealed at the time of publication

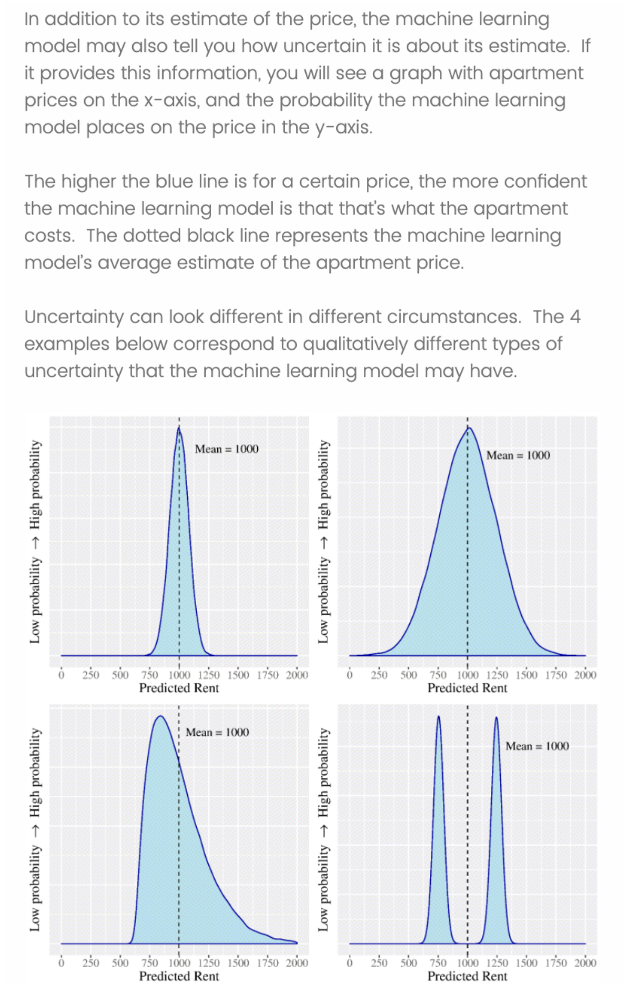

Figure 1: The uncertainty tutorial page given to participants before completing the task. The tutorial shows 4 examples of uncertainty distributions that correspond to the 4 types of uncertainty shown in different conditions–this same page is shown regardless of the condition the participant is randomized to. The text tells participants how to interpret the x and y axis, and the dotted black line marking the mean of each distribution.

3.2.3 Procedure

At the beginning of the study, all participants were trained on how to interpret posterior distributions with a tutorial page (See Figure 1). This shows examples of different distributions of uncertainty (corresponding to the 4 conditions where participants are shown uncertainty in our experiment), and describes how to read the figures showing the uncertainty distribution. Participants are shown the same tutorial regardless of which condition they are randomized to.

To allow participants to calibrate their estimates of apartment rental prices and become familiar with the performance of the model, all participants were then taken through a practice run. Each participant was shown 3 example apartments and asked to predict their prices. They were then shown the true rental price of each of those 3 apartments. Of these apartments, the model had overestimated the true rental price in at least 1 case, and underestimated the true rental price in one case, so the participant could not easily decide that the model always over or underestimated the true value.

After finishing this practice run, participants then completed the following task for 7 new apartments. For each apartment, the participant was asked for two estimates of rental price. Specifically, the participant was first presented with the square footage and number of rooms corresponding to each apartment, and then asked to make a prediction about the monthly rental cost of the apartment. Then, the participant was presented with the prediction of the machine learning model, and the corresponding posterior predictive distribution wherever applicable. They were then asked to update their original prediction based on this information. They were reminded of their original prediction when completing this part. This design allowed us to quantify the impact of showing participants the model output by the difference between the participant's first and second estimates; This was suggested as good practice by prior work (Poursabzi-Sangdeh et al., 2021). Participants were not given feedback on whether or not they were correct after any of these trials.

At the end of the survey, participants were asked to rate their trust in and the usefulness of the model using a 5-point Likert scale (Likert, 1932). They were also asked to rate their familiarity with machine learning by choosing one of the following options – "No understanding or little understanding", "Worked with it a few times", "Worked with it many times or on a larger project". Participants were also asked if they ever lived in off-campus housing in City A or surrounding areas. The response to this question conveys participants domain expertise in City A rental listings, since people who have lived off campus in the City A area are more likely to be familiar with apartment rental prices in the area than those who have not. Finally, we asked participants whether they had a background in probability on the following scale – "No understanding or little understanding", "Some understanding", and "Strong understanding". This allowed us to establish that all participants had a minimum level of probability expertise to understand the uncertainty information (all but one of the participants said they had some or strong understanding of probability).

Participants were paid \$3 for completing the study and could earn up to an additional \$30 based on their performance in the study. Specifically, for each participant, we computed the average distance between each apartment's true price (i.e. the model's prediction) and the participant's first and second estimates. 3 participants with the lowest average distance received the \$30 bonus.

3.2.4 Interface

The interface (see Figure 2 for an example depicting the **Normal, High Variance** condition) contains three important elements to aid participants with the study. First, it contains the model's estimate of the house price in bold (A). Second, if a participant was randomly assigned to an uncertainty group, the interface shows the corresponding uncertainty distribution in color (B). Third, participants are reminded of their previous estimate in bold so they may choose to condition on it (C). It also contains the 2 features of the apartment used in the model (D).

3.2.5 Participants

We recruited 95 participants for our user study[2] during July–September 2020. Participants were mainly students and researchers spanning various fields (machine learning, biology, physical sciences, history etc.) across multiple universities. We recruited these participants by advertising in multiple mailing lists and reaching out to various faculty across multiple universities.

3.3 Analysis

We look at 3 metrics to understand the extent to which seeing predictive uncertainty affects people's decisions: the distance (in the L_1 norm) between the first estimate (before seeing any model outputs) and the second estimate (after seeing the prediction and uncertainty wherever applicable)–we call this the "update"; the distance between the second estimate (after seeing the model's prediction) and the model's prediction–we call this the "final disagreement"; and the distance between the first estimate (before seeing the model's prediction) and the model's prediction–we call this the "initial disagreement". The update allows us to determine the extent to which participants updated their estimate based on the model's prediction and uncertainty information. The final disagreement measures the extent to which people agree with what

[2]This study was approved by our institution's IRB.

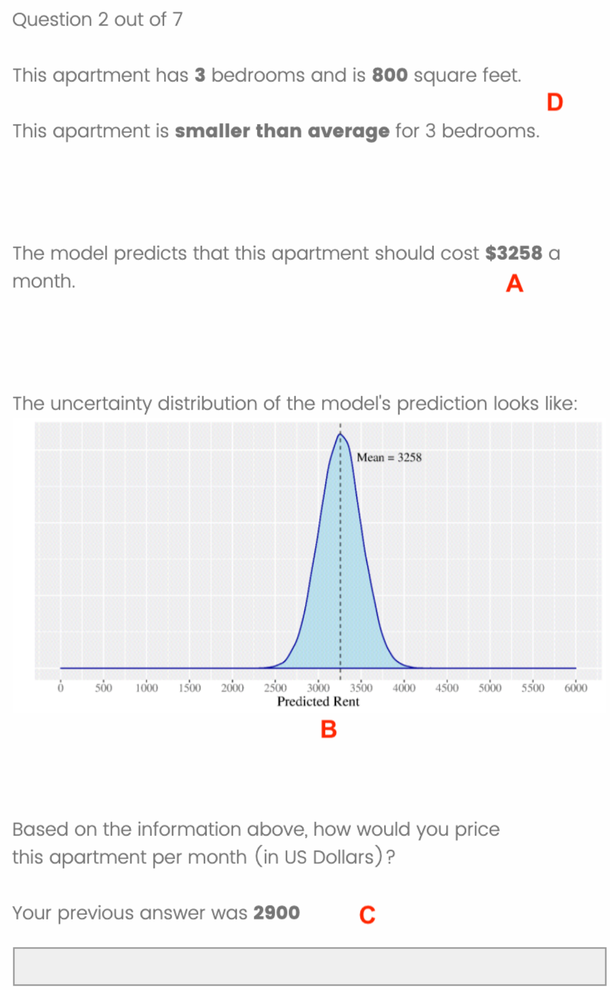

Figure 2: Example of the interface for a participant in the Normal, High Variance group after providing an initial estimate of the monthly rent of the apartment. A shows the model's prediction, B shows the visualization of the model's prediction uncertainty, C shows the participant's first response so they can reference it, and D shows the features used in the model's prediction.

the model predicted. The initial disagreement allows us to understand any baseline differences between participants of varying levels of expertise before they were shown the model's prediction and uncertainty information.

Our primary analyses use mixed-effects regression models for these three metrics. Unlike regression models with only fixed effects, mixed effects models allow one to properly account for correlation between repeated measurements from participants in the study (e.g., some participants may consistently change their estimates a large amount while others may consistently change their estimates by a small amount) (Bates et al., 2015). A random intercept was included for the participant ID in all models. All the relevant factors including the type of predictive uncertainty, background in ML, experience living in City A, self-reported usefulness and self-reported trust in the model were included as fixed effects in their respective analyses. All p-values reported are derived from a one-way ANOVA based on these models. We report p-values that are less than 0.05 as *significant* and p-values in the interval $[0.05, 0.10]$ as *marginal.* The bar plots illustrate the sample mean of the outcome in the relevant group of participants ± 1 standard error based on the fitted model.

4 Results & Insights

In this section, we describe in detail the results we observe. Specifically, we analyze the data along 2 dimensions: (1) how decision making differed between experts (either in the domain or machine learning) and nonexperts, and (2) how decision making differed when shown different types of model uncertainty. We describe each of these in turn below, then run a qualitative analysis of the effect of model uncertainty on decision making stratified by participant expertise. In Appendix B, we analyze the association between subjective metrics–confidence and trust, and how users update their estimates to agree with the machine prediction.

4.1 Comparing decision making between experts and non-experts

Our first set of research questions are about how people's willingness to update their estimates differed based on their familiarity with either the domain (apartment rentals in City A) or machine learning models. Table 1 captures the details of ML expertise and domain expertise of all the study participants. Note that we collected similar information about probability to verify that participants had the background to understand the uncertainty types, and confirmed that all but one participant had at least some understanding of probability. Figure 3 illustrates the magnitude of the update (left), final disagreement (middle), and initial disagreement (right) among participants in each of the two categories of expertise in City A apartment prices (top) and each of the three categories of ML expertise (bottom). For clarity, we relabelled the category of 'No understanding or little understanding" to "No ML background / Weak ML background", relabelled "Worked with it a few times" to "Some ML background", and relabelled "Worked with it many times or on a larger project" to "Strong ML background" in the figure.

Table 1: Number of participants with each level of expertise for the domain, and machine learning. We have a reasonable number of participants with each level of expertise.

	N. participants (%)
Experience living in City A	
Yes	67/95 (70.53%)
No	28/95 (29.47%)
Experience with ML	
Worked with it many times or on a larger project	49/95 (51.58%)
Worked with it a few times	33/95 (34.74%)
No understanding or little understanding	13/95 (13.68%)

Participants with domain expertise in City A apartment prices had lower initial disagreement and smaller updates compared to those without such expertise. However, the final disagreements of both groups were similar. Participants who have previously lived in or around City A made initial estimates that were closer to the model's prediction (smaller initial disagreement) even before having seen the prediction compared to those who never lived in City A (significant: $F(1, 93) = 4.8390$, $p = 0.0303$). This likely reflects their increased familiarity of rental prices in City A. However, we also see that domain experts have smaller updates (significant: $F(1, 93) = 8.3402$, $p = 0.0048$), and there was no significant difference between domain experts and non-experts in their final disagreement with the model ($F(1, 93) = 0.15094$, $p = 0.6985$). This suggests that the machine learning model equalizes differences in the initial disagreement between experts and non-experts by bringing participants' final disagreement with the model into the same range, regardless of expertise.

Participants with varying degrees of machine learning expertise updated their estimates differently after seeing the model's prediction, but this relationship was non-monotonic (marginally significant: $F(2, 92) = 2.4434$, $p = 0.0925$). Interestingly, this trend suggests that people who have worked with machine learning a few times update their estimates the most, while people who have no or basic understanding of machine learning do so the least, with experts having worked with ML many times falling

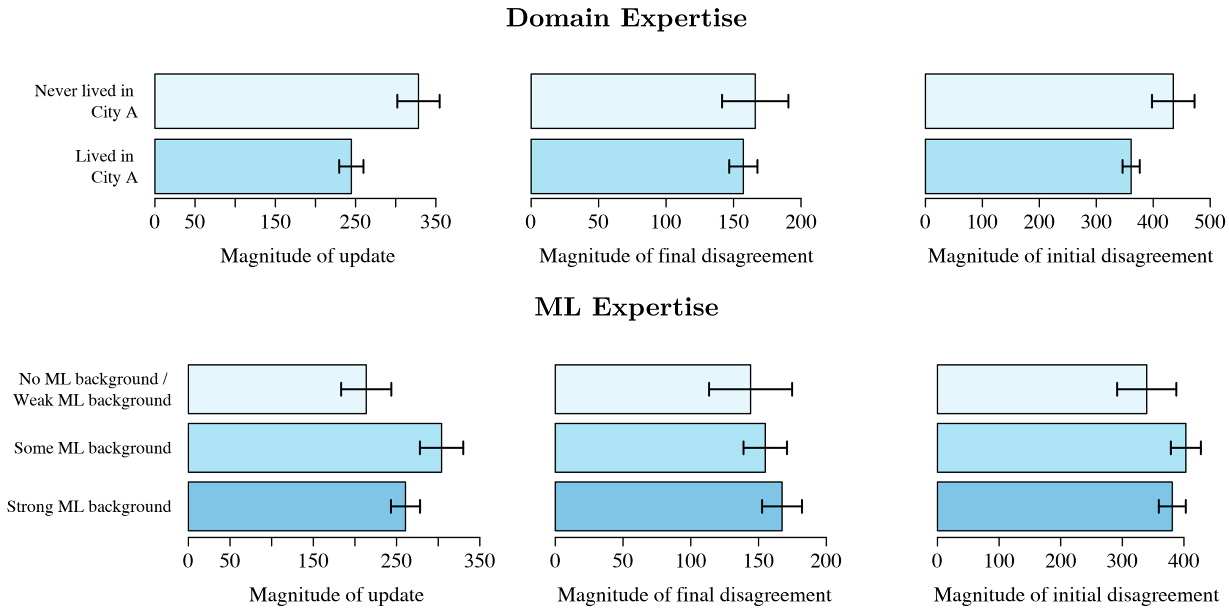

Figure 3: Magnitude of update (left), final disagreement (center) and initial disagreement (right) stratified by expertise in domain (top) and ML (bottom). Bar width corresponds to the mean, and standard errors are shown in black. Those without domain expertise have lower initial disagreement, and update their predictions more, but they have similar final disagreement to those with domain expertise. There appears to be a non-monotonic relationship between ML expertise and the size of the update.

somewhere in between. Similar to the analysis comparing domain experts in City A apartment prices to non-experts, the final disagreement did not significantly differ between the groups defined by level of ML expertise, nor did the initial disagreement.

4.2 Impact of type of uncertainty on decision making

Our second set of research questions are about the extent to which showing uncertainty influences the decisions people make, and how the type of uncertainty modulates that effect. To this end, we plot the magnitude of the update (left), final disagreement (middle), and initial disagreement (right) for each of the five conditions corresponding to a different kind of uncertainty or no uncertainty (Figure 4).

The type of uncertainty affects participants' final disagreement with the model. Furthermore, showing no uncertainty results in final estimates that are farthest from model predictions. The type of uncertainty had an effect on how closely participants' second estimate agreed with the machine's prediction (size of the final disagreement) (marginally significant: $F(4, 90) = 2.23436$, $p = 0.0715$). It can be seen from Figure 4 that the condition where participants were shown a normal distribution with low variance moved people closest to the model's prediction, however even in cases like the normal distribution with high variance where the model is clearly uncertain, or the bimodal distribution where there is very little probability mass on the model's prediction, people are still moving their second estimates at least as close to the model's prediction as when they are shown no uncertainty. We do not see statistically significant differences in the magnitude of the update ($F(4, 90) = 1.6571$, $p = 0.1669$). This could be tied to high variance in the original estimate before being shown the prediction, although these are also not statistically different ($F(4, 90) = 1.2553$, $p = 0.2935$), suggesting that our randomization worked.

4.3 Impact of type of uncertainty on decision making stratified by expertise

Finally, we present a qualitative analysis of the impact of uncertainty type stratified by participant expertise. The impact of the type of uncertainty on agreement with the prediction may be modulated by expertise.

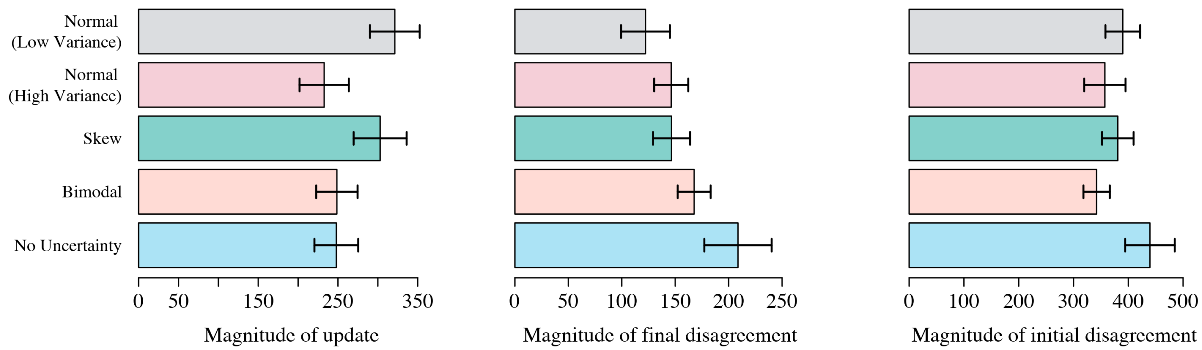

Figure 4: Magnitude of update (left), final disagreement (center) and initial disagreement (right) stratified by uncertainty condition. Bar width corresponds to the mean, and standard errors are shown in black. The final disagreement differs between conditions, with highest agreement for normal-low variance, and lowest agreement for no uncertainty.

Table 2 shows the number of participants in each uncertainty group, stratified by different levels of domain and ML expertise. While the sample sizes are small to draw strong conclusions, this analysis helps us gain further context on how expertise and uncertainty type interact, and suggests areas for future research. Figure 5 illustrates the magnitude of the update (left), final disagreement (middle), and initial disagreement (right) for each uncertainty type/no uncertainty among participants in each of the two categories of domain expertise. Figure 6 shows analogous results w.r.t. ML expertise. In this figure we show participants with high ML expertise and some expertise. Figure 7 (see appendix) shows results with participants having no ML expertise, but the sample size for this is quite small.

Table 2: Number of participants in each uncertainty group, stratified by expertise for the domain and for machine learning. Most groups except the No / Weak-ML-expertise group (below dashed line) have more than 5 participants.

Expertise	No Uncertainty	Normal, Low Variance	Normal, High Variance	Skew	Bimodal
Domain					
Yes	15	15	13	11	13
No	6	3	6	7	6
ML					
Strong	15	8	8	10	8
Some	5	7	7	8	6
No / Weak	1	3	4	0	5

Uncertainty type by domain expertise: The impact of the type of uncertainty (or none) on the magnitude of the update is much higher for people without domain expertise, but relative trends are similar for the two groups. For the normal low-variance, skew and no uncertainty conditions, the average magnitude of the update (column one in Figure 5) is on average 107 larger for people without domain expertise than domain experts. This difference is not as marked for the normal high-variance and bimodal conditions. However, the relative ordering of the different conditions is relatively consistent between the 2 groups with normal low-variance, then the skew conditions resulting in the largest updates for both groups. This is likely due to domain experts' lower initial disagreement, as noted in Section 4.1.

Uncertainty type by domain expertise: Participants in the no uncertainty condition have largest final disagreement, but this may be driven by larger initial disagreement. The final difference between the participants' estimates and the model's predictions are relatively consistent between the different types of uncertainty for both groups, with the normal low-variance condition appearing to result

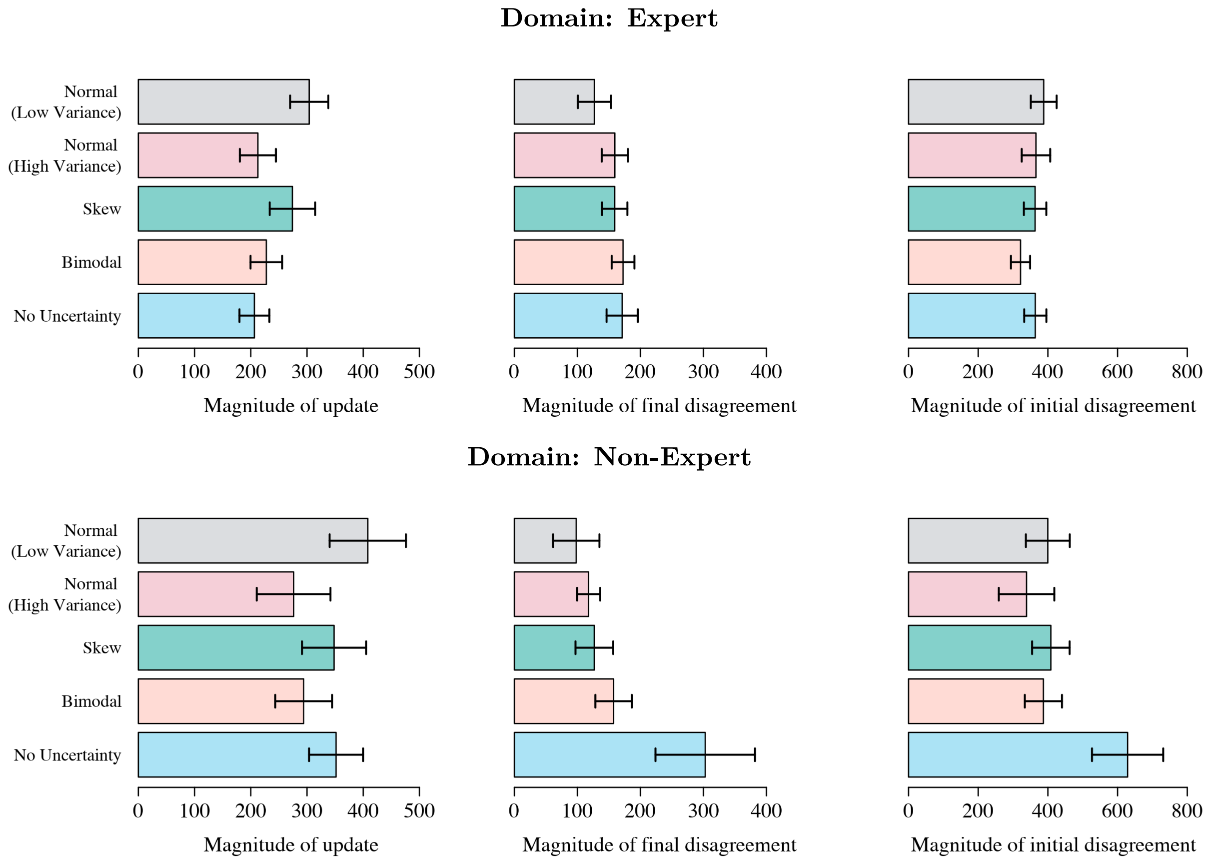

Figure 5: Magnitude of update (left), final disagreement (center) and initial disagreement (right) stratified by uncertainty condition and by domain experts (top) and non-experts (bottom). Bar width corresponds to the mean, and standard errors are shown in black. Trends in the magnitude of update across uncertainty types are similar for domain experts and non-experts, however the magnitude is quite different; non-experts have much larger updates across conditions. Non experts have final estimates much further from the model than experts when shown no uncertainty, however this appears to be largely driven by initial estimates.

in the smallest difference between the final estimate and the model's prediction. However for non-domain-experts, the final disagreement is larger in the no uncertainty condition than for any of the uncertainty conditions. This is not true for domain experts. The trend appears largely driven by initial estimates that are much further from the model's prediction for participants randomized to this condition. Whether this is the result of the small sample sizes in this qualitative analysis, or whether the different types of uncertainty have different impacts on learning effects across apartments is an interesting open question (note that participants were not shown correct answers at any point during the main experiment).

Uncertainty type by ML expertise: ML experts update their original estimates more with the normal low variance condition than for any other, while there are no clear trends for non-ML-experts. In the left column of Figure 6, the normal low variance condition has much larger updates for experts (top) than those with only some ML experience (bottom). For the experts, the updates appear much larger for this condition than for any of the others, while the updates look quite similar for all of the other uncertainty conditions and the no uncertainty condition. Perhaps ML experts are used to working with normal distributions and trust the prediction when the ML model is relatively certain, but do not agree to the same extent ML models that do not present their uncertainty, or present uncertainty estimates that have high variance and possibly hard-to-interpret distributions. For the non-experts, the trends are not as clear, and the updates appear relatively similar for the different uncertainty conditions. Perhaps participants with less experience with ML do not consider the variance of the distribution as much and have less of a prior on

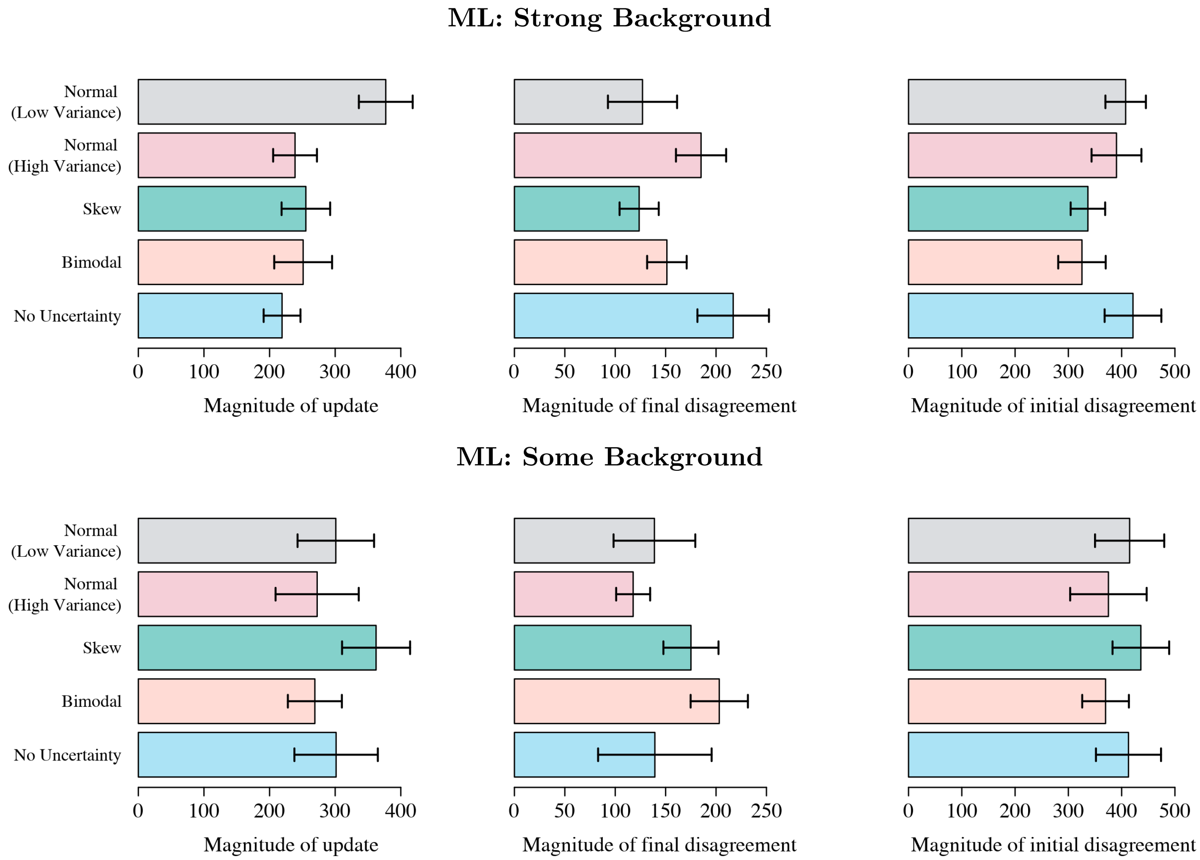

Figure 6: Magnitude of update (left), final disagreement (center) and initial disagreement (right) stratified by uncertainty condition and by background in ML. Bar width corresponds to the mean, and standard errors are shown in black. See Figure 1 in the appendix for no / weak ML background; sample sizes in this condition were much smaller. ML experts appear to update their estimates the most when shown the normal low-variance uncertainty, while all other types of uncertainty have a similar effect to not showing any at all. There are no comparable marked differences for the non-experts.

which uncertainty displays from the model are reliable. Further exploring this hypothesis is interesting future work, as it suggests that non-ML experts are perhaps less likely to respond appropriately to uncertainty.

5 Discussion and Conclusions

We studied how conveying predictive uncertainty to end users impacts decisions in the context of ML assisted decision making. In particular, we explored the extent to which different properties of the posterior predictive distribution affected participant agreement with the ML model's prediction. We also explored how participant agreement with the ML model's prediction differed between experts and nonexperts. Better understanding when humans agree with machine predictions, and how human expertise and machine uncertainty influence this agreement, lays the foundation for successful human-ai teamwork.

We also presented new insights about how uncertainty can affect ML assisted decisions. Most interestingly, showing posterior predictive distributions (regardless of the shape and variance of the distribution) increased participant agreement with model's predictions compared to showing no uncertainty estimates at all. However we also found that people respond most to predictions with accompanying uncertainty information when that information is Gaussian and has low variance, and this effect is particularly marked for people with

ML expertise. This suggests that the level of human expertise modulates the effect of showing posterior predictive distributions, and therefore should be taken into account.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Syed Z Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the annual meeting of the Australian special interest group for computer human interaction*, pp. 352–360, 2015.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via model extraction. *arXiv:1706.09773*, 2017.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48, 2015. ISSN 1548-7660. doi: 10.18637/jss.v067.i01. URL `https://www.jstatsoft.org/v067/i01`.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 2020.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 401–413, 2021.

Jacob Bien and Robert Tibshirani. Classification by set cover: The prototype vector machine. *arXiv:0908.2284*, 2009.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces*, pp. 454–464, 2020.

Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–24, 2019.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015a.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. ACM, 2015b.

Michael Correll and Michael Gleicher. Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE transactions on visualization and computer graphics*, 20(12):2142–2151, 2014.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *arXiv:1906.07983*, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3173718. URL `https://doi.org/10.1145/3173574.3173718`.

Heather M. Giannini, Jennifer C. Ginestra, Corey Chivers, Michael Draugelis, Asaf Hanish, William D. Schweickert, Barry D. Fuchs, Laurie Meadows, Michael Lynch, Patrick J. Donnelly, Kimberly Pavan, Neil O. Fishman, C. William III Hanson, and Craig A. Umscheid. A machine learning algorithm to predict severe sepsis and septic shock: Development, implementation, and impact on clinical practice*. *Critical Care Medicine*, 47(11), 2019. URL `https://journals.lww.com/ccmjournal/Fulltext/2019/11000/A_Machine_Learning_Algorithm_to_Predict_Severe.2.aspx`.

Miriam Greis, Passant El. Agroudy, Hendrik Schuff, Tonja Machulla, and Albrecht Schmidt. Decision-making under uncertainty: How the amount of presented uncertainty influences user behavior. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450347631. doi: 10.1145/2971485.2971535. URL `https://doi-org.ezp-prod1.hul.harvard.edu/10.1145/2971485.2971535`.

Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1): 1–26, 2020.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pp. 5092–5103, 2016.

Susan S. Kirschenbaum, J. Gregory Trafton, Christian D. Schunn, and Susan B. Trickett. Visualizing uncertainty: The impact on performance. *Human Factors*, 56(3):509–520, 2014. doi: 10.1177/0018720813498093. URL `https://doi.org/10.1177/0018720813498093`. PMID: 24930172.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293, 08 2017. ISSN 0033-5533. doi: 10.1093/qje/qjx032. URL `https://doi.org/10.1093/qje/qjx032`.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1885–1894, 2017.

Morgane Koval and Yvonne Jansen. Do you see what you mean? using predictive visualizations to reduce optimism in duration estimates. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*, HCOMP '19, New York, NY, USA, 2019. AAAI Press.

Vivian Lai and Chenhao 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*, pp. 29–38, 2019.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020a.

Himabindu Lakkaraju and Osbert Bastani. “How do I fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the 3rd AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020b.

Himabindu Lakkaraju, Stephen Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684, 2016a.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684, 2016b.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2nd AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, pp. 131–138, 2019.

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

Dirk Leffrang and Oliver Müller. Should i follow this model? the effect of uncertainty visualization on the acceptance of time series forecasts. In *2021 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 20–26. IEEE, 2021.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 2015.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–158, 2012.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774. 2017.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–52, 2021.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 1527–1535, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, pp. 1135–1144, 2016.

Mark S. Roulston, Gary E. Bolton, Andrew N. Kleit, and Addison L. Sears-Collins. A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting*, 21(1): 116–122, 2006. doi: 10.1175/WAF887.1. URL `https://doi.org/10.1175/WAF887.1`.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206, 2019.

Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.

Ramprasaath Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

LINDA J. Skitka, KATHLEEN L. Mosier, and MARK Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991 – 1006, 1999. ISSN 1071-5819. doi: https://doi.org/10.1006/ijhc.1999.0252. URL `http://www.sciencedirect.com/science/article/pii/S1071581999902525`.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP: adversarial attacks on post hoc explanation methods. In *Proceeding of the 3rd AAAI/ACM Conference on AI, Ethics, and Society,*, pp. 180–186, 2020.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems*, 34:9391–9404, 2021.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda B. Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*, 2017.

David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011. ISSN 0036-8075. doi: 10.1126/science.1191181. URL `https://science.sciencemag.org/content/333/6048/1393`.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *arXiv preprint arXiv:1610.00168*, 2016.

Fulton Wang and Cynthia Rudin. Falling rule lists. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 1013–1022, 2015.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pp. 279:1–279:12, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300509. URL `http://doi.acm.org/10.1145/3290605.3300509`.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

Jianlong Zhou, Syed Z Arshad, Simon Luo, and Fang Chen. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *IFIP Conference on Human-Computer Interaction*, pp. 23–39. Springer, 2017.

A Additional results on the impact of type of uncertainty on decision making stratified by expertise

We present additional results from our qualitative analysis of the impact of uncertainty type stratified by participant expertise. Figure 7 illustrates the magnitude of the update (left), final disagreement (middle), and initial disagreement (right) for participants with little or no ML expertise. These results are presented in the appendix because sample sizes are much smaller for this level of ML expertise (See Table 2). We show only those conditions for which we have at least 3 participants.

Participants in the no uncertainty condition have higher initial disagreement and higher final disagreement, despite also having larger updates. Participants appear to disagree with the model much more in the second estimate when shown no uncertainty, however this appears to be driven by higher initial disagreement, even despite participants in this condition updating their estimates the most. This is similar to the trend we see for people without domain expertise (Main document Figure 5), although in that case, the updates were of a similar magnitude to those in other conditions. This again raises questions about whether people learn to react differently over time in different conditions, or whether the sample size is simply too small to draw robust conclusions.

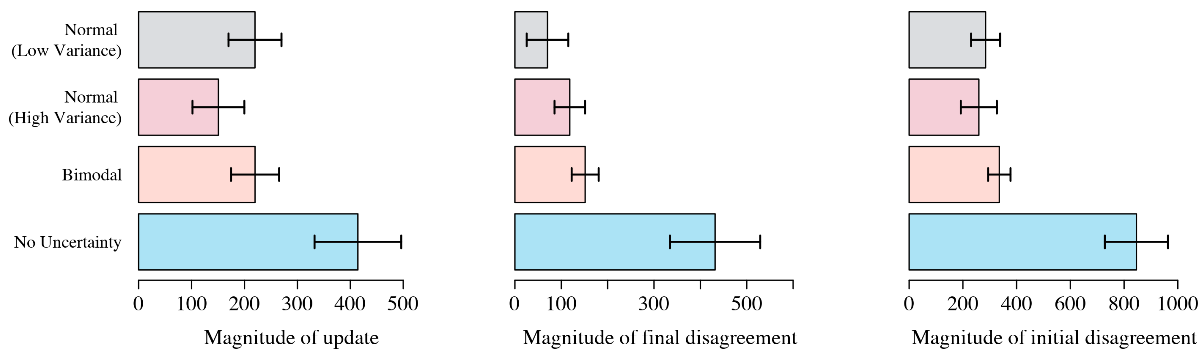

Figure 7: Magnitude of update (left), final disagreement (center) and initial disagreement (right) stratified by uncertainty condition. Bar width corresponds to the mean, and standard errors are shown in black for participants with little or no ML expertise; sample sizes in this condition were much smaller. Participants appear to disagree with the model much more in the second estimate when shown no uncertainty, however this appears to be driven by higher initial disagreement, even despite participants in this condition updating their estimates the most.

B Correlations between self-reported measures and decision making

The final question we explored was to what extent self-reported measures of trust and usefulness correlated with people's actual decisions. Table 3 captures details of participant ratings indicating how much they trust the model and how useful they find it. In Figure 8, we give the average distance between first and second estimates (left panel), average distance between the second estimate and model prediction (middle panel), and average distance between the first estimate and model prediction (right panel) among participants in each category of self-reported trust. Figure 8 gives the analogous results for self-reported usefulness.

Self-reported trust correlates with the magnitude of the update between the first and second estimates, but the trend is less clear for usefulness. The amount of self-reported trust has a marginally significant effect on the difference between the first and second estimate (significant at the p<0.10 level: $F(3, 91) = 2.3361$, $p = 0.079$). The magnitude of the update largely increases with trust, as expected. We do not observe similar trends in case of the difference between the second estimate and the model's prediction. This suggests that people alter their original predictions more when they trust the model more, but this does not necessarily result in updated estimates that are closer to the model's prediction. We do not observe any statistically significant differences in any of the three considered metrics between the groups defined by self-reported usefulness of the model.

Table 3: Number of participants with each level of self-reported trust, and perceived usefulness of the model. Most participants trusted the model moderately or a lot, and most participants found the model either somewhat useful or very useful.

Self-reported measure	**N. participants (%)**
Trust in the model	
None at all	0/95 (0.00%)
A little	13/95 (13.68%)
A moderate amount	47/95 (49.47%)
A lot	30/95 (31.58%)
A great deal	5/95 (5.26%)
Usefulness of the model	
Very useless	0/95 (0.00%)
Somewhat useless	6/95 (6.32%)
Neither useful nor useless	10/95 (10.53%)
Somewhat useful	55/95 (57.89%)
Very useful	24/95 (25.26%)

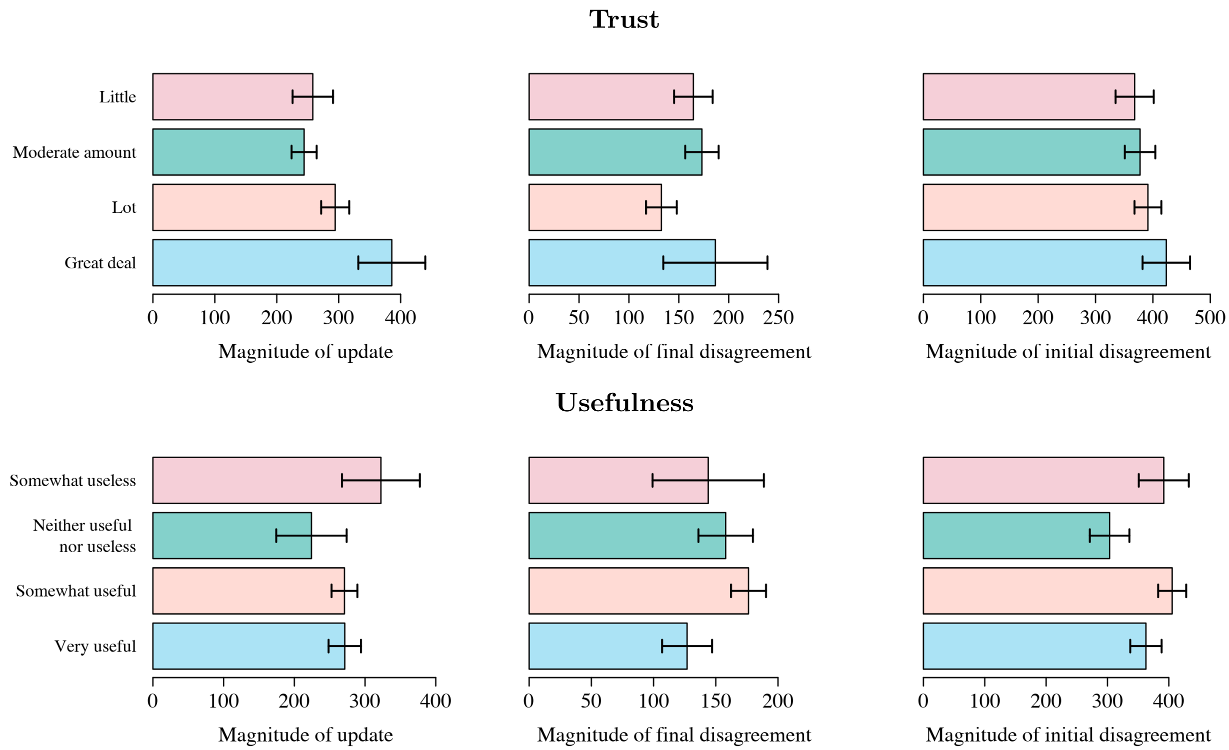

Figure 8: Magnitude of update (left), final disagreement (center) and initial disagreement (right) stratified by self-reported trust (top) and self-reported usefulness (bottom). Bar width corresponds to the mean, and standard errors are shown in black. The magnitude of update is significantly different between the groups defined by self-reported trust. Self reported usefulness does not have statistically significant effects in any of these cases.